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

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

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 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.