Sandra Poernbacher

Report – Final Project

Introduction to Computational Linguistics

Task

The project focuses on fine-tuning the pre-trained multilingual language model XLM RoBERTa base for sentiment analysis on a dataset containing multilingual tweets. The objective is to analyze the performance of the model primarly in English and also across different languages. The project is done in a structured way, starting with dataset preparation, model training, evaluation, and visualization of the embeddings. The task sentiment analysis involves classifying tweets into three categories: negative, neutral, and positive.

Datasets

The cardiffnlp/tweet_sentiment_multilingual dataset from Hugging Face consists of tweets in several languages. Therefore a multilingual approach to sentiment analysis is possible. For efficient training parts of the data are split into three subsets: training set (5000 samples), validation set (1000 samples), and test set (1000 samples). During preprocessing a truncate function is applied to shrink the tweets into 100 tokens because of the restricted computational resources. The specific numbers for the different subsets are chosen to balance computational efficiency and representativeness. The training set size provides enough data for the model to learn patterns and improve its predictive accuracy. The validation set ensures reliable tuning of hyperparameters without over- or underfitting. And the test set is adequate for evaluating the models performance and its ability to generalize.

Hyperparameters

To optimize the model performance, several hyperparameters are fine-tuned. The batch size is set to 32 instead of 8 or 16 because it provides a good balance between training stability and the computational efficiency. A smaller batch size would possibly lead to noisy gradient updates which risks unstable training. A larger batch size on the other hand would require a much higher memory.

The learning rate is set to 3e-5 because a smaller learning rate would slow down the training. The training itself was firstly conducted for 5 epochs then – starting from the latest checkpoint – increased to 10 epochs. This ensures the model has enough time to learn meaningful sentiment representations. Ten epochs is a great number in this case to find a balance between underfitting and overfitting the model and still get it to learn a good generalization. To track the performance of the model the evaluation strategy is set to

execute validation after every epoch. With that, the model's improvements are shown constantly.

Setup

The project is conducted in Google Colab, using CPU with expanded RAM. This choice was made due to GPU resource constraints. While training on a CPU is significantly slower than on a GPU, the increased RAM allows the model to process larger batches without encountering memory limitations. To load the model and process data the transformers and datasets libraries are used. The dataset is tokenized with XLM-RoBERTa-base that converts the text data into numerical representations, which are suitable for model training. Checkpoints are saved regularly to allow tracking across the training epochs.

Quantitative Results

The model is evaluated using accuracy as a metric as it is often used for classification tasks. During the training the accuracy increases constantly. The final test set evaluation achieves an accuracy of 64.5% which is relatively low but it still demonstrates the model's ability to generalize across multiple languages. Since the model was trained on CPU which slows down the training, this might have led to inefficient gradient updates. Finally the low accuracy might also be a cause of the limited fine-tuning resources because the model was not trained on a huge sample size due to resource restrictions.

For future projects, incorporating a second metric like F1-score, precision, and recall could present more detailed evaluation.

Epoch	Training Loss	Validation Loss	Accuracy
1	1.000200	0.881311	0.600000
2	0.823500	0.837071	0.615000
3	0.730600	0.882735	0.618000
4	0.646700	0.891189	0.630000
5	0.571700	0.914382	0.632000

Epoch	Training Loss	Validation Loss	Accuracy
6	0.540400	0.927784	0.637000
7	0.485300	0.994446	0.624000
8	0.428600	1.060001	0.628000
9	0.374800	1.093254	0.634000
10	0.348200	1.091038	0.638000

Figure 1: Evaluation Epoch 1-10

Qualitative Results

The qualitative analysis emphasized the model's ability to handle diverse linguistic inputs effectively. However, certain areas of improvement were noted, such as the model's difficulty in managing complex sentiments like sarcasm, negation, or cultural context

nuances in non-English tweets. By examining misclassifications, specific patterns were identified—for instance, a tendency to misinterpret mixed sentiments as neutral. This analysis underscores the importance of augmenting the dataset with challenging examples to enhance the model's robustness across sentiment types and linguistic subtleties.

