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

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

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 features from the pre-processed dataset. In this phase, the underlying pattern of the dataset can be identified by LLMs. Additionally, DBSCAN clustering model will be implemented into the retrieved features to classify 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 this phase.

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

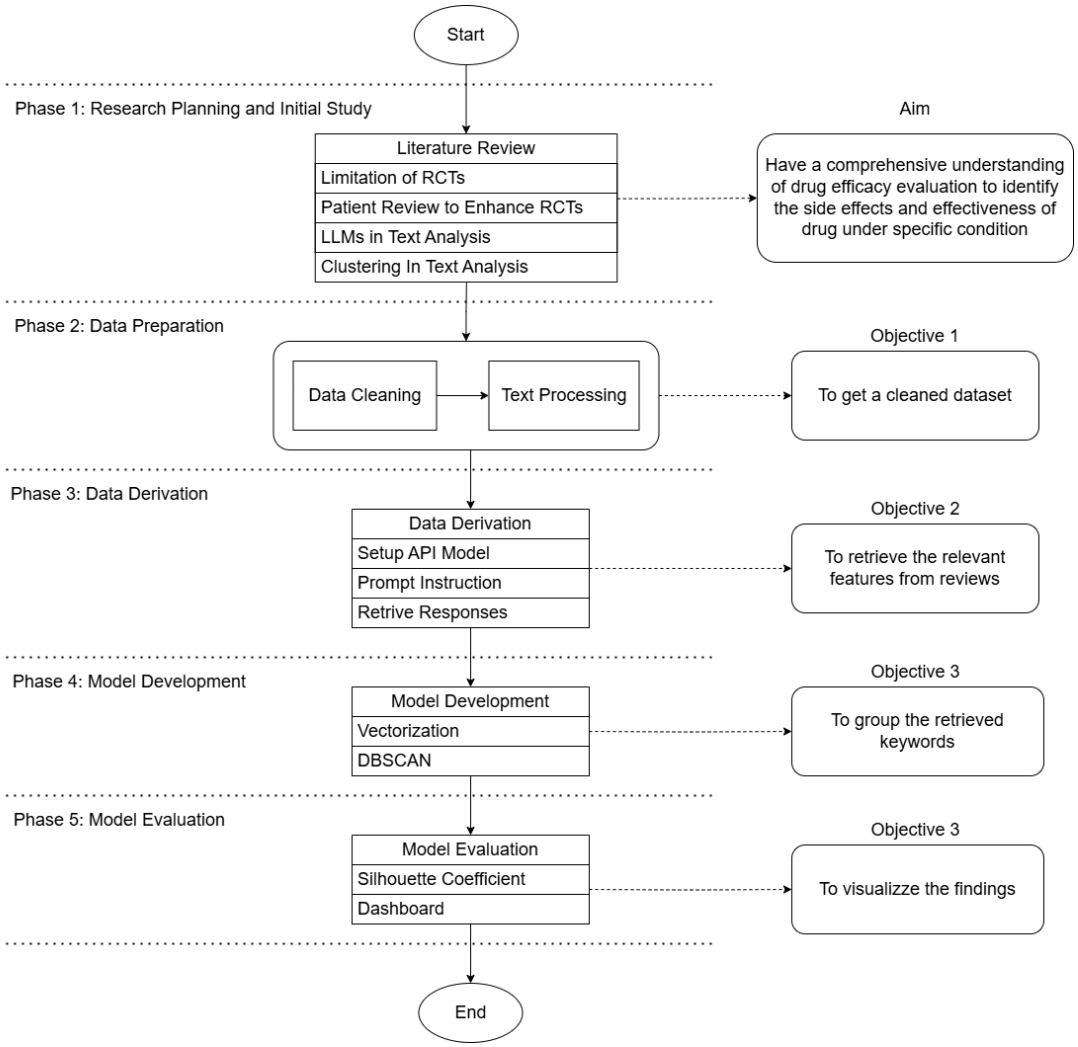

Figure 3.1: Overall Research Framework

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. 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 ChatGPT to focus on the meaningful words.

3.2.3 Phase 3: Data Derivation

Data derivation process will be carried out using LLMs 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 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 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.

3.2.4 Phase 4: Model Development

DBSCAN was chosen as the clustering technique to identify the clusters on the retrieved keywords. The ability of DBSCAN to exclude the outliers from the clusters and the capability to find the arbitrary shape further enhanced the accuracy of the output. The algorithm started by defining the parameters such as the Ɛ and MinPts. Then, DBSCAN will classify the data points based on the density of clusters and the minimum number of points that are required for the data points to form cluster.

3.2.5 Phase 5: Model Evaluation

The last phase was evaluating the effectiveness of clustering model. After retrieving a set of cluster labels from DBSCAN, 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.

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 3.2: 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.