CS-4420: Natural Language Processing

Assignment 2

LSTM and Transformer Models for Classification

Objectives

The goal of this assignment is to implement and compare different deep learning models for text classification and explore explainability techniques. You will choose one of two datasets and train both LSTM and Transformer-based models. You will then analyse and interpret the model predictions.

Datasets

You can choose one of the two following datasets (attached in the Classroom assignment):

1. **Mental Health Dataset**: Tweets labeled with mental health diagnoses (e.g., depression, anxiety, bipolar disorder, etc.).
2. **Biomedical Research Papers Dataset**: Excerpts from biomedical research papers classified into cancer types (lung, thyroid, or colon cancer).

Tasks

1. **Exploration**

   Conduct data exploration, and gather information which may be useful for later tasks. For example, you may want to check if there are class imbalances which can affect training performance.

2. **Classification using LSTM and Transformer**

   (a) Train/fine-tune and evaluate **LSTM-based** (e.g., LSTM, BiLSTM) and **Transformer-based** (e.g., BERT, BioBERT) models for document classification.

   (b) Use standard metrics (accuracy, F1-score, precision, recall) for comparison.

3. **Generation of Explanations**

   Generate explanations for **both** the models' classifications using **at least two** methods, including:

   - **LIME** (more info here) generates an approximation of your "black-box" model by perturbing the input and observing changes in the output.
   - **SHAP** (more info here): calculates feature importance scores based on Shapley values (game theory-based approach). Note that this method is computationally expensive.

- **Attention-based scores**: use the attention weights from the model to identify which parts of the input the model pays attention to.
- **Other metrics**: you can identify and implement other methods for explainability from your own research.

4. **Identification and Interpretation of Class Signatures**
   - (a) Identify patterns/keywords/features that define each class.
   - (b) Compare the outputs from the mechanisms you implemented in Q2.
   - (c) Discuss how reliable and interpretable these explanations are.

Submission

Due date: March 28, 2025

1. **Implementation**: Jupyter Notebook(s)/Python script(s) with **well-documented** code.
2. **Report (PDF)**:

   Your report must include:

   - Dataset choice
   - Insights from data exploration, if applicable
   - Preprocessing steps
   - Model choice, architectures, training details
   - Explainability analysis (question 2), with visualizations and justification for choice of methods
   - Observations on class signatures (question 3)
   - Challenges faced, potential improvements

Please zip these two items together and submit following the naming convention: **First_Last_A2.zip**.

Grading Criteria

Component	**Points**
Model Implementation & Performance Comparison	40
Explanation Generation	25
Interpretation of Class Signatures	20
Data Exploration, Report, and Code Documentation	15
Total	**100**