IDENTIFYING PATTERNS IN DRUG EFFICACY BY ANALYZING DRUG REVIEWS THROUGH A CLUSTERING APPROACH

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 4

RESEARCH DESIGN AND IMPLEMENTATION

4.1 Introduction

In this chapter, the steps to obtain the clusters based on side effects and effectiveness ad been introduced. Exploratory data analysis (EDA) was carried out to investigate the dataset. Then, data preparation was done by handling the missing data and duplicates data that was presented in the dataset. After obtaining a cleaned dataset, the review column in the dataset was further retrieved by LLM to identify the side effects and effectiveness of the drug.

4.2 Exploratory Data Analysis (EDA)

EDA was carried out to understand the data patterns. The rating column illustrated the ratings of patients when experienced with the drug. The rating value was starting from 1 to 10 in which 1 represented the patient was dissatisfied with the drug meanwhile 10 represented the patient was satisfied with the drug. From the figure below, the rating 10 showed highest frequency indicated that majority of patients had the better drug experience. Meanwhile, rating 4 showed the lowest frequency indicated that few of patients had a bad experience with drug. The distribution of rating illustrated that most patients were satisfied with the drug taken.

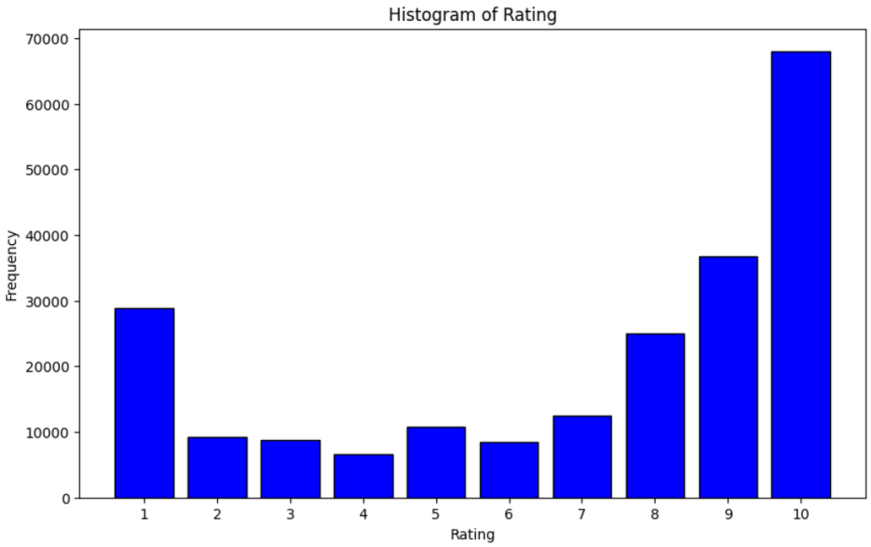

Figure 4.1: Histogram of Rating Feature

The usefulCount column described the number of people that found the review is useful. The box plot of usefulCount allowed the visualization on the distribution of the number of people found the review is useful. Based on the figure below, majority of count was considered as outliers. The interquartile range of box plot showed that most of the reviews had the low number of people found that the reviews is useful. The presence of outliers represented that small number of reviews obtained higher count of people that found the review is useful. However, the useful count of reviews did not indicate the low performance of drug. Instead, the useful count reflected the quality of review that help people in choosing their drug.

