Analysis of Political Orientation and Power Classification in Parliamentary Debates in the Grand National Assembly of Turkey

Meriç Buğra Haliloğlu[1]

[1]*Middle East Technical University, Turkey*

1. Approach

In this study, I address two classification tasks: identification of political ideologies and classification of power orientation in parliamentary debates, specifically in the Grand National Assembly of Turkey. The first task involves determining whether a speaker's political party leans left or right, while the second focuses on identifying whether a speaker belongs to the governing party (or coalition) or the opposition.

I used FacebookAI's "XLM-RoBERTa" multilingual model for both classification tasks, using its ability to handle text in multiple languages. XLM-RoBERTa is a multilingual version of "RoBERTa". It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. RoBERTa is a transformer model pre-trained on a large corpus in a self-supervised fashion. This means it was pre-trained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. In addition, for inference purposes, I used "llama3.2:1b", an open-source LLM, via "Ollama". I tested the model's capability of classification on both the Turkish and the English texts.

Each task was treated as a binary classification problem, with dedicated training and evaluation datasets. I utilized the "Transformers" library by Hugging Face to perform the fine-tuning, and used "SKLearn" for model evaluation purposes. Furthermore, for inference tasks, I used "ollama" library for creating responses and "SKLearn" for evaluation purposes. Additionally, during fine-tuning tasks, I used the "Weights and Biases" library to track the training processes of my models. WandB library gives useful insights on how the training occurs by showing changes in loss, gradient, and learning rate in real time during training. Also it shows how evaluation statistics change during evaluation phase.

Further explanations on the experiment's scripts and discussion is available on the GitHub repo: https://github.com/merichoglu/CLEF2025.

meric.haliloglu@ceng.metu.edu.tr (M. B. Haliloğlu)

CEUR Workshop Proceedings (CEUR-WS.org)

1.1. Training Configuration

The fine-tuning process was configured using the `TrainingArguments` class. The model was trained for 3 epochs with a learning rate of 2×10^{-5}, a batch size of 16 for both training and evaluation, and a weight decay of 0.01. Evaluation and checkpoint saving were performed at the end of each epoch, with only the best model retained. Logs were recorded every 10 steps and monitored using Weights & Biases under the `ParliamentaryDebates` workspace.

1.2. Inference Configuration

Inference was performed using the `ollama` library for generating model predictions and `SKLearn` for evaluation. Custom prompts were designed for each task: identifying political ideology (Left/Right) and power classification (Coalition/Opposition). A structured logging system was implemented using Python's `logging` module to track progress and excluded samples.

The evaluation pipeline iterated through test datasets, generating predictions via LLaMA-3.2:1b and handling unrecognized outputs by logging and exclusion. Metrics such as accuracy and classification reports were computed using `SKLearn`. Dataset samples were optionally reduced using stratified sampling for testing, and progress was monitored with the `tqdm` library.

2. Dataset Statistics and Splitting

The dataset contains a selection of speeches from ParlaMint corpora (version 4.0) as the training set for the shared task on *Ideology and Power Identification in Parliamentary Debates* in CLEF 2024.

The number of training instances and the class imbalance differ for each training set. A fixed validation split wasn't provided. I decided to proceed with the datasets related to Turkey and trained my models on these datasets.

The power and orientation dataset for Grand National Assembly of Turkey has the following characteristics:

- Orientation dataset contains 16138 entries.
- Power dataset contains 17384 entries.

As requested, I performed a 9:1 stratified split of both datasets, and used the relevant parts for training and testing of my models.

2.1. Class Balance

- **Orientation Training Set**: Class 0 (left) accounted for 42%, and Class 1 (right) accounted for 58%.
- **Orientation Testing Set**: Class 0 (left) accounted for 42%, and Class 1 (right) accounted for 58%.
- **Power Training Set**: Class 0 (coalition) accounted for 49%, and Class 1 (opposition) accounted for 51%.
- **Power Testing Set**: Class 0 (coalition) accounted for 49%, and Class 1 (opposition) accounted for 51%.

Considering the political state of Turkey, it makes sense to have the imbalance in orientation task dataset because Grand National Assembly of Turkey consists mostly of right-wing parties. Since the class imbalance is not huge, I decided not to do anything to handle it. Possible ways of addressing this imbalance:

- Under sampling the majority class
- Over sampling the minority class

3. Results

3.1. Model Performance

Both fine-tuned XLM-RoBERTa models and Llama-3.2:1b were evaluated for both tasks using accuracy, precision, recall, and F1-score metrics. The evaluation graphs and tables for all tasks are given below.

