Assignment 4

Report

Name: Jonas Schrade
Student Number: 01/1080887
Course (Instructor): Deep Learning for Social Science (Giordano Di Marzo)

1 Introduction

This assignment aims to pre-process a large corpus of parliamentary speeches from the Austrian Nationalrat, implement a baseline Large Language Model (LLM), perform domain-specific fine-tuning using LoRA, and apply both models to classify speeches by the speaker's party affiliation and political orientation.

2 Data Analysis

This study utilizes speech data from the Austrian subset of ParlaMint 4.1 (Erjavec et al., 2024), which includes both speech transcripts and metadata.

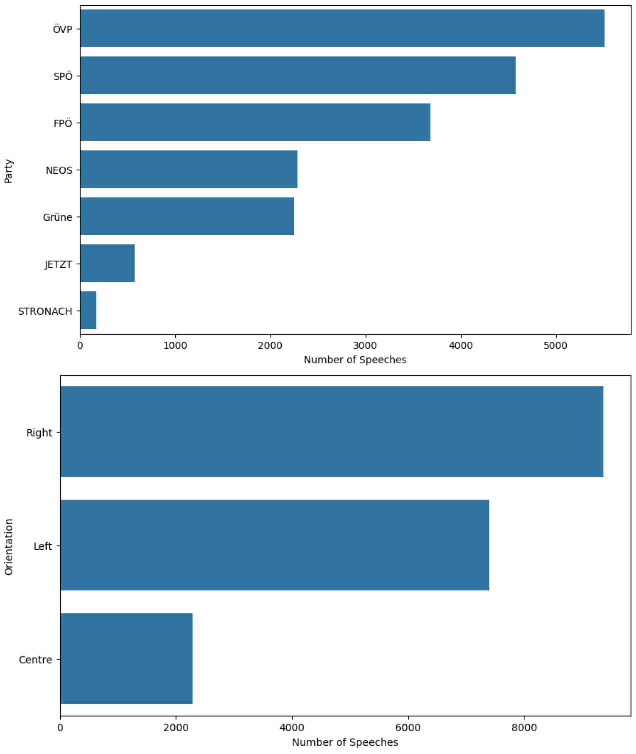

Figure 1: Target variables.

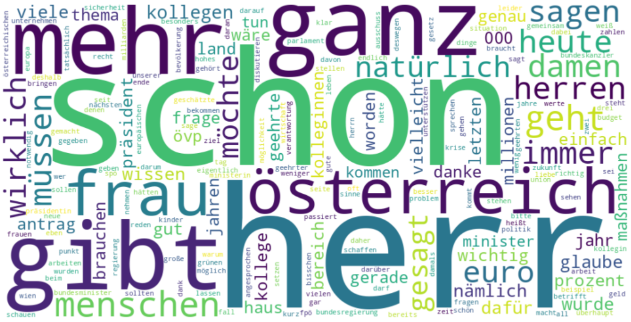

Figure 2: Vocabulary composition.

The analysis is limited to speeches delivered between 2017 and 2022 that contain at least 50 words. Procedural and administrative statements by chairpersons are excluded, and interruptions are removed from the text. The resulting corpus comprises 19,053 unique speeches, annotated with metadata, spanning a vocabulary of 161,910 words and an average length of 555.53 words per speech. Fig 2 presents a wordcloud of the overall corpus. For model fine-tuning and evaluation, the original five-point party orientation scale is recoded into three categories: "left," "centre," and "right," with left- and right-leaning orientations merged accordingly. Figure 1 displays the distribution of target categories in the final dataset. The vocabulary composition largely reflects typical patterns of (political) discursive language. However, analysis by party and ideological orientation indicates a rightward tilt in the Austrian parliament, with the ÖVP playing a comparatively dominant role in parliamentary discourse. Despite this imbalance, all major parties are represented with at least 2,000 speeches, while only the short-lived, activist-oriented parties JETZT and STRONACH exhibit clear underrepresentation.

3 Methods

Both classification task are approached using the Qwen 3 model, a 0.6B-parameter causal decoder-only transformer pretrained on diverse multilingual data, available via HuggingFace.

To adapt Qwen 3 to the parliamentary domain, parameter-efficient fine-tuning is applied using Low-Rank Adaptation (LoRA). The LoRA configuration uses a rank of 8, scaling factor $\alpha = 16$, and dropout of 0.1. Adapters are inserted into the attention projections (q_proj, k_proj, v_proj, o_proj) and feed-forward layers (gate_proj, up_proj, down_proj). All other model weights remain frozen. The LoRA-augmented model is trained using the HuggingFace Trainer-API.

Instruction-style prompts are constructed to align with the generative objective of the model. Each prompt comprises an expert instruction specifying the classification task and permissible label set, a cleaned speech segment, and a label completion slot. During fine-tuning, the correct label is appended as a single-token answer followed by the EOS token. The same prompt format is used at finetune inference.

Tokenization is performed using the Qwen tokenizer with a maximum sequence length of 512 tokens. All examples are padded to this length. To prevent padding tokens in the label sequence from affecting the loss, these positions are masked with -100. The EOS token is included to mark completion.

Fine-tuning is executed over two epochs with a per-device batch size of 2 and gradient accumulation over 4 steps, resulting in an effective batch size of 8. A fixed learning rate of 5×10^{-5} is used, with 20 warm-up steps. Training employed bfloat16 mixed-precision and gradient checkpointing to reduce memory consumption. Checkpoints are saved every 200 steps, and ultimately, the model with the lowest validation loss is selected. Both training and validation losses decreased steadily throughout, suggesting stable convergence and further optimization potential. Improved performance may be achieved through longer training, increased batch size, longer token sequences, or broader hyperparameter tuning, but such extensions were constrained by limited memory and (run-)time resources.