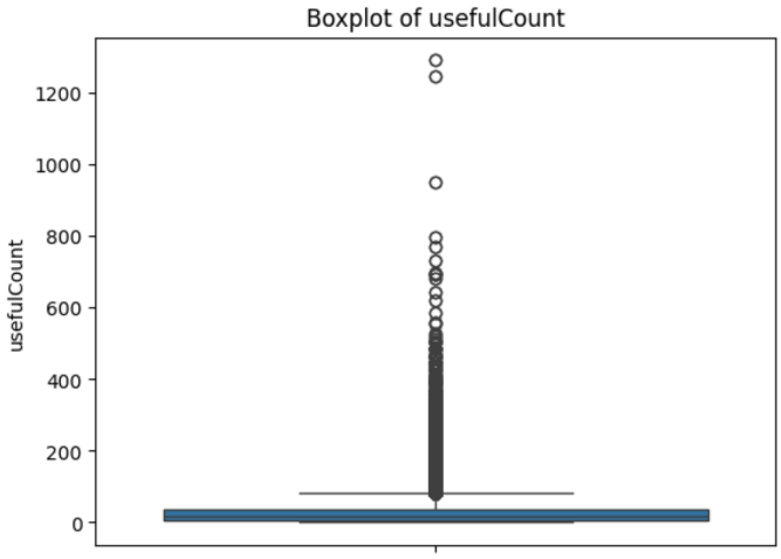

Figure 4.2: Box Plot of usefulCount

The scatter plot below illustrated positive relationship between rating and useful count. The increasing of rating of drug performance, the increasing of the number of people that found the reviews is useful. Therefore, reviews with higher ratings tend to help others to select a better treatment plan.

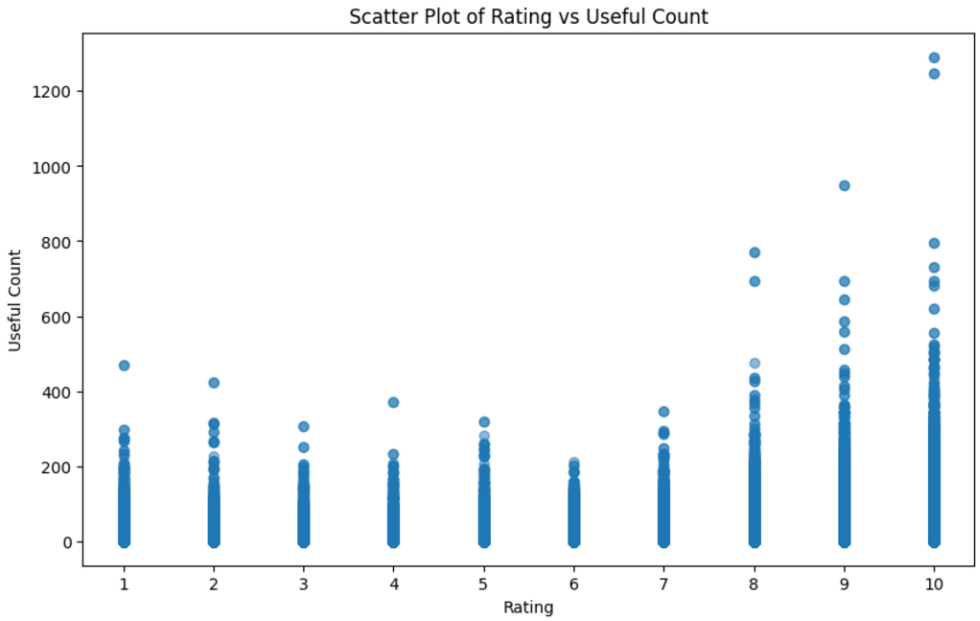

Figure 4.3: Scatter Plot between Rating and usefulCount

There are total of 916 conditions that had been involved in the dataset. The conditions represented as the specific health issues that the drug being used by the patient. The figure below shows the top 10 conditions that had been commented on by patients. Among 916 conditions, birth control achieved the highest frequency at 38436 while there still existed with the conditions that only discussed once. Figure 4.5 also showed that the presence of irrelevant data in the condition. Therefore, data preparation will be carried out to handle the misinformation.

