Final Project Assignment Questions

1 Part 0: Dataset Selection

- **Objective**: Choose a unique dataset for text classification from the HuggingFace hub.
- **Task**:
  - Email me, with your group in cc, to confirm your dataset choice to ensure no overlap with other groups. First come, first served.

2 Part 1: Setting Up the Problem (1.5 points)

- **Objective**: Understand and establish the baseline for your chosen dataset.
- **Tasks**:
  - **a. Bibliography and SOA (0.25 points)**:
    Present briefly your task by researching and documenting the main objective, a potential business case and the current state of the art for your dataset's task. Include relevant benchmarks and methodologies. You can look at google scholar, NLP index or papers with code.
  - **b. Dataset Description (0.5 points)**:
    Provide a brief overview of your dataset, including size, class distribution, and any peculiar characteristics. Include basic descriptive statistics.
  - **c. Random Classifier Performance (0.25 points)**:
    Calculate the expected performance of a random classifier for your dataset to set a benchmark. The calculation should include an implementation.
  - **d. Baseline Implementation (0.5 points)**:
    Develop a rule-based classifier as a baseline. Discuss its performance in the context of the dataset's complexity and compare it with human-level performance if available.

3 Part 2: Data Scientist Challenge (3.5 points)

- **Objective**: Explore different techniques to enhance model performance with limited labeled data. You will be limited to 32 labeled examples in your task. The rest can be viewed as unlabelled data.
- **Tasks**:

- **a. BERT Model with Limited Data (0.5 points)**:
  Train a BERT-based model using only 32 labeled examples and assess its performance.
- **b. Dataset Augmentation (1 point)**:
  Experiment with an automated technique to increase your dataset size without using LLMs (chatGPT / Mistral / Gemini / etc...). Evaluate the impact on model performance.
- **c. Zero-Shot Learning with LLM (0.5 points)**:
  Apply a LLM (chatGPT/Claude/Mistral/Gemini/...) in a zero-shot learning setup. Document the performance.
- **d. Data Generation with LLM (1 point)**:
  Use a LLM (chatGPT/Claude/Mistral/Gemini/...) to generate new, labeled dataset points. Train your BERT model with it + the 32 labels. Analyze how this impacts model metrics.
- **e. Optimal Technique Application (0.5 points)**:
  Based on the previous experiments, apply the most effective technique(s) to further improve your model's performance. Comment your results and propose improvements.

4 Part 3: State of the Art Comparison (2 points)

- **Objective**: Benchmark your model against the SOA with the full dataset now available.
- **Tasks**:
  - **a. Full Dataset Training (0.25 points)**:
    Incrementally train your model with varying percentages of the full dataset (1%, 10%, 25%, 50%, 75%, and 100%). Record the results.
  - **b. Learning Curve (0.25 points)**:
    Plot a learning curve based on the training data percentages.
  - **c. Technique Comparison (0.5 points)**:
    Incorporate the techniques tested in Part 2 into your training schema for comparison.
  - **d. Methodology Analysis (1 point)**:
    Analyze and compare all methods employed. Discuss the effectiveness and limitations observed.

5 Part 4: Model Distillation/Quantization (3 points)

- **Objective**: Reduce the computational requirements for deploying your model.
- **Tasks**:
  - **a. Model Distillation/Quantization (1.5 point)**:
    Distill/Quantize your best-performing model into a lighter model. Document the

process and tools used.

- **b. Performance and Speed Comparison (0.5 points)**:
  Evaluate the distilled model's performance and inference speed compared to the original. Highlight key findings.
- **c. Analysis and Improvements (1 point)**:
  Analyze deficiencies in the student model's learning. Suggest potential improvements or further research directions.

6 Submission Guidelines

- Document your analysis, code, and findings in one or several Jupyter notebooks.
- Host all your results on a github repo.
- Make an executive summary of your objective, main findings and results. Don't abuse on LLMs!
- Add a random seed at the beginning of your code so I can reproduce the code.
- Use markdown cells to elaborate on your decision-making process and insights at each step.
- Final presentation will take place on June 17th and 18th with a quick presentation of 10 minutes maximum to discuss the main findings and 5 minutes of questions.

Total Points: 10 points