CENG463 fall 2024: Homework 2

Ayşe Aysu Cengiz 2580371

1. Approach

1.1. Environment

The environment I used was free Google Colab with T4 GPU. I had 12.7GB system RAM and 15GB GPU RAM.

1.2. Data

Turkey was chosen as the country due to its large data size, which gives a more robust result than the other datasets. Moreover, the class distribution was nearly even. For both tasks, the distribution of test dataset was as follows: 51.4% label 0, 49.6% label 1. The test dataset was not used as it did not include labels.

For splitting the data, these two guides were used which were present in the homework pdf: Guide for sklearn.model_selection.train_test_split, guide for stratify. Using a pandas dataframe was the easiest for applying splitting function. Applying stratification ensured that the even class distribution for the overall dataset was preserved both in train and test dataset. The data was splitted as the train being 90% and test being 10%

1.3. Masked LM

As a masked LM, XLMRoberta was chosen to be able to compare different languages without changing much in the code. Though other languages will not be discussed in this report nor they are present in the finalized code, the topic will be mentioned slightly in the discussion section.

For fine-tuning, again the guide given in the homework pdf was followed: Fine-tune a pretrained model.

The followed steps can be seen below:

- Tokenize
- Turn pandas dataframe into a tensor and then to dataloader
- Determine the device
- Fine-tune
- Evaluate

CENG463 fall 2024: Homework 2

1.4. Causal LM

For causal language models, choosing a model was more tricky. The resources at hand resulted in a very limited selection of causal models. Models such as Bloom and Llama crashed while trying to get their classifiers due to the limited RAM at hand. Because of this factor, distilgpt2, which is not primarily a multilanguage model was chosen. Since the model is not inherently trained for Turkish, it may have some disadvantages.

Moreover, in order to see the difference between fine-tuning and zero-shot clearer, four types of evaluations were made:

- Original language zero-shot
- Original language fine-tuning
- English zero-shot
- English fine-tuning

The steps are similar to Masked LM for fine-tuned models. For evaluation pipeline function from transformers was used. Again, Pipelines for inference guide given in the pdf was used. As the task name "text-classification" was chosen for the fine-tuned models and "zero-shot-classification" for zero-shot ones.

2. Process

3. Task 1: Orientation

In the orientation task Turkish orientation dataset was used. While fine-tuning the masked LM, the preffered label was "text", meaning the fine-tuning was made with the text in the original language, Turkish.

3.1. Results

1. Masked LM
    - Original Language - fine-tuned: 86.55 %
2. Causal LM
    - Original Language - zero-shot: 58.05%
    - Original Language - fine-tuned: 66.48%
    - English - zero-shot: 46.03%
    - English - fine-tuned: 79.74%

4. Task 2: Power

In the power task, again, Turkish orientation dataset was used. While fine-tuning the masked LM, the preffered label was "text_en", meaning the fine-tuning was made with the text in English.

Masked LM Original Language - fine-tuned: 88.72 Causal LM Original Language - zero-shot: 54.05Original Language - fine-tuned: 68.14English - zero-shot: 51.01English - fine-tuned: 80.33

4.1. Results

1. Masked LM
   - English - fine-tuned: 88.72%
2. Causal LM
   - Original Language - zero-shot: 54.05%
   - Original Language - fine-tuned: 68.14%
   - English - zero-shot: 51.01%
   - English - fine-tuned: 80.33%

5. GitHub Repository

You can reach the repository via aysucengiz/ceng463-hw2. Resulting files are in main branch while task1 branch includes some trials I did with other languages and other approaches. Llama branch has the attempt of importing Llama-2-7B-hf model, which resulted in a crash due to the limited RAM.

6. Discussion

The results mostly align for both of the tasks. Below are the comparisons for various results.

1. **Masked (English and Turkish):** For both languages, the result is quite close, with English being slightly higher. This slight difference might be due to the dataset difference or English being a higher resource language than Turkish.
2. **Causal (Zero-Shot and Fine-Tuned):** As expected, when the causal model is fine-tuned its results go from random to robust. While the zero-shot results differed greatly every time the same model was evaluated, such variation was not present in the fine-tuned models.
3. **Causal (English and Turkish):** For both tasks, English fine-tuned model performs better than the Original Language fine-tuned model. Since gpt2 is mainly an English model, unlike XLMRoberta, the difference between the accuracies when the models are fine-tuned is as expected and the modal is unable to reach high accuracies in the Turkish dataset. An inherently multilingual causal model would be more accurate in labeling Turkish data. However, for zero-shot models such a pattern cannot be observed as zero-shot models label the data in a rather randomized fashion.
4. **Masked and Causal:** In the Turkish dataset, it is clear that the masked model performs better than the causal model even when the causal model is fine-tuned. However, for smaller datasets such as Spanish, masked model sometimes performed very low, dropping down to 55% accuracy. Which shows the importance of the data size.

As a result, it can be said that the masked multilingual model performs very good when fine-tuned, independent from the language and the zero-shot evaluation results in about a 50% accuracy which is as good as guessing.