# Fine-Tuning and Inference on Multilingual Models

Report for the Assignment-2 of Introduction to Natural Language Processing course

Ahmed Said Gençkol [1]

[1] *Middle East Technical University, Ankara, Turkiye*

## 1. Approach

### 1.1.Objectives

There are two main objectives in this assignment. First we need to Fine tune a base large language model for and evaluate it. Secondly we need to evaluate an instruction-tuned large language model. Both objectives are done on two given tasks whose descriptions are given in the pdf.

### 1.2.Setup

I have chosen the Turkish parliament speeches dataset for both tasks. General setup of objectives are given below:

Fine-Tuning Setup:

1. Tasks:
   a. Task 1: Speaker's Ideology (Binary Classification: Left vs. Right)
      i. Data: text_en (English-translated text)
   b. Task 2: Speaker's Party Status (Binary Classification: Governing vs. Opposing)
      i. Data: text (Original language)
2. Model: XLM-RoBERTa (Masked Language Model)
3. Preprocessing:
   a. Tokenization: Used AutoTokenizer with padding and truncation (max_length=512).
   b. Class Balance: No balancing was done as the dataset was fairly balanced..
   c. Data Splits: Stratified splits into training, validation, and test sets.
      i. Training: 80%, Validation: 10%, Test: 10%.

Inference Setup:

1. Model: LLaMA-3.1-8B (Causal Language Model)
2. Evaluated cross-lingual capability using:
   a. text_en: English-translated text
   b. text: Original language
3. Prompt Engineering: Task-specific prompts designed for clarity and expected binary output (0 or 1).
4. Decoding Strategy: Used deterministic settings (do_sample=False) for consistent outputs.
5. Preprocessing:
   a. Small datasets that have equal label proportions(50:50) were created from the original ones and evaluation was done on those datasets(for performance reasons).

### 1.3.Datasets

#### 1.3.1. Original dataset

Original dataset has two files, one for ideology and one for power status.(Note that labels are relatively balanced)

Dataset for ideology:

- Size:16140
- Label Distribution: 58% Right (1), 42% Left (0).

Dataset for power:

- Size:17384

- Label Distribution: 51% Governing (0), 49% Opposing (1).

## 1.4.Experimental Setup

### 1.4.1. Fine-Tuning

Fine-Tuning:

- Batch Size: 32
- Epochs: 3
- Optimizer: AdamW
- Learning Rate: 2e-5 with linear warmup (10% of total steps)
- Scheduler: Linear decay
- Weight Decay: 0.01
- Evaluation Metrics: Accuracy and F1-Score
- Early Stopping: Triggered if no improvement in validation loss for 3 epochs.
- Distribution: 51% Governing (0), 49% Opposing (1).

3 epochs with the given learning rate seems to be the sweet spot hence the model improves steadily up until around 2.7 epochs. There was no overfitting but one thing I could not resolve was training loss was higher than evaluation loss. Either there is a bug or some misconfiguration.

### 1.4.2. Inference

Inference:

- Decoding Strategy: Deterministic (do_sample=False)
- Max New Tokens: 1
- Well structured prompt

A more deterministic approach (i.e. no temperature or top_p,top_k etc.)used here to avoid unexpected outputs. The model tries to classify the speeches hence there is no point increasing randomness.

# 2. Results

Results are given in a table format for each objective

## 2.1.Fine-Tuning

**Table 1**

Fine-Tuning for Ideology(Based on test dataset)

| Task Name | Loss | Accuracy | f1-score |
|---|---|---|---|
| Ideology( original) | 0.33 | 0.86 | 0.88 |
| Power(english) | 0.34 | 0.85 | 0.86 |

## 2.2.Inference

**Table 2**

Evaluation of inference(Ideology original)

| Type | Precision | Recall | f1-score |
|---|---|---|---|
| label 0 | 0.56 | 0.90 | 0.69 |
| label 1 | 0.75 | 0.30 | 0.43 |
| macro avg | 0.66 | 0.6 | 0.56 |

**Table 3**

Evaluation of inference(Ideology English)

| Type | Precision | Recall | f1-score |
|---|---|---|---|
| label 0 | 0.67 | 0.74 | 0.70 |

| | | | |
|---|---|---|---|
| label 1 | 0.71 | 0.63 | 0.67 |
| macro avg | 0.69 | 0.69 | 0.68 |

**Table 4**

Evaluation of inference(Power original)

| Type | Precision | Recall | f1-score |
|---|---|---|---|
| label 0 | 0.87 | 0.62 | 0.72 |
| label 1 | 0.70 | 0.90 | 0.79 |
| macro avg | 0.78 | 0.76 | 0.76 |

**Table 5**

Evaluation of inference(Power English)

| Type | Precision | Recall | f1-score |
|---|---|---|---|
| label 0 | 0.86 | 0.43 | 0.58 |
| label 1 | 0.62 | 0.93 | 0.75 |
| macro avg | 0.74 | 0.68 | 0.66 |

### 2.3.Comparison

Based on the results, the fine tuned model works better than the zero-shot inference model. The inference model gives different results for both languages, surprisingly in ideology it performs better on English and in power it performs better on original(Turkish).

## 3. Discussions

Fine-tuning a multilingual masked language model and evaluating a causal language model provided valuable insights into their performance. For the **Ideology Task**, fine-tuning achieved an F1-score of 0.88 on the original language dataset, significantly outperforming zero-shot inference with LLaMA-3.1-8B, which yielded F1-scores of 0.56 (original text) and 0.68 (English-translated text). Similarly, for the **Power Task**, the fine-tuned model showed strong performance with an F1-score of 0.86, while LLaMA achieved 0.76 and 0.66 for original and English-translated text, respectively. These results highlight the importance of fine-tuning for task-specific training and robust performance across languages.

The datasets were relatively balanced (e.g., 58:42 for Ideology and 51:49 for Power), so no additional balancing techniques were applied. Fine-tuned models handled both datasets well, benefiting from balanced label distributions and effective stratified splits.

Zero-shot inference revealed limitations in LLaMA's cross-lingual capability. For the Ideology Task, it performed better on English text, likely due to pretraining biases favoring English. Conversely, in the Power Task, it performed better on the original language, suggesting that governing and opposing language patterns were easier for the model to generalize. However, LLaMA struggled with short, deterministic outputs, and its performance was sensitive to decoding strategies and prompt engineering.

While fine-tuned models consistently outperformed zero-shot inference, their computational cost and sensitivity to hyperparameter tuning are notable limitations. To improve zero-shot inference, incorporating few-shot prompting, dataset augmentation, and advanced decoding strategies such as beam search could enhance consistency and cross-lingual generalization.

## 4. Github Link

You can find the project files here.