Assignment 4: Multilingual Large Language Models (15 points)
CS 410/510 Natural Language Processing, Fall 2023

In this assignment, you will explore multilingual large language models (LLM) for text classification in multiple languages. The goal is to build yet another text classifier that will predict whether a piece of text is "positive" or "negative", but this time, instead of using monolingual PLMs such as BERT, you will explore a multilingual LLM that can process text in several languages.

1. Instructions

Data Preparation: ~~You are provided with a dataset containing customer reviews of products. The dataset is the "Multilingual Amazon Reviews Corpus" [1]. It includes reviews in six languages: **English, Japanese, German, French, Spanish, and Chinese**. For this assignment, we will be using a subset of the original dataset in **all these six languages** and are concerned with the following two columns:~~

~~(a) "review_title" (the title of the review), and~~

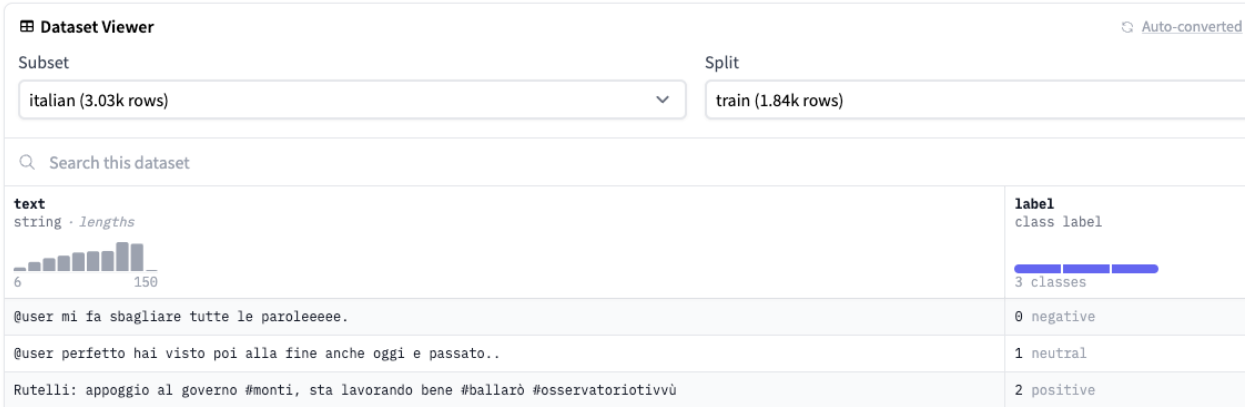
~~(b) "stars" (an integer 1 or 5 indicating the number of stars, which is considered as the sentiment label, with 1 indicating ``negative" and 5 indicating ``positive").~~

It appears that the Multilingual Amazon Reviews Corpus data is defunct since a few days ago. So, we will pivot to another multilingual dataset called "Unified Multilingual Sentiment Analysis Benchmark" [1]. This dataset contains tweets labeled with sentiment in eight different languages – Arabic, English, French, German, Hindi, Italian, Portuguese, and Spanish. The dataset contains labels for 'positive', 'negative', and 'neutral'.

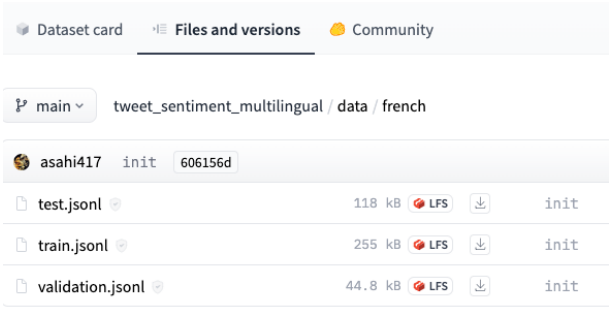
Warning: *As is the case with most 'natural language' text, this data was collected from public websites and is mostly unfiltered. Therefore, it is possible that some text may be disturbing or you may not agree with it.*

You may access the dataset through the Hugging Face library[1]. The dataset is pre-split into a training set and a test set. For this assignment, you will only need to use the test set. You may explore the dataset using the "Dataset Viewer":

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



You may download the dataset by navigating to "Files and versions":



Remember, you only need to experiment with ~50 instances from the test set for each language, and only the 'positive' and 'negative' classes (so ignore the instances with 'neutral' label).

Multilingual Large Language Models for Text Classification: As part of this assignment, you will experiment with two multilingual LLMs:

- one is an open-source model Llama 2 [2-4] available through HuggingChat[2], and
- the other is a proprietary model OpenAI's ChatGPT[3] [3].

You will need to design appropriate and effective prompts to classify your text input. No training is required; hence, you will only use the test set in this assignment. Considering that you'll be using these models through a graphical user interface rather than an API, you may choose to experiment with a small subset of the test set (e.g., 50 instances). Whatever size you choose, remember to keep the test set balanced across both the classes. Then, evaluate the model's performance on this test set using appropriate evaluation metrics (e.g., precision, recall, F1-score).

---

[2] https://huggingface.co/chat

[3] https://openai.com/blog/chatgpt

Results and Analysis: Provide a detailed analysis of your model's performance including the F1-score and other relevant metrics. Answer the following questions:

(i) How do the two LLMs perform? Which one is better? Any possible explanation?

(ii) Comparing the results across the six different languages, what do you observe? Any possible explanation?

(iii) What challenges did you face?

2. Submission Guidelines and Grading

Submit a PDF report of your work showing all the steps on Canvas, under "Assignment 4". Include any additional files or resources used. Ensure your process is well-documented and organized.

Your assignment will be assessed based on the following criteria:

- data preparation, prompt engineering (5 points),
- model evaluation, analysis, and discussion (8 points),
- and overall clarity and organization of the assignment (2 points).

Good luck!

References

[1] Barbieri, Francesco, Luis Espinosa Anke, and Jose Camacho-Collados. "Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond." arXiv preprint arXiv:2104.12250 (2021).

[2] Meta, 'Introducing LLaMA: A foundational, 65-billion-parameter large language model,' Meta AI, 2023. https://ai.meta.com/blog/large-language-model-llama-meta-ai/.

[3] Meta, 'Meta and Microsoft Introduce the Next Generation of Llama,' Meta AI, 2023. https://ai.meta.com/blog/llama-2/.

[4] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[5] OpenAI. "GPT-4 Technical Report." ArXiv abs/2303.08774 (2023).