UTM
UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA

MDS Project

Identifying Patterns in Drug Efficacy by Analysing Drug Reviews Through A Clustering Approach

Done by: Gui Yu Xuan (MCS241003)
Supervised by: Dr Chan Weng Howe

INTRODUCTION

Why Analyze Drug Reviews

- Consumers increasingly **relying on the reviews** to make purchasing decisions
- Reviews provide **useful information** about drug's effectiveness, side effects and overall patient experience

Online review about drugs consist of two major parts:
Ratings: Numerical assessment of patient's overall experience
Text Review: Details insight into drug's effectiveness and side effects

Optimize Medication Therapy

Medication errors can be minimized as drug reviews **providing the patterns of patient's medical conditions**

Improved Clinical Decision

Select a better treatment plans by **understanding the drug performance across various conditions**

PROBLEM BACKGROUND

RCTs considered as gold standard in evaluating drug efficacy due to the monitored procedure that aim to eliminate bias

RCTs **fail to provide comprehensive understanding** of drug effectiveness in the real world

Traditional text processing on patient feedback have the **limitations in recognizing the patterns** in words.

Application of LLMs and clustering extract insightful information from unstructured data

PROBLEM STATEMENT

“

The Limited Generalizability of RCTs in Real World Scenarios Restrict The Ability To Capture the Side Effects and Effectiveness of Drug Consumed under Specific Condition

RCTs frequently conducted with **controlled conditions** and select individuals with specific disease, health status and characteristic, hence **limiting its generalizability** for a more diverse population. Thus, drug reviews commented by patient provide the real-world experiences with unreported side effects and the efficacy results.

”

Research Goal

To identify patterns in drug efficacy to enhance the understanding of drug performance across diverse patient populations by utilizing LLMs and clustering techniques in patient drug reviews

Research Question

RQ: What preprocessing steps to carry out for the analyzing of drug efficacy from drug reviews dataset?

RQ: What relevant keywords can be identified and retrieved by LLMs from drug reviews dataset?

RQ: What insights can be drawn from drug reviews clusters by the retrieved keywrods?

Research Objectives

RO: To **conduct a preprocessing** of the drug reviews datasets for drug efficacy analysis

RO: To **retrieve relevant keywords** from the preprocessed dataset by using LLMs

RO: To implement clustering techniques to **identify the similarities** of retrieved keywords and **visualize** the findings

LITERATURE REVIEW

Drug Efficacy Evaluation In RCTs

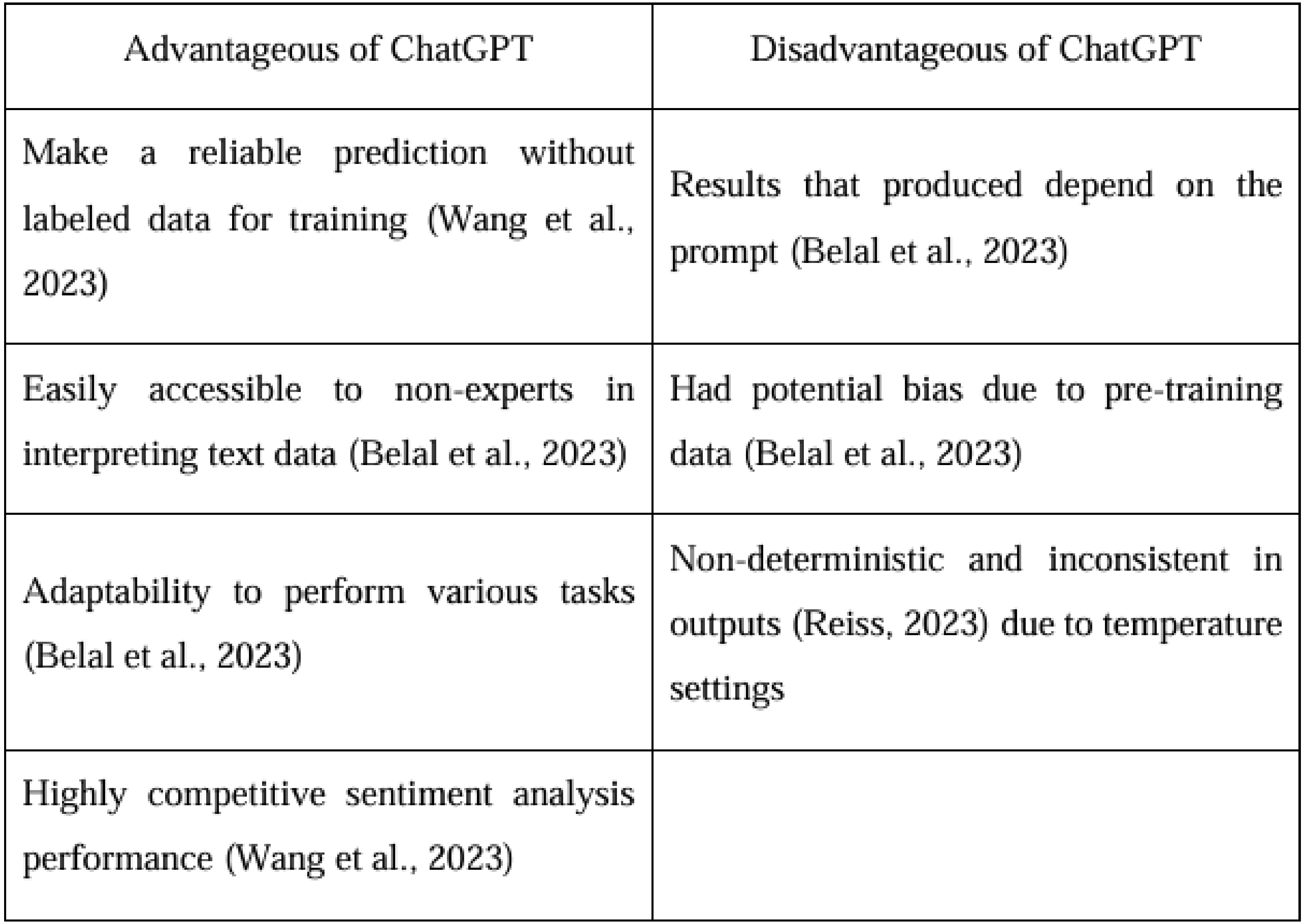
- **RCTs** are important for medication regulatory approval and the development of medical knowledge and policy
- The **volunteers are randomly distributed** to various groups
  - Randomization process ensure the even distribution of age, gender and health status
  - But there are **strict requirements** to follow and will exclude individuals with various health issues
- The differences between RCT and real-world populations does exist
  - Characteristics in real-world populations hard to measure cause the RCT result **cannot accurately represent** the features in real-world scenarios