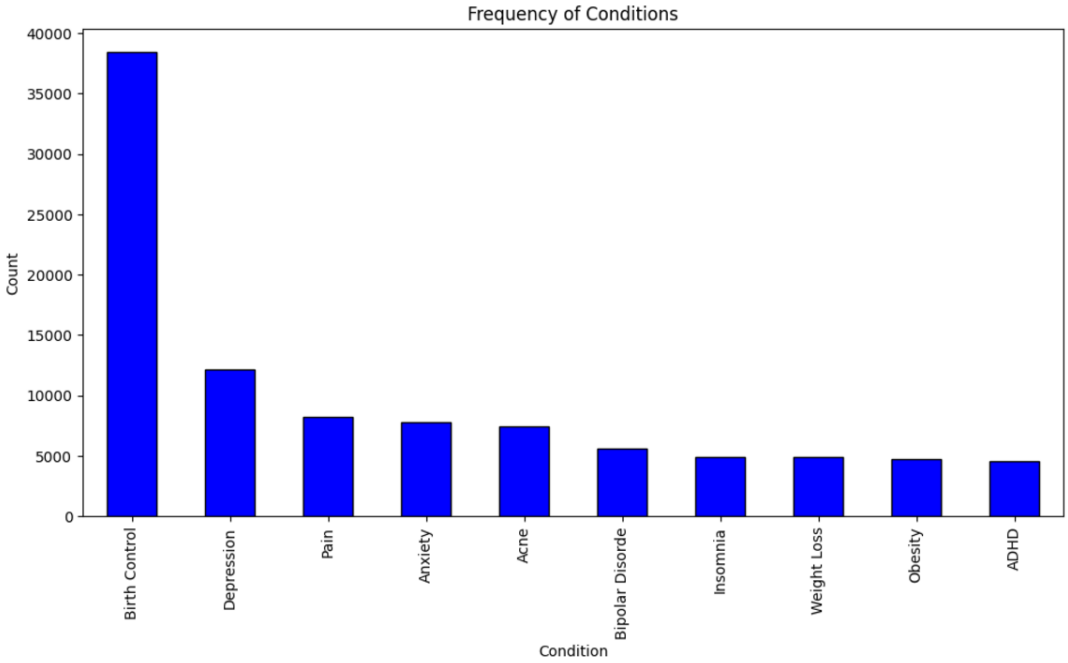

Figure 4.4: Top 10 Conditions

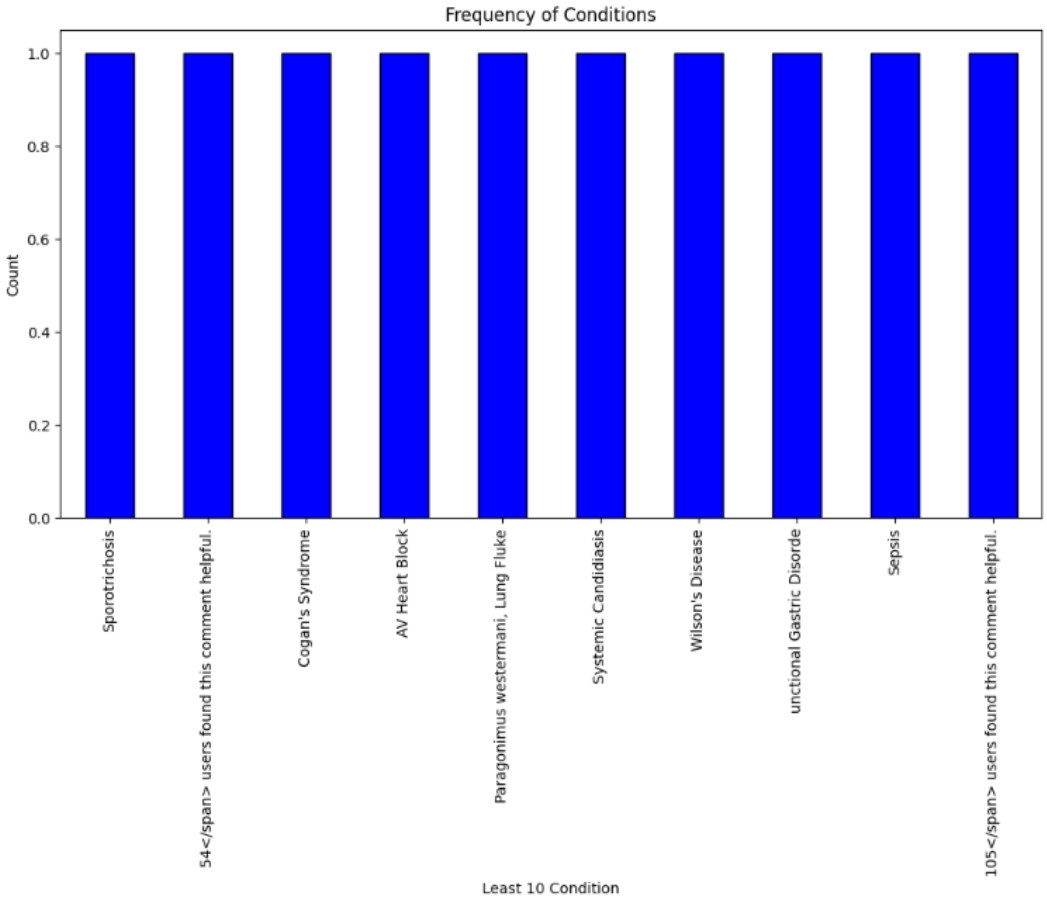

Figure 4.5: Least 10 Conditions

There are total of 3671 drugs that had been involved in the dataset. The drug represented as the drug being used by the patient. The figure below showed the top 10 drug that had been commented by patients. Among 3671 drugs, levonorgestrel achieved the highest frequency at 4930 while there still existed with the drugs that only discussed once. Figure 4.7 illustrated the least 10 drugs that had been discussed in the reviews. The limited discussed on the drug can be due to the low usage, limited availability of the drugs and the drugs that are new to the market.

