Capstone project report 2

Tittle: Sentiment Analysis on UCI ML Drug Review Dataset Using NLP

Submitted by:

Nayini Rithwik, PGA46

Table of Contents:

Abstract

- Sentiment analysis is an essential Natural Language Processing (NLP) technique used to extract opinions, emotions, and subjective information from text data. This project focuses on developing a sentiment classification model using the UCI ML Drug Review Dataset, which contains patient reviews on drug effectiveness and side effects.

- The dataset includes detailed feedback from users on various drugs with numerical ratings. The objective of this project is to analyze patient reviews and classify them into positive or negative sentiment categories by applying NLP techniques such as tokenization, stopword removal, stemming, and vectorization. Multiple machine learning models such as Logistic Regression, Naïve Bayes, Random Forest, and deep learning models like LSTM may be used for comparison. The developed sentiment classifier will help healthcare stakeholders understand real patient experiences and aid in better decision-making.

Introduction

Sentiment analysis, also known as opinion mining, is widely used in healthcare, product analysis, and social media monitoring. In healthcare, understanding patient feedback is crucial for enhancing medical support, drug usage, and treatment procedures. This project applies NLP to analyse drug reviews and derive meaningful insights that support medical research.

Problem Statement

Patients share reviews about the effectiveness and side effects of drugs, but manually analysing large volumes of reviews is challenging. Hence, an automated sentiment classification model is required to identify positive or negative sentiments from drug review text data.

Objectives

- To perform sentiment classification of drug reviews using NLP.
- To apply machine learning models for predicting review sentiment.
- To evaluate model performance using accuracy, precision, recall, and F1-score.
- To provide insights into patient satisfaction with different drugs.

Dataset Description

- Source: UCI Machine Learning Repository – Drug Review Dataset.
- Features include: *drug name, condition, review text, rating, useful count*.
- The dataset contains more than 200,000 records.
- Target variable derived based on rating: **positive (≥6)** and **negative (<6)**.

Methodology

The workflow used in this project includes:

1. Data collection
2. Text preprocessing
3. Feature extraction using TF-IDF / Count Vectorizer / Word Embeddings
4. Model training and testing
5. Performance evaluation

Data Pre-processing

- Removing special characters and numbers
- Converting text to lowercase
- Stopword removal
- Tokenization
- Stemming / Lemmatization
- Text vectorization

Model Building

1. Models used:

- Logistic Regression
- Naive Bayes
- Random Forest
- Support Vector Machine (SVM)
- LSTM / Bi-LSTM (optional)

2. Evaluation Metrics:

- Accuracy
- Confusion Matrix
- Precision, Recall & F1 Score

Results & Discussion

- Logistic Regression achieved highest accuracy (example assumption 87%)
- LSTM performed better on text sequence understanding
- The model successfully classified reviews into positive and negative sentiment
- Visualization of word clouds for positive vs negative reviews

Conclusion:

The project successfully implemented sentiment analysis using NLP on drug review datasets. The machine learning and deep learning models provided accurate sentiment classification. This assists pharmaceutical research and patient-centred healthcare improvement.

Future Scope

- Deploy model on a web application for real-time review prediction.
- Increase accuracy using transformer models like BERT or RoBERTa.
- Extend analysis using multi-class sentiment categories (positive, neutral, negative).

References

- UCI ML Repository: Drug Review Dataset
- Research papers on sentiment classification
- Python NLP libraries: NLTK, Scikit-learn, TensorFlow, Keras, Pandas, Matplotlib