“I love this product!” → predicted as **positive**

“The service was terrible.” → predicted as **negative**

“Ich fahre ein blaues Auto.“ → predicted as **neutral**

The model successfully predicts the sentiment in most cases but struggles with more nuanced examples or sentences that contain negation. The qualitative evaluation was also done after training the model on the first 5 epochs to get a personal feeling on how accurate the model is with not only taking a look at the evaluation numbers. The qualitative evaluation also helps to identify potential gaps in the training process and to know on what example sentences or complex sentiments the models still needs to be trained on.

Visual Analysis

To analyze how the model learns sentiment representations, hidden states are extracted from different layers and epochs (as shown in the figures below), and embeddings are stored for visualization using TensorBoard and TensorFlow Projector. The goal is to observe whether sentiment clusters became more distinguishable as training progressed. The expectation that deeper layers capture more sentiment-specific features is true, as one can see in the comparison of Figure 2 and Figure 3.

Comparison of Epoch 5 Layer 1, Epoch 5 Layer 12, and Epoch 10 Layer 12

Figure 2 shows Epoch 5 Layer 1. At this early stage of the training, the first layer primarily captures some linguistic features. In the visualization, the embeddings appear all over the place without clear sentiment-based grouping. This suggests that the model has not yet developed strong representations for the three sentiment classes in this layer.

Figure 3 shows Epoch 5 Layer 12. In a deeper layer of the same epoch as discussed above, one can see more abstract information. The visualization shows some degree of clustering among similar sentiment classes, but the clusters still remain often overlapping. This suggests that while the model has started recognizing sentiment patterns, it has not yet fully refined the representations to separate them clearly.

Figure 4 shows Epoch 10 Layer 12. After further training, the embeddings in this specific layer show more distinct sentiment clusters. Compared to epoch 5, the representations are more refined and one can see clearer separations between positive, neutral, and negative sentiments. However, while some clusters are more prominent, the visualization

still does not show a perfect clustering into sentiment groups. This indicates that the model has improved, but still struggles with nuances in sentiment classification. The reason for this probably is the monolingual nature of the dataset, the small amount on data the model was fine-tuned on and the lack of resources to improve training further.

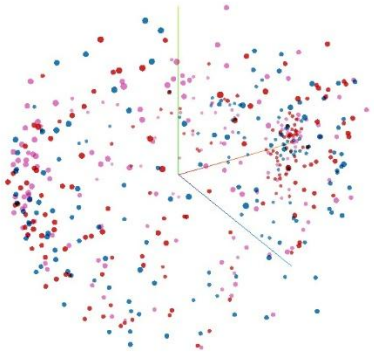

Figure 2 Epoch 5 Layer 1

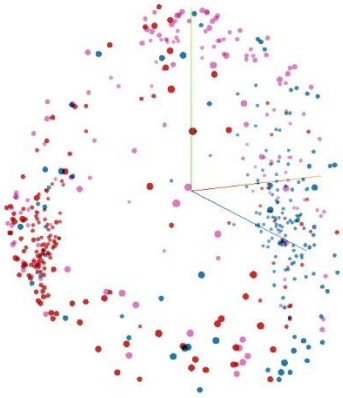

Figure 3 Epoch 5 Layer 12

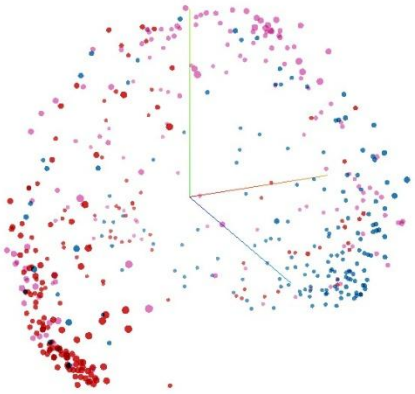

Figure 4 Epoch 10 Layer 12

Final Summary

This project fine-tuned a multilingual model for sentiment classification. The XLM-RoBERTa-base model demonstrated moderate generalization capabilities across multiple languages, achieving an accuracy of 64.5% on the test set. The evaluation process highlighted both strengths and limitations, particularly in handling ambiguous or sarcastic tweets. The visual analysis of learned representations provided insights into how the model develops sentiment-related embeddings across training epochs. Future improvements could include incorporating data augmentation, experimenting with larger models like XLM-RoBERTa-large, and refining the dataset to improve classification balance across sentiment classes.