NLP Homework 2: Multilingual Classification of Parliamentary Debates

Beyza Nur Koç

Task1: Power Task

Fine-Tune a Multilingual Masked Language Model - Multilingual BERT

Dataset Exploration

I choose the Latvia (lv) dataset [1] for this homework, due to size of the other datasets. In the power set, the dataset is not balanced. The training set contained 944 samples for Class 0 and 466 samples for Class 1. So, to address this issue, oversampling was applied to Class 1, resulting in a balanced dataset with 944 samples in each class. The dataset was then split into 90% training and 10% validation subsets using stratified sampling [2].

Approach Description

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

Experimental Setup:

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 used as the loss function, optimized with AdamW.

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 77.8%, with a precision of 77.8%, recall of 77.8%, and an F1-score of 77.8%. 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 evaluation was conducted using the fine-tuned causal language model, `xlm-roberta-base` [6]. The performance of the model was assessed on both the original Latvian text column (`text`) and the English-translated column (`text_en`).

Results The model achieved an accuracy of **69.3%** on the `text` column and **87.3%** on the `text_en` column. The better performance on `text_en` compared to `text` emphasize that the model's architecture might be better optimized for English-based inputs rather than the original Latvian text.

- For **Class 0**, the model demonstrated moderate recall (0.77 for `text` and 0.88 for `text_en`) with consistent precision (0.67 for `text` and 0.87 for `text_en`). This indicates that most of the true `Class 0` instances were correctly identified, with fewer false positives in the `text_en` column.
- For **Class 1**, the recall improved significantly for `text_en` (0.86) compared to `text` (0.62). This highlights that the model could identify more positive examples in the English-translated text. Precision also showed an increase from 0.72 (`text`) to 0.87 (`text_en`).

While the oversampling technique balanced the dataset, the model's precision and recall for `Class 1` still lag behind `Class 0`, particularly for `text`. This indicates that additional methods, such as task-specific fine-tuning or data augmentation, are necessary to bridge this gap.

Recommendations for Improvement

- **Refine Preprocessing:** Improving preprocessing steps for non-English text, such as employing language-specific tokenization or embeddings, could enhance performance on `text`.
- **Data Augmentation:** Expanding the dataset with synthetically generated or translated examples for `Class 1` may help improve recall and F1 scores.
- **Change the Model:** Exploring alternative models, could yield better results.

In conclusion, while the `xlm-roberta-base` model demonstrates a strong performance on English-translated text, additional efforts are required to optimize its performance for the original Latvian dataset.

Task2: Orientation Task

Fine-Tune a Multilingual Masked Language Model - Multilingual BERT

Dataset Exploration

For the orientation task, the Latvia (lv) dataset [1] also exhibited class imbalance. The training set included 170 samples for Class 0 and 628 samples for Class 1. Oversampling was applied to Class 0, balancing the dataset with 628 samples in each class. Similar to Task1, the dataset was split into 90% training and 10% validation subsets using stratified sampling [2].

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 orientation 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 92.8%, with a precision of 92.9%, recall of 92.8%, and an F1-score of 92.9% 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 fine-tuned causal language model, `xlm-roberta-base` [6]. The performance of the model was assessed on both the original Latvian text column (`text`) and the English-translated column (`text_en`).

Results The model achieved an accuracy of **94.4%** on the `text` column and **96.8%** on the `text_en` column. The better performance on `text_en` compared to `text` suggests that the model's architecture is better optimized for English-based inputs rather than the original Latvian text.

- For **Class 0**, the model demonstrated strong recall (0.94 for `text` and 0.95 for `text_en`) and precision (0.95 for `text` and 0.98 for `text_en`). This indicates that the majority of true `Class 0` instances were correctly identified, with a slight advantage for the `text_en` column.
- For **Class 1**, the recall improved for `text_en` (0.98) compared to `text` (0.94). Precision for `Class 1` also increased slightly from 0.94 (`text`) to 0.95 (`text_en`).

The F1-scores were consistently high across both columns, with `text` achieving 0.94 and `text_en` reaching 0.97, reflecting the model's strong ability to balance precision and recall.

Discussion

The results demonstrate that the causal language model performs exceptionally well on both `text` and `text_en`, with a clear improvement in the English-translated column.

The high precision and recall for both classes indicate that the model is well-suited for the orientation task. However, the slight differences between `text` and `text_en` highlight the importance of considering language-specific preprocessing or fine-tuning to further enhance performance on non-English datasets.

Even though these results are quite well, below are the general recommendations for improvements;

Recommendations for Improvement

- **Refine Preprocessing:** Incorporating language-specific tokenization or embeddings could further improve performance on `text`.
- **Data Augmentation:** Adding more synthetic examples, particularly for `text`, may enhance the model's generalizability.

In conclusion, while the `xlm-roberta-base` model demonstrates excellent performance, particularly on English-translated text, additional optimizations could further enhance its applicability across both language columns.

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] F. AI, XLM-RoBERTa Base, 2025. URL: https://huggingface.co/FacebookAI/xlm-roberta-base.

1. Online Resources

- GitHub