NLP Homework 2: Multilingual Classification of Parliamentary Debates

Beyza Nur Koç

Task1: Orientation Task

Fine-Tune a Multilingual Masked Language Model - Multilingual BERT

Dataset Exploration

For the orientation task, the France (fr) dataset [1] also exhibited class imbalance. The training set included 1093 samples for Class 0 and 2525 samples for Class 1. Oversampling was applied to Class 0, balancing the dataset with 2525 samples in each class. Similar to Task1, the dataset was split into 90% training and 10% validation subsets using stratified sampling [2]. And after that, split 90% training and 10% validation.

Approach Description

I choose to fine-tune multilingual BERT [3] model for this task using the Hugging Face `Trainer` API [4].

Experimental Setup:

Evaluation was conducted at the end of each epoch. Initially, training was conducted on the free version of Google Colab [5] with a single GPU. However, this setup was time-consuming and often interrupted due to resource limitations. To address these issues, the training process was migrated to Google Colab Pro, utilizing an Nvidia A100 GPU with 100 compute units. This upgrade significantly improved the training speed and ensured uninterrupted execution.

Results

The model achieved a validation accuracy of 77.2%, with a precision of 77.5%, recall of 77.2%, and an F1-score of 77.2% as a result of the 3rd epoch. The consistent increase in these performance metrics indicates that the model is learning effectively and generalizing well. Moreover, since the performance metrics are consistent and close to each other, it can be concluded that this fine-tuned model effectively addressed the class imbalance in this task.

Causal Model Results for the Orientation Task

For the orientation task, the evaluation was conducted using the causal language model, `meta-llama/Llama-3.2-1B` [6] in a zero-shot setting. The model's performance was assessed on both the original language text column (`text`) and the English-translated column (`text_en`) to evaluate its cross-lingual capabilities.

Results:

- On the `text` column, the model achieved an accuracy of **53.47%**.
- On the `text_en` column, the accuracy dropped slightly to **50.1%**.

Analysis: The results suggest that the causal model has limited ability to generalize for the orientation task, even when evaluated on English-translated data. The following factors may have contributed to this performance:

- **Zero-shot Setting:** As the model was not fine-tuned on the task-specific dataset, its understanding of task nuances was limited.
- **Language Complexity:** Although the model performed slightly better on the original language, the nuances of the Bosnian language and the complexity of political texts may have impacted the results.
- **Dataset Preparation:** The dataset may require further refinement, such as more balanced examples or additional linguistic features, to better support causal model inference.

Recommendations for Improvement:

- **Fine-tuning:** Training the causal model on the orientation dataset could improve its ability to classify data more accurately.
- **Enhanced Preprocessing:** Employing language-specific preprocessing techniques, such as stemming or lemmatization, might help the model understand non-English text better.
- **Data Augmentation:** Increasing the dataset size by generating synthetic examples or translating additional data can enhance generalizability.
- **Alternative Models:** Experimenting with more recent or specialized causal models could yield better results.

Task2: Power Task

Fine-Tune a Multilingual Masked Language Model - Multilingual BERT

Dataset Exploration
I choose the France (fr) dataset [1] for this homework, due to language support of the LLama model. In the power set, the dataset is not balanced. The training set contained 6178 samples for Class 0 and 3635 samples for Class 1. So, to address this issue, undersampling was applied to Class 0, resulting in a balanced dataset with 3635 samples in each class. The dataset was then split into 90% (training+validation) and 10% validation subsets using stratified sampling [2]. And after that, split 90% training and 10% validation.

Approach Description
I use the same methodology as in Task1 to fine-tune the multilingual BERT [3] model for binary classification.

Experimental Setup:
The hyperparameter configuration was identical to Task1: For the power task, the model was trained using a **learning rate** of 2×10^{-5}, a **batch size** of 16, and for 3 **epochs**. Cross-entropy loss was utilized as the loss function, and optimization was performed using AdamW (including weight decay) to ensure efficient parameter updates. Same as above, evaluation was conducted at the end of each epoch. Training was performed on Google Colab Pro [5] using an Nvidia A100 GPU.

Results
The model achieved a validation accuracy of 72.5%, with a precision of 72.6%, recall of 72.5%, and an F1-score of 72.5%. The consistent increase in these performance metrics indicates that the model is learning effectively and generalizing well. Moreover, since the performance metrics

are consistent and close to each other, it can be concluded that this fine-tuned model effectively addressed the class imbalance in this task.

Causal Model Results for the Power Task

For the power task, the causal language model `meta-llama/Llama-3.2-1B` [6] was used in a zero-shot setting to assess its cross-lingual capabilities. Evaluation was performed on both the original language column (`text`) and the English-translated column (`text_en`).

Results:

- On the `text` column, the model achieved an accuracy of **50.1%**.
- On the `text_en` column, the accuracy increased slightly to **51.6%**.

Analysis: The low performance of the causal model can be attributed to the following factors:

- **Zero-shot Setting:** The model was not fine-tuned for the task-specific dataset, which limited its ability to make accurate predictions.
- **Cross-Lingual Challenges:** Despite multilingual capabilities, the model struggled with understanding nuances in non-English text.

Recommendations for Improvement:

- **Fine-tuning:** Fine-tuning the causal model on the task-specific dataset could significantly improve performance.
- **Enhanced Preprocessing:** Incorporating advanced tokenization and handling language-specific peculiarities may improve model understanding.

In conclusion, while the causal model demonstrates basic cross-lingual capabilities, its zero-shot performance highlights the need for fine-tuning and additional optimizations to handle the complexities of political text classification effectively.

References

[1] Zenodo record 10450641, 2025. URL: https://zenodo.org/records/10450641.

[2] Scikit-learn, Train-Test Split Documentation, 2025. URL: https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.train_test_split.html.

[3] G. Research, BERT Base Multilingual Cased, 2025. URL: https://huggingface.co/google-bert/bert-base-multilingual-cased.

[4] H. Face, Transformers Training Documentation, 2025. URL: https://huggingface.co/docs/transformers/training.

[5] G. Research, Google Colaboratory, 2025. URL: https://colab.research.google.com/.

[6] M. AI, Llama 3.2 - 1B, 2025. URL: https://huggingface.co/meta-llama/Llama-3.2-1B.

1. Online Resources

- GitHub