CENG463 - HW2 Report

Furkan Numanoğlu[1,*,†]

[1]*Middle East Technical University (METU), Ankara/Türkiye*

Abstract

A clear and well-documented LaTeX document is the report document on HW2 in the METU CENG463 course. Based on the "ceurart" document class, this article presents and explains most of the common variations as well as most of the formatting elements an author can use in preparing documentation of their work.

1. Approach

This study implements a comprehensive analysis of parliamentary speeches using two distinct approaches. For fine-tuning tasks, I employed XLM-RoBERTa-base as multilingual masked language model, specifically chosen for its robust performance in handling complex political discourse. The implementation uniquely leverages both original Turkish texts and their English translations, allowing for cross-lingual performance evaluation. For ideology classification, I utilized English translations (text_en), while power identification was trained on the original Turkish texts.

For zero-shot inference, I employed the Llama-3.1-8B model, configured with 4-bit quantization for efficient inference. The zero-shot approach was implemented through a carefully designed pipeline using custom prompts tailored to each classification task. This dual-model strategy enables to evaluate both traditional fine-tuning and emerging zero-shot approaches in political text analysis.

2. Dataset Statistics

The analysis utilized the Turkish parliamentary dataset from ParlaMint, revealing the following distributions:

For ideology classification dataset:

- Total samples: 16,138 speeches
- Training set: 14,524 samples (90%)
- Validation set: 1,614 samples (10%)

*Corresponding author.

†These authors contributed equally.

furkan.numanoglu@metu.edu.tr (F. Numanoğlu)

https://github.com/furkannumanoglu/ (F. Numanoğlu)

CEUR Workshop Proceedings (CEUR-WS.org)

- Class distribution maintained through stratified splitting
- Imbalanced class distribution observed (58% - 42%)

For power identification dataset:

- Total samples: 17,384 speeches
- Same 90-10 stratified split strategy applied
- Balanced class distribution observed

To achieve stratified splitting, I utilized the train_test_split function from the sklearn.model_selection module. By setting the stratify parameter to df['label'], I instructed the function to split the data in a stratified manner based on the values in the 'label' column of the DataFrame. This ensures that the class proportions are preserved in the resulting training and validation sets.

By applying stratified splitting, I aimed to create a more reliable and unbiased evaluation of the model's performance, especially in scenarios where class imbalance is present.

3. Experimental Setup

Fine-tuning Configuration:

- Batch size: 8 (When I selected 16 or more, I received an error due to my system specifications. For this reason, I selected the highest limit that did not cause any problems.)
- Training epochs: 10
- Learning rate: 1e-5
- Warmup ratio: 0.1
- Weight decay: 0.05
- Evaluation strategy: Per epoch
- Model saving: Best model based on F1 score

Zero-shot Configuration:

I tried various numbers and possibilities for configuration. You can see some of the attempts in the code, while some remain as my own attempts.

4. Results and Discussion

4.1. Fine-tuned Model Results

Ideology Classification (English text):

- Peak accuracy: 81.91% (Epoch 7)
- Best F1 score: 0.845 (Epoch 7)
- Best Precision: 0.842 (Epoch 7)
- Best Recall: 0.849 (Epoch 7)
- Training loss: Steady decrease from 0.557 to 0.075

Power Classification (Turkish text):

- Peak accuracy: 82.5% (Epoch 9)
- Best F1 score: 0.833880 (Epoch 9)
- Best Precision: 0.815 (Epoch 9)
- Best Recall: 0.853 (Epoch 9)
- Training loss: Steady decrease from 0.58 to 0.063

4.2. Zero-shot Results

The zero-shot approach using Llama-3.1-8B demonstrated consistent performance across both tasks and languages:

- Ideology Task:
  - English text accuracy: Around 46%
  - Turkish text accuracy: Around 45%
- Power Task:
  - English text accuracy: Around 50%
  - Turkish text accuracy: Around 50%

4.3. Comparative Analysis

The fine-tuned XLM-RoBERTa model demonstrated superior performance compared to the zero-shot approach, achieving higher accuracy and more stable predictions. While the fine-tuned model reached 81.91% accuracy for ideology classification, the zero-shot approach achieved around 45% accuracy. This performance difference suggests that task-specific training remains valuable for political text analysis, but that zero-shot extraction offers a viable alternative when training resources are limited.

My zero-shot approach may be yielding low accuracy due to complex task, ineffective prompt, suboptimal model choice, lack of fine-tuning, and evaluation metrics. To improve, I'll experiment with prompts, expand the dataset, try different models, consider fine-tuning, and use multiple evaluation metrics.

Key observations:

- Fine-tuned models showed more consistent performance
- Zero-shot approach provided unsatisfactory results without task-specific training
- Cross-lingual performance remained stable across both approaches

Implementation details and complete source code are available at: [https://github.com/furkannumanoglu/CENG463]

References

A. Resources

- HW2 GitHub Link,
- Google Colab