Patient Review as A Real-World Data Source

- Large amounts of data on Internet allows comprehensive evaluations and pattern identification.
- Drugs review can be considered as statistical data that enable medical professionals in collecting medical data.
- The advanced analytics techniques can extract valuable insights from unstructured data.
  - **Sentiment analysis:** investigate patient review by understanding patient experiences
  - **ML:** classify text data based on disease states, patient references
- Real-world data allow **inclusion of diverse patient populations**
- Analyzing real-world data able to enhance patient outcomes and quality of life.

LLMs In Text Analysis

Table 2.1: Summarization of Advantage and Disadvantage of ChatGPT



Advantageous of ChatGPT	Disadvantageous of ChatGPT
Make a reliable prediction without labeled data for training (Wang et al., 2023)	Results that produced depend on the prompt (Belal et al., 2023)
Easily accessible to non-experts in interpreting text data (Belal et al., 2023)	Had potential bias due to pre-training data (Belal et al., 2023)
Adaptability to perform various tasks (Belal et al., 2023)	Non-deterministic and inconsistent in outputs (Reiss, 2023) due to temperature settings
Highly competitive sentiment analysis performance (Wang et al., 2023)	

Clustering Techniques in Text Analysis

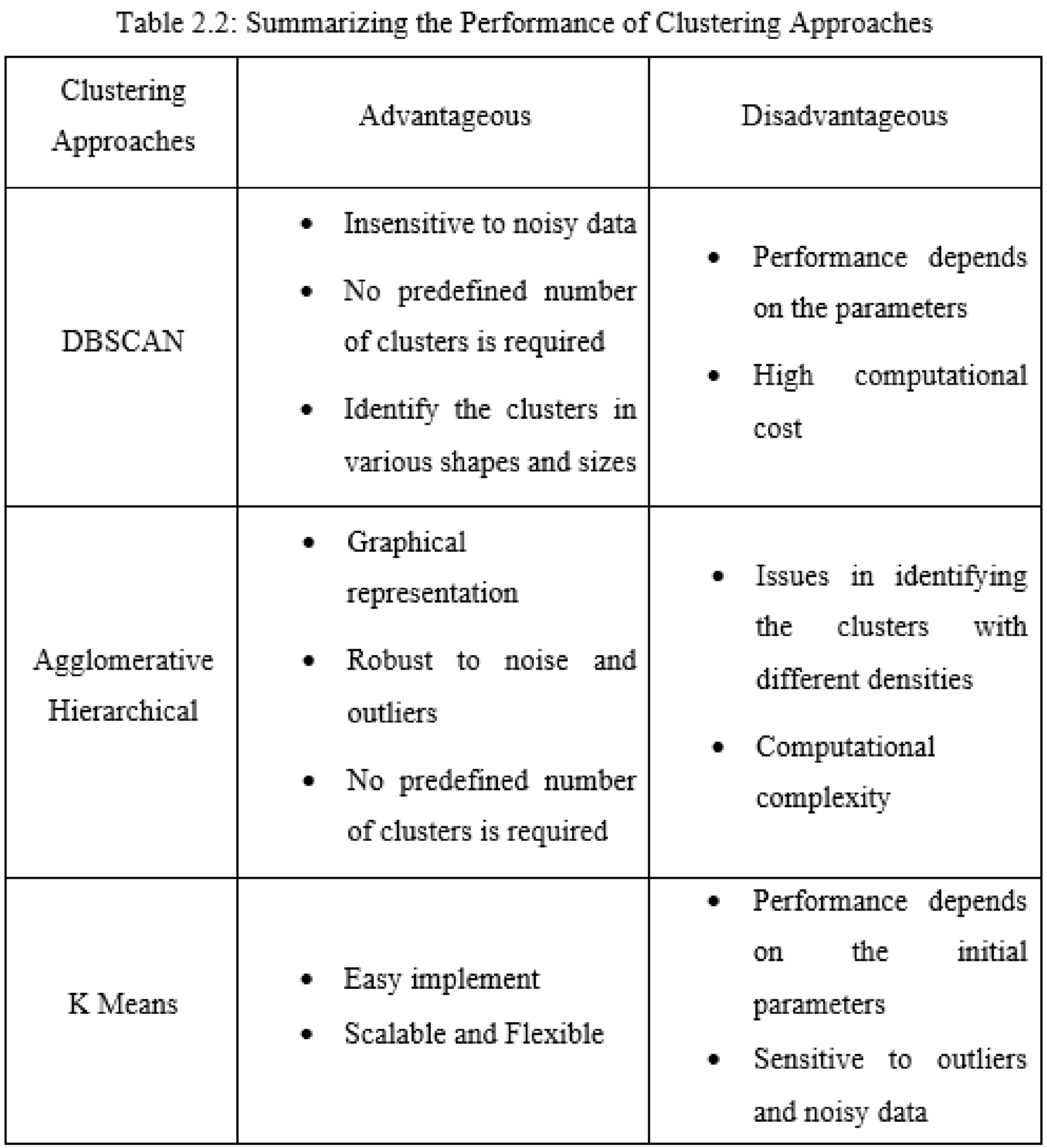
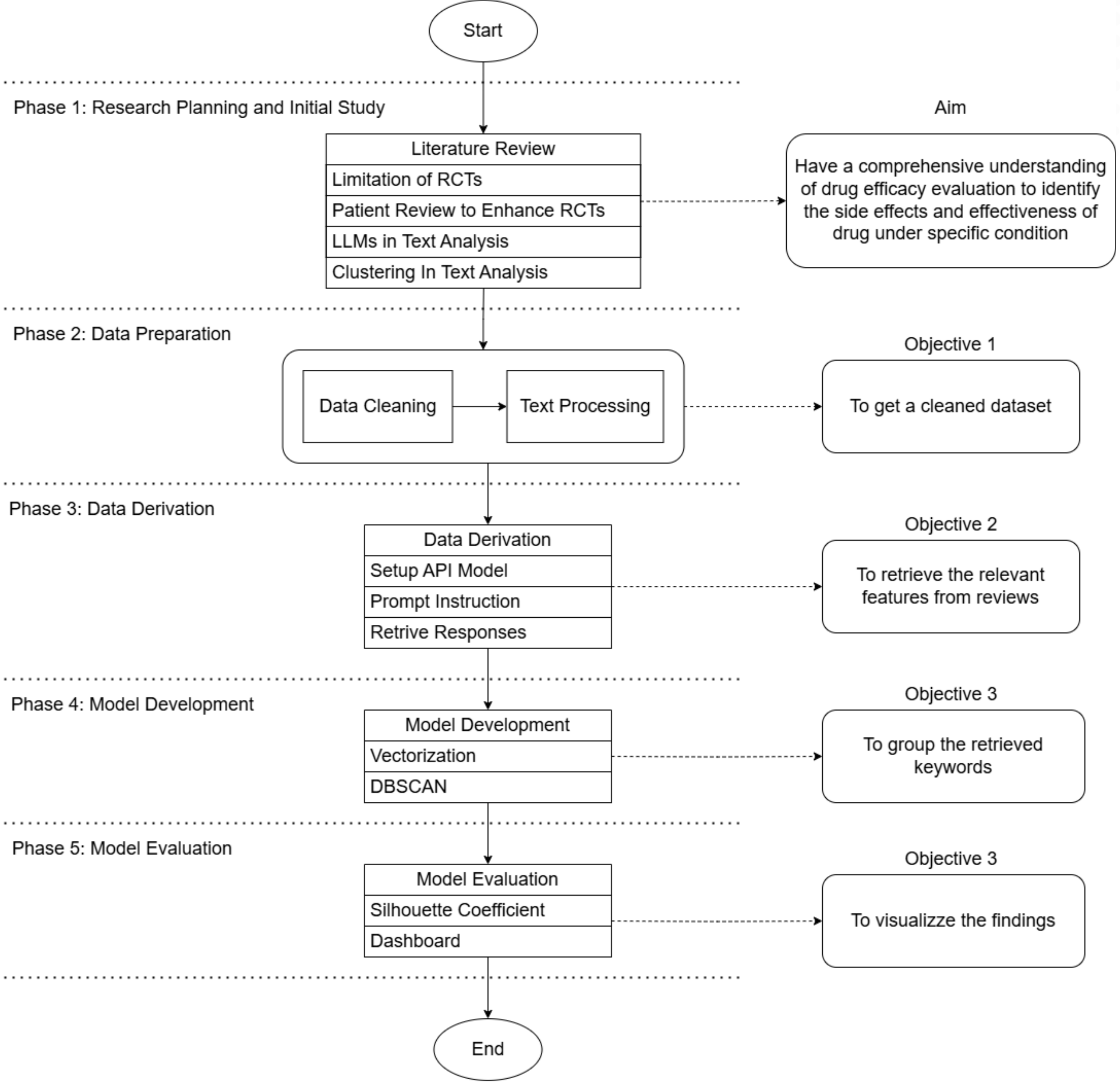


Table 2.2: Summarizing the Performance of Clustering Approaches

Clustering Approaches	Advantageous	Disadvantageous
DBSCAN	• Insensitive to noisy data • No predefined number of clusters is required • Identify the clusters in various shapes and sizes	• Performance depends on the parameters • High computational cost
Agglomerative Hierarchical	• Graphical representation • Robust to noise and outliers • No predefined number of clusters is required	• Issues in identifying the clusters with different densities • Computational complexity
K Means	• Easy implement • Scalable and Flexible	• Performance depends on the initial parameters • Sensitive to outliers and noisy data

RESEARCH METHODOLOGIES



Start

Phase 1: Research Planning and Initial Study

Literature Review
- Limitation of RCTs
- Patient Review to Enhance RCTs
- LLMs in Text Analysis
- Clustering In Text Analysis

Aim

Have a comprehensive understanding of drug efficacy evaluation to identify the side effects and effectiveness of drug under specific condition

Phase 2: Data Preparation

Data Cleaning → Text Processing

Objective 1

To get a cleaned dataset

Phase 3: Data Derivation

Data Derivation
- Setup API Model
- Prompt Instruction
- Retrive Responses

Objective 2

To retrieve the relevant features from reviews

Phase 4: Model Development

Model Development
- Vectorization
- DBSCAN

Objective 3

To group the retrieved keywords

Phase 5: Model Evaluation

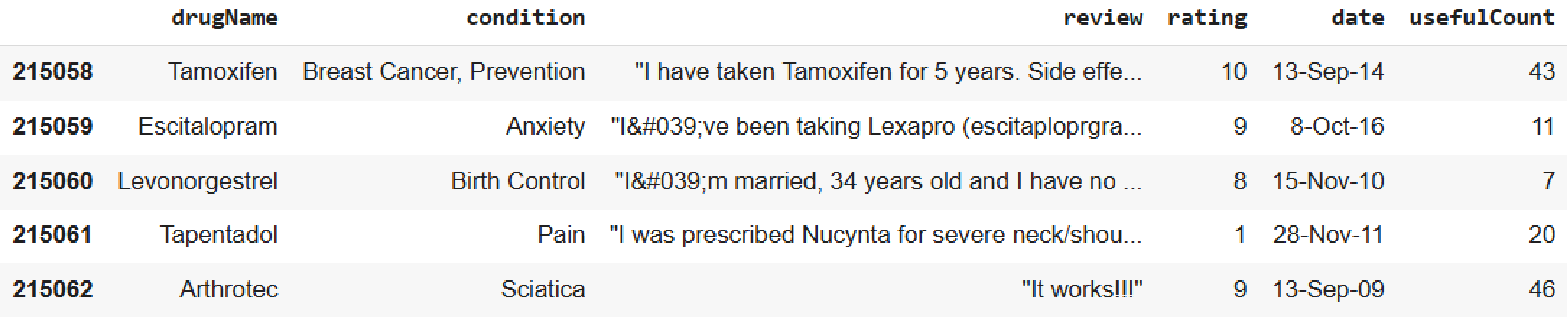
Model Evaluation
- Silhouette Coefficient
- Dashboard