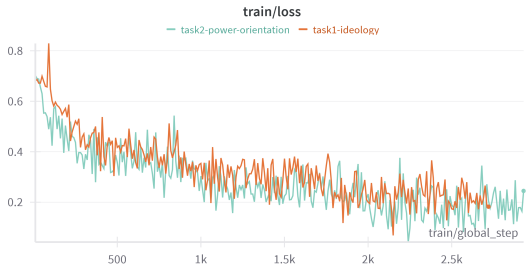

Figure 1: Graph displaying training loss for both tasks

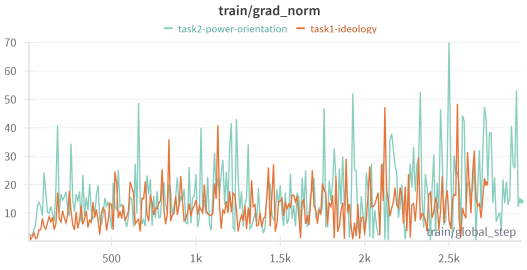

Figure 2: Graph displaying changes in gradient for both tasks

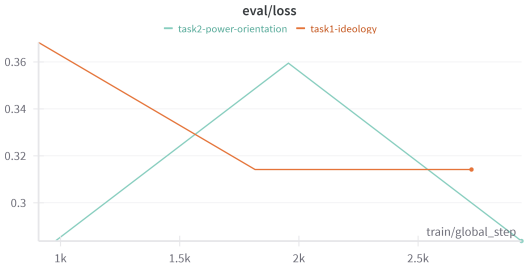

Figure 3: Graph displaying evaluation loss for both tasks

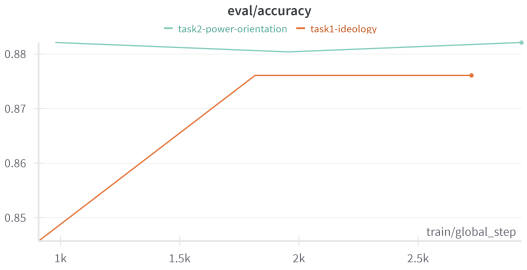

Figure 4: Graph displaying accuracy for both tasks

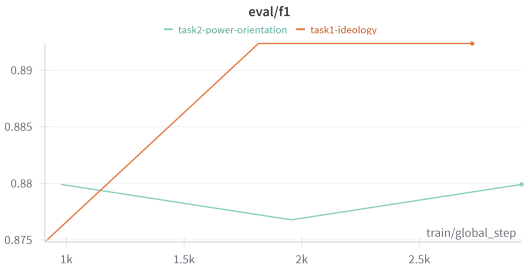

Figure 5: Graph displaying F1 Score for both tasks

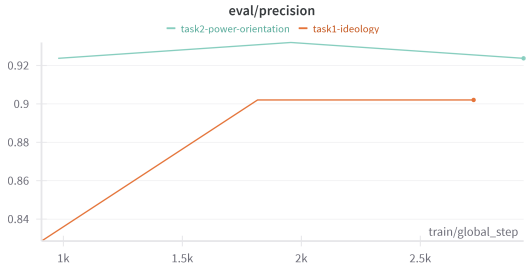

Figure 6: Graph displaying Precision for both tasks

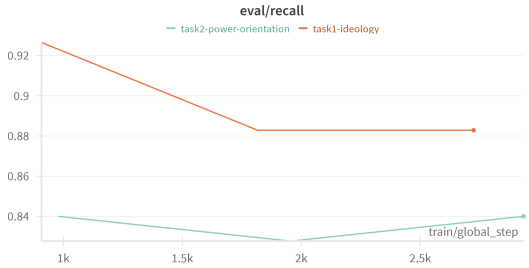

Figure 7: Graph displaying Recall for both tasks

Class	Precision	Recall	F1-Score	Support
Left	0.42	0.82	0.56	630
Right	0.63	0.22	0.32	894
Accuracy			0.47	1524
Macro avg	0.53	0.52	0.44	1524
Weighted avg	0.54	0.47	0.42	1524

Table 1
Performance metrics for inference on Turkish text for political orientation.

Class	Precision	Recall	F1-Score	Support
Left	0.42	0.76	0.54	668
Right	0.58	0.24	0.34	933
Accuracy			0.46	1601
Macro avg	0.50	0.50	0.44	1601
Weighted avg	0.52	0.46	0.42	1601

Table 2
Performance metrics for inference on English text for political orientation.

4. Discussion

The datasets were relatively balanced, reducing the need for additional class-balancing techniques. The XLM-RoBERTa model demonstrated robust performance in both tasks. Task 2 yielded slightly better results in terms of accuracy, likely due to fewer ambiguities in defining

Class	Precision	Recall	F1-Score	Support
Coalition	0.58	0.10	0.17	760
Opposition	0.51	0.93	0.66	775
Accuracy			0.52	1535
Macro avg	0.55	0.51	0.42	1535
Weighted avg	0.55	0.52	0.42	1535

Table 3
Performance metrics for inference on Turkish text for power in the parliament.

Class	Precision	Recall	F1-Score	Support
Coalition	0.51	0.05	0.10	832
Opposition	0.51	0.95	0.67	878
Accuracy			0.51	1710
Macro avg	0.51	0.50	0.38	1710
Weighted avg	0.51	0.51	0.39	1710

Table 4
Performance metrics for inference on English text for power in the parliament.

power orientation compared to political ideology. However, the F1 score of first task was higher due to significantly higher recall. Since XLM-RoBERTa is a state-of-the-art multilingual model, it is not surprising that it performs well on both English and Turkish texts.

The surprising result of the experiments was the performance of Llama-3.2:1b. It performed extremely poor for a large language model, a strong one, with at most 0.52 F1 Score over all tasks. First of all, the model declined to answer most of the classification questions, stating that it is unethical for a language model to assign political labels to people. As a result of this, most of the test instances were dropped, so Llama was tested with fewer instances than XLM-RoBERTa. Another shock for me was that Llama performed better on Turkish data than on English data. LLaMA-3 appears to have stricter ethical constraints when dealing with English political texts, as seen in its refusal to classify political labels in English. These refusals resulted in fewer test samples which led to skewed evaluation metrics. Also, the Turkish dataset contains original texts, while the English dataset involves machine-translated versions of Turkish parliamentary speeches. Translation artifacts, idiomatic loss, or contextual distortions in the English texts probably contributed to LLaMA-3's underperformance on English data.

For further improvements, possible steps to take:

- Experiment with advanced prompting techniques, such as chain-of-thought (CoT) prompting.
- Experiment with stronger LLMs that have more parameters to improve classification accuracy.