CSCE 580/ Project B/ Fall 2025 **Srivastava, B (TA: Pallagani, V., Lakkaraju, K)**

Project: Why and When to Use LLMs for Classification using IMDB Movie Review Dataset
Keywords: Classification, DistilBERT, GPT2, Fine-Tuning, Text processing, AI test cases

Due date: Thursday, Nov 20, 2025

Objective:

The goal of this project is to understand the benefits and trade-offs in using different forms of AI. We will do this by building and testing different models for sentiment analysis (a) fine-tune a pre-trained transformer model, DistilBERT, on the IMDB movie review dataset, and compare its performance with both its base version (without fine-tuning) and traditional machine learning algorithms for sentiment classification. We will

Steps and Tasks:

1. **Dataset Overview**:
    - Use the IMDB Movie Review Dataset (https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews), which contains 50,000 movie reviews labeled as either positive or negative.
2. **Preprocessing**: **[30 points]**
    - Load and preprocess the dataset (tokenization, padding, etc.) suitable for input into DistilBERT and classical models.
    - Discuss any data splitting (e.g., train/test split, stratified sampling).
3. **Fine-Tuning DistilBERT**: **[50 points, using TensorFlow or PyTorch]**
    - Fine-tune the pre-trained DistilBERT model on the training portion of the IMDB dataset.
    - Monitor the training process (e.g., training loss, validation loss).
4. **Base Model Comparison**: **[60 points]**
    - Use the pre-trained DistilBERT model **without fine-tuning** to classify the test set.
    - Compare the results of the fine-tuned DistilBERT with the base model.
    - Compare the results of the fine-tuned DistilBERT with any GPT model. E.g., base GPT-2 model which can be accessed here - https://huggingface.co/openai-community/gpt2. (Or GPT from Univ)
5. **Classical Machine Learning Model**: **[30 points]**
    - Train at least one classical machine learning algorithm (e.g., Logistic Regression, Support Vector Machine, or Random Forest) on the same IMDB dataset, using TF-IDF features or Bag-of-Words representation.

- Compare its performance with DistilBERT models (base and fine-tuned).

Analysis and Graphs:

Students should provide the following analyses in their report:

1. **AI test cases: [30 points]**
   - Using the testcase template [5], create at least 3 testcases. They would have input (movie information) of known sentiment and of different complexity of inputs, e.g., input word length, sentence length, sentence structures.
   - Report on results for all the four models (statistical, DistilBERT, finetuned DISTILBERT, GPT) on the three test cases using GAICO. Use statistical model's performance as the baseline.
2. **Accuracy and Loss Curves**: **[30 points]**
   - Plot the training and validation accuracy and loss over epochs for the fine-tuned DistilBERT model.
   - Comment on how the model is learning and any potential overfitting or underfitting.
3. **Confusion Matrix**: **[30 points]**
   - Generate confusion matrices for the fine-tuned DistilBERT, base DistilBERT, and the classical machine learning model.
   - Discuss misclassifications (e.g., false positives, false negatives) and possible reasons.
4. **Precision, Recall, and F1-Score**: **[30 points]**
   - Report and compare the precision, recall, and F1-scores of all models.
5. **Performance Comparison**: **[30 points]**
   - Create a comparative table with the evaluation metrics (accuracy, precision, recall, F1-score) for:
     - GPT model
     - Fine-tuned DistilBERT
     - Base DistilBERT
     - Classical Machine Learning Model(s)
6. **Time Complexity**: **[30 points]**
   - Discuss and compare the time taken for training and inference in all models. Which model is more efficient in terms of time and resources?

Questions: [50 points]

1. What do the accuracy and loss curves tell you about the fine-tuning process?
2. How does the fine-tuned DistilBERT model compare to the classical ML model? What advantages or limitations do transformers present over classical algorithms?
3. What insights can you draw from the confusion matrix? Are there any patterns in the misclassifications?
4. Why might the fine-tuned model outperform the base model?
5. Which model would you recommend for deployment in a real-world scenario, and why? Consider both performance and efficiency in your answer.

Deliverables:

- Data: Testcases and their results
- Code (Jupyter Notebook/Python script) for fine-tuning DistilBERT, training the classical model, evaluation
- A report (PDF/Markdown) that includes:
    - Plots of accuracy and loss curves.
    - Confusion matrices.
    - Comparative tables for evaluation metrics.
    - Answers to the provided questions.

References:

1. Fine-tuning DistilBERT for binary classification tasks(tensforflow) - https://towardsdatascience.com/hugging-face-transformers-fine-tuning-distilbert-for-binary-classification-tasks-490f1d192379
2. Smart Expert System: Large Language Models as Text Classifiers - https://arxiv.org/html/2405.10523v1
3. Example notebook of fine-tuning DistilBERT and classical machine learning models on IMDB dataset - IMDB sentiment analysis - EDA, ML, LSTM, BERT (https://www.kaggle.com/code/ducnm030303/imdb-sentiment-analysis-eda-ml-lstm-bert)
4. Finetuning sentiment analysis, DistilBERT (pytorch), https://ai.plainenglish.io/fine-tuning-sentiment-analysis-model-db088da71879
5. AI testcase template - https://github.com/biplav-s/book-trustworthy-chatbot/blob/main/ai-testcases/testcase-template.md