Objective 3

To visualizze the findings

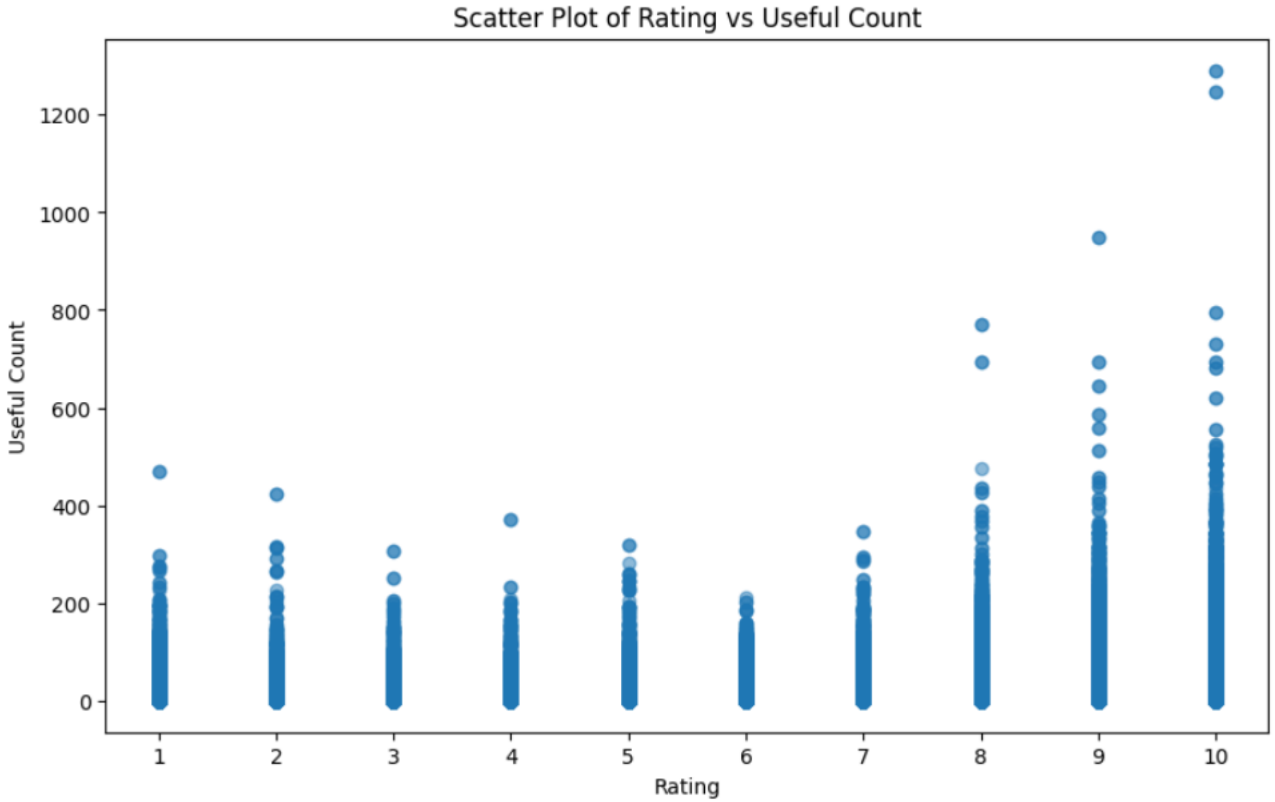
End

Dataset

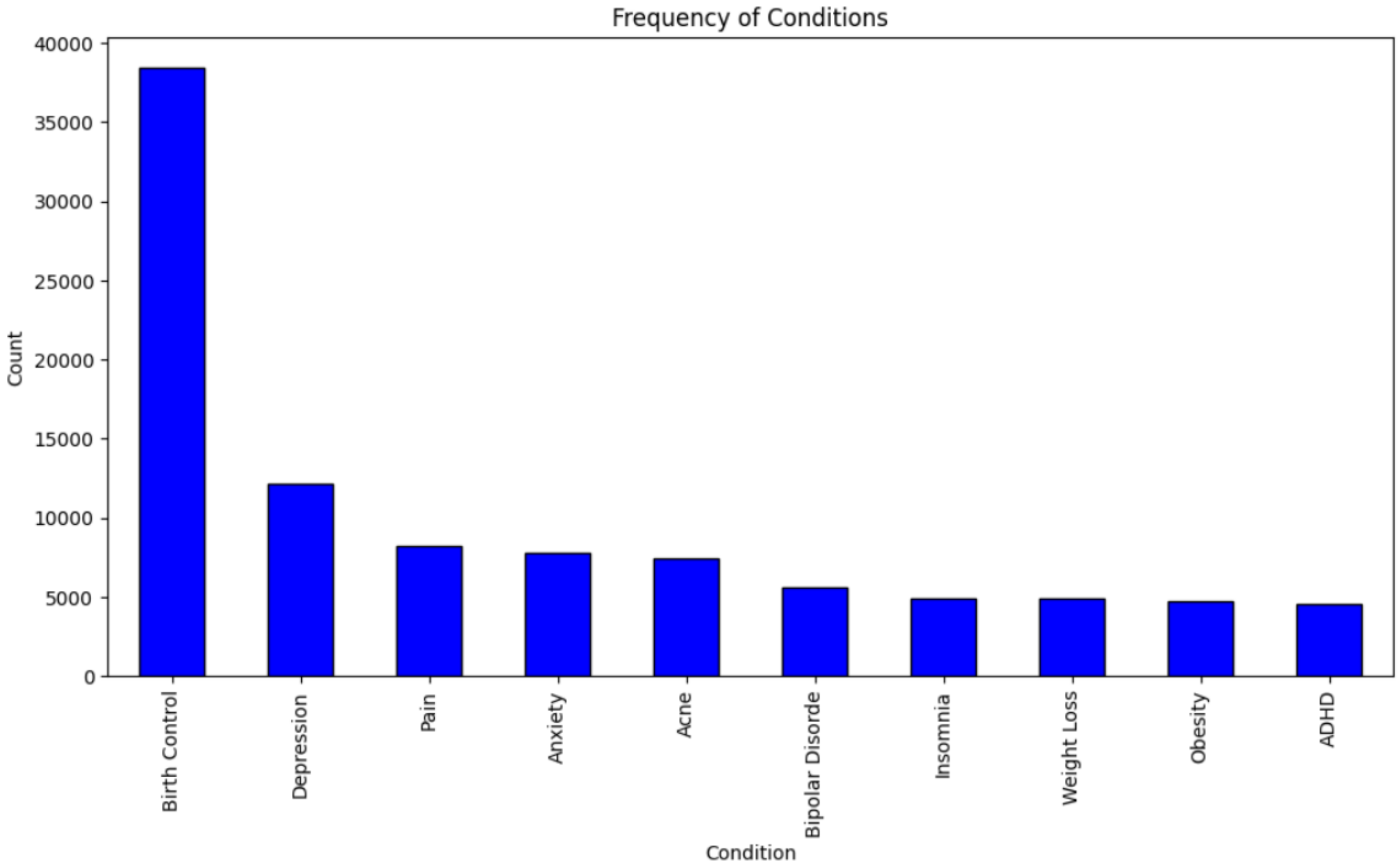


	drugName	condition	review	rating	date	usefulCount
215058	Tamoxifen	Breast Cancer, Prevention	"I have taken Tamoxifen for 5 years. Side effe...	10	13-Sep-14	43
215059	Escitalopram	Anxiety	"I&#039;ve been taking Lexapro (escitaploprgra...	9	8-Oct-16	11
215060	Levonorgestrel	Birth Control	"I&#039;m married, 34 years old and I have no ...	8	15-Nov-10	7
215061	Tapentadol	Pain	"I was prescribed Nucynta for severe neck/shou...	1	28-Nov-11	20
215062	Arthrotec	Sciatica	"It works!!!"	9	13-Sep-09	46

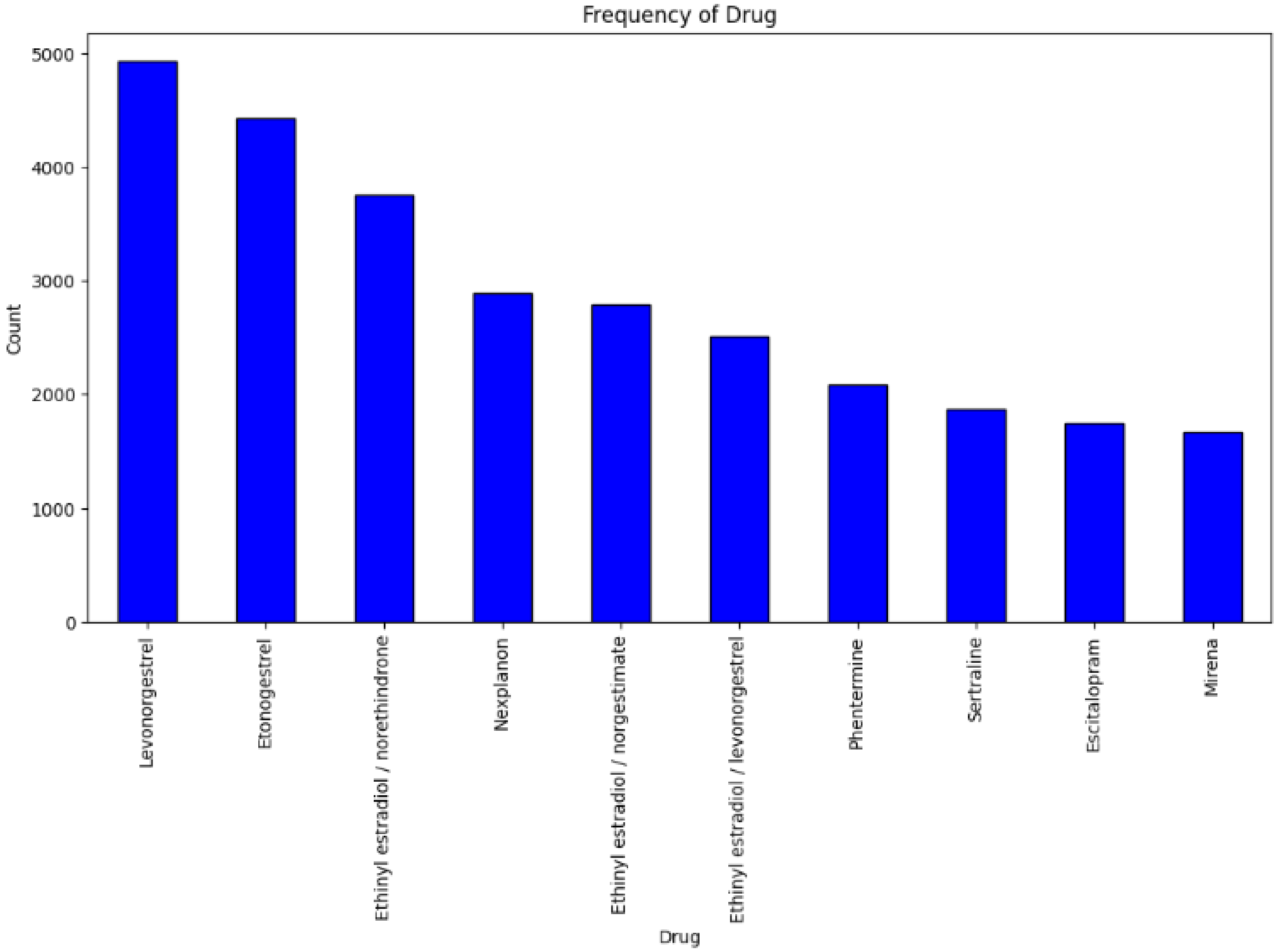
RESEARCH DESIGN AND IMPLEMENTATION



RESEARCH DESIGN AND IMPLEMENTATION



RESEARCH DESIGN AND IMPLEMENTATION



RESEARCH DESIGN AND IMPLEMENTATION

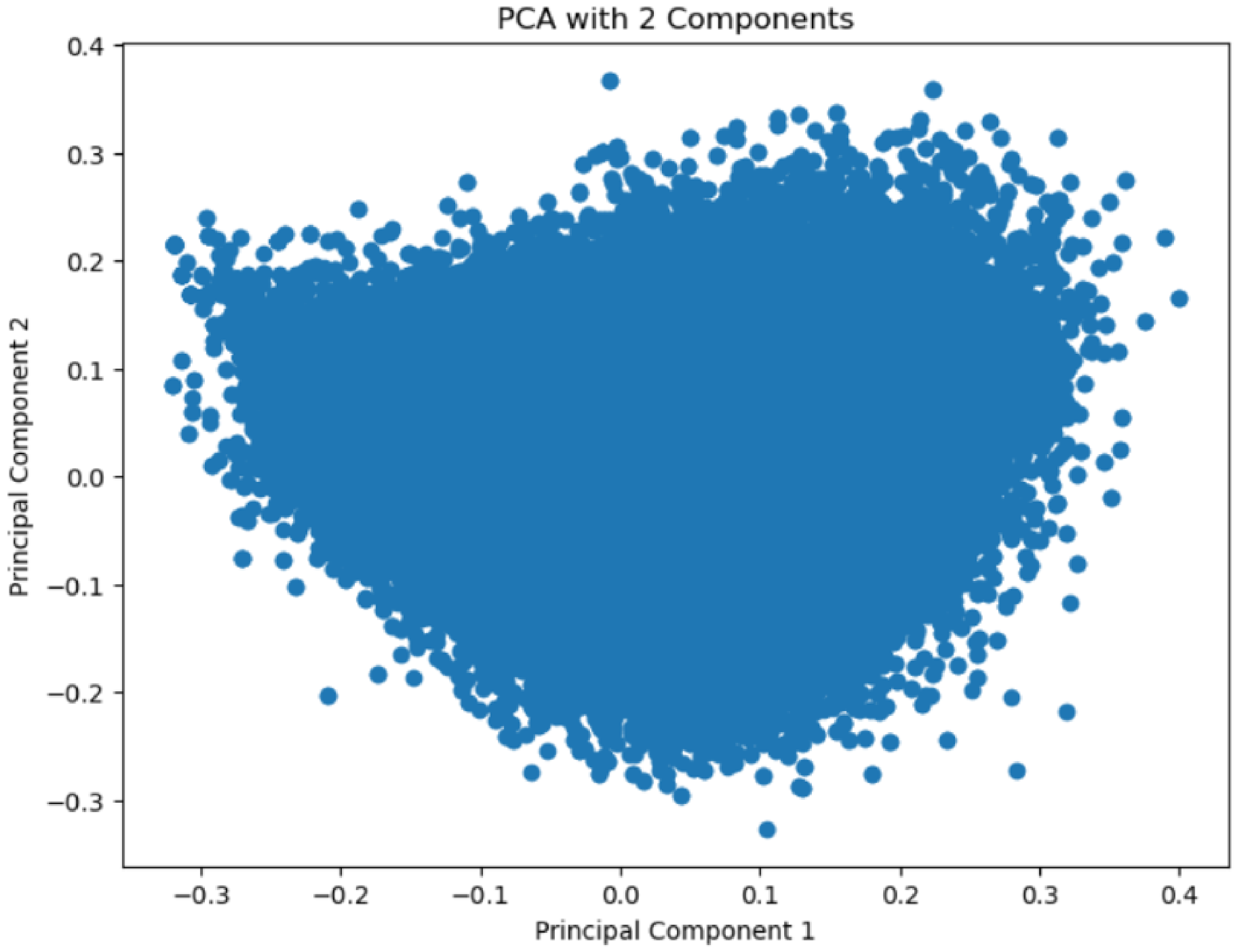
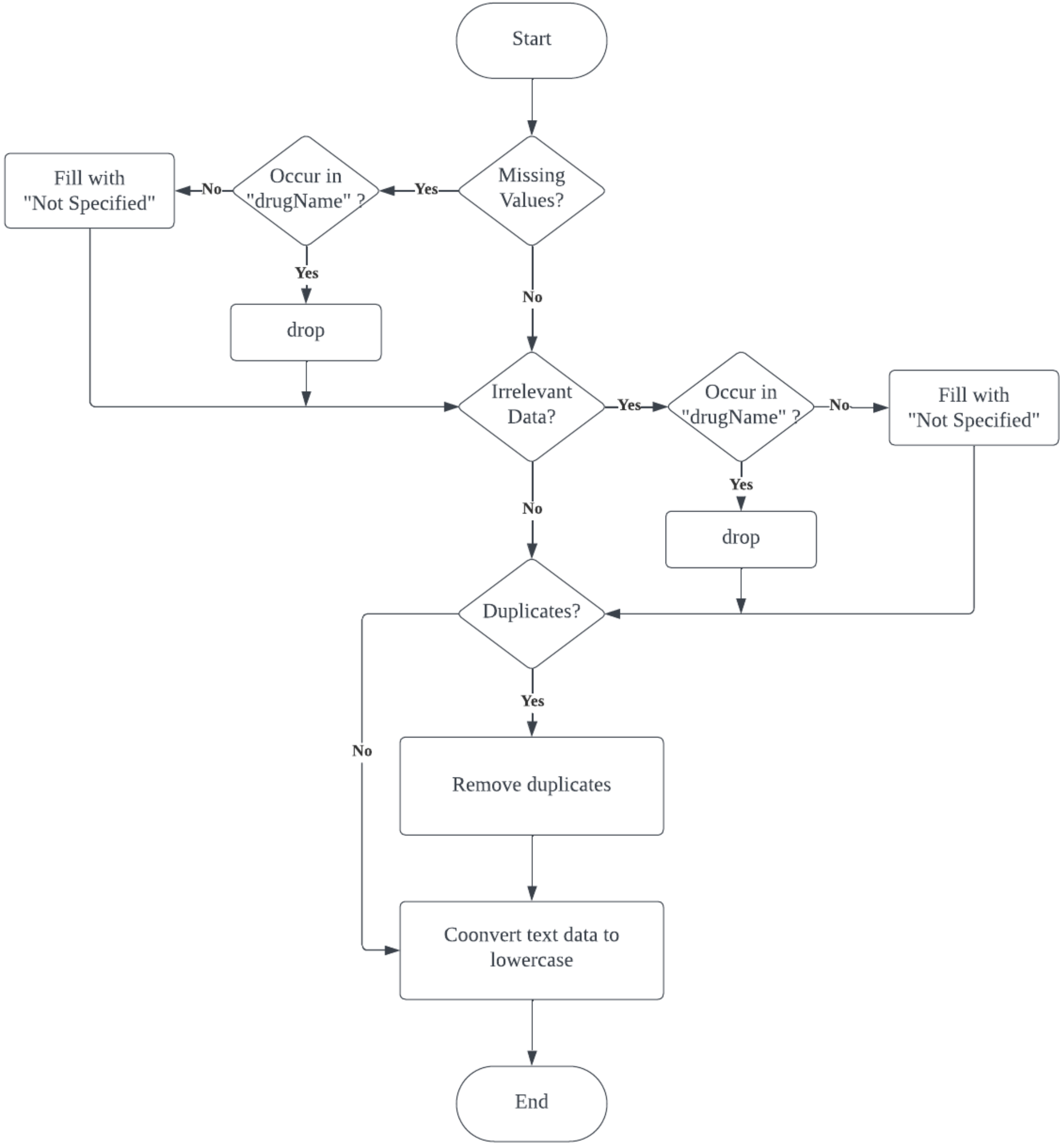
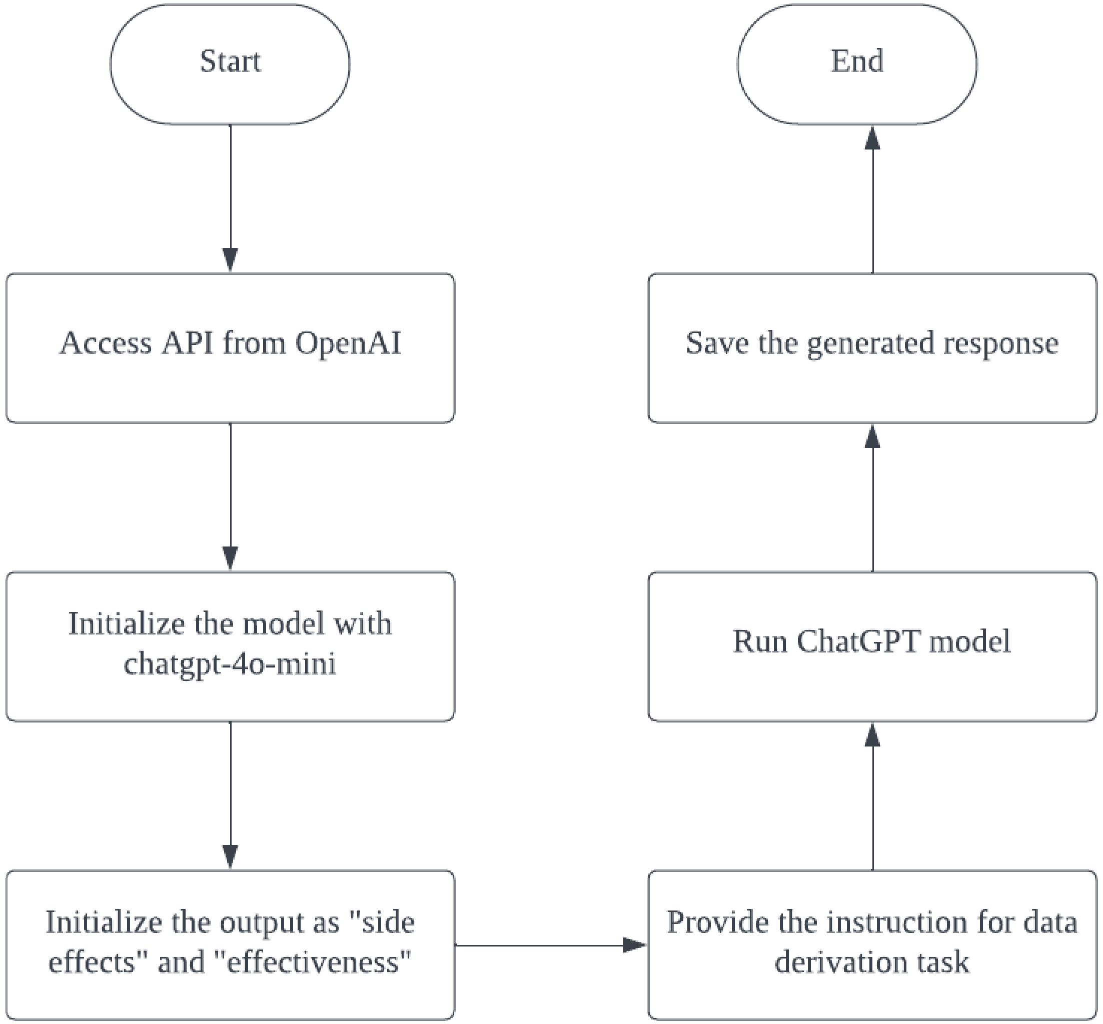


Figure 4.11: Outlier Detection with PCA

DATA PREPARATION FLOW



DATA DERIVATION FLOW



Start

Access API from OpenAI

Initialize the model with chatgpt-4o-mini

Initialize the output as "side effects" and "effectiveness"

Provide the instruction for data derivation task

Run ChatGPT model

Save the generated response

End

Output after Data Derivation

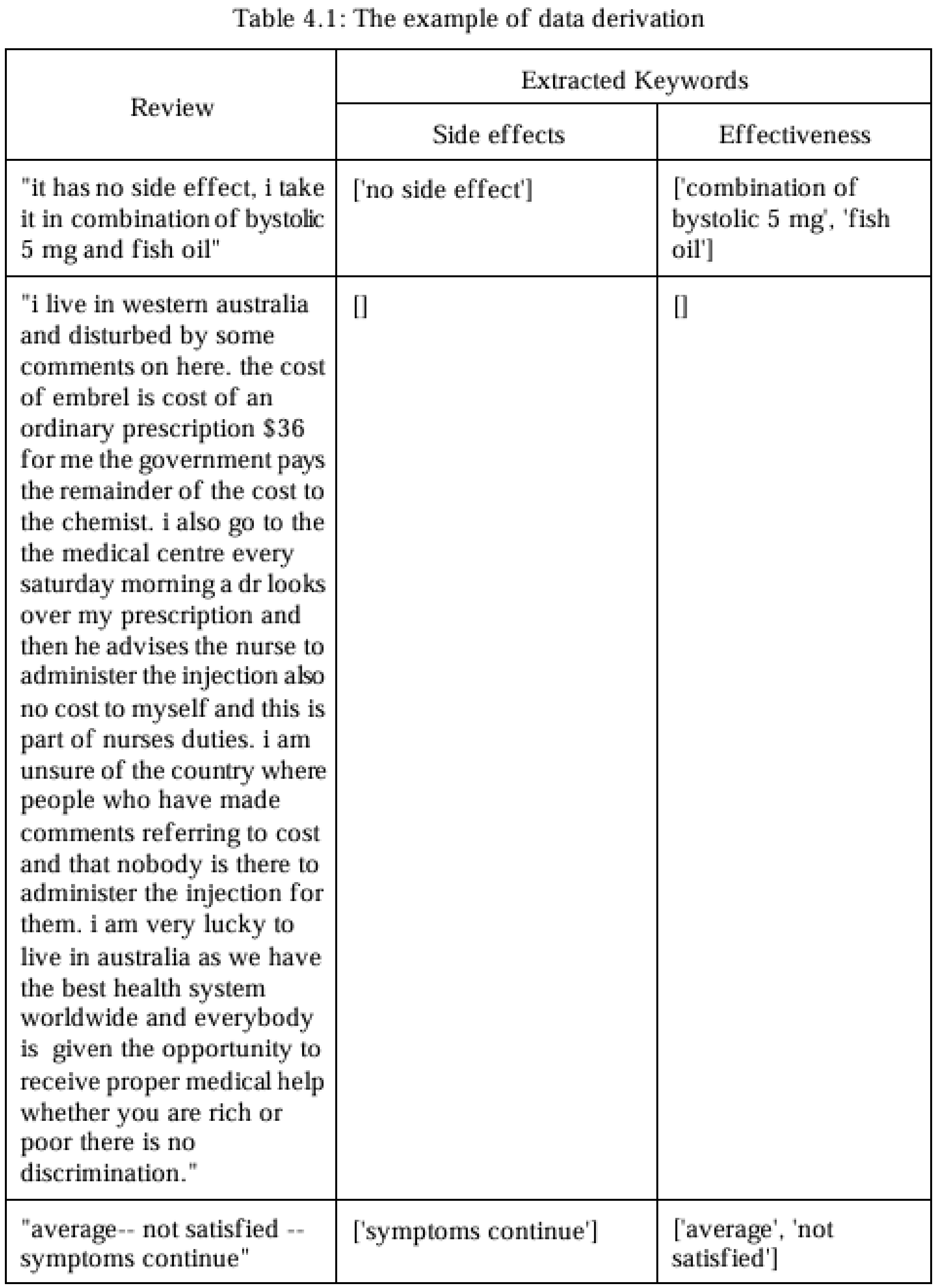


Table 4.1: The example of data derivation

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']

Discussion and Future Works

Achievements for this research are:

- A **cleaned dataset can be retrieved** by **removing the duplicates** and **handling the missing values** and **irrelevant information** that occurred in the dataset.
- **ChatGPT API was called to retrieve the features** that wish to analysis further.

Future works in MDS Project II are:

- **Successfully retrieved** the features of **side effects and effectiveness** from drug reviews.
- Clustering techniques such as **DBSCAN** were implemented and **formed different groups of clusters for the side effects and effectiveness**.
- Validation the clusters formed with silhouette coefficients.

THANK YOU

univteknologimalaysia utm.my utmofficial