****

CSC 590 ─ Graduate Project Design and Analysis Report

Project Title: Medical Text Classification Using LLMs

Student Name: Medha Maisa

****

1. Architecture/Component Diagram

The project architecture involves multiple stages, including data preprocessing, traditional NLP-based models, and LLM-based models. The system is designed to classify medical text data into 'Cancer' and 'Non-Cancer' categories using the TCGA Pathology Reports Dataset. The architect ture is represented in the following block diagram:

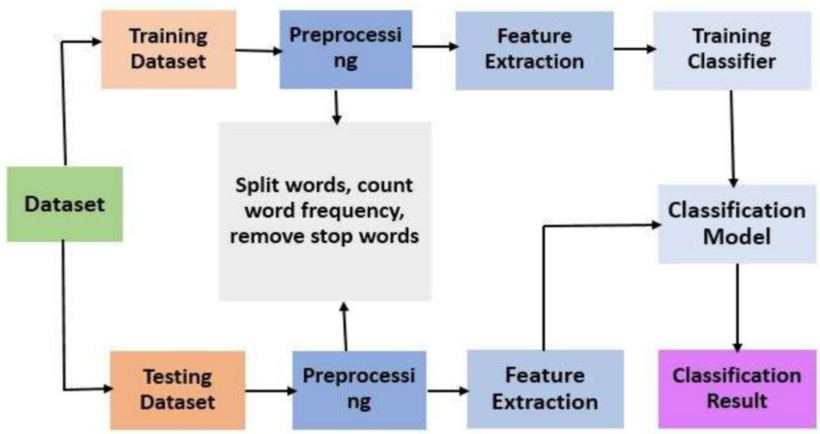

Figure: 1 Block diagram of text classification

2. Detailed Functions of Components

2.1 Data Preprocessing & Augmentation

- Removing duplicates, null values, and extremely short texts
- Text normalization (lowercasing, removing special characters)
- Handling class imbalance using SMOTE (Synthetic Minority Oversampling Technique)

2.2 Feature Engineering

- TF-IDF Vectorization for traditional models
- Word embeddings (BERT embeddings) for LLM-based models

2.3 Baseline Model (Traditional NLP)

- Logistic Regression
- Random Forest Classifier
- Performance Metrics: Accuracy, F1-Score, Precision, and Recall

2.4 LLM-Based Model

- Fine-tuning BERT or GPT models on the TCGA Pathology Reports Dataset

- Implementing Few-shot learning techniques for enhanced performance

2.5 Evaluation Metrics

- Accuracy
- F1 Score
- Precision and Recall
- Confusion Matrix Visualization

3. Mathematical Formulation and Analysis

1. **TF-IDF Vectorization**

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

2. **Logistic Regression Hypothesis Function**

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

3. **LLM Fine-Tuning Loss Function (Binary Cross Entropy)**

$$L = -\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

4. **SMOTE for Handling Class Imbalance**

$$x_{new} = x_i + \lambda \times (x_j - x_i)$$

where λ is a random number between 0 and 1, and xi and xj are minority class samples.

4. Expected Outcomes

Baseline Model Performance (Traditional NLP)

- TF-IDF + Logistic Regression: 87.32% accuracy

LLM-Based Model Performance

- Improved accuracy and F1 Score (expected 90%+ accuracy) through fine-tuning BERT

Few-shot Learning Performance

- Enhanced performance with limited data

5. Future Work

5.1 Dataset Expansion and Model Efficiency

- Future work will focus on exploring additional medical datasets for training and testing the model. This will help build a more robust and efficient model capable of handling a variety of medical domains.

5.2 Extending the Model to Multi-Class Classification

- The model will be extended from binary to multi-class classification, enabling it to classify a wider range of medical conditions, such as different types of cancers.

5.3 Custom User Input and Feedback Loop Integration

- The model will allow users to input custom medical text and generate predictions. It will learn from user feedback to continuously improve its predictions, incorporating reinforcement or active learning techniques.