UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING
UTM Johor Bahru

Project
Proposal Form
MCST 1215

Sem: 1 Session: 20242025

SECTION A: Project Information.

Program Name: **Masters in Data Science**

Subject Name: **Project 1 (MCST 1215)**

Student Name: GUI YU XUAN

Metric Number: MCS241003

Student Email & Phone: guixuan@graduate.utm.my & 0166291961

Project Title: Identifying Patterns in Drug Efficacy by Analyzing Drug Reviews through A Clustering Approach

Supervisor 1: DR CHAN WENG HOWE

Supervisor 2 / Industry Advisor(if any):

SECTION B: Project Proposal

Introduction:

According to Dinh, Chakraborty and McGaugh (2020), healthcare professionals and consumers were able to gain medical information from drug reviews. Healthcare professionals can optimize their medication therapy and reduce medication errors from the interpretation of drug reviews (Sridharan and Sivaramakrishnan, 2024). Besides that, drug reviews provide useful information that contributes to the pharmacology domain, which help to optimize drugs, provide optimal medical outcomes and improve drug marketing and sales (Liu et al, 2020). In addition, understanding patients' medical conditions will help patients to choose a better medicine, especially when medical advice resources are limited (Zeroual et al, 2020). Thus, sentiment analysis of patient reviews can show the critical insights into drugs effects or benefits that are not captured in clinical trials. Analyzing patient experiences offered more informed clinical decisions by healthcare professionals and enabled more tailored treatment options to patient. Therefore, this project aims to utilize Large Language Models and clustering techniques on drug reviews to identify patterns in drug efficacy. By extracting meaningful insights from patient feedback will recognize a better understanding on the drugs perform across diverse conditions.

Problem Background:

Randomized clinical trials (RCTs) are the foundation for evaluating the effectiveness and safety of medical interventions. Even though clinical trials are crucial for establishing preliminary information under controlled conditions, they face limitations when applied to real world scenarios. RCTs include issues such as poor external validity, inaccurate statistical inference and publication bias (Kostis and Dobrzynski, 2020). According to Kostis and Dobrzynski (2020), the trial populations often consist of relatively healthy individuals with a given condition. Therefore, limiting the generalizability of findings to a more diverse patient population (Kostis and Dobrzynski, 2020). Patient drug reviews have the potential to address the limitations by giving real world feedback on efficacy and safety of a drug across diverse groups. This diversity makes patient reviews useful for understanding how drugs perform under typical usage conditions and also help identify any potential side effects (Oyebode and Orji, 2023). Therefore, the use of Large Language Model and clustering technique further enhance the sentiment analysis of patient reviews by extracting and categorizing the insights about drug efficacy.

Problem Statement:

Despite being a benchmark for evaluating medical efficacy and safety, randomized clinical trials (RCTs) have significant drawbacks when used in real-world situations. RCTs frequently use highly controlled conditions and select relatively healthy individuals with specific diseases, limiting the generalizability of findings to a more diverse patient population (Kostis & Dobrzynski, 2020). This controlled environment ignored long-term effects, side effects, or variations in drug efficacy that occur in our everyday life. Furthermore, RCTs are limited in scope (Collet, 2000). The results frequently lack information regarding the quality-of-life effects and real-world adverse medication reactions (Collet, 2000). As a result, healthcare professionals and patients lacked comprehensive knowledge about drug performance in various demographics and health status. However, patient drug reviews can fill these gaps by collecting a wide range of real-world experiences that provide unreported side effects and varying efficacy results. Thus, the problem statement of this study is the lack of comprehensive information from randomized clinical trials limit the understanding of drugs efficacy in diverse patient populations.

Aim of the Project:

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 Large Language Models and clustering techniques in patient drug reviews.

Objectives of the Project:

1. To conduct a preprocessing of the drug reviews dataset for drug efficacy analysis
2. To extract relevant keywords from the preprocessed dataset by using Large Language Models
3. To implement clustering techniques to categorize the extracted keywords and visualize the findings

Scopes of the Project:

1. The data will be collected from UCI Irvine Machine Learning Repository
2. The programming languages used is Python
3. Concentrate on the sentiment analysis of the patient drug review, aiming to extract insights related to drug efficacy, side effects and overall patient satisfaction

Expected Contribution of the Project:

The expected contribution of this project is to enhance the clinical decision-making process by healthcare professionals. The meaningful patterns and important features that were discovered from the analysis of drug reviews can be visualized and interpreted. To have a comprehensive understanding of the drug reviews dataset, Large Language Models (LLMs) and clustering techniques were applied to the dataset and the drug efficacy was identified. In conclusion, this project provides the real-world scenarios to medical professionals regarding the drug performance and help patients to choose their treatment plans.

Project Requirements:

Software:	Python

Hardware:	Intel i5 CPU, RAM 20GB
Technology/Technique/ Methodology/Algorithm:	Large Language Models, Clustering

Type of Project (Focusing on Data Science):

[/] Data Preparation and Modeling

[/] Data Analysis and Visualization

[ ] Business Intelligence and Analytics

[/] Machine Learning and Prediction

[ ] Data Science Application in Business Domain

Status of Project:

[/] New

[ ] Continued

If continued, what is the previous title?

SECTION C: Declaration

I declare that this project is proposed by:

[ ] Myself

[/] Supervisor/Industry Advisor (DR CHAN WENG HOWE)

Student Name: GUI YU XUAN

01/11/2024

Signature **Date**

SECTION D: Supervisor Acknowledgement

The Supervisor(s) shall complete this section.

I/We agree to become the supervisor(s) for this student under aforesaid proposed title.

Name of Supervisor 1: Dr Chan Weng Howe

Signature **Date**

Name of Supervisor 2 (if any):

Signature **Date**

SECTION E: Evaluation Panel Approval

The Evaluator(s) shall complete this section.

Result:

[ ] FULL APPROVAL

[ ] CONDITIONAL APPROVAL (Minor)

[ ] CONDITIONAL APPROVAL (Major)*

[ ] FAIL*

***** Student has to submit new proposal form considering the evaluators' comments.

Comments:

Name of Evaluator 1:

Signature **Date**

Name of Evaluator 2:

Signature **Date**