CP2-Project Proposal:
Sentiment Analysis on UCI ML Drug Review Dataset Using NLP

1.Executive Summary:

This project aims to develop a sentiment analysis model that classifies patient reviews on drug effectiveness, side effects, and satisfaction using Natural Language Processing (NLP) techniques. The project uses the UCI ML Drug Review Dataset to extract insights into public opinions and evaluate sentiment trends across different medications. The developed system can assist pharmaceutical companies and healthcare professionals in understanding patient experiences more effectively.

2.Problem Statement:

Background:

Online drug reviews have become a crucial source of feedback for healthcare providers, pharmaceutical companies, and patients. However, the vast amount of unstructured textual data makes manual analysis difficult and time-consuming. There is a growing need for automated systems that can efficiently classify and interpret sentiments expressed in these reviews.

Objective:

The objective of this project is to build an NLP-based sentiment analysis system that can accurately classify user reviews of drugs into sentiment categories. The model will utilize text preprocessing, tokenization, and deep learning algorithms to enhance accuracy and reliability in sentiment prediction.

Scope:

The project will cover data preprocessing, feature extraction, model training, and evaluation using machine learning and deep learning models such as Logistic Regression, Random Forest, LSTM, or BERT. The application of this system extends to healthcare recommendation engines, pharmaceutical analysis, and digital health review monitoring systems.

3.Data Sources:

The dataset used for this project is the UCI ML Drug Review Dataset, which contains user reviews, drug names, conditions treated, ratings, and timestamps. It will be used to train and test the sentiment classification models.

4.Methodology:

- **Data Preprocessing:** Cleaning, tokenization, and removal of stop words.
- **Feature Extraction:** TF-IDF vectorization and Word Embeddings (Word2Vec/BERT).
- **Modeling:** Training supervised learning models (Logistic Regression, Random Forest, and LSTM).
- **Evaluation:** Accuracy, Precision, Recall, and F1-Score.
- **Deployment:** Stream lit-based interactive web dashboard for sentiment prediction.

5.Expected Outcomes:

- An NLP-based sentiment classifier trained on drug reviews.
- Visual analytics on sentiment distribution and review trends.
- A deployed interface for real-time sentiment prediction.
- Actionable insights for pharmaceutical and healthcare stakeholders.

6.Risks and Challenges:

- Handling imbalanced sentiment classes in the dataset.
- Dealing with ambiguous or sarcastic language.
- Ensuring generalization across diverse drug types.
- Computational requirements for deep learning model training.

7.Conclusion:

This project successfully demonstrates the application of NLP in healthcare analytics by performing sentiment analysis on patient drug reviews. It provides valuable insights into patient satisfaction and drug effectiveness, promoting data-driven decision-making for the pharmaceutical industry. The system can be extended for broader healthcare feedback analysis in the future.