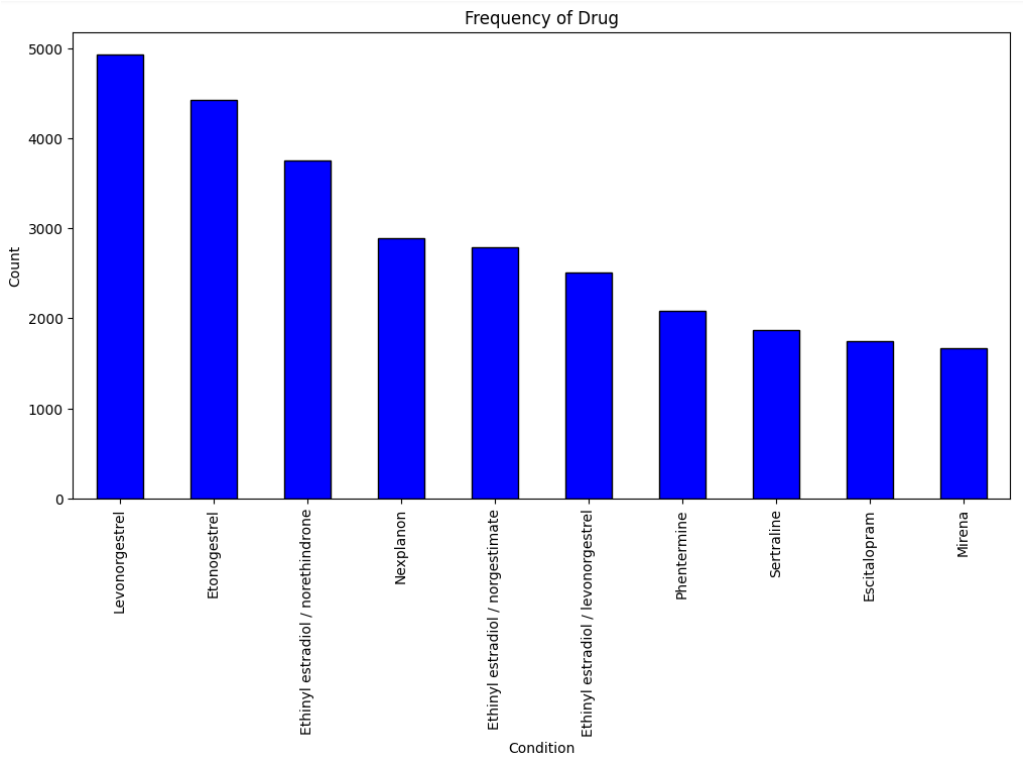

Figure 4.6: Top 10 Drug

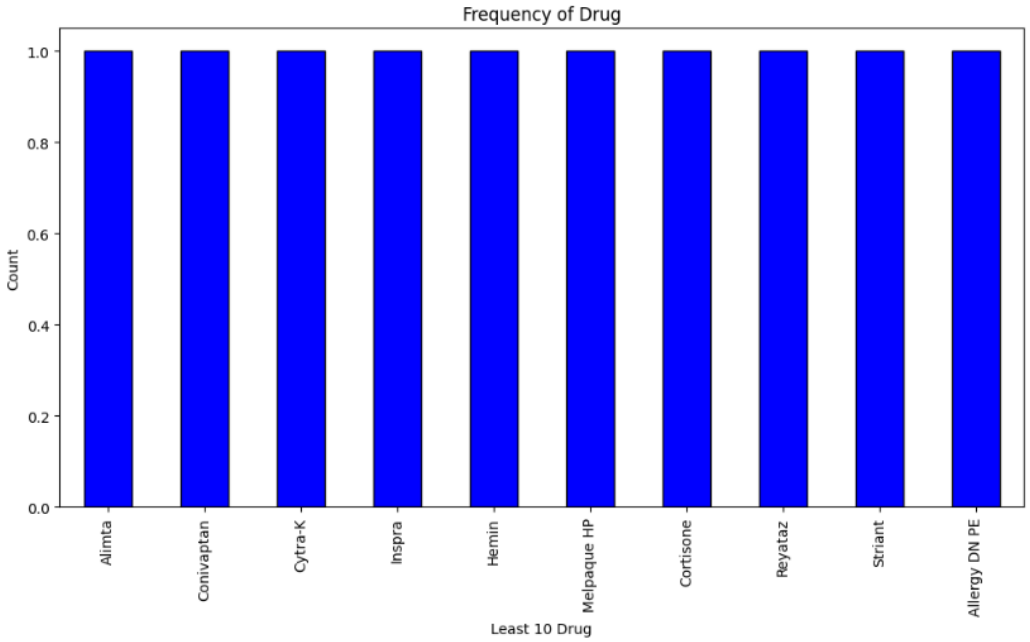

Figure 4.7: Least 10 Drug

The frequency was calculated after text processing. The text in the review column was cleaned by removing stop words and punctuation. Then, the words had been tokenized and lemmatized.

Figure below showed the word cloud for positive sentiment reviews. Word cloud analysis illustrated that “right”, “new”, “high”, “hard”, “live”, “calm” and “strong” were the most frequent used words in the review. The word “right” represented that the drug was working with their condition. The word “high” represented the high effectiveness of drug. Meanwhile, the word “live” and “calm” represented the improvement of life quality with the drug.

Figure 4.8: Word Cloud of Positive Sentiment

Figure below showed the word cloud for negative sentiment reviews. Word cloud analysis illustrated that “hard”, “sick”, “high”, “right”, “new”, “live”, “mean”, “calm” and “serious” were the most frequent used words in the review. The word “hard”, “sick” and “serious” represented the side effects of drug. The word “right”, “new”, “live” and “calm” that presented in the negative sentiment review can represented as the drug is not working as expected.

Figure 4.9: Word Cloud of Negative Sentiment Review

Figure below showed the word cloud for neutral sentiment reviews. Word cloud analysis illustrated that “pain”, “day”, “back”, “lik”, “year”, “non”, “nausea”, “mouth” and “back” were the most frequent used words in the review. These words can represent that the drug did not produce with a better experience and worse experience to the patient.

Figure 4.10: Word Cloud of Neutral Sentiment Review

4.3 Data Preparation

Data preparation is a crucial step in ensuring the data is clean and formatted before further processing. Therefore, the drug reviews dataset had been investigated to handle the data-related issues. The process involved checking for missing values, irrelevant data and duplicates to improve dataset quality. If the missing values and irrelevant data had occurred in the “drugName” column, then the row will be removed. However, if the missing values and irrelevant data had occurred in other columns, then will be replaced with “Not Specified”. This is because “drugName” is the primary feature in the dataset. The lack of drug name information limits the understanding of the side effects and effectiveness. Lastly, the text data such as drug name, condition and review were converted into lowercase to ensure the uniformity of the word and avoid the duplication of words existing in the dataset.

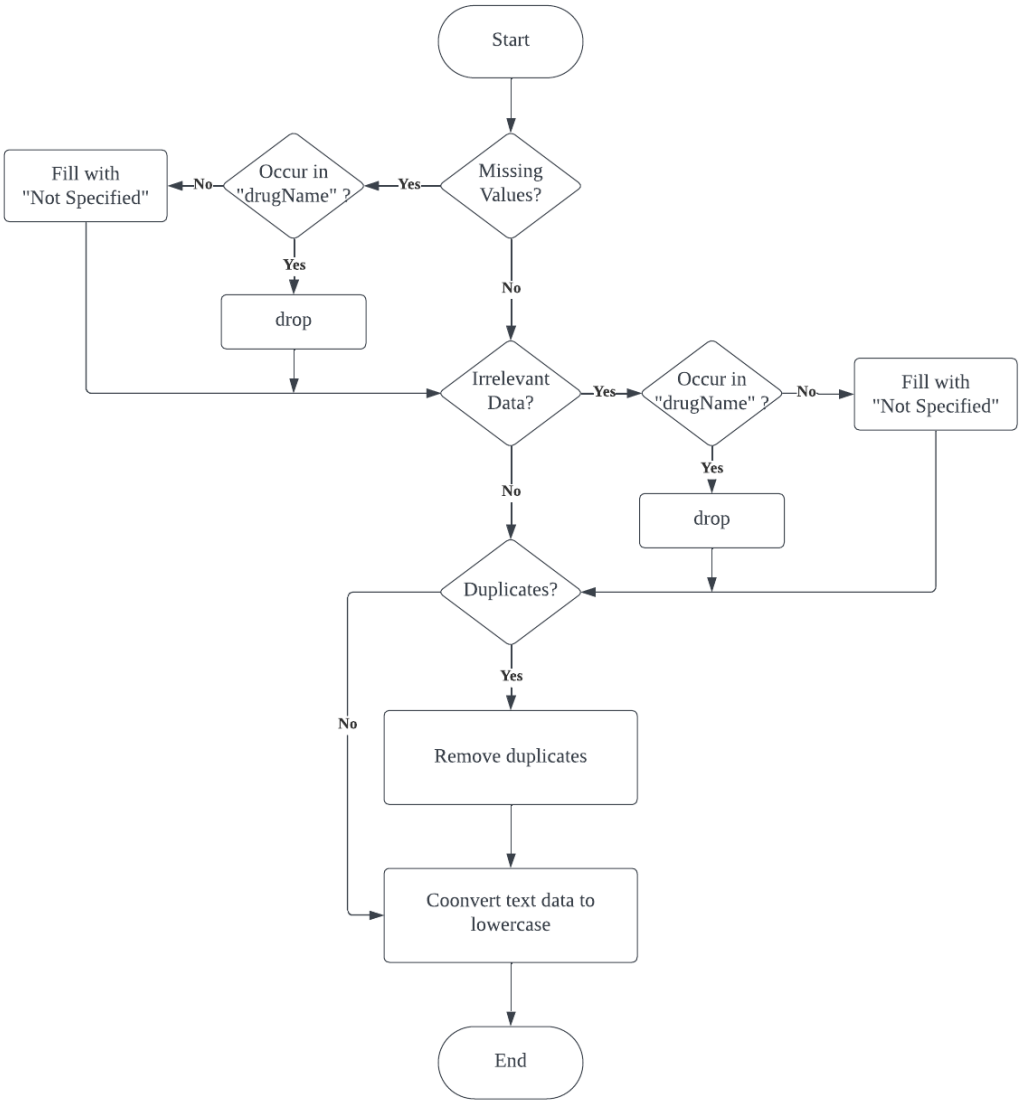

Figure 4.11: Data Preparation Flow

4.4 Data Derivation

As mentioned before, patient reviews always provide the information regarding to the experience when consuming drug. The side effects and effectiveness of the drug can be recorded and found from the patient reviews. Therefore, it is important to identify the drug efficacy from the reviews. From the literature review, LLMs were able to understand the relationship between words and phrases which make it more advanced than the traditional methods in text processing. Therefore, the

ChatGPT 4o mini model was used to derive the features such as side effects and effectiveness from the review.

OpenAI company provided the platform to allow users to use their products. By creating the application programming interface (API) key, users were allowed to use the services that OpenAI had provided. After inputting the API key, the chatgpt-4o-mini model was able to use, the temperature parameter was set as 0 to ensure the consistency of the output. The prompt question that had been defined in this research is *"Analyze the following drug review and extract keywords that are specifically related to side effects and the effectiveness of the drug. Provide the output as a JSON object with two keys: 'side_effects' and 'effectiveness'."*. The prompt will guide the model to derive the relevant information from the review.

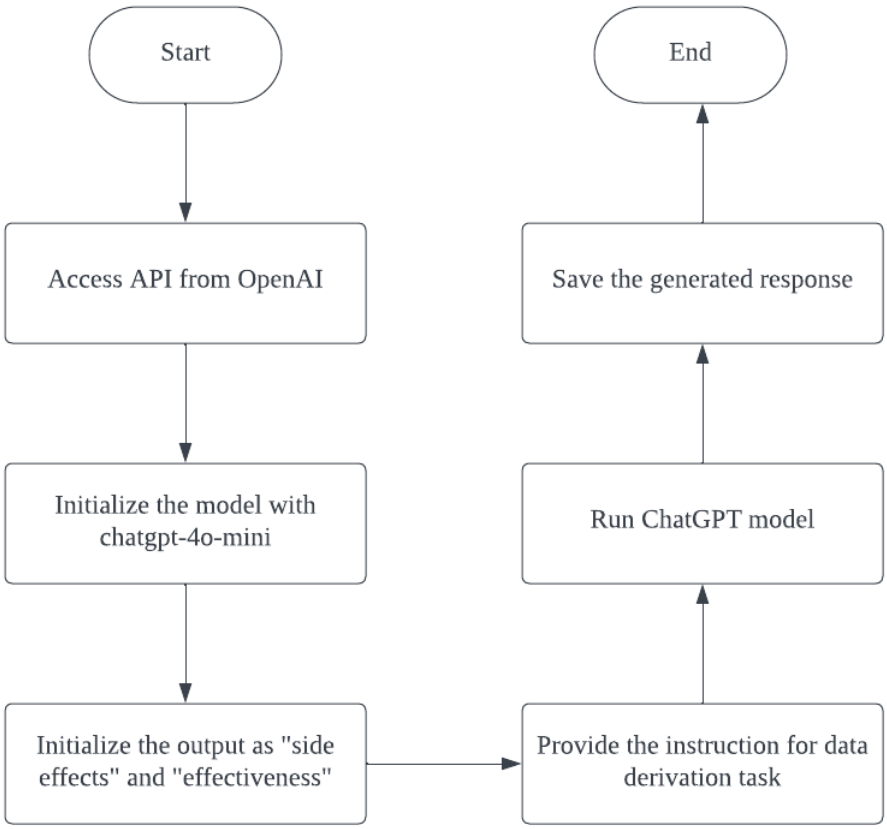

Figure 4.12: Data Derivation Flow

Table 4.1 showed the example of data extraction. The model will observe the review and retrieve the keywords from the review based on the instructions given. If

there are no features detected, then the model will return blank. Meanwhile, if the model detected the related features, then will return the information.

