Exