CENG463-THE2

Aybüke Aksoy

January 5, 2025

1 Approach

The solution involves fine-tuning multilingual masked language models and using multilingual causal language models during inference for binary classification on parliamentary debates. The workflow includes:

- **Data Loading:** Retained relevant columns (`text`, `text_en`, `label`).
- **Data Splitting:** Stratified split with 95% training and 5% testing to preserve label distribution.
- **Model Selection:** BERT (masked language model) was fine-tuned for orientation (`text`) and power (`text_en`) classification. Llama-3.1-8B (causal language model) was used in a zero-shot setting for both tasks.
- **Evaluation:** Metrics (Accuracy, Precision, Recall, F1-Score) were used for task comparison.

2 Dataset and Balancing

The dataset, derived from the ParlaMint corpus, contains parliamentary debates from multiple European countries. For this assignment, debates from the **Netherlands** were selected. The dataset includes the following important fields:

- `text` and `text_en`: The speech in the original language and its English translation.
- `label`: A binary value indicating political ideology (0 for left, 1 for right) or political orientation (0 for governing, 1 for opposition).

The datasets for both tasks are not balanced, as indicated by the label distributions:

- **Orientation Task:** The dataset contains 61.57% samples labeled as `1` (right) and 38.43% as `0` (left), with a maximum class imbalance of 23.15%.
- **Power Task:** The dataset contains 58.50% samples labeled as `0` (government) and 41.50% as `1` (opposition), with a maximum class imbalance of 16.99%.

I have not implemented any technique to overcome class imbalance, yet I will be using F1 Score as an evaluation metric to account for the imbalance issue in the next sections.

3 Experimental Setup

The experimental setup includes details about the hyperparameters, training configuration, and inference strategy for both masked and causal language models.

3.1 Fine-tuning of a multilingual masked language model

Model:	Multilingual BERT
Training-Validation Split:	95% training, 5% testing (stratified)
Loss Function:	Binary cross-entropy loss for both tasks
Learning Rate:	2×10^{-5}
Batch Size:	16 (train and evaluation)
Number of Epochs:	10
Weight Decay:	0.01
Best Model Selection Metric:	Validation loss (`eval_loss`), lower is better

3.2 Inference with a multilingual causal language model

Model ID:	`meta-llama/Meta-Llama-3.1-8B-Instruct`	
Decoding Strategy:	Greedy decoding	
Tokenizer:	`AutoTokenizer` from Hugging Face	
Pipeline Parameters:	**Task:**	Text generation
	Max New Tokens:	50
	Temperature:	0.0 (Greedy)
	Sampling:	Disabled (`do_sample=False`)
	Return Full Text:	Disabled (`return_full_text=False`)

4 Results

4.1 Multilingual BERT

The **text** column was used for the orientation model and **text_en** column was used for the power model. The performance results after fine-tuning can be seen on tables below, however they cannot be compared as they utilize different columns.

Class / Avg	Precision	Recall	F1-Score	Support
0	0.44	0.50	0.47	109
1	0.66	0.60	0.63	174
accuracy	—	—	0.56	283
weighted avg	0.57	0.56	0.57	283

Table 1: Classification report for `orientation model`. Test Accuracy = 56.18%.

Class / Avg	Precision	Recall	F1-Score	Support
0	0.65	0.67	0.66	232
1	0.51	0.48	0.49	164
accuracy	—	—	0.59	396
weighted avg	0.59	0.59	0.59	396

Table 2: Classification report for `power_model`. Test Accuracy = 59.09%.

4.2 Llama-3.1-8B

While for orientation task, the inference on text_en performed better, for the power task, usage of the original language Dutch gave slightly better accuracy.

Overall, it can be observed that BERT models have higher F1-scores than Llama-3.1-8B for

Task	Accuracy (%)	Precision	Recall	F1-Score
Orientation (text)	42.76	0.73	0.11	0.19
Orientation (text_en)	47.35	0.90	0.16	0.27
Power (text)	65.66	0.57	0.66	0.62
Power (text_en)	63.38	0.58	0.41	0.48

Table 3: Comparison of performance metrics for both tasks with text and text_en.

both of the classification tasks with 0.57, 0.59 and 0.19, 0.48 respectively (The graphs can be seen on jupyter notebook that is provided at the end of the report)

5 Discussion

5.1 Performance Analysis

The fine-tuned BERT models achieved higher F1-scores for both tasks compared to the zero-shot Llama-3.1-8B.
Fine-tuning BERT allows the model to learn task-specific patterns, such as the vocabulary associated with political ideology or power. This task-specific optimization is crucial for better performance, as reflected in the consistently higher F1-scores.
On the other hand, the Llama-3.1-8B model's lower performance indicates difficulty in identifying all relevant instances, potentially due to its lack of exposure to labeled task-specific data. Although zero-shot models are convenient, fine-tuning Llama-3.1-8B on task-specific data could significantly enhance its performance by enabling it to learn patterns unique to the dataset.

Moreover, the orientation task (left vs. right) appears more challenging than the power task (government vs. opposition), possibly due to the nuanced and context-dependent nature of political ideology compared to the clearer linguistic signals in political roles.
While a stratified split was used to keep balanced label distributions, the small test size (5% of the data) could limit the reliability and generalizability of the results. Larger test sets might provide more stable and representative metrics.

5.2 Language-Specific Observations

The results also highlight the impact of using `text` (original language) versus `text_en` (English translation):

- For the **orientation task**, the Llama-3.1-8B model performed slightly better on `text_en` than on `text`. This might be due to the pre-trained causal language model being more optimized for English text and struggling with the linguistic structures of Dutch.
- For the **power task**, the original `text` in Dutch yielded significantly better F1-score compared to `text_en`. This suggests that certain linguistic markers indicating political roles (e.g., opposition or government) might be potentially lost in translation.

However, as there is no one specific pattern, those conclusions are not reliable and are dependent on the dataset features. The quality of the English translations in `text_en` could vary, potentially introducing noise that affects the model's ability to learn and make accurate predictions. Improving the quality of translations in `text_en` or incorporating language-specific tokenization techniques might improve performance.

6 Repository Link

The source code for fine-tuning and inference, along with a detailed README file, is available at: https://github.com/aybukeaksoy/CENG463-THE2