Table 4.1: The example of data derivation

Review	Extracted Keywords	
	Side effects	Effectiveness
"it has no side effect, i take it in combination of bystolic 5 mg and fish oil"	['no side effect']	['combination of bystolic 5 mg', 'fish oil']
"i live in western australia and disturbed by some comments on here. the cost of embrel is cost of an ordinary prescription $36 for me the government pays the remainder of the cost to the chemist. i also go to the the medical centre every saturday morning a dr looks over my prescription and then he advises the nurse to administer the injection also no cost to myself and this is part of nurses duties. i am unsure of the country where people who have made comments referring to cost and that nobody is there to administer the injection for them. i am very lucky to live in australia as we have the best health system worldwide and everybody is given the opportunity to receive proper medical help whether you are rich or poor there is no discrimination."	[]	[]
"average-- not satisfied -- symptoms continue"	['symptoms continue']	['average', 'not satisfied']

4.5 Model Development

After retrieving the important features from the reviews, DBSCAN was applied to the features. Figure 4.13 illustrated the process flow of DBSCAN. DBSCAN will start with defining the epsilon, the density of the neighborhood and the MinPts, minimum number of points to form a cluster. Then, it will randomly select the points and check with the requirements. If the number of points in the neighborhood is greater or equal to the minimum number of points defined, then will form a cluster. In contrast, the point either mark with noise or boarder point. The DBSCAN will end when there are no points that have not yet been processed.

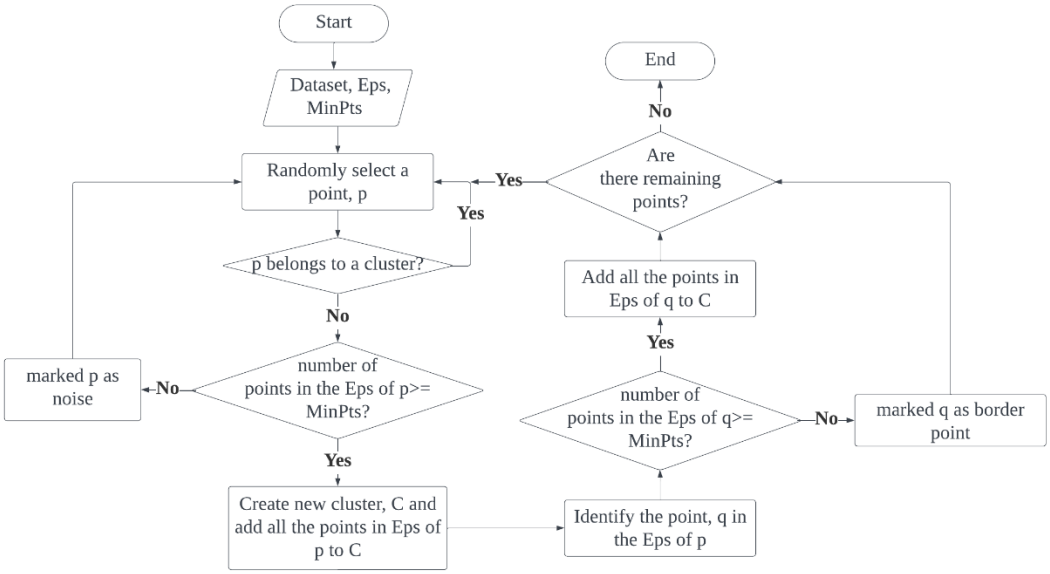

Figure 4.13: DBSCAN Flow

4.6 Model Evaluation

The silhouette coefficient will be used to measure the similarities between the data points within the cluster and between the clusters. By computing each of the data points, the clustering quality can be evaluated. Therefore, the effectiveness of models in grouping similar information together can be investigated.

4.7 Summary

In conclusion, there were five phases involved in the identification of drug efficacy. The ChatGPT model is used to extract the relevant information regarding the side effects and effectiveness of drug from the review. Then, DBSCAN was implemented to cluster the side effects and effectiveness based on their similarities. Lastly, the silhouette coefficient was used to measure the effectiveness of the model in clustering the data.