# Assignment 2: Large Language Models for Text Classification (15 points)

## CS 410/510 Large Language Models, Fall 2024

You will explore open-source Large Language Models (LLMs) including Llama and Phi models. The task is to build a text classifier and assess the strengths and limitations of these models across various settings. By comparing their performance, you will gain insight into how different LLMs handle sentiment analysis.

Your deliverable for this assignment is a PDF report generated from a Google Colab notebook. The report should include your observations, supported by figures, tables, and relevant analysis. Submit the PDF on Canvas. If converting your Colab notebook to PDF cuts off critical content, please directly share the Colab link with the TA (yutao@pdx.edu) and mention the link in the submission comments.

## Instructions

The goal is to build a text classification model that will predict whether a piece of text is *positive*, *negative* or *neutral*.

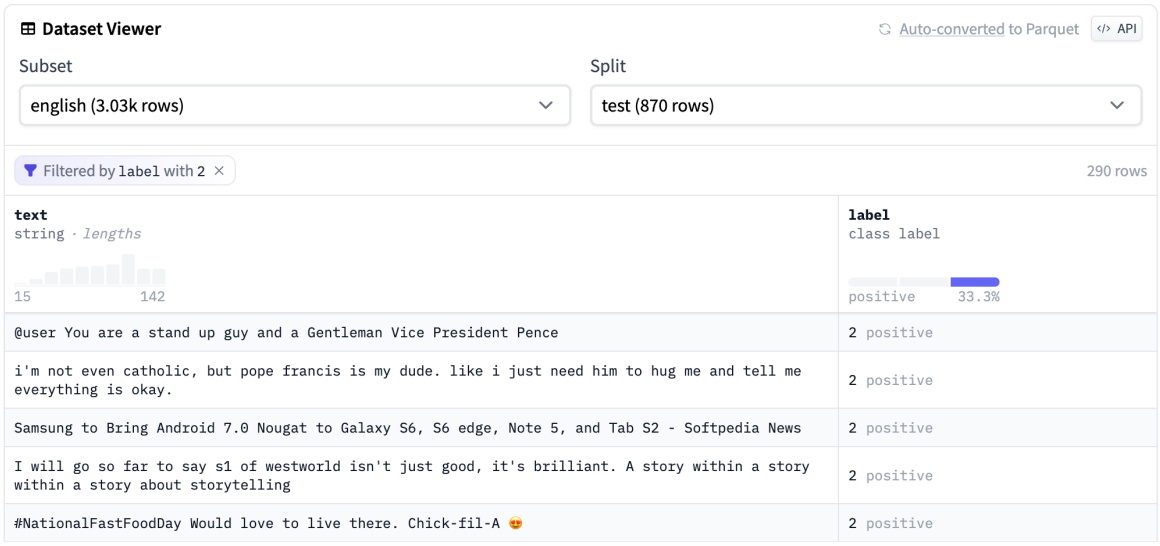
**Models:**
You will continue exploring the following three models in this assignment (same as the previous assignment):

1. Llama-3.2-1B: https://huggingface.co/meta-llama/Llama-3.2-1B
2. Llama-3.2-3B: https://huggingface.co/meta-llama/Llama-3.2-3B
3. Phi-3.5-mini-instruct: https://huggingface.co/microsoft/Phi-3.5-mini-instruct

**Dataset:**
For this task, you'll use only the **English** subset from the Unified Multilingual Sentiment Analysis Benchmark dataset [1], which includes labeled tweets (positive, negative, neutral) in various languages. The dataset is pre-split into:
training set: 1,840 instances
validation set: 324 instances (**use this as your test set**)
test set: 870 instances.

**Important Note**: Since the dataset comes from public websites and is unfiltered, it may contain sensitive or inappropriate content. If you feel uncomfortable with the data, pause the assignment and contact your instructor.

You can access this dataset via Hugging Face's library[1] and explore it through their "Dataset Viewer".




## Experiments and Analysis

For each experiment, use the validation set (324 instances) as your test set and evaluate the model performance using suitable metrics such as precision, recall, and F1-score.

**Experiment 1. Zero-shot inference** Design appropriate and effective prompts to classify the sentiment (positive, neutral, or negative) of the text. Recall that in zero-shot setting no training is required; hence, you will use only the validation set in this experiment.

**Experiment 2. Few-shot in-context learning** Investigate how the model's performance changes when a few demonstrations are provided to the model. Consider setting k = 1 and k = 2. Recall that k = 1 indicates 1 sample *per class.* Use training samples from the train split of the dataset. Remember to use the same validation set for testing for your results to be comparable.

**Experiment 3. Advanced prompting technique (zero-shot inference)** Apply advanced prompting techniques of your choice (chain-of-thought, emotion-based prompting, etc.) under zero-shot setting to see if performance improves. Use the same validation set as in previous experiments for consistency in comparison. Feel free to get creative and explore additional prompting techniques or variations in model settings beyond what's outlined here!

---

[1] https://huggingface.co/datasets/cardiffnlp/tweet_sentiment_multilingual

## Analysis [4 points]

For each experiment, provide:

- A detailed explanation of your approaches (e.g., the prompts used, the hyperparameter settings) [2 points for each experiment].
- Model performance metrics (precision, recall, F1-score) [3 points].

Additionally, plot the results to compare:

- How model family and size affect performance [3 points].
- The differences between zero-shot, few-shot, and advanced prompting [3 points].

Discuss any challenges encountered while conducting the experiments.

All the best!

## References

[1] Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. Barbieri et al. arXiv preprint arXiv:2104.12250 (2021).