Assignment 3: Multilingual Sentiment Analysis (15 points)

CS 410/510 Large Language Models, Fall 2024

In this assignment, you will explore any proprietary multilingual large language model (LLM) of your choice. The goal is to classify text in diverse languages, analyze outputs, and assess the strengths and limitations of these models. By comparing their performance, you will gain insight into how different LLMs handle sentiment analysis across various languages.

Your deliverable for this assignment is a PDF report discussing your observations, supported by figures, tables, and relevant analysis. Submit the PDF on Canvas.

Instructions

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

Models:
You will explore any one proprietary LLM of your choice in this assignment (e.g., OpenAI's ChatGPT or Google's Gemini).

Dataset:
For this task, you'll use tweets from **six languages** (French, German, Hindi, Italian, Portuguese, and Spanish) from the Unified Multilingual Sentiment Analysis Benchmark dataset [1], which includes labeled tweets (positive, negative, neutral) in various languages. You can access this dataset via Hugging Face's library[1]. The dataset is pre-split into:

- training set: 1,840 instances
- validation set: 324 instances (use for testing)
- 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.

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

Experiments and Analysis

For each experiment, use **50 samples (25 from the 'positive' class and 25 samples from the 'negative' class) from the validation set for each language** and evaluate the model performance using metrics such as precision, recall, and F1-score. Note that we are not using the 'neutral' class in this assignment.

1. **Dataset annotation [3 points]**: Before diving into model evaluation, let's first understand our dataset thoroughly. Describe how the dataset was created and identify any weaknesses in the annotation process. Provide examples to illustrate your points.

2. **Multilingual sentiment analysis**:
   a) **Multilingual prompting [6 points]**: Investigate methods for adapting the LLMs to various languages. Determine whether English prompting or language-specific prompting yields better results across languages. Report results for all six languages. Describe what you observe and discuss your findings.
   b) **Explainable AI [3 points]**: Design your system such that it not only identifies sentiment but also explains its reasoning behind the classification by highlighting specific words or phrases that influence the sentiment prediction, or generating explanations in different languages. Do you find the explanations to be satisfactory? Conduct this qualitative analysis on a handful of samples from any one language of your choice.
   c) **"Smaller" open-source LLMs accessed through an API vs. larger proprietary LLMs accessed through a UI [3 points]**: Reflect on the differences in performance, ease of use, and customization when comparing the proprietary model used in this assignment to those used in assignment 2 (Llama-3.2 1B/3B and Phi-3.5 mini). Consider aspects like flexibility in adapting prompts, computational requirements, and model interpretability. Did you notice significant trade-offs in accuracy or functionality between the two approaches? How might the access method (API vs. UI) impact the workflow and insights for different types of sentiment analysis tasks?

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).