Evaluation conducts next-token logit scoring. For each element in the test set, the model produced logits for the token following the prompt. Candidate labels are pre-tokenized, and their scores are computed as the sum of logits over their initial tokens, enabling multi-token label scoring. The label with the highest score is selected as the model prediction. This scoring procedure is applied identically to both the pretrained and fine-tuned models.

4 Results

Model	Exact Match Accuracy	Avg. Precision	Avg. Recall	F1 Score
Party Classification				
Fine-tuned	0.259	0.191	0.169	0.149
Baseline	0.133	0.159	0.130	0.091
Orientation Classification				
Fine-tuned	0.397	0.339	0.344	0.326
Baseline	0.439	0.330	0.328	0.303

Table 1: Classification performance.

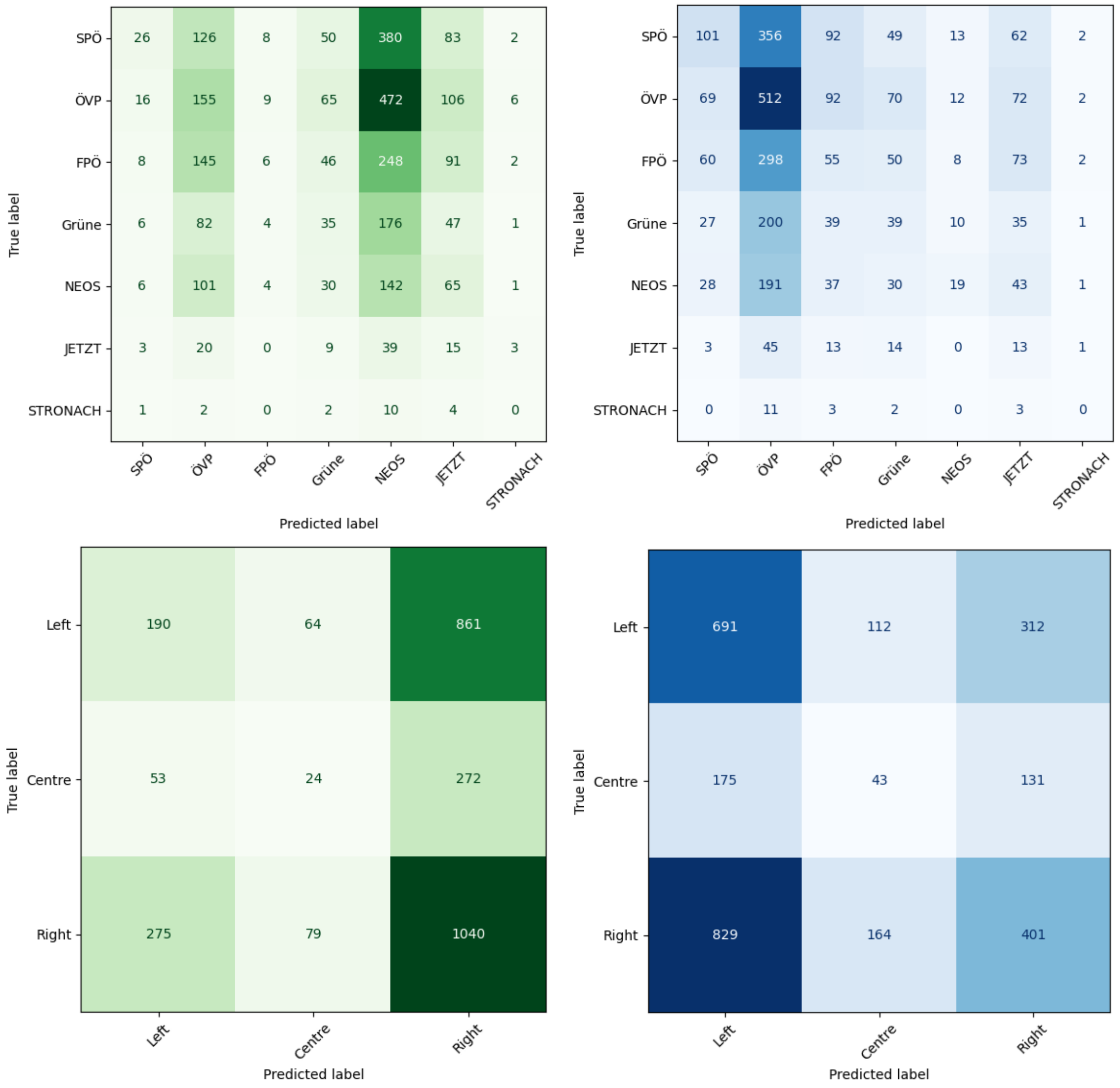

Figure 3: Confusion Matrix: Baseline (Green) vs. Fine-tuned (Blue).

5 Discussion

The overall rather weak classification performance is indicative of the LLM's struggle to exhaustively capture abstract concepts of the political/parliamentary sphere such as party association and political orientation only through speech patterns.

While remaining low, fine-tuning on party memberships improved the baseline's strong overemphasis on NEOS by generally spreading classifications over a broader range of classes and also taking into account the strong parliamentary representation of the ÖVP. A similar pattern, may be responsible for better precision and recall when fine-tuning for orientation classification. While the baseline has a strong leaning towards political right classifications (possible reflecting the NEOS bias), the fine-tuned model again distributes classifications across a broader range. While this comes at a price of accuracy, it benefits precision and recall.

Ultimately, these findings underscore the potential of targeted fine-tuning to reduce systemic biases and improve the model's sensitivity to nuanced political cues embedded in parliamentary language. While performance remains modest, the observed gains in precision and recall suggest that domain adaptation enables large language models to better approximate abstract political attributes especially when classes are limited.