**Project Proposal for Drug Review Sentiment Analysis**

The focus of this project is to perform sentiment analysis on patient reviews for various drugs. The goal is to predict the sentiment (positive or negative) of reviews based on textual data and related features such as drug name, condition treated, and patient ratings. The sentiment prediction will help pharmaceutical companies and healthcare professionals gain insights into patient satisfaction, identify potential issues with drugs, and ultimately improve patient outcomes.

The **Drug Review Dataset** from the UCI Machine Learning Repository is utilized for this analysis. It contains 161,297 entries with the following 7 features:

• drugName: The name of the drug.

• condition: The medical condition for which the drug was prescribed.

• review: The text review provided by the patient.

• rating: A 10-star rating reflecting overall patient satisfaction.

• date: The date the review was posted.

• usefulCount: The number of users who found the review useful.

**Target Variable**: The sentiment of the review, which needs to be derived from the rating. A review rating above a certain threshold (e.g., 7/10) will be classified as positive, while a rating below this threshold will be classified as negative.

**Problem Statement and Algorithms**

The problem is to build a system that can analyze the textual reviews and related features to predict the sentiment of patient reviews. We plan to leverage advanced Natural Language Processing (NLP) techniques, focusing on deep learning algorithms like Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT).

**Algorithms to Investigate**:

• **LSTM (Long Short-Term Memory)**: To handle sequential data and capture long-term dependencies in the text reviews.

• **BERT (Bidirectional Encoder Representations from Transformers)**: For its ability to understand context and word semantics bidirectionally, leading to improved accuracy in sentiment analysis.

These deep learning models will be compared against traditional machine learning classifiers (e.g., LightGBM, Logistic Regression) to highlight their strengths in handling complex NLP tasks.

**Related Course Topics**

This project relates to the following course topics:

• **Natural Language Processing (NLP)**: Tokenization, text preprocessing, and

embeddings.

• **Deep Learning**: Neural networks, LSTM, and Transformer architectures.

• **Classification**: Binary classification techniques for sentiment analysis.

• **Data Preprocessing**: Handling missing data, tokenization, and feature engineering.

**Expected Behavior and System Capabilities**

The system is expected to:

• Analyze the textual content of patient reviews to predict sentiment (positive or

negative).

• Leverage drug ratings to assign sentiment labels to the reviews, which will serve as

ground truth for training.

• Predict the sentiment of new, unseen reviews based on patterns learned from

historical data.

• Provide insights into patient satisfaction for healthcare stakeholders.

**Key Issues to Focus On**

The main challenges expected in this project include:

• **Text Preprocessing**: Cleaning the review text, handling special characters, stop

words, and stemming/lemmatization.

• **Imbalanced Data**: Since patient satisfaction is often skewed toward positive

ratings, the dataset might be imbalanced, requiring techniques like SMOTE or class

weighting.

• **Feature Engineering**: Extracting meaningful features from the text, including n-

grams and embeddings.

• **Model Selection**: Balancing model complexity with interpretability and

performance. LSTM and BERT models require significant computational resources

but may offer better accuracy than simpler classifiers.

**Justification for Using LSTM and BERT Over Simpler Classifiers**

While simpler classifiers like Logistic,**LightGBM** are effective for tabular data, they have limitations in handling the sequential nature of text data. Here’s why **LSTM** and **BERT** are better suited for this task:

1. **Sequential and Contextual Understanding**:

**LSTM** networks are designed to capture long-term dependencies in sequences,

making them ideal for text data where the order of words matters.

**BERT** uses a bidirectional approach to understand the context of words in both

directions, leading to more accurate sentiment predictions by considering the full

context.

2. **Handling Complex Language Nuances**:

Reviews often contain nuances, sarcasm, and complex expressions that are difficult

for traditional classifiers to understand.

**BERT** excels in understanding context, synonyms, and word sense disambiguation,

which are critical for accurate sentiment analysis.

3. **State-of-the-Art Performance**:

LSTM and BERT have consistently outperformed traditional machine learning

models on NLP tasks like sentiment analysis, text classification, and question

answering.

BERT’s pre-trained embeddings can be fine-tuned on your specific dataset, allowing

it to adapt to the nuances of the drug reviews.

4. **Feature Extraction from Raw Text**:

Traditional classifiers require extensive feature engineering, whereas **deep learning**

**models** like LSTM and BERT can automatically extract high-level features from raw

text data, reducing the need for manual feature engineering.

5. **Scalability for Large Datasets**:

With a dataset of over 160,000 entries, LSTM and BERT models can efficiently scale

and handle large volumes of text, whereas simpler models may struggle with feature

extraction and scalability.

**Team Contributions**

The team members will contribute equally to:

**Data Preprocessing and Cleaning**: Handling missing values, text tokenization, and feature extraction.

**Model Development and Evaluation**: Implementing LSTM, BERT, and baseline models like LightGBM.

**System Deployment and Reporting**: Creating a user-friendly interface and presenting findings in a final report.

**Resources and References**

A list of resources that will be referred to throughout this project includes:

Research papers on sentiment analysis using LSTM and BERT.

Documentation and examples from the Hugging Face Transformers library.

Courses and tutorials on deep learning for NLP, such as *Deep Learning Specialization* by Andrew Ng.

Github repository used to collaborate with team members,

https://github.com/PSswathi/aai-501-group7-sentiment-analysis-drug-reviews