IDENTIFYING PATTERNS IN DRUG EFFICACY USING LARGE LANGUAGE MODEL AND CLUSTERING TECHNIQUES FOR ANALYZING DRUG REVIEWS

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF THESIS

Author's full name	:	GUI YU XUAN			
Student's Matric No.	:	MCS241003	Academic Session	:	20242025-01
Date of Birth	:	7 APRIL 2000	UTM Email	:	guixuan@graduate.utm.my
Thesis Title	:	IDENTIFYING PATTERNS IN DRUG EFFICACY USING LARGE LANGUAGE MODEL AND CLUSTERING TECHNIQUE FOR ANALYZING DRUG REVIEWS			

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the thesis belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature :

Full Name : GUI YU XUAN
Date : 31 JANUARY 2025

Approved by Supervisor

Signature of Supervisor I:

Full Name of Supervisor I
DR CHAN WENG HOWE

Date : 31 JANUARY 2025

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

“I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of the degree of Master in Data Science”

Signature	:	_____
Name of Supervisor I	:	DR CHAN WENG HOWE
Date	:	31 JANUARY 2025

IDENTIFYING PATTERNS IN DRUG EFFICACY USING LARGE LANGUAGE MODEL AND CLUSTERING TECHNIQUE FOR ANALYZING DRUG REVIEWS

GUI YU XUAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master in Data Science

Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2025

DECLARATION

I declare that this thesis entitled *"Identifying Patterns in Drug Efficacy Using Large Language Models and Clustering Technique For Analyzing Drug Reviews"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	……
Name	:	GUI YU XUAN
Date	:	31 JANUARY 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Dr Chan Weng Howe, for encouragement, guidance, critics and friendship. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my master study. Librarians at UTM also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

The randomization process in randomized controlled trials (RCTs) restrict the involvement of diverse population with various health status and demographic information. Even though RCTs able to provide a reliable outcome of the drug performance, but it failed to represent the drug performance in real-world scenarios. The increase in the use of social media lead to the increase in the people to share their experience or opinion regarding the services and products that they had utilized on the Internet. Therefore, the drug efficacy from various perspectives can be derived from the drug reviews and act as the additional resources that enable the healthcare professionals in treatment planning. This project aims to identify patterns in drug efficacy to enhance the understanding of drug performance across diverse patient populations by utilizing large language models (LLMs) and clustering techniques in patient drug reviews. In this research, the drug reviews will be undergoing the data preparation process to handle the irrelevant information, duplicates and missing values that present in the dataset. Then, LLMs such as ChatGPT 4o mini will be called to run the data derivation process. The information of side effects and effectiveness of drugs will be identified from the reviews made by consumers. Then, clustering techniques such as density-based spatial clustering of applications with noise (DBSCAN) will be implemented to cluster the retrieved keywords. Lastly, the performance of the model will be evaluated by silhouette coefficients. The insights of clustering results will then be further visualized by dashboard. In conclusion, the used of LLMs and clustering techniques in this research tends to narrow the gap between RCTs and real-world scenarios. The relevant keywords obtained from the drug reviews allowed the healthcare professionals to have a better understanding of drug performance in a diverse population. Therefore, the treatment strategies can be further improved due to the valuable insights derived from the drug reviews.

ABSTRAK

Process percubaan terkawal rawak (RCT) telah mengehadkan penglibatan populasi dengan pelbagai status kesihatan dan maklumat demografi. Walaupun RCT memberikan hasil yang boleh dipercayai mengenai prestasi ubat, tetapi RCT gagal mewakili prestasi ubat dalam situasi kehidupan sebenar. Peningkat penggunaan media sosial telah menyebabkan penglibatan ramai individu untuk berkong pengalaman atau pendapatan mereka tentang perkhidmatan dan produk yang telah digunakan di Internet. Oleh itu, keberkesanan ubat dari pelbagai perspektif dapat diperolehi melalui ulasan ubat dan sebagai sumber tambahan yang dapat membantu kepakaran dalam menentukan rawatan. Projek ini bertujuan mengenal pasti corak dalam keberkesanan ubat untuk meningkatkan pemahaman tentang prestasi ubat merentasi populasi pesakit yang pelbagai dengan menggunakan model Bahasa besar (LLMs) dan teknik pengelompokan dalam ulasan ubat. Dalam kajian ini, ulasan ubat akan melalui proses penyediaan data untuk menyelakkan maklumat yang tidak relevan, pendua dan nilai yang hilang dalam set data. Seterusnya, LLMs seperti ChatGPT 4o mini akan digunakan untuk menjalankan proses derivasi data. Maklumat tentang kesan sampingan dan keberkesanan ubat akan didapatkan daripada ulasan ubat. Selepas itu, teknik pengelompokan seperti pengelompokan berbasis densitas (DBSCAN) akan dilaksanakan untuk klastering ciri-ciri yang telah diperolehi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RCTs	-	Randomized Controlled Trials
LLMs	-	Large Language Models
NLP	-	Natural Language Processing
DSPs	-	Disease Specific Programs
ANN	-	Artificial Neural Network
LDA	-	Latent Dirichlet Allocation
DBSCAN	-	Density-Based Spatial Clustering
EDA	-	Exploratory Data Analysis
BERT	-	Bidirectional Encoder Representations from Transformers
CNN	-	Convolutional Neural Network
LSTM	-	Long Short-Term Memory
T5	-	Text to Text Transfer Transformer
PEGASUS	-	Pre-training with Extracted Gap Sentences for Abstractive Summarization
TF-IDF	-	Term Frequency – Inverse Document Frequency
PCA	-	Principal Component Analysis

LIST OF SYMBOLS

D, d	-	Distance

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The increase in the use of social media was lead to the increase in the number of people relying on the reviews before making purchasing (Dinh et al., 2020). This is because people were able to share and comment their opinions that regarding to the products or services that they had experiences. These comments and opinions by other people helped the user to have an insight on the products or services. According to Qiu and Zhang (2024), there was 95 percent of people were read on the comments or reviews before purchase the products. Biswas et al. (2022) found that there was 270 percent of chance that the product will be brought if it has at least five reviews. Therefore, these studies demonstrated that customer review was important as they helped people to do the purchasing decision.

As the product reviews able to share the satisfaction to others, then drug reviews able to provide useful information about drug performance. The side effects and overall patient experience can be gained from the drug review and indirectly allowed healthcare professionals making a better treatment plan. The main part of drug review is ratings and text reviews. Ratings is a numerical data that showed the satisfaction of patient while text reviews are textual data that represent the overall experiences such as the drug effectiveness to the patient.

According to Sridharan and Sivaramakrishnan (2024), drug review able to enhance the therapy and reduce the error in medical field. Besides that, drug performance can be further enhanced by understanding the drug review completely (Liu et al., 2020). Additionally, understanding patients' medical conditions through drug reviews can help patients to choose a better medicine when medical advice is limited (Zeroual et al., 2020). Thus, analysis of drug reviews offers valuable insights

into drug effects and benefits that may not be fully addressed in clinical trials. Therefore, this thesis tends to apply Large Language Models and clustering techniques to analyse drug reviews and identify patterns in drug efficacy. By extracting meaningful insights from patient feedback, the project aims to enhance the understanding about the performance of drugs across various conditions.

1.2 Problem Background

Randomized controlled trials (RCTs) are usually considered as the gold standard in evaluating drug efficacy due to the monitored procedures that aim to eliminate bias (Hariton & Locascio, 2018). The randomization process in RCTs ensures that distribution of age, gender and health status are evenly distributed across different group of treatments. This strengthened the ability of experiment to determine that observed effects were due to the drug itself as the bias had been minimize. Hence, the controlled structured make RCTs reliable for the treatment outcomes. However, despite their quality, RCTs fail to provide a complete overview of a drug's effectiveness in the real world because limitations exist in its generalizability and applicability to the larger patient population.

RCTs frequently include strict eligibility requirements that exclude individuals with various health issues. As a result, the outcomes cannot fully generalize in a more diverse population (Kostis & Dobrzynski, 2020). RCTs results not accurately reflect real world situations of drug efficacy regarding long-term side effects and the improvement of symptoms across different patient groups. Hence, although RCTs offer valuable information about drug effectiveness in optimum conditions, it’s still have drawbacks in evaluating its efficacy in uncontrolled situations (Kaul et al., 2021). In this context, patient reviews are critical in bridging the gap between RCT findings and real-world scenarios.

RCTs were recognized as a benchmark for evaluating drug efficacy results under optimum conditions. However, the controlled environments and eligibility requirements that aligned with the RCTs exclude diverse patient populations with

various conditions. Even though this study did not involve with the RCTs, but the patient reviews had been utilized to address the limitations of RCTs result in real-world scenarios. The resulted outcomes of this research by focusing on the drug reviews reflected the real-world scenarios in drug experience. Therefore, instead of referencing the RCTs result in providing medicine suggestions to the patient, this research tends to provide an extra information of drug performance that allow the healthcare professionals in planning the treatment strategies.

Several studies have been conducted to address this gap in understanding. Kumar and Shekhar (2024) showed that the implementation of supervised learning and unsupervised learning in classifying the patient text into different health segment achieved high accuracy at 98 percent. The traditional text processing such as remove punctuations, stop words, covert to lower case and data cleaning was done on the review data before applying with k-means clustering method. The k-means clustering successful to group the review based on prevention, symptoms, treatment, news and other.

While traditional clinical trials give valuable information on drug performance in controlled conditions, they fail to concern on the wide range of patient experiences and long-term effects that may occur with actual use in real world. According to Gruber and Votta (2024), traditional text processing had the limitation in recognising the relationship between words which can lead to unreliable result or oversimplified the analysis. However, LLMs can explore the various aspects in reviews and understand the relationship between the words and sentences. Therefore, utilizing LLMs to identify the important features that presented in the reviews can further improved the interpretation of the drug efficacy. Clustering drug reviews able to discover underlying patterns of drug efficacy by analysing the personal experiences of consumers. Hence, by detecting patterns of reviews will help in clinical decision-making.

1.3 Problem Statement

As RCTs frequently conducted with controlled conditions and select the individuals with specific disease, hence limiting its generalizability for a more diverse patient population. Medical professionals lack the comprehensive knowledge about drug performance in various scenarios. However, patient drug reviews with the real-world experiences provide unreported side effects and varying efficacy results. The resulted outcome from this study by analysing the drug reviews tends to offer the extra information for healthcare professionals to make informed decisions about treatment plans. Thus, the problem statement of this study is the limited understanding of drug efficacy in real world scenarios restrict the ability to capture the side effects and effectiveness of drug consumed in a more diverse population.

1.4 Research Questions

This thesis aims to visualize the patterns in drug efficacy based on side effects and effectiveness by utilizing Large Language Models (LLMs) in retrieving the relevant keywords and clustering patient reviews based on the keywords. There are a few of steps to be carried out for visualizing the patient reviews pattern. Therefore, a few of research questions have been identified for the experiment.

The research questions are:

(a) What preprocessing steps to carry out for the analyzing of drug efficacy from drug reviews dataset?

(b) What relevant keywords can be identified and retrieved by LLMs from drug reviews dataset?

(c) What insights can be drawn from the drug review clusters by the retrieved keywords?

1.5 Research Objectives

The aim of the project is 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.

The objectives of the research are:

(a) To conduct data preparation on the drug reviews datasets for a cleaned dataset which essential for efficient drug efficacy analysis

(b) To retrieve relevant keywords which focus on the side effects and effectiveness from the unstructured data of cleaned dataset by using Large Language Models

(c) To analyse the clustering outcomes from retrieved relevant keywords by clustering technique for general drug performance pattern and display the findings through the dashboard

1.6 Scope

The scopes of the research are:

(a) The drug reviews dataset will be collected from UCI Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com)

(b) The programming languages used is Python

(c) Utilize pretrained language model to retrieved relevant keywords that are related to drug efficacy such as side effects and effectiveness

(d) Implement clustering technique to group the retrieved relevant keywords

(e) Method used for clustering quality evaluation is silhouette coefficient

(f) Power BI was used to construct dashboard

1.7 Significance of Research

The use of LLMs and clustering techniques in this research tends to narrow the gap of the understanding of drug efficacy in real-world scenarios. The relevant keywords in drug reviews can be analyzed by LLMs. Then, the meaningful pattems will be shown by applying clustering techniques to categorize the relevant keywords that had been identified by LLMs. The controlled conditions in RCTs had the limitation in representing the real-world scenarios in analyzing drug efficacy. In contrast, drug review provides insight from diverse populations who share their experiences regarding drug efficacy. As the characteristics of individual is hard to measure for study, but the involvement of diverse populations in reviews does help in identifying the patterns of drug efficacy in general.

LLM that is utilized in this study tends to retrieve the relevant keywords that are regarding to the side effects and effectiveness of drug which are essential for drug efficacy evaluation. The ability of LLM in understanding the complex language pattern further improves the process of handling the complex unstructured data and avoids the loss of information. The retrieved relevant keywords by LLM can ensure that the analysis consists of the information as much as possible.

Clustering technique that is implemented in this study tends to group the retrieved relevant keywords based on their similarities. The clustering process ensures the discovery of hidden patterns in reviews across different groups of populations. This technique helps to group the vast amounts of data into coherent themes which further enhance the investigation on drug efficacy evaluation.

In conclusion, the involvement of diverse populations in drug reviews allows healthcare professionals to gain comprehensive understanding on drug performance.

Instead of relying on the RCTs results to suggest treatment plans, the analysis of drug reviews can act as an additional resource for healthcare professionals. Thus, healthcare professionals can make better clinical decisions and choose better treatment plans for patients. Therefore, treatment plans and therapy strategies are improved due to the valuable insights that have been obtained by healthcare professionals.

1.8 Summary

This chapter illustrates the significance of analyzing drug reviews to enhance the understanding of drug efficacy. Even though RCTs is a standard method in identifying and approving the drug performance and the use of drug in real-world, drug reviews further improve the result in a more diverse population.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will discuss the related issues and the previous studies that have been done. The limitations of RCTs were determined and how the drug reviews can enhance the findings of RCTs was discussed in this chapter. Lastly, the use of LLMs and clustering techniques in text analysis were illustrated in this chapter.

2.2 Drug Efficacy Evaluation in Randomized Controlled Trials (RCTs)

RCTs are important for allowing the medicine to be used in the real-world (Liakos et al., 2024). Besides that, RCTs also provide the framework and structure to describe the policy of the medicine and the way to consume the medicine. RCTs can avoid bias in findings and can generate reliable results. This is because selected volunteers are randomly separated into different groups. Therefore, the distribution of people with certain characteristics can be evenly distributed to test the drug performance. The research done stated that RCTs are complex and expensive due to the strict rules to conduct them. These rules and standards aligned with the study of RCTs are to ensure there is no harm to the patient when the drug is in used. Thus, when RCTs are conducted correctly, then healthcare professionals can determine the safety of drug. Indirectly, healthcare professionals can make clinical decisions.

