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, Chakraborty, & McGaugh, 2020). This is because people was able to share and comment their opinions that regarding to the products or services that they had experiences (Dinh et al., 2020). These comments and opinions by other people helped the user to have an insight on the products or services (Dinh et al., 2020). According to Qiu and Zhang (2024), there was 95 percent of people were read on the comments or reviews before purchase the products. Biswas, Sengupta, and Ganguly (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 (Dinh et al., 2020). Ratings is a numerical data that showed the satisfaction of patient while text reviews is textual data that represent the overall experiences such as the drug effectiveness to the patient (Dinh et al., 2020). 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 (J. Liu, Zhou, Jiang, & Zhang, 2020). Additionally, understanding patients' medical conditions through drug reviews can help patients to choose a better medicine when medical advice is limited (Zeroual, Harrou, Dairi, & Sun, 2020). Thus, sentiment 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, Bose, Kumar, Ilahi, & Garg, 2021). In this context, patient reviews are critical in bridging the gap between RCT findings and real-world scenarios.

Several studies have been conducted to address this gap in understanding. Shahid, Singh, Gupta, and Sharma (2022) proposed a deep learning-based medical recommendation system with N-Gram and patient review data. They apply sentiment analysis to recommend treatments based on patient feedback (Shahid et al., 2022). Furthermore, Rathod, Patel, Goswami, Degadwala, and Vyas (2023) studied the application of machine learning methods in sentiment analysis of drug reviews to extract insightful information from unstructured data collected on the Internet. Their study examined the performance of different machine learning algorithms in sentiment analysis and feature engineering techniques, and can accurately capture the sentiment of drug reviews, achieving high accuracy and F1 scores (Kostis & Dobrzynski, 2020; Rathod et al., 2023). 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. Clustering drug reviews able to discover underlying patterns of drug efficacy by analysing the personal experiences of consumers. Hence, by detecting positive or negative 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. Thus, the problem statement of this study is the lack of comprehensive information from RCTs limit the understanding of drug efficacy in diverse patients’ populations.

1.4 Research Questions

This thesis aims to visualize the patterns in drug efficacy by utilizing Large Language Models (LLMs) in extracting the relevant keywords and clustering patient reviews based on the keywords.

The research questions are:

(a) How do LLMs extract meaningful information from drug reviews?

(b) How do clustering techniques categorize the extracted keywords from drug reviews?

(c) What can be concluded from the drug review clusters?

1.5 Research Goal

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.

1.5.1 Research Objectives

The objectives of the research are:

(a) To conduct a preprocessing of the drug reviews datasets for drug efficacy analysis

(b) To extract relevant keywords from the pre-processed dataset by using Large Language Models

(c) To implement clustering techniques to categorize the extracted keywords and visualize the findings

1.6 Scope

The scopes of the research are:

(a) The data will be collected from UCI Irvine Machine Learning Repository

(b) The programming languages used is Python

(c) Concentrate on the sentiment analysis of the patient drug review, aiming to extract insights related to drug efficacy, side effects and overall patient satisfaction

1.7 Significance of Research

The use of LLMs and clustering techniques in this research tends to narrow the gap between RCTs and real-world scenarios. The important features in drug reviews able to be analyzed by LLMs. Then, the meaningful patterns will be shown by applying clustering techniques to categorize the important features. The involvement of diverse populations in drug reviews allows healthcare professionals to gain comprehensive understanding on drug performance. 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 had been obtained by healthcare professionals.