RCTs is a key indicator to analyze drug performance as it provides the guidance in defining the experiment for new treatment. Besides that, RCTs provides information of drug efficacy in optimum conditions highlight its ability to determine disease severity, drug usage and possible outcomes. However, the procedure of RCTs by selecting a group of people with requirements exclude the involvement of real

populations which limits its capability to make the assumptions on the possible outcomes when the drug is non-adherence consumed.

Therefore, instead of investigating the findings of RCTs, drug reviews dataset will be utilized for enhancing the understanding the drug efficacy based on real-world scenarios. The findings obtained from this project are believed to capture the general drug efficacy patterns that will be found in the diverse populations. Indirectly, the findings of the research can serve as extra information that can be referenced by the healthcare professionals to suggest suitable medicine.

2.3 Patient Review as A Real-World Data Source

Customers share their opinions about experienced drugs on internet review sites (Dinh et al., 2020). As a result, drugs reviews can be considered as statistical data that enable medical professionals in collecting medical data before making clinical decisions. This is because drug reviews that are commented on by patients provide insights on their experiences with medicine, including its efficacy and side effects. The Internet offered lots of information to enable the analysis on pattern recognition. Research can understand the pattern of the topics by analyzing the data on the Internet. The study stated that the valuable insights from multimodal data can be visualized by using machine learning algorithms. Thus, healthcare professionals can utilize the tools to categorize drug reviews based on their effectiveness and side effects. Online platforms allow all people to share and comment on their experience. The involvement of different populations with different health status and demographic information in online drug reviews are important to gain insight of a drug performance across diverse populations.

There was a sentiment analysis that had been done to investigate the patient's experiences by studying patient review from an online medical platform. According to the authors, patient reviews help to gain the understanding of drug when used in patients with different diseases. (Cimino et al., 2024). The authors utilized natural language processing (NLP) and machine learning algorithms to analyze patient

reviews. NLP was used to process and analyze text data. The patterns in reviews were identified and assigned text data according to their sentiment emotions. Then, the text data was classified by the support vector machines (SVM) and random forest to validate the performance of the model. In conclusion, the study highlighted drug reviews are important because they provide information that involved patients with different disease states. Sentiment analysis that categorizes patient experiences in positive, negative and neutral allow healthcare professionals to identify the drug efficacy effectively.

A study had been done to interpret the real-world data in the disease specific programs (DSPs) analysis. Real-world data allow the healthcare professionals to understand the disease management which help to make clinical decisions (Anderson et al., 2023). DSPs is a multi-perspective real-world data source that gathers the information from patients, caregivers and physicians into treatment patterns, patient reported outcomes and the patient experience. Real-world data allow the inclusion of diverse patient populations compared to traditional clinical trials. As mentioned before, real-world data captured the experiences and outcomes of patient which are important to analyze the effects of drugs. Besides that, the study also stated that the differences between groups of patients able to be identified with the real-world data and the results from analyzing process will be further enhancing the patient outcomes and quality of life.

Table 2.1 below summarizes the previous studies that had been done on exploring the use of real-world data in drug efficacy. The researchers did indicate that consumers have difficulty going through all comments due to the unstructured text data (Dinh et al., 2020). Therefore, to ensure that the drug reviews that done by patients able to be understandable by others and assist medical professionals in improving the performance and effectiveness of drugs, some models and algorithms will be carried out to classify the text data into meaningful insights. Text mining, supervised learning and sentiment analysis that implemented on drug reviews offered insightful information and reflected the experiences and opinions of diverse patient populations. Even though the machine learning approach that used to discuss the drug effects help

the people to understand the drug performance, they exist with limitations where the underlying patterns in drug reviews cannot be identified completely.

Table 2.1: Summary of The Study Done on Drug Review

References	Experiment	Strength	Limitation
(Dinh et al., 2020)	Utilize the text mining and supervised learning to discuss the online drug reviews	• Can considered as statistical data • Involvement of different populations in sharing their experiences	• A deeper analysis is necessary to improve the results
(Cimino et al., 2024)	Sentiment analysis on patient experiences	• Gain understanding of drug when used in different conditions	• NLP can't understand the meaning of words in different context
(Anderson et al., 2023)	Interpret the involvement of real-world data source with DSP	• Real-world data from DSP provide insight into disease management • Involve with patients, caregivers and physicians	• DSPs do not represent the real population • Include bias due to the eligibility of volunteers

2.4 Large Language Models (LLMs) in Text Analysis

With the growth of technology, there is an increasing number of textual datasets that have been available from digital sources. There is lots of information that can be obtained from social media posts to online review platforms. However, analyzing the vast amounts of unstructured data to discover the underlying patterns is a complex task because unstructured data do not have a standardized format that enables analysis in a simple way. In addressing these issues, LLMs were recognized for their effectiveness in classification, summarization and generation task. LLMs are advanced deep learning models that are pre-trained on large amounts of textual data to capture the complex language patterns (Ampel et al., 2024). The pre-training allowed

LLMs to perform well on a variety of downstream tasks. For your information, downstream task is a task that depends on previous output.

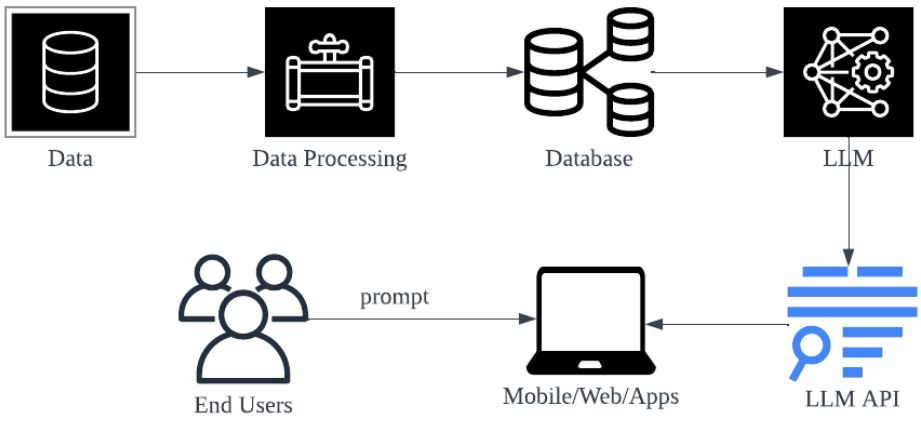


Figure 2.1: LLM Flow

According to Yao et al. (2024), LLM is a language model with a large number of parameters that has been pre-trained for tasks including self-supervised learning to produce and predict the text. The ability of LLMs that help in decision making and problem solving was due to the comprehensive understanding of natural language context, capability in producing human-like text, strong awareness of context and powerful problem-solving skills. ChatGPT, Gemini, Mixtral and Claude are examples of LLMs (Rangapur & Rangapur, 2024). The evaluation of accuracy, fluency and coherence of the generated responses by the LLMs model had been conducted by. By evaluating the potential of LLMs in performing the conversational question answering task, ChatGPT showed higher accuracy, relevance and consistency in generating the relevant response compared to others. Thus, the ability of ChatGPT in producing relevant and accurate responses makes it became the first option to select as a conversational AI.

LLMs consists of large computing system that take textual data as an input into the Artificial Neural Network (ANN) to transform data into numerical format (Tai et al., 2024). The ANN will become more powerful and able to produce more reliable information when there is lots of data been inputted into LLMs. LLMs such as ChatGPT which developed to learn from human feedback to conduct conversations

and solve mathematical problems. LLMs have been used in various fields which included writing clinical information about patients in medical field and summarizing text from academic paper in academic field. From the experiment done by the authors in enhancing the coding in qualitative research, they found that LLMs can perform the checking on codes and providing additional knowledge that help authors to understand the steps. With extensive training data for LLMs to learn the pattern, the accuracy of code identification and interpretation can be further improved.

ChatGPT is a new chatbot that developed by OpenAI to answer questions on various topics (Belal et al., 2023). GPT model can write code, generate phrases and sentences and perform arithmetic solutions. According to the study that had be done to analyze the use of GPT model for data labeling compared to lexicon-based approach such as VADER and TextBlob, it showed that GPT model able to perform better and achieved 20 percent and 25 percent higher accuracy than other lexicon-based unsupervised methods in Tweets dataset and Amazon Reviews Dataset respectively. The advantages of ChatGPT include user-friendly interface, easily accessible to non-experts in interpreting text data and the adaptability to perform various tasks. However, the results that produced by ChatGPT were dependent on the prompt used in the analysis and had potential bias. The bias was due to the training of ChatGPT with vast amounts of data that available on the internet.

The research in analyzing the sentiment analysis ability of GPT model had been conducted (Wang et al., 2023). Sentiment analysis is used to learn the expression patterns in the text. The authors use GPT model for the evaluation of language understanding ability because of its performance and low cost. The experiment was started by giving the instruction for each task and evaluate the performance by accuracy and F1 score. The findings illustrated that GPT model is highly competitive sentiment analysis performance and able to make a reliable prediction without labeled data for training. Meanwhile, a study on investigating the reliability and consistency of GPT model had been carried out (Reiss, 2023). This study was based on the ability of GPT model in classifying websites into News or not News. There are total of 234 websites that had been randomly selected, and the website texts were obtained to

transform into plain text. Krippendorff's Alpha was used to measure consistency by evaluating the output generated from the same input.

To ensure the consistency and reliability of the classification results, there are several scenarios that were introduced to GPT model. The scenarios included using various parameters such as temperature settings, changing the words in provided instruction and repeating the inputs multiple times. Even though there are advantages from GPT model, the experiment concluded that GPT model is non-deterministic and inconsistent in outputs. This is due to the temperature settings that had been assigned to control the randomness of generated output (Reiss, 2023). Lowering temperature settings will reduce the randomness of generated text and produce a deterministic output. The study also demonstrated that pooling output by obtaining the important features from previous features map can improve the reliability of GPT model.

Table 2.2: Summarization of Advantage and Disadvantage of GPT model

Advantageous of GPT Model	**Disadvantageous of GPT Model**
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)	

Table 2.2 summarizes the advantages and disadvantages of the GPT model. GPT model is an advanced language model that offers a few advantages that make it perform well in text related tasks. However, the drawbacks of GPT model highlight its

limitations in producing consistency and reliable results. Therefore, consideration should be given when implementing GPT model to perform the text related task.

2.5 Drug Review Processing for Machine Learning

Machine learning, especially deep learning models, has become one of the most used approaches in analyzing textual data such as online reviews (Jayapradha et al., 2024). The ability of LLMs such as bidirectional encoder representations from transformers (BERT) in capturing the complexity of language patterns outperformed traditional sentiment analysis models. The author implemented BERT to classify the drug reviews based on sentiment analysis to determine the recommendations of a drug. Drug review dataset had been undergone data cleaning process by removing duplicates, handling missing values and replacing irrelevant information. Besides that, the rating value had been transformed to binary level according to the rating value given by patient. Textual data had been preprocessed by BERT tokenizer with attention masks. Attention masks allow the model to determine the crucial parts of the text. Then, the features such as rating (binary level), review, drug name and condition (textual data after tokenization) were applied to the classification to make the prediction on the drug recommendations.

The effectiveness of deep learning models such as convolutional neural networks (CNN), long short-term memory (LSTM) networks and BERT model with bidirectional LSTM networks had been explored by analyzing the sentiment analysis of drug review (Colón-Ruiz & Segura-Bedmar, 2020). The main features used by the author are review and ratings to classify the reviews based on the sentiment. The sentiment of drug reviews had been categorized into three categories which are positive, negative and neutral. The sentiment of the reviews was identified by rating. The author converted the ratings into three categories where positive sentiment indicated higher ratings, negative sentiment indicates lower rating while neutral sentiment indicated middle rating scale. The review data had been tokenized and lemmatized by NLTK lemmatizer. Then, the tokenized text data had been converted into a matrix of word embeddings by using word2vec model. The comparison between

three models showed that BERT achieved the highest performance in classifying the review according to the sentiment.

A previous study had been done to explore the text summarization and sentiment analysis on drug review by transfer learning (Abuka, 2023). The author had compared the NLP techniques such as text to text transfer transformer (T5), pre-training with extracted gap sentences for abstractive summarization (PEGASUS), BERT and LSTM to evaluate the performance for sentiment analysis. The authors reduced the reviews by selecting 10 most useful reviews for each 500 drugs. The top 10 reviews with the most useful count of the reviews will be retrieved for further analysis. The retrieved review data had been converted into lowercase, removed the punctuation, special characters and stop words. Then, the cleaned reviews were combined to undergone tokenization. The rating in the dataset was performed to do the sentiment analysis by grouping them into three categories which are positive, negative and neutral according to their rating value. The features used in this study are review, ratings and useful count.

Most of the previous studies had done the classification of drug reviews based on the sentiment. The sentiment studies mainly utilize deep learning approaches such as CNN, LLM models such as BERT, T5 and PEGASUS and neural network techniques such as LSTM to perform the classification task. The rating in the dataset will serve as the guideline for authors to categorize the reviews in three sentiments. The classification task will classify the review (textual data) based on the sentiments. Therefore, to ensure the classification is accurate, data processing and feature selection were performed.

Previous studies used CNN, BERT, T5, PEGASUS and LSTM to perform feature extraction on reviews (textual data). These models have the advantage compared to the traditional text processing methods. The ability of these models in capturing the complex patterns in unstructured data enables the understanding of sentiments and relationships of words and context.

Table 2.3 below summarizes the data processing step and features used for the classification.

Table 2.3: Summarization of Features Processing

References	Experiment	Data Preprocessing	Features Used
(Jayapradha et al., 2024)	Classify the drug reviews based on sentiment analysis	• Removing duplicates • Handling missing values • Replacing irrelevant information • Rating transformed to binary • Review been tokenized with attention masks	Rating, Review, Drug Name and Condition
(Colón-Ruiz & Segura-Bedmar, 2020)	Analyzing the sentiment analysis of drug review	• Convert the ratings into three categories • Review data had been tokenized and lemmatized • Tokenized review data had been converted into a matrix of word embeddings	Review and Ratings
(Abuka, 2023)	Evaluate the performance for sentiment analysis	• Reduced the reviews by selecting 10 most useful reviews • Retrieved review data had been converted into lowercase, removed the punctuation, special characters and stop words • Tokenized the cleaned reviews • Rating converts into three sentiments	Review, Ratings and Useful Count

Even though the classification of drug reviews allows the people to understand the quality of drug consumed in certain conditions, the focus on sentiment of reviews unable to provide an insight into the drug efficacy. This research aims to identify patterns in drug efficacy to enhance the understanding of drug performance across diverse patient populations. Therefore, the classification of review based on sentiment exist with the limitations in identifying the drug efficacy in a more detailed view such as the effectiveness and side effects of drug.

GPT model used instead of using BERT, CNN and LSTM as the previous studies due to the ability of GPT model in understanding the entire content and analysing to identify the important features. The strength of GPT model in capturing the relationships between words and context does exclude the BERT, CNN and LSTM in data derivation process.

Clustering allows the discovery of the pattern without the needs of target data. By utilizing clustering techniques in the study, the similar characteristics of side effects and effectiveness can be identified. This further improved the healthcare professionals and patients in determining the drug consumed for certain conditions. Therefore, using clustering techniques instead of classification is due to the ability of clustering in offering the deeper insights into the diverse populations.

2.6 Text Vectorization Techniques

Text vectorization is a crucial step to enable the clustering on textual data. Textual data represent in the unstructured format make it unsuitable for undergo clustering directly. Therefore, text vectorization which convert the textual data into numerical data type allow the clustering methods to capture the meaningful patterns.

Sentiment classification had been studied to identify the performance and effectiveness of text vectorization techniques (Abubakar et al., 2022). Four text vectorization techniques such as bag of words, term frequency inverse document frequency (TF-IDF), Word2Vec and Doc2Vec had been examined in this previous

study. Author applied these text vectorization techniques to the textual data and summarized that TF-IDF performed the best for sentiment classification task. This result is due to the ability of TF-IDF in identifying the informative and unique words in the data. Bag of words had the potential to lose the contextual relationship as it focuses on the word occurrence only. Word2Vec had the limitation in capturing document level semantics while Doc2Vec with limited context always have a slightly drawback compared to TF-IDF.

A review on text vectorization techniques to the Arabic text classification had been carried out to determine the performance of classification algorithms with different text vectorization techniques (Sabri et al., 2022). There are total of three techniques used in this previous study which are word count, TF-IDF and Word2Vec. These three techniques transform textual data into numerical data type. Then, classification algorithms had been implemented to each of the dataset. The result illustrated that TF-IDF perform the best among another two techniques. The classification algorithm with TF-IDF able to result in higher accuracy. This is due to word count is a straightforward representation of text without the understanding of the relationship between words and the ability of Word2Vec in generating a low dimension vector that might not capture the real patterns in textual data. Meanwhile, TF-IDF focuses on calculating the frequency of words by capturing the importance of words and handling the common words.

The effectiveness of TF-IDF and bag of words had been discussed and examined with classification algorithms (Özdemir & Ortakcı, 2024). According to author, bag of words had the advantage of simplicity and disadvantages of treat the similar meaning of words as different features. Meanwhile, TF-IDF is the enhancement of bag of words. It calculates the frequency of words based on their occurrences and rarity in the data. The textual data had been applied with these two text vectorization techniques. Then, classification algorithms had been applied to the vectorized dataset. This previous study found that TF-IDF can perform better than bag of words with higher metric value. However, the implementation of artificial neural network with bag of words did further improve the classification performance.

As the clustering techniques cannot implement directly to the unstructured data, text vectorization is important to transform textual data into numerical data type. To ensure the clustering techniques in capturing the meaningful patterns and grouping the similar words, the text vectorization techniques used in this study must be carefully selected. Each of the text vectorization techniques had their own advantages and drawbacks. Previous studies examined the performance of different text vectorization techniques. Among the techniques, TF-IDF performed the best due to its ability to identify the important and unique of words. This ability of TF-IDF is important for textual data. When the meaningful patterns in textual data can be captured completely, the performance of clustering can be further improved. Indirectly, the clustering results can be reliable.

Table 2.4: Summarize the Use of Text Vectorization Techniques

References	**Experiment**	**Techniques Used**	**Best Performance**
(Abubakar et al., 2022)	Sentiment classification with different text vectorization techniques	• bag of words • TF-IDF • Word2Vec • Doc2Vec	TF-IDF: ability in identifying the informative and unique words in the data
(Sabri et al., 2022)	Examine performance of classification algorithms with different text vectorization techniques	• word count • TF-IDF • Word2Vec	TF-IDF: capturing the importance of words and handling the common words
(Özdemir & Ortakcı, 2024)	Effectiveness of text vectorization techniques in classification	• TF-IDF • bag of words	TF-IDF: enhancement of bag of words

Table 2.4 summarized the use of different text vectorization techniques. TF-IDF had been proven the ability in capturing the relationship between the words by calculating their importance and frequency from previous studies. Therefore, TF-IDF

will be utilized in this study to convert the textual data into numerical features for clustering purposes.

2.7 Clustering Techniques in Text Analysis

Clustering is a technique that groups the unlabelled data into different class without training and the grouping process was conducted by measuring the similarity between the features (Oyewole & Thopil, 2023). The training of clustering is by analysing the patterns and relationship between features in the dataset. Identification of patterns, measurement of similarities, grouping of data and the outcomes were the process in clustering algorithms. The authors suggested that pattern representation was referred as feature selection where only the useful information that will be recognized. The similarity between two data had been computed in clustering process to group the data into different groups. Furthermore, according to authors, optimum number of clusters was important and made impact on the output of data. In general, Euclidean distance was the most used methods to obtain the similarity between two data while sum of squared error and Silhouette index are the methods that had been used to obtain the optimum number of clusters. Today, clustering techniques had been used in several field including manufacturing, energy and healthcare. Clustering techniques in healthcare field assisted in identifying the diseases, understanding the patterns of data and predicting health issues.

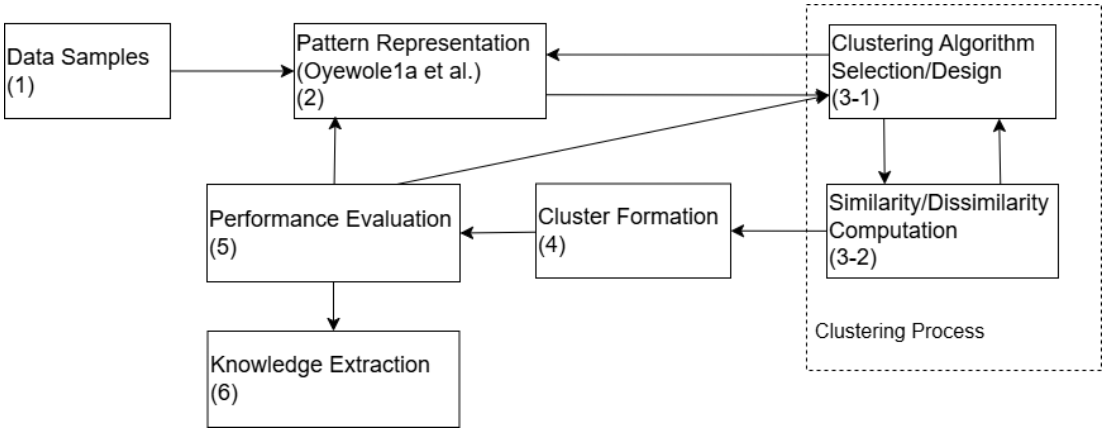


Figure 2.2: Clustering Steps (Oyewole & Thopil, 2023)

Clustering techniques was playing a crucial role in the data mining. This is because clustering techniques can discover the valuable information in the dataset (Hu et al., 2024). Clustering techniques partition data into different groups based on their characteristics. Clustering algorithms generate the clusters by understanding the relationship between data points. Clustering techniques can be interpreted in three ways, in-clustering, pre-clustering and post-clustering. Pre-clustering focused on the feature extraction and feature selection to ensure the capture of significant characteristics in the dataset. Meanwhile, in-clustering was illustrated the clusters with the selecting models that applied to the features. Lastly, post-clustering was the interpretation of the generated outcomes. The interpretation of in clustering and post-clustering was based on the applied models which are decision tree, rules, prototype, convex polyhedral and description. Decision tree model demonstrated the derived process from dataset into clusters along the path; rules-based model generated rules based on the features; prototype model utilized prototype as the representative of each clusters and group the data points if closely to the prototype; convex polyhedral model defined the boundaries planes to capture the cluster group while description model represented the key features as a description and grouped the features based on the specific concept. The authors believed that interpreting clusters were important to ensure the reliable and consistent result. Therefore, interpreting the generated clusters by understanding the context of models is crucial in the decision-making process.

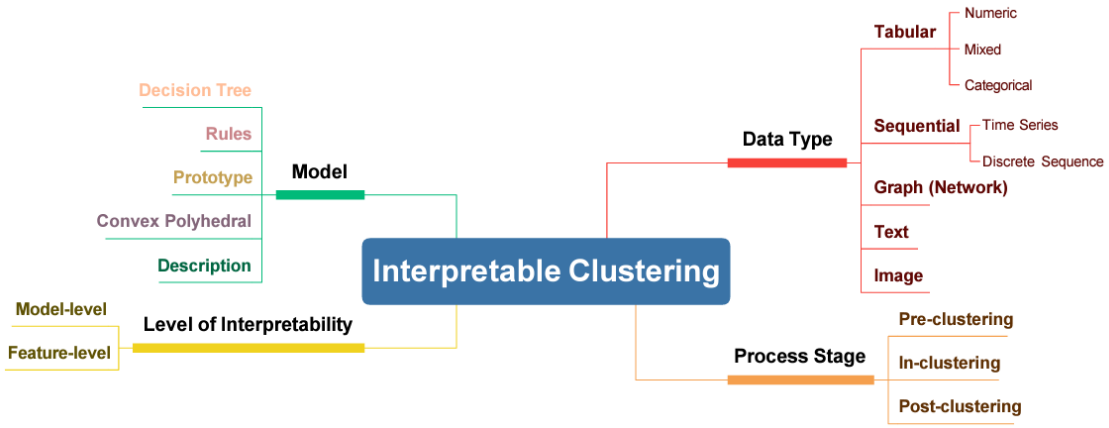


Figure 2.3: Criteria in Interpreting Clustering (Hu et al., 2024)

There is an experiment that utilized deep learning-based text clustering framework to analyse the accuracy and efficiency of text clustering. Clustering of text

is a method the grouping text data based on similarity and extracting the important features from unstructured data by classifying the similar text data into same categories (Xu et al., 2024). According to the author, the process of text clustering is mapping texts into a feature vector space and employ the clustering techniques to categorize texts based on their similarity. At the beginning of experiment, several steps had been carried out in the data preprocessing such as tokenization, stop words removal and text normalization. Second, pre-train models (or called as LLMs) were implemented to understand the pattern of information in the text data. Third, deep embedded clustering based on autoencoders was used to extract the meaningful features and apply clustering algorithm to cluster the data. The result showed that with the deep learning-based text clustering framework, the accuracy and efficiency of text clustering can be further improved and a more reliable results can be generated. According to the author, clustering the patient reviews able to classify patient according to diseases or help in analysing drugs performance. Thus, the clustering results able to assist medical professionals in the diagnosis and develop a new drug.

Besides that, there is another research had been studied to improve the drug repositioning performance. Drug repositioning is the investigation of existing drugs for new discovery strategy based on the analysing of clinical data (Lee et al., 2022). Authors highlighted that applying text mining approach in biomedicine field can analysed the large amounts of biomedical data effectively. Thus, the authors used the word2vec algorithm to generate embedded word vectors for the diseases and drugs to represent the relationship between diseases and drugs. Then, hierarchical clustering method had been applied to the word vectors to group the data based on their similarities. According to authors, the experiment successfully extracting the meaningful features from the dataset where there are 4,163 diseases and 3,930 drugs were extracted from 17,606,652 MEDLINE abstracts. Then, clustering techniques was grouping the extracted features into nine clusters. Therefore, the study that enabled the identification of potential drugs for discovery enhance drug selection process.

In conclusion, the ability of clustering techniques in discovering the underlying patterns of the text data and grouping the data based on their similarities enhancing the medical process. By grouping the similar reviews into a cluster, the experiences and

opinions of patient regarding to the drug can be visualized. The ability of clustering without relying on the independent variable to discover the performance allow the identification of hidden patterns in the dataset. Thus, clustering approach offer a more data driven decision process to make the treatment decisions. The popular clustering algorithms were discussed, and the suitable approach will be chosen for the analysis.

2.7.1 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-based clustering techniques have the ability in capturing the arbitrary shape of clusters (Hahsler et al., 2019). Thus, the data points will be grouped by this ability. The authors also suggested that noisy data will be excluded and would not group together with other data points. According to the author, density-based clustering started by defining density of the dataset. There is no predefined number of clusters needed in density-based clustering techniques. This is because density-based clustering techniques captured the clusters by density. Therefore, unlike other clustering techniques that required the predefined parameters, density-based clustering techniques assigned the data points according to the density. The commonly used density-based clustering technique is DBSCAN. DBSCAN identified all data points as core points, border points or noise data and clustered the core points by measuring the density. For your information, the algorithm started with assigning random data points as the central and defining the data points that were closer to the central point. The algorithm stops when there are no more data points linked as the density reachable points and a cluster will be formed.

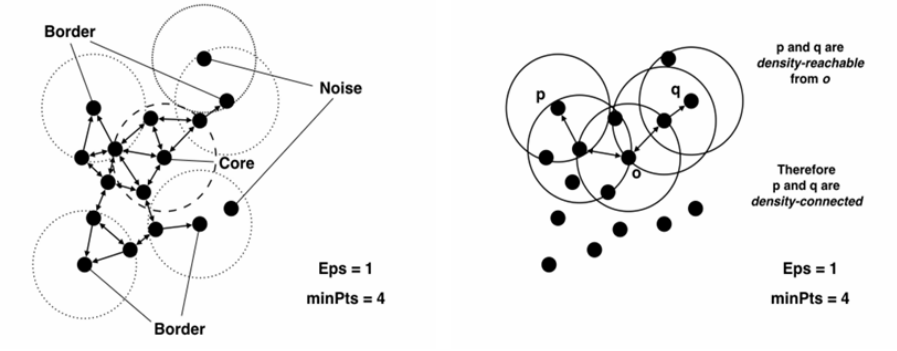


Figure 2.4: Concept of DBSCAN (Hahsler et al., 2019)

The figure 2.4 was further explained as below. There are two important parameters in DBSCAN, Ɛ (radius of neighbourhood) and minPts (minimum number of points to form a cluster) (Hahsler et al., 2019). Let considered there is a dataset to be clustered. The Ɛ-neighbourhood will be assigned with the value to identify to the data points within the radius of the assigned central point. The data points will be classified into core points, density reachable points and outliers. If data point had a distance with the minimum value of minPts will be considered as core point. Meanwhile, density reachable point referred as a data point that was reachable to the core point and is with the assigned radius. Lastly, the data point that does not meet the conditions of core points and density reachable points was clustered as outliers.

It was defined as:

$$N\epsilon(p) = [q \in D \mid d(p,q) < \epsilon] \quad (2.1)$$

Where:

$N\epsilon(p)$: set of points within the radius

$d(p,q)$: measurement of distance

D: dataset

The DBSCAN had the advantage in identifying clusters by effectively removing noise and outliers, do not require prior knowledge of the number of clusters and able to identify the clusters in various shapes and sizes (Bushra & Yi, 2021; Hahsler et al., 2019). However, the performance of DBSCAN depended on the parameters which can lead to misleading results when not specified the parameters correctly and the computational cost was high for distance measurement (Bhardwaj et al., 2022; Bushra & Yi, 2021; Ji & Wang, 2021). Therefore, a few steps on the selection of parameters should be considered to improve the clustering results and optimize the performance of DBSCAN.

2.7.2 Agglomerative Hierarchical

Agglomerative Hierarchical clustering is an unsupervised technique that build a binary merge tree that started to store the data into leaves and merge the two closest sets until reach the root of tree (Nielsen & Nielsen, 2016). Hierarchical clustering approach was introduced to have a large number of partitions and each partitions had its own dendrogram. (Murtagh & Contreras, 2017). Dendrogram is the graphical representation of the tree. The agglomerative hierarchical algorithm started by assigning each of the data points as a cluster. Then, for each iterative, the distance between two clusters was calculated and merged the closest pair of clusters to one cluster until single cluster was left. There are three strategies to define the good linkage distance which are single linkage, complete linkage and average linkage. Single linkage calculated the minimum distance between two data points, complete linkage calculated the maximum distance between two data points while average linkage calculates the average distance between all data points in two clusters.

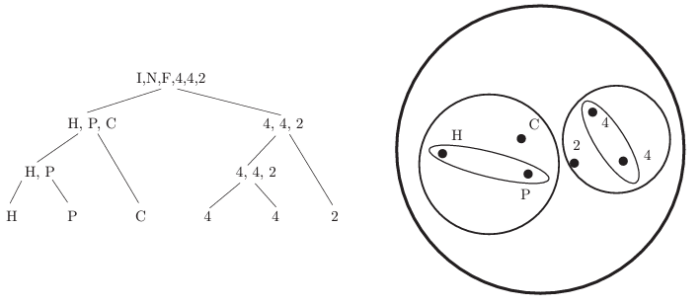


Figure 2.5: Dendrogram (left) and Venn Diagram (right) for Visualization (Nielsen & Nielsen, 2016)

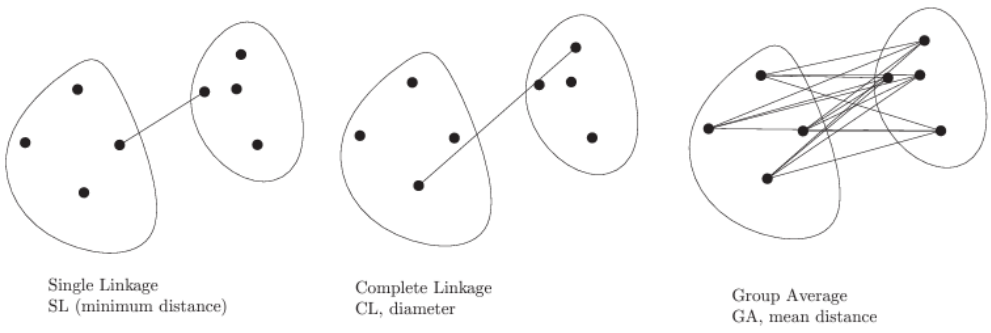


Figure 2.6: Linkage Strategies to Define Distance (Nielsen & Nielsen, 2016)

Single linkage defined as:

$$L(R,S) = \min\big(D(i,j)\big)\,, i \in D\,, j\ \in D] \tag{2.2}$$

Where:

$L(R,S)$: linkage between two cluster

$\min(D)$: minimum distance between data

$D(i,j)$: distance between two data points

Complete linkage defined as:

$$L(R,S) = \max\big(D(i,j)\big)\,, i \in D\,, j\ \in D] \tag{2.3}$$

Where:

$\max(D)$: maximum distance between data

Average linkage defined as:

$$L(R,S) = \frac{1}{n_R \times n_S} \sum_{i=1,j=1}^{n_R,n_S} D(i,j)\ \,, i \in R\,, j\ \in S \tag{2.4}$$

Where:

$\sum_{i=1,j=1}^{n_R,n_S} D(i,j)$: sum distance of clusters

Agglomerative hierarchical clustering techniques offered several advantages than other clustering algorithms. First, the dendrogram provided graphical representation of the hierarchical structure of data allow the understanding of relationships between clusters (Oti & Olusola, 2024). This is because the graphical allow the researchers to gain the insights into data at various levels. Besides that, agglomerative hierarchical clustering also robust to noise and outliers (Benatti & Costa, 2024). This characteristic allowed the agglomerative hierarchical clustering to perform in high dimensional dataset. Lastly, as agglomerative hierarchical clustering did not require a predefined number of clusters, thus it was flexible in clustering data as the clusters were formed naturally without controlling. However, this situation did rise the issues in identifying the clusters with different densities. Besides that, agglomerative

hierarchical clustering was computational complexity because the distance between all data points needed to be calculated.

2.7.3 K Means

K Means is partitional clustering algorithm that partitions dataset into smaller groups based on the distance between the centroid point (Ikotun et al., 2023). With the increasing number of clusters, K means algorithm able to achieve the decreasing in the square error. The minimum squared error between data points and the mean of the cluster will be found and assigned the data points to the nearest cluster. The step in K Means algorithm began by randomly selecting a few centroids from dataset. Then, the distance of data points with centroids will be calculated and assigned the data points to the nearest centroids. Lastly, the new centroid value was calculated for the next iteration. There were three parameters that should be considered in the K means algorithm which are the number of clusters to be formed, the centroid points and the distance metric to be used in the experiment. This is because the performance of the clustering depends on the number of clusters while different initial centroids can produce different resulted clusters.

K Means algorithm defined as:

$$D(C_k) = \sum |x_i - \mu_k|^2 \tag{2.5}$$

Where:

C_k: data points of Cluster k

$\sum |x_i - \mu_k|^2$: distance of data points and centroids

According to Chong (2021), K means clustering was straightforward algorithm that enabled non expert users to partition dataset into the desired number of clusters. The implementation and interpretation of K means approach was easy and widely used for clustering tasks (Liu, 2022; Pratama et al., 2023). Furthermore, the scalability and flexibility of K means algorithms enabled it to perform well in large dataset and work with various type of data such as numerical data and categorical data. However, K

means algorithm was sensitive and needed to be carried out carefully at the initial stage. This is because the performance of the K means algorithm was determined by the number of generated clusters, the initial centroids and the outliers or noisy data that was presented in the dataset.

Table 2.5: 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

Table 2.5 summarizes the advantages and disadvantages of clustering approaches. Each of the clustering approaches had its own advantages and disadvantages. In this research, the unstructured data in the dataset will be further investigated to identify the most suitable clustering techniques to be implemented for a better result.

2.8 Clustering on Related Studies with Similar Dataset

There was a study done on applying the k-means clustering algorithm to develop drug recommender system (Posch & Tiwari, 2021). The authors believed that drug recommender system can assist healthcare professionals to select the most suitable drugs based on the factors such as conditions. The dataset used in this previous study is the drug reviews dataset (similar dataset as this research). Text processing such as removing stop words and lemmatization had been done to remove noise and standard the reviews (textual data). Then TF-IDF was applied to convert the text data into numerical data. Then, principal component analysis (PCA) was applied to reduce the dimension of dataset. K-means clustering had been implemented to identify the personas based on the reviews. The personas referred to the groups of patients who had similar side effects and effectiveness with the same drugs under same conditions. The outcome of this previous study was to provide personalized drug recommendations by identifying the personas of patients. However, the clustering approach in this previous study was grouping the group of patients with similar side effects and effectiveness. This outcome limits the generalizability of identifying the side effects and effectiveness across diverse patients as the k-means clustering had assume the clusters result are spherical and in equal size which limits the capturing of the uneven and non-linear result.

A method that combine of unsupervised learning and supervised learning was carried out to investigate the impact of drug reviews in predicting medical preferences (Allenbrand, 2024). Analyzing patient review had the benefit in understanding patient satisfaction and addressing medication non-adherence. Therefore, this previous study had been carried out to improve the healthcare outcome. The raw data of reviews had been processed and extracted to consider only meaningful information for supervised and unsupervised learning. The methods of feature extraction were started by bag-of words and associated with n-grams to identify the relationship between words and context. Then, the n-grams results and reviews were used as the input data for the formation of matrix data by document term matrix. The output of document term matrix was served as the input data for the supervised and unsupervised learning. Then, topic modeling with Latent Dirichlet Allocation (LDA) was conducted to identify the

topics for each review by considering the TF-IDF results that had been calculated for the weight of the words.

After that, clustering methods such as k-means, agglomerative and DBSCAN were utilized to cluster the review based on identified topics by LDA. The elbow and silhouette method had been used for the determination of optimum clusters. After obtaining the clustering results, each review was assigned with their corresponding cluster. The rating in the dataset was used as the target for the classification task by categorizing the rating value into positive, negative and neutral categories. Lastly, the classification process was carried out to classify the reviews by ratings with support vector machine (SVM), random forest (RF) and logistic regression (LR). Confusion matrix had been used to evaluate the performance in predicting the rating based on the clusters. According to the author, clustering techniques and topic modeling were implemented to further enhance the performance of supervised learning in distinguishing the pattern of reviews. The combination supervised and unsupervised learning allow the prediction of patient satisfaction with the drug consumed as the experiment showed that the benefits and side effects found from the feedback was essential to develop the personalized medicine. However, the feature extraction by using bag of words, n-grams and TF-IDF had the limitation in capturing the patterns in reviews.

Breast cancer is the most common cancer among woman in worldwide (Nilashi et al., 2024). Patient generated data such as reviews on the drug consumed to cure breast cancer allow the enhancement of patient decision making as the reviews provide with the detailed opinion of the patient into the side effects and effectiveness. Therefore, clustering, forward entry regression analysis and LDA had been utilized in the study to discover the knowledge on breast cancer drugs by drug reviews. LDA was used in this study to obtain side effects aspects of patient experiences. Then, clustering techniques were used to group the reviews based on overall satisfaction. Lastly, forward entry regression analysis was implemented to predict the possible outcome by examining the relationship between side effects and effectiveness. In conclusion, LDA allows the identification of key topics of reviews and eight side effects have been successfully identified. The Gaussian mixture model clustering group reviews three

clusters by examining their similarities in side effects. Forward entry regression analysis is used to predict the effectiveness of the drug. The limitations in this study are the author claimed that patient reviews can be subjective, and the implementation of advanced machine learning techniques in the interpretation of reviews can further identify the underlying patterns in patient reviews.

Besides that, there was a study on applying clustering to the online patient medication reviews to discover the underlying knowledge of patient information and side effects (Yildirim & Kaya, 2019). The authors claimed that the risk of obesity had been increased significantly. Therefore, the analysis of medication for obesity had been crucial to inform healthcare professionals in managing obesity. In this study, author discussed the side effects of using the drugs for obesity through patient reviews. The patient reviews were collected from medical websites as the reviews offer diverse perspectives for healthcare professionals to improve the decision making. K-means clustering approach had been utilized to discover the patterns of side effects and satisfaction of patient into the drugs. In this study, clustering was implemented to examine the relationship between side effects and patient demographic. The results obtained from this research showed the potential side effects on different demographic patient groups. For example, the younger females had larger chances to experience the side effects such as dry mouth when consuming the drugs. However, the author did not justify the number of clusters chosen as varying number of k can produce different clustering outcomes.

The findings from previous studies illustrated the clustering in discovering the underlying insights of side effects and effectiveness of drug by analyzing drug reviews. Grouping of similar side effects and effectiveness based on the drug review allow the understanding of the groups of people who share similar experience with specific drugs and conditions. Categorizing drug reviews based on the topics that related to side effects and effectiveness, discovering the relationship between demographic and side effects of drug and predicting the ratings based on the patient experiences clusters did demonstrated the benefit to understand the relationship between drug efficacy and patient satisfaction. However, there are considerations to be carefully considered due to the limitations on the discovering of drug performance.

Table 2.6: Summarize Clustering Approach Done of Similar Dataset

References	Experiment	Clustering Process	Limitations
(Posch & Tiwari, 2021)	Applying the k-means clustering algorithm to develop drug recommender system	• Text processing • TF-IDF • PCA • Clustering – K-means	• Generalizability of K-means that assume the input data is in linear relationship
(Allenbrand, 2024)	Predicting medical preferences with the combination of unsupervised and supervised learning	• Text processing • Feature extraction by bag of words, n-grams and DTM • LDA • Clustering – DBSCAN, K-means, Agglomerative • Classification – SVM, RF, LR	• Feature extraction by using bag of words, n-grams and TF-IDF had the limitation in capturing the patterns in reviews
(Nilashi et al., 2024)	Discover the knowledge on breast cancer drugs by drug reviews	• Text processing • LDA • Clustering – Gaussian Mixture Model • Classification – Forward Entry Regression Analysis	• Patient reviews can be subjective • Traditional text processing limits the capture of complex pattern in textual data
(Yildirim & Kaya, 2019)	Applying clustering to discover the underlying knowledge of patient information and side effects	• Do not specify how processing textual data • Clustering – K-means	• Did not specified the limitations • Did not justified the number of optimum clusters chosen

Table 2.6 summarizes the explore of clustering approaches in identifying the patterns in drug performance. Different clustering approaches had been used for the examination of drug efficacy in terms of side effects and effectiveness. Previous studies also illustrated the text processing and feature extraction that had been carried out for improving the accuracy of clustering results. However, the limitations still existed during the feature extraction and clustering process. As textual data often contain complex patterns, previous studies faced with the challenges in capturing the accurate drug performance especially when using bag of words, n-grams and TF-IDF to capture the meaningful pattern from the raw reviews data directly. Advanced natural language processing methods such as pre-trained language model was used in this research rather than traditional text processing methods such as bag of words, n-grams and LDA to visualize the pattern in the textual data. Besides that, clustering approach will be implemented in this research also due to the ability in identifying the non-linear relationship between the data points.

2.9 Research Gap

Previous studies utilized traditional text processing such as tokenization, lemmatization and transformation of text to identify the important features for clustering. BERT, LSTM and CNN were used to capture the pattern in reviews. Before the implementation of these models, textual data must undergo data cleaning and standardizing process to ensure the important features and underlying patterns of unstructured data can be analyzed correctly. These three models had the advantage in detecting the important parts of text. Therefore, the implementation of these models allows the feature extraction process to be efficiently carried out. By utilizing these models, the important features in reviews can be extracted and further improved the clustering and classification task.

Instead of using traditional text processing and previous studies model in data extraction, LLMs such as pre-trained language model had been utilized to retrieve the side effects and effectiveness of drug from reviews. This is due to the ability of pre-trained language model in understanding the entire context of reviews. The capability

in identifying the complex relationship between unstructured data and the capability in handling raw data directly prevents the loss of information and enables the clustering approach performed on the detailed dataset. Pre-trained language model that can understand the relationship between words allows it to identify the features regarding side effects and effectiveness completely from the reviews. Therefore, the information loss during the feature extraction can be further minimized. Indirectly, the clustering model can perform more accurately than the models used in previous studies.

Previous studies focused on the sentiment classification by categorizing drug reviews into three sentiments. The clustering methods used in previous studies are mainly for grouping and understanding the reviews with similar structured and patterns. The clustering results helped the author to identify the diversity and variation in patient reviews and reduce the dimensions of reviews by focusing on the important features that are present in the dataset. Sentiment classification implemented after the clustering approach is to provide a detailed insight into patient satisfaction of drug performance. By predicting the sentiment on reviews, healthcare professionals can monitor the drug performance and identify whether the drug can be consumed by the patient with specific cluster.

The use of clustering techniques in previous studies was different from this study. The clustering technique used in this study tends to group the reviews based on their similarities in side effects and effectiveness. The clustering technique used in this study determines the insights into side effects and effectiveness which influence in the drug performance rather than act as a support tool to predict the drug performance. Previous studies that focused on the sentiment classification limited the understanding of drug performance. However, the clustering results based on the side effects and effectiveness of drug in this study allow the deeper understanding of drug efficacy in a diverse population generally as the patterns in side effects and effectiveness can be identified.

2.10 Discussion

RCTs were conducted with a controlled environment by selecting volunteers with specific demographics limiting the involvement of large populations. Though it can evaluate the drug efficacy without bias but exist with the drug performance gap in real-world scenarios. Patient reviews offer valuable information that captures the experience of patients when consuming the drug. Patient reviews can be acted as the additional resources that aligned with the RCTs to help healthcare professionals in conducting treatment plans. From the drug review, the side effects and effectiveness of the drug when applied to different groups of people can be derived from the patient review. However, bias can occur in the review based on the patient's preferences. To address this issue, text analysis methods such as LLMs were used to retrieve reliable insights from patient reviews.

LLMs such as pre-trained language model provide with the techniques that able to understand the sentiment of the words, phrases and sentences. The ability of LLM in analyzing the variations of text data allows the identification of important features in the reviews. Besides that, LLM can perform better than traditional text processing such as LDA due to the ability of LLM in understanding the context and meaning of the reviews. However, LLMs have limitations in terms of identifying the similarities between words. Indirectly, it limits the grouping of similar side effects and effectiveness.

Therefore, clustering techniques had been carried out to enhance this study. The grouping of similar features from reviews into different groups enhanced the overall findings of the results. As it is hard for human to directly identify the information regarding to side effects and effectiveness of the drug from the textual data, therefore the clustering approach that implemented in this study helps to reduce the complexity of textual data and allow the individual to visualize the drug efficacy immediately.

In conclusion, the combination of LLMs and clustering techniques in this research allows the evaluation of drug efficacy in a more comprehensive way. LLMs

with the ability to recognize and analyze the relationship between the characters, words and sentences help in retrieving the meaningful insights of drug review. Meanwhile, clustering techniques with the ability to group similar data points reduce the complexity of the dataset and enable the identification of patterns effectively.

2.11 Summary

As discussed before, RCTs still had limitations in considering diverse patient population in analyzing the drug performance. Thus, drug reviews generated by patients provide valuable information into the patient experience and side effects that can support the clinical results of RCTs. Most of the previous studies implement the traditional text processing such as LDA to identify the relevant features in drug review. However, with the advancement in technology, LLMs can provide a more comprehensive understanding of the drug reviews pattern. Therefore, instead of using traditional text processing, the experiment started with applying LLM, pre-trained language model to retrieve the keyword from the drug reviews. Then, clustering technique will be implemented to group the keywords due to its ability in handling vary shapes and densities of clusters, able to remove outliers effectively and no predefined number of clusters was required.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter illustrated the overall framework in conducting the research for identifying patterns in drug efficacy by analysing drug reviews through a clustering approach. The research process from initial study of topic to model evaluation will be further discussed. The utilized dataset and performance measurement will be identified and demonstrated in this chapter.

3.2 Research Framework

There were five phases of research to identify the drug efficacy. Each phase contributed to a milestone. Phase one is research planning and initial study which contributed to problem formulation and background research. A milestone of an overview of point of interest can be identified and enable an insight into the whole project. Besides that, data preparation fell into phase two in which a cleaned dataset that was ready for further analyzation was well-prepared. Furthermore, phase three is to retrieve the relevant keywords from the cleaned dataset. In this phase, the underlying pattern of the unstructured data in dataset can be identified by LLMs. Additionally, DBSCAN clustering model will be implemented into the retrieved relevant keywords to group the data based on their similarities. In this case, a milestone of drug categories on their effectiveness can be illustrated. Lastly, silhouette coefficient was applied for model evaluation. The relationship between drugs and its performance can be visualized in phase 5.

Figure 3.1 illustrates the overall research framework. Each phase will be discussed in detail in this chapter.

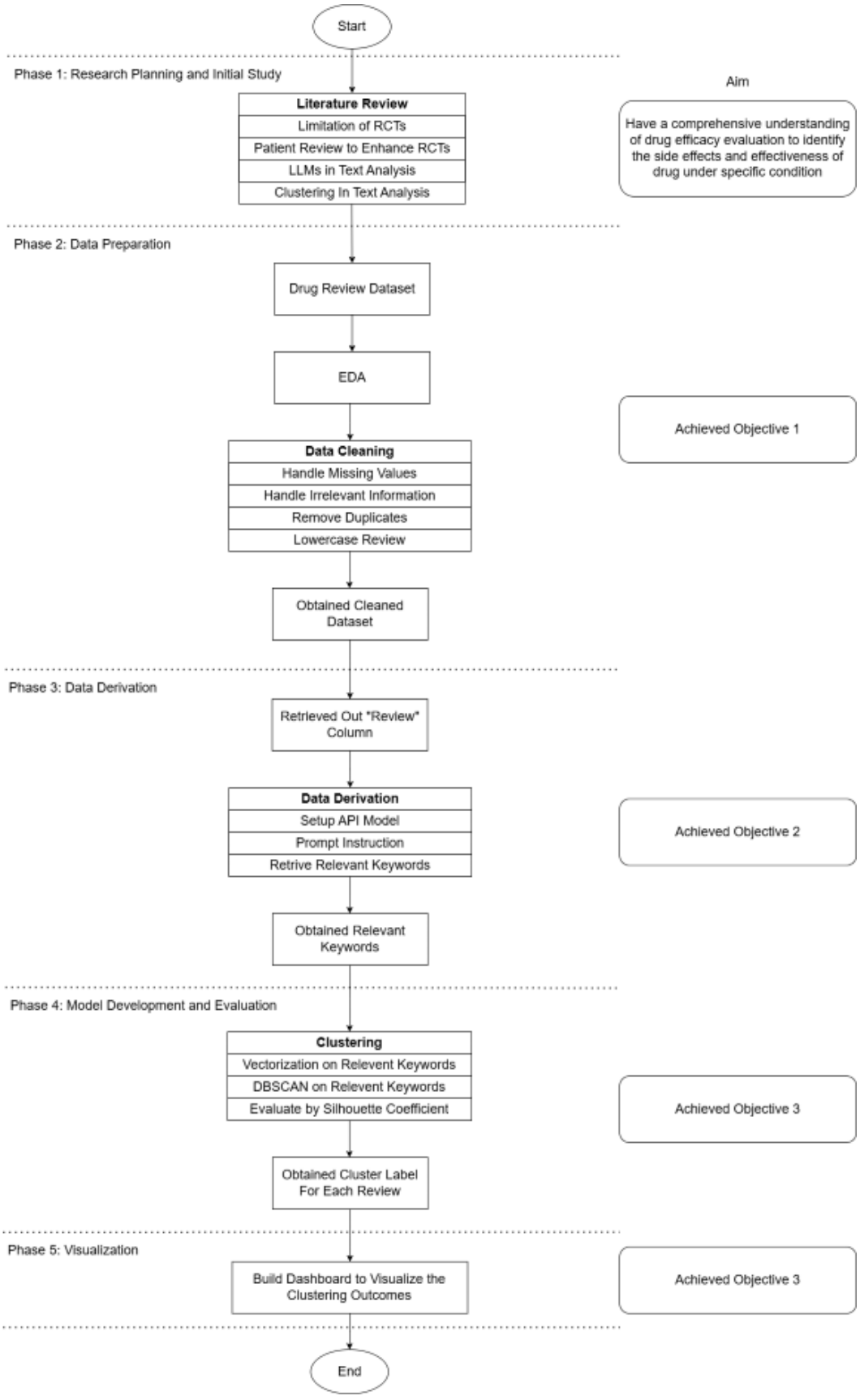


Figure 3.1: Overall Research Framework

3.2.1 Phase 1: Research Planning and Initial Study

This phase was the foundation of conducting the research study. In this phase, there are three parts that will be focused on to ensure that an overall understanding of the interesting topic. The first part was background study. Literature reviews on related topics had been conducted to gain an understanding of the research domain. For example, the previous experiments and theories that had been used for the analysis on the topics were explored to identify the historical development and current issues of the field. In this way, the current issues were identified, and the point of interest was able to be formulated.

Even though RCTs provided the record of drug performance, RCTs had the limitation in generalization comprehensive insight into the drug efficacy across diverse populations. This is because there were requirements that limit everyone to be involved in the RCTs. Therefore, conducting text analysis on drug review was helpful to determine the specific effectiveness as it involved different patient groups with a wide range of health issues and characteristics. Besides that, applying machine learning to the dataset enabled the author to learn the relationship among data points. In this way, the underlying patterns of data had been discovered and indirectly facilitated the grouping and categorizing process for further analysis.

3.2.2 Phase 2: Data Preparation

Data preparation is crucial to ensure the dataset is reliable and consistent. Before preprocessing the dataset for further analysis, exploratory data analysis (EDA) was carried out to understand the pattern of the dataset. Visualization such as scatter plot and bar chart will be used to determine the insights into the data structure. Therefore, anomalies and trends of dataset can be understood.

There were several steps to enhance the quality of the dataset which are data cleaning and text processing. At the data cleaning stage, missing data and duplicates that appeared in the dataset were identified and addressed. To handle the missing data

and duplicates, there are a few considerations needed to consider. For example, the type of missing data should be defined before implementing the cleaning process. This is because missing data either can be removed directly if it was irrelevant for further analysis or replaced with other values when it was carried the important features in the dataset.

Besides that, text preprocessing was required to prepare the textual data for analysis. In this process, converting the text data into lowercase to eliminate the sensitivity and allow the pre-trained language model to focus on the meaningful words.

3.2.3 Phase 3: Data Derivation

Data derivation process will be carried out using LLM to retrieve meaningful insights from the preprocessed data. The steps in this phase are setting up API, providing prompt instruction and processing API response. ChatGPT 4o mini was chosen as the LLM to integrate with the system. Then, an API key will be obtained from OpenAI to establish the connection. The instructions will be provided to the LLM for the operations. The instructions will outline the task that the ChatGPT 4o mini needed to perform such as *"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'."*. Lastly, the generated responses by the API were captured and processed for clustering approach.

To allow the retrieval of relevant keywords that regarding to side effects and effectiveness of drug, the "Review" column in the dataset will be used to undergo data derivation process. OpenAI provides the platform for individuals to use the services. Thus, the API key was obtained from the OpenAI platform for setting up the API model. The API setup for the GPT model involves the model that will be used in processing and generating the instruction given, the interaction in which the model to generate the response to a readable format and the temperature setting that control the

deterministic of generated output. Then, instructions were provided to allow the model to generate the relevant keywords that wish to analyze further.

3.2.4 Phase 4: Model Development and Evaluation

The clustering technique will be applied to identify the clusters on the retrieved keywords. Each of the clustering techniques that had been discussed in previous studies had their advantages and limitations. The performance of the clustering process can be affected by the pre-defined parameters. Therefore, the investigation of the unstructured data will be made to identify the most suitable clustering technique to be implemented in this research.

The following task was evaluating the effectiveness of clustering model. After retrieving a set of cluster labels from clustering technique, silhouette coefficient will be applied to evaluate the quality of clustering. Cluster label indicates which cluster that the data point belongs to, and each label will have unique value. The overall performance of the clustering approach was interpreted.

The relevant keywords that were obtained in phase 3 were in textual format. Unstructured data unable to directly undergo the clustering process. Hence, vectorization will be carried out to transform the text into numerical format. As TF-IDF has been proven effective in capturing the relationship between words by calculating their importance and frequency, therefore it will be used in this study to ensure that the text data can be treated appropriately for clustering process.

3.2.5 Phase 5: Visualization

After obtaining the cluster label of each review for side effects and effectiveness, dashboard is conducted to visualize the clustering outcome. This is because dashboard provided with the tool that enables the representation of data for

understanding the patterns in dataset. For example, the relationship between side effects and effectiveness to the drug and condition can be determined. Therefore, the stakeholder can utilize the dashboard for exploring the data and making informed decisions.

Figure 3.2 illustrates the example of visualization in dashboards. Healthcare professionals can use the filter box to visualize the patterns of dataset. This dashboard can provide the details of the drug and condition that corresponds to the cluster that had been selected by the user. Rating and useful count also allow the user to understand the patient preferences and satisfaction to the drug consumed.

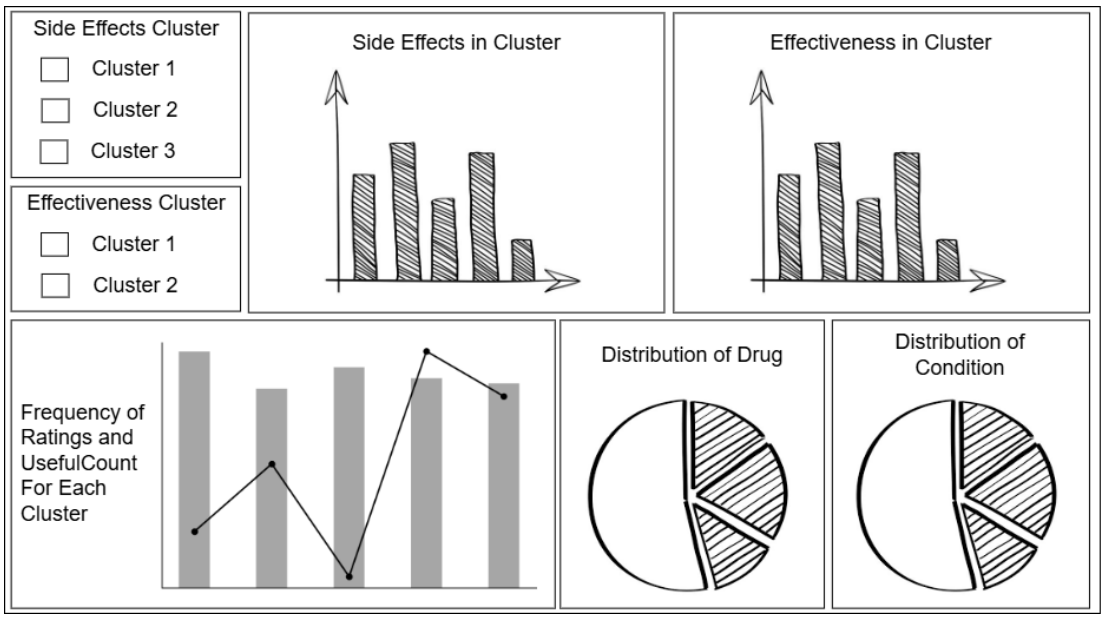


Figure 3.2: The Example of Dashboard Visualization

3.3 Dataset

The drug reviews dataset was obtained from UCI Machine Learning Repository. It consists of 215,063 rows of data representing the individual review and 6 columns that contain drug name, health condition, patient's comment and ratings, date of reviews and the number of users who found the review helpful. It provided an insight of overall patient satisfaction with patient reviews on specific drugs along with related conditions and a ten-star patient rating.

	drugName	condition	review	rating	date	usefulCount
0	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

Figure 3.3: Drug Review Dataset

3.4 Performance Measurement: Silhouette Coefficient

Silhouette coefficient was a method that evaluated the cohesion within clusters and the distance between clusters (Gui et al., 2024). Silhouette coefficient had the value between -1 and 1 (Řezanková, 2018) while higher value indicated that the quality of clustering is high. According to Shahapure and Nicholas (2020), the data point is correctly cluster when the score near to 1 while the data point is wrongly cluster when the score near to -1. A silhouette score with 0 indicated that the data point belongs to some other clusters (Shahapure & Nicholas, 2020).

Silhouette coefficient was defined as:

$$S = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{3.1}$$

Where:

b_i, a_i: features for calculation

S: average distance between the features

3.5 Summary

In a nutshell, this chapter explained the research framework as well as the steps that needed to be carried out to ensure the research process is smooth. The objectives and goal of the research had been considered to conduct the research framework. Next chapter will discuss the research design and implementation.

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. 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.

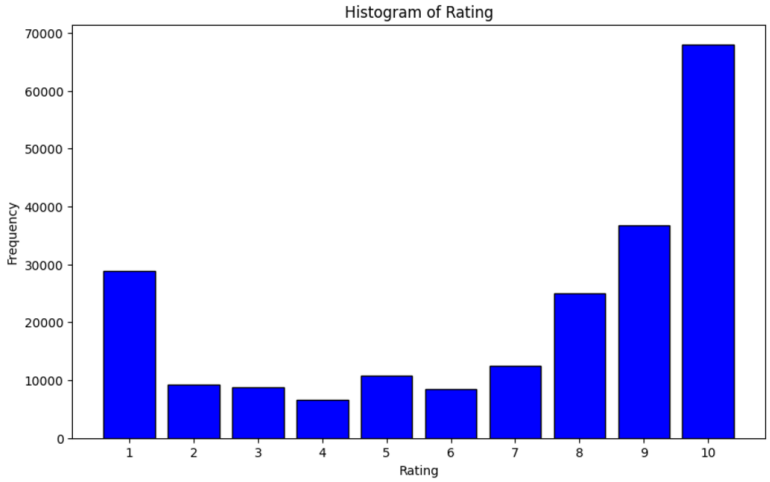


Figure 4.1: Histogram of Rating Feature

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.

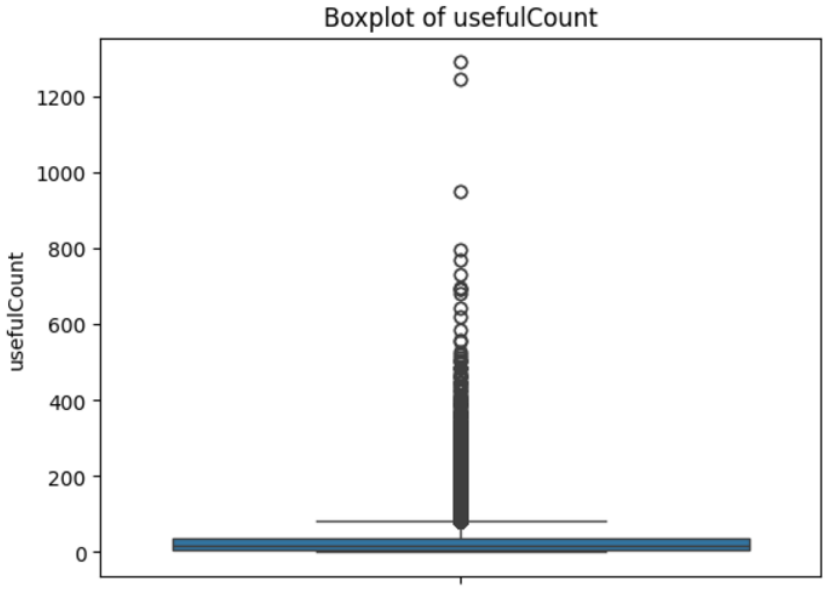


Figure 4.2: Box Plot of usefulCount

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.

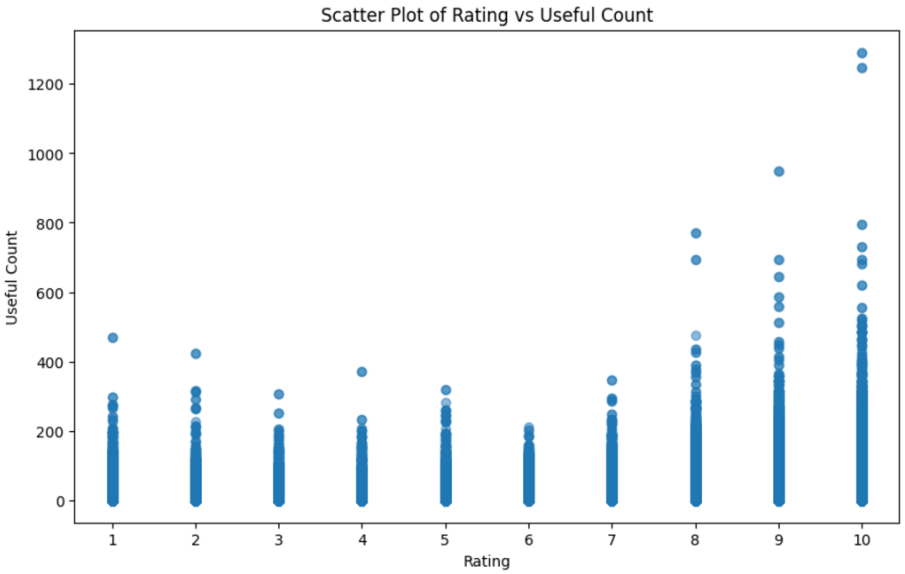


Figure 4.3: Scatter Plot between Rating and usefulCount

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. Besides that, among 916 conditions, there are 82 conditions only discussed once in the dataset. For example, systemic candidiasis and uveitis posterior are the conditions that only discussed one time in the dataset.

The observation into the conditions that only appeared only once in the dataset found that the conditions are rare or uncommon in medical scenarios such as systemic candidiasis and rat bite fever. Besides that, the typo issues also indicated that the conditions did not record accordingly. For example, ungal pneumonia that found in the dataset should refer as fungal pneumonia.

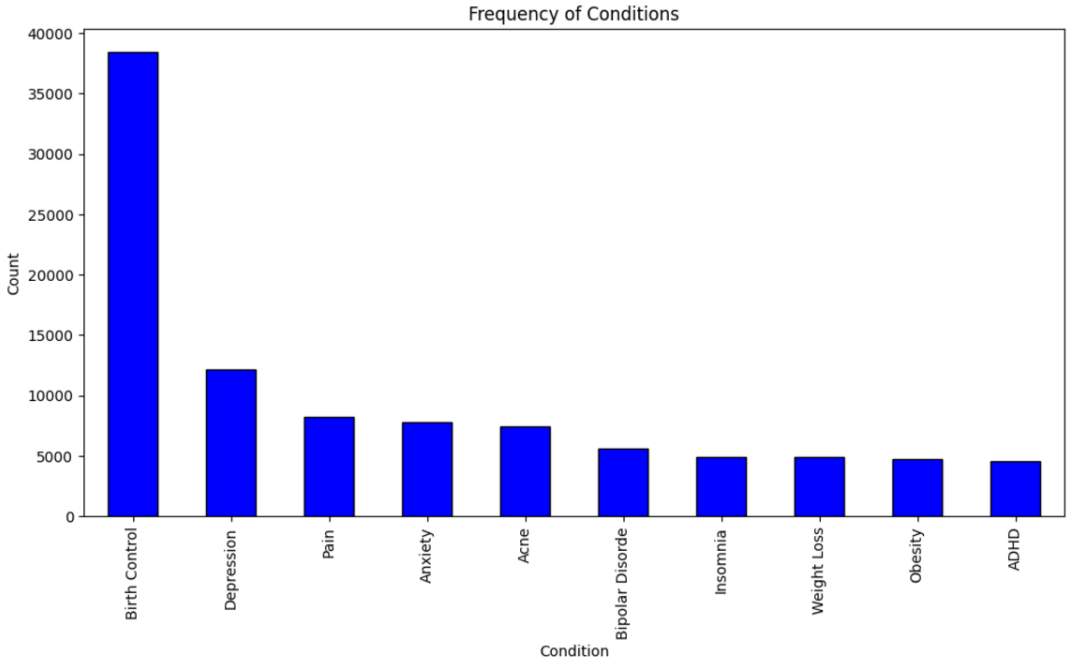


Figure 4.4: Top 10 Conditions

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. Besides that, among 3671 drugs, there are 798 drugs only discussed once in the dataset. For example, acetic acid/hydrocortisone and phenylephrine/pyrilamine only appeared one time in the dataset.

The observation had been made on the drugs that only discussed once in the dataset. The drugs such as Rilonacept and Elosulfase Alfa are considered as the rare or uncommon drugs been consumed by individuals due to rare diseases aligned with it. Besides that, the drug such as Bremelanotide has not actively in used due to approval was made recently based on the reviews date. Furthermore, the drug such as Rofecoxib had been banned due to the cardiovascular toxicity.

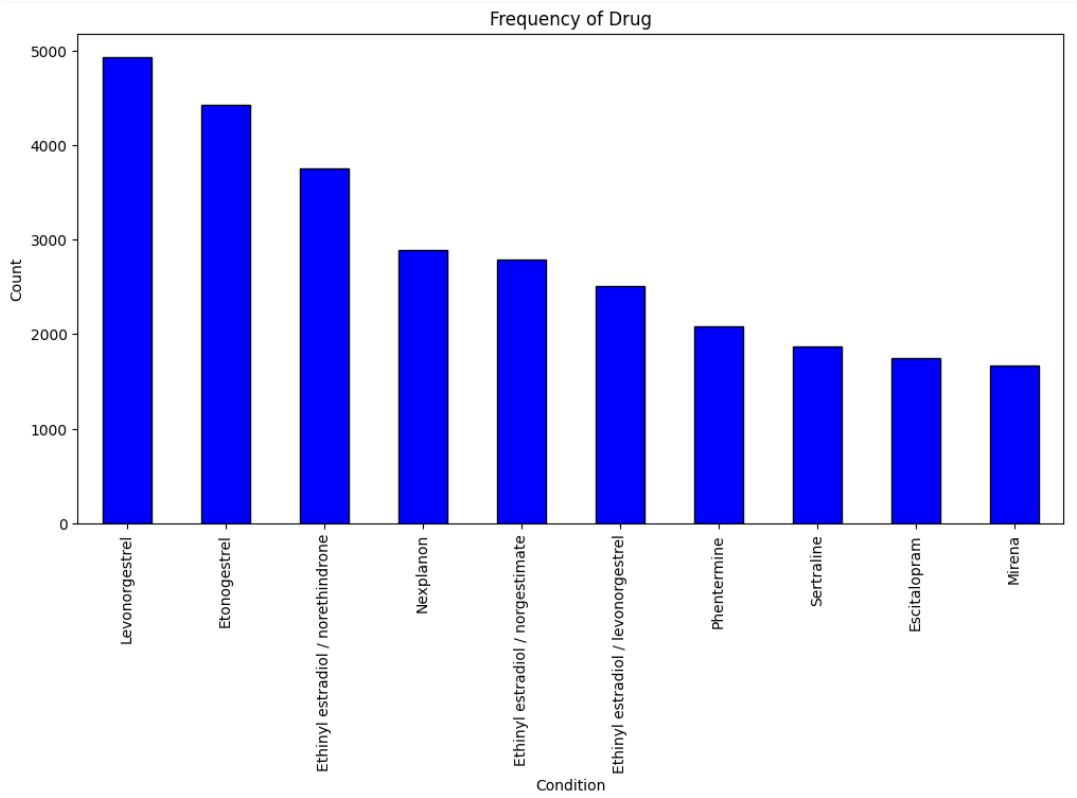


Figure 4.5: Top 10 Drug

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 4.6: 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 4.7: 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 4.8: Word Cloud of Neutral Sentiment Review

Figure 4.11 illustrated the presence of outliers in the reviews. Majority of data points were packed together in the central region while there exist with some data points that is spread away from the central region. These data points which distribute from central region can be considered as the outliers. Outliers in textual data can be due to the unusual terms that used by patient and uncommon drug performance that experience by patient. These outliers that represent the unique or rare drug performance can introduce noise when identifying the general patterns in drug performance. Thus, DBSCAN used in this study can exclude the noise data points and focus on the features that are more commonly found in patient experience.

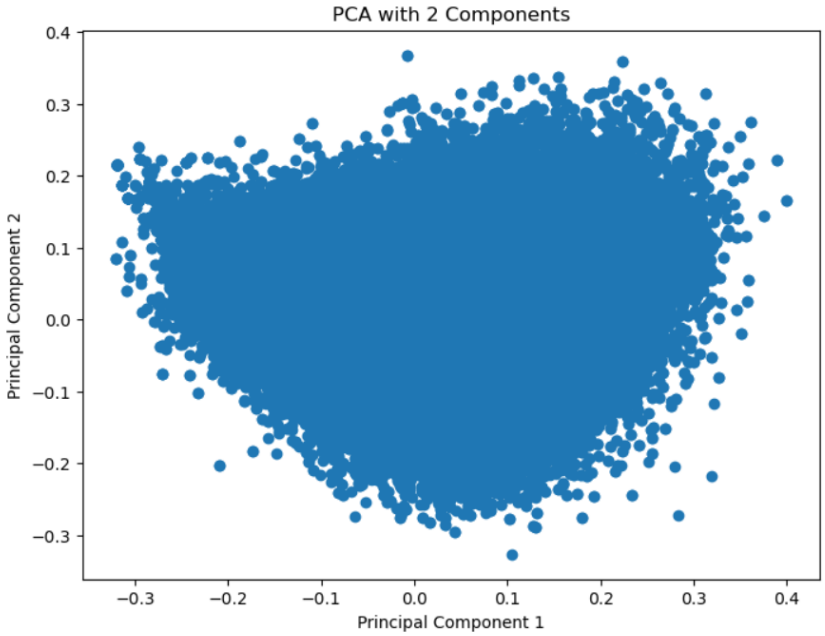


Figure 4.9: Outlier Detection with PCA

4.2.1 Comparison Between Clustering Techniques

There are total of three clustering techniques that had been observed in literature review. To determine the most suitable clustering approach to be used for further analysis, three clustering techniques were implemented on a random sample of dataset. The standardized review column had been used to retrieve 10 percent of the dataset randomly. Then, three clustering techniques were applied to the random sample

dataset to observe the performance of each technique. Figure 4.10 illustrates the flow in determining the most suitable clustering techniques to be used in this study.

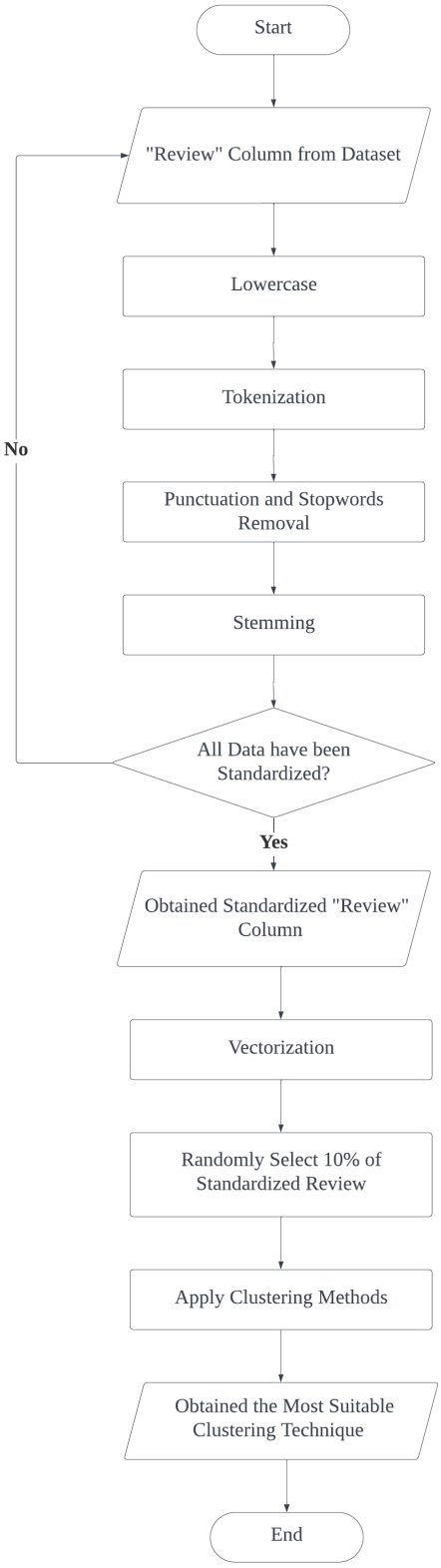


Figure 4.10: Flow to Determine the Most Suitable Clustering Techniques

Table 4.1 showed the silhouette score of each clustering technique. DBSCAN outperformed k-means and agglomerative hierarchical. Therefore, DBSCAN will be used in this study to cluster the retrieved keywords. These initial findings provide an initial study for determining the clustering approach that will be applied after data derivation.

Table 4.1: The Performance of Clustering Techniques

Clustering Techniques	**Silhouette Score**
K-Means	0.0001
Agglomerative Hierarchical	0.0042
DBSCAN	0.0142

The silhouette score obtained from the clustering is low because the clustering technique did not just focus on the important features but considered all information that was presented in the dataset to undergo the clustering process. However, the raw textual data will be further processed to retrieve the relevant keywords for clustering purposes.

4.3 Data Preparation

Data preparation is a crucial step in ensuring the data is clean and formatted before further processing. Therefore, 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.

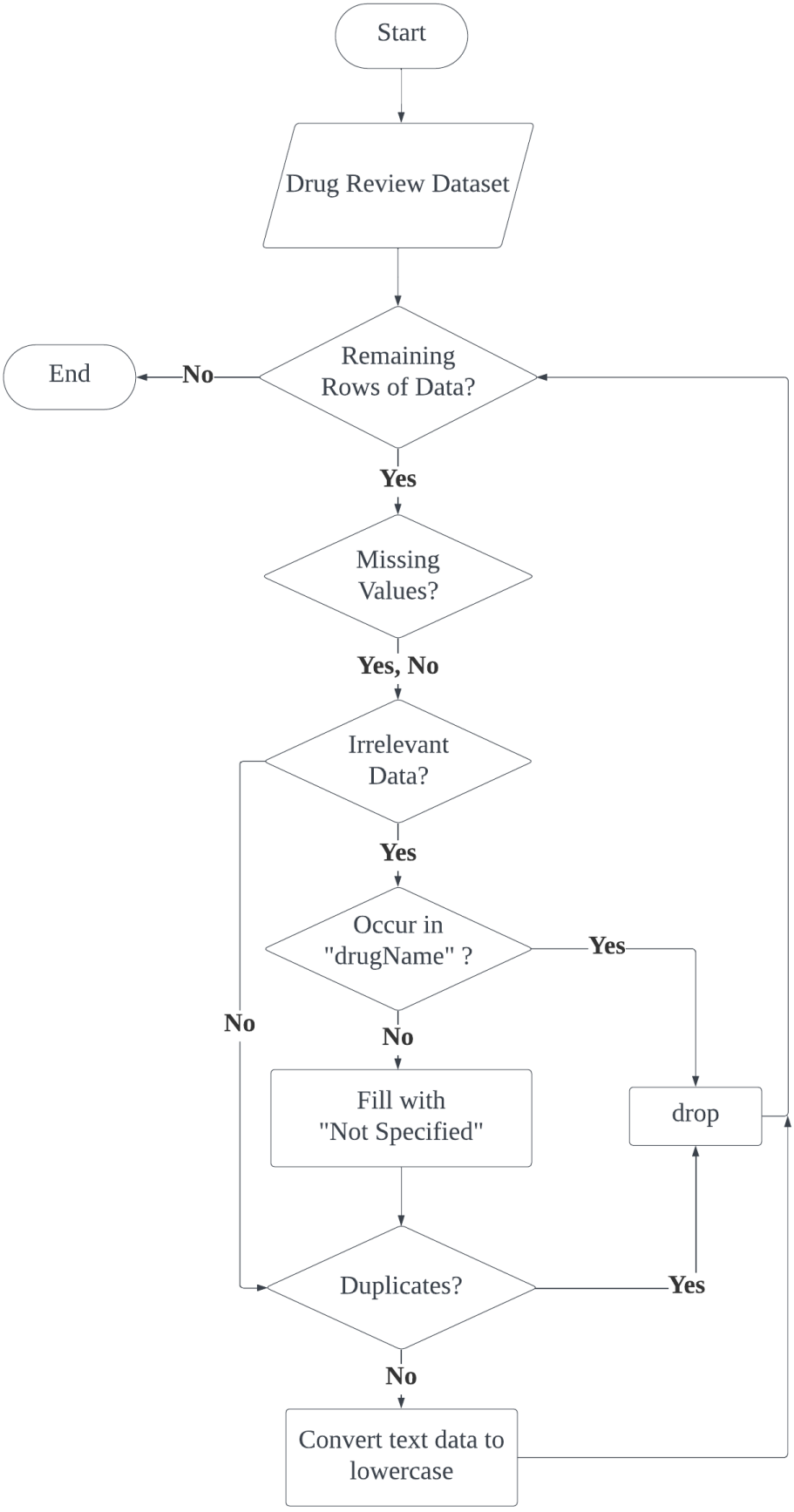


Figure 4.11: Data Preparation Flow

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.

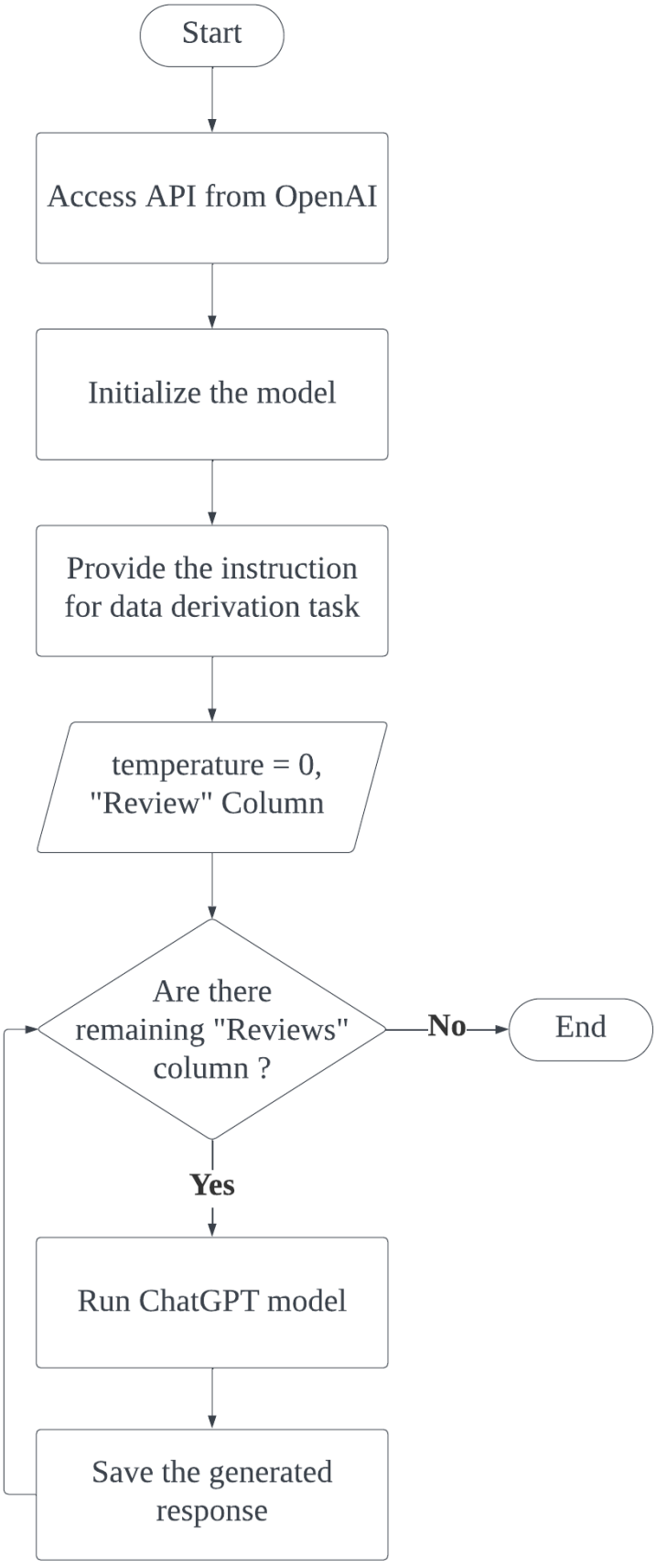


Figure 4.12: Data Derivation Flow

Table 4.2 shows the example of data derivation. 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 4.2: 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 and Evaluation

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.

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.

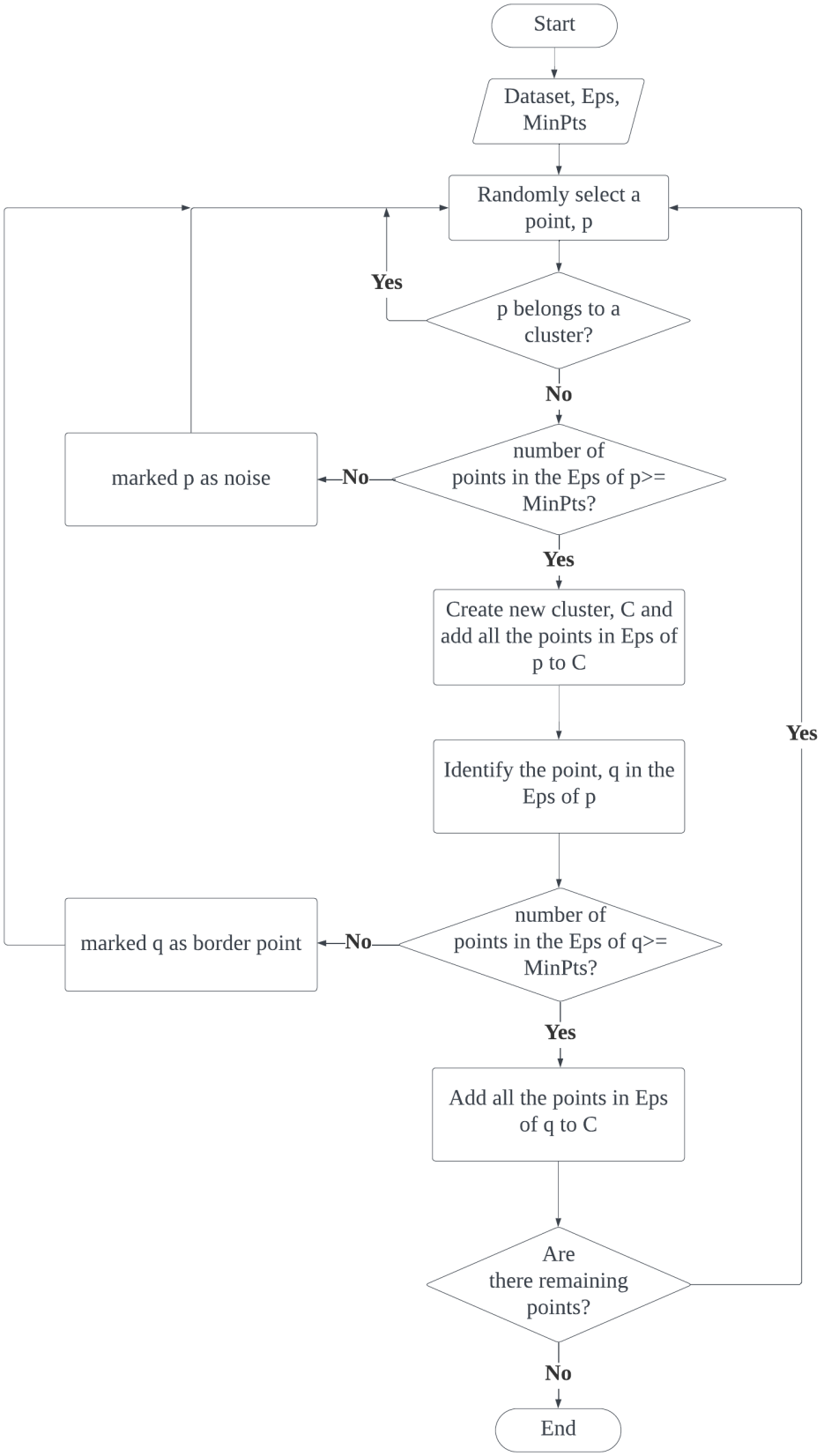


Figure 4.13: DBSCAN Flow

4.6 Visualization

After obtaining the cluster label for side effects and effectiveness based on the relevant keywords, dashboard will be constructed to visualize the clustering outcomes. It provides an overview of the drug and condition for their respective side effects and effectiveness. Therefore, healthcare professionals can make the clinical decision immediately.

4.7 Summary

In conclusion, there were five phases involved in the identification of drug efficacy. The GPT 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. Then, the clustering outcomes will be visualized in dashboard.

CHAPTER 5

DISCUSSION AND FUTURE WORKS

5.1 Introduction

This project aims to identify patterns in drug efficacy to enhance the understanding of drug performance across different patient populations. To ensure the result obtained from the analysis is accurate and reliable, a few steps have been taken to improve the data quality. Preprocessing steps such as EDA and data preprocessing were carried out to retrieve the input data for further analysis. This preprocessing ensures that the dataset is consistent and complete. Then, ChatGPT 4o mini model was called to analyze reviews (textual data) and produce the output that regarding to side effects and effectiveness of drugs. The ability of GPT model in understanding the relationship between words and context allows the important features to be derived. Thus, clustering techniques can focus on those important features by excluding the noise data presented in the textual data. The grouping of side effects and effectiveness into similar characteristics further enhances the visualization of drug performance when consumed in certain conditions.

5.2 Achievements

Data preparation was carried out to ensure the input data used for data derivation and model development was consistent and formatted. Data derivation by ChatGPT model allowed the identification of important features regarding side effects and effectiveness of drug from the review. Then, the implementation of DBSCAN to group the side effects and effectiveness to the cluster was believed to visualize the drug efficacy effectively. This research was predicted to achieve a silhouette score that was near to 1.

Achievements for this research are:

(a) A cleaned dataset can be retrieved by removing the duplicates and handling the missing values and irrelevant information that occurred in the dataset.

(b) ChatGPT-4o-mini model API was called to retrieve the features that wish to analysis further.

5.3 Future Works

In this project, the research framework only achieved the halfway of phase 3. Research planning and initial study that conducted during phase 1 provide with the comprehensive understanding of drug efficacy evaluation. Meanwhile, data preparation that is carried out during phase 2 allows a cleaned dataset to be collected by cleaning process and normalizing text data. Among the total of 215,063 drug reviews available in the dataset, 57,000 drug reviews were successfully processed in phase 3 to derive the side effects and effectiveness of drug.

Future works in MDS Project II are:

(a) Successfully retrieved the features of side effects and effectiveness from drug reviews.

(b) Clustering techniques such as DBSCAN were implemented and formed different groups of clusters for the side effects and effectiveness.

(c) Validation the clusters formed with silhouette coefficients.

(d) Visualization of the insights from drug review clusters through dashboard.

REFERENCES

Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, *4*(1), 27-33.

Abuka, G. (2023). *Text Summarization and Sentiment Analysis of Drug Reviews: A Transfer Learning Approach* Middle Tennessee State University].

Allenbrand, C. (2024). Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews. *Healthcare Analytics*, *5*, 100288.

Ampel, B., Yang, C.-H., Hu, J., & Chen, H. (2024). Large language models for conducting advanced text Analytics Information Systems Research. *ACM Transactions on Management Information Systems*.

Anderson, P., Higgins, V., Courcy, J. d., Doslikova, K., Davis, V. A., Karavali, M., & Piercy, J. (2023). Real-world evidence generation from patients, their caregivers and physicians supporting clinical, regulatory and guideline decisions: an update on Disease Specific Programmes. *Current Medical Research and Opinion*, *39*(12), 1707-1715.

Belal, M., She, J., & Wong, S. (2023). Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*.

Benatti, A., & Costa, L. d. F. (2024). Agglomerative clustering in uniform and proportional feature spaces. *arXiv preprint arXiv:2407.08604*.

Bhardwaj, A., Pandey, A., & Dahiya, S. (2022). Review based on Variations of DBSCAN algorithms. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS),

Biswas, B., Sengupta, P., & Ganguly, B. (2022). Your reviews or mine? Exploring the determinants of "perceived helpfulness" of online reviews: a cross-cultural study. *Electronic Markets*, *32*(3), 1083-1102.

Bushra, A. A., & Yi, G. (2021). Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. *IEEE Access*, *9*, 87918-87935.

Chong, B. (2021). K-means clustering algorithm: a brief review. *vol*, *4*, 37-40.

Cimino, A., Culbertson, C., Watkins, E., Li, J., & Wangeshi, S. (2024). RWD119 A Methodological Approach Using Sentiment Analysis of Online Medical Platforms As a Real-World Data Source of Patient Experiences. *Value in Health*, *27*(6), S381.

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, *110*, 103539.

Dinh, T., Chakraborty, G., & McGaugh, M. (2020). Exploring Online Drug Reviews using Text Analytics, Sentiment Analysis and Data Mining Models. SAS 2020 Global Forum,

Gruber, J. B., & Votta, F. (2024). Large Language Models.

Gui, C., Han, D., Gao, L., Zhao, Y., Wang, L., Xu, X., & Xu, Y. (2024). Application of Enhanced K-Means and Cloud Model for Structural Health Monitoring on Double-Layer Truss Arch Bridges. *Infrastructures*, *9*(9), 161.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*, 1-30.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, *125*(13), 1716.

Hu, L., Jiang, M., Dong, J., Liu, X., & He, Z. (2024). Interpretable Clustering: A Survey. *arXiv preprint arXiv:2409.00743*.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178-210.

Jayapradha, J., Kulkarni, Y., Naveen, P., & Anaam, E. A. (2024). Treatment Recommendation using BERT Personalization. *Journal of Informatics and Web Engineering*, *3*(3), 41-62.

Ji, Z., & Wang, C.-L. (2021). Accelerating DBSCAN algorithm with AI chips for large datasets. Proceedings of the 50th International Conference on Parallel Processing,

Kaul, P., Bose, B., Kumar, R., Ilahi, I., & Garg, P. K. (2021). The strength of a randomized controlled trial lies in its design—randomization. *Supportive Care in Cancer*, 1-3.

Kostis, J. B., & Dobrzynski, J. M. (2020). Limitations of randomized clinical trials. *The American journal of cardiology*, *129*, 109-115.

Kumar, A., & Shekhar, S. (2024). Hybrid model of unsupervised and supervised learning for multiclass sentiment analysis based on users' reviews on healthcare web forums. *J. Artif. Intell*, *7*(4).

Lee, D.-g., Kim, M., & Shin, H. (2022). Drug Repositioning with Disease-Drug Clusters from Word Representations. 2022 IEEE International Conference on Big Data and Smart Computing (BigComp),

Liakos, A., Pagkalidou, E., Karagiannis, T., Malandris, K., Avgerinos, I., Gigi, E., Bekiari, E., Haidich, A.-B., & Tsapas, A. (2024). A Simple Guide to Randomized Controlled Trials. *The International Journal of Lower Extremity Wounds*, 15347346241236385.

Liu, J., Zhou, Y., Jiang, X., & Zhang, W. (2020). Consumers' satisfaction factors mining and sentiment analysis of B2C online pharmacy reviews. *BMC medical informatics and decision making*, *20*, 1-13.

Liu, R. (2022). Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*, *2022*(1), 3762431.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), e1219.

Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195-211.

Nilashi, M., Ahmadi, H., Abumalloh, R. A., Alrizq, M., Alghamdi, A., & Alyami, S. (2024). Knowledge discovery of patients reviews on breast cancer drugs: Segmentation of side effects using machine learning techniques. *Heliyon*, *10*(19).

Oti, E. U., & Olusola, M. O. (2024). OVERVIEW OF AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS. *Technology*, *7*(2), 14-23.

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*(7), 6439-6475.

Özdemir, M., & Ortakcı, Y. (2024). TEXT CLASSIFICATION WITH FREQUENCY-BASED TEXT VECTORISATION METHODS FOR ENHANCING CALL CENTRE EFFICIENCY. *Current Trends in Computing*, *1*(2), 122-138.

Posch, A., & Tiwari, P. (2021). Persona-based drug recommender system using online reviews. In.

Pratama, M. A. Y., Hidayah, A. R., & Avini, T. (2023). Clustering K-Means untuk Analisis Pola Persebaran Bencana Alam di Indonesia. *Jurnal Informatika Dan Tekonologi Komputer (JITEK), 3*(2), 108-114.

Qiu, K., & Zhang, L. (2024). How online reviews affect purchase intention: A meta-analysis across contextual and cultural factors. *Data and Information Management, 8*(2), 100058.

Rangapur, A., & Rangapur, A. (2024). The Battle of LLMs: A Comparative Study in Conversational QA Tasks. *arXiv preprint arXiv:2405.18344*.

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Řezanková, H. (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. 21st international scientific conference AMSE applications of mathematics and statistics in economics,

Sabri, T., El Beggar, O., & Kissi, M. (2022). Comparative study of Arabic text classification using feature vectorization methods. *Procedia Computer Science, 198*, 269-275.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. 2020 IEEE 7th international conference on data science and advanced analytics (DSAA),

Sridharan, K., & Sivaramakrishnan, G. (2024). Unlocking the potential of advanced large language models in medication review and reconciliation: a proof-of-concept investigation. *Exploratory Research in Clinical and Social Pharmacy, 15*, 100492.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods, 23*, 16094069241231168.

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

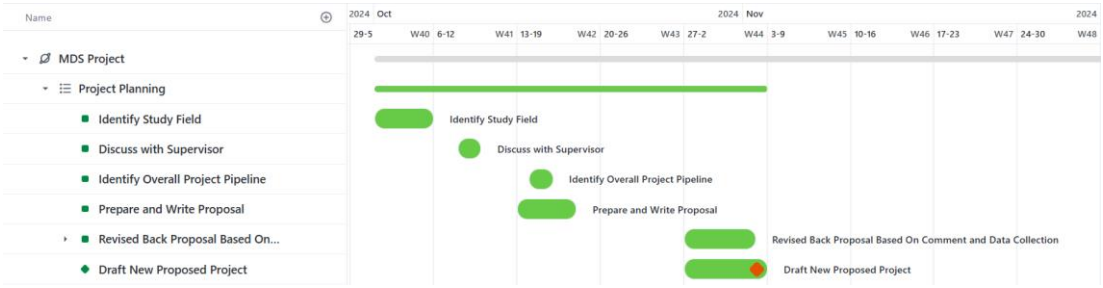
Xu, Q., Gu, H., & Ji, S. (2024). Text clustering based on pre-trained models and autoencoders. *Frontiers in Computational Neuroscience, 17*, 1334436.

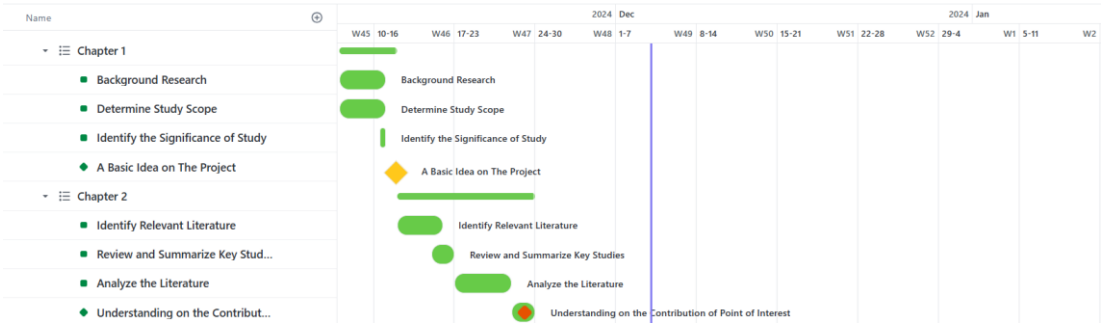
Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

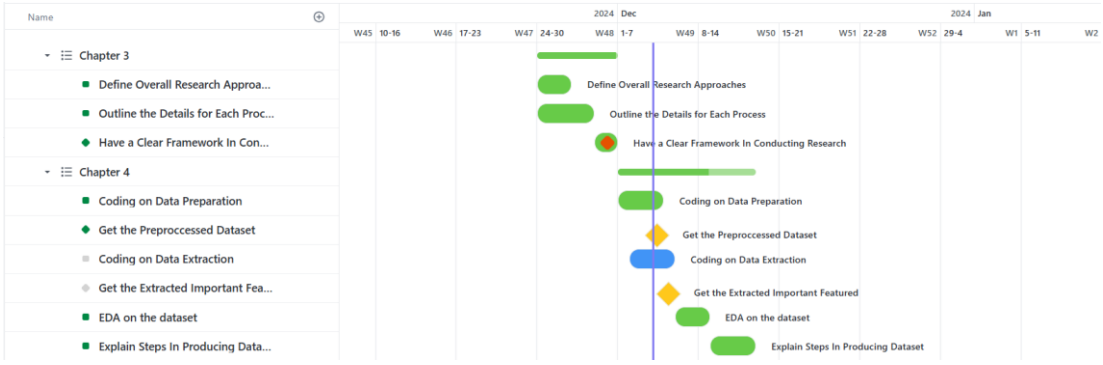
Yildirim, P., & Kaya, A. (2019). Clustering of Phentermine HCL Drug from Online Patient Medication Reviews. *Procedia Computer Science*, *151*, 1146-1151.

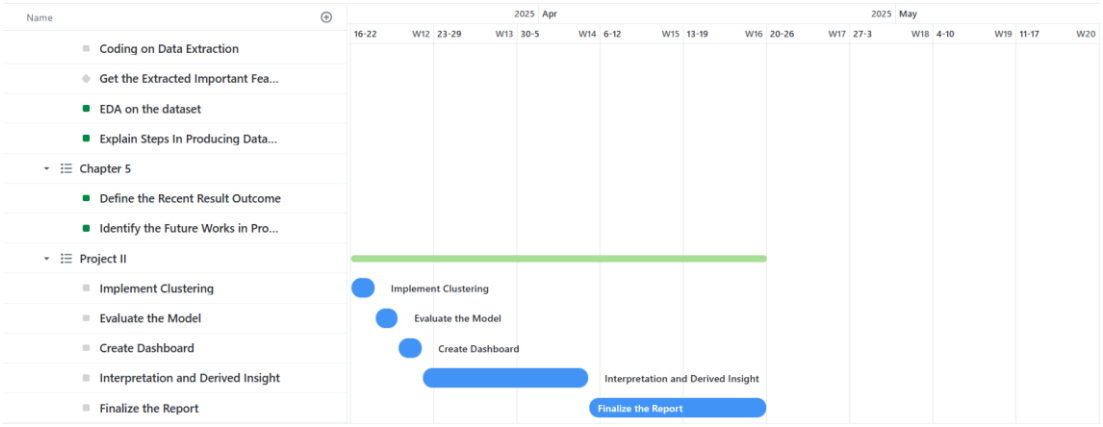
Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, solitons & fractals*, *140*, 110121.

Appendix A Gantt Chart








Name
2025 Apr
2025 May
16-22
W12
23-29
W13
30-5
W14
6-12
W15
13-19
W16
20-26
W17
27-3
W18
4-10
W19
11-17
W20
Coding on Data Extraction
Get the Extracted Important Fea...
EDA on the dataset
Explain Steps In Producing Data...
Chapter 5
Define the Recent Result Outcome
Identify the Future Works in Pro...
Project II
Implement Clustering
Evaluate the Model
Create Dashboard
Interpretation and Derived Insight
Finalize the Report
Implement Clustering
Evaluate the Model
Create Dashboard
Interpretation and Derived Insight
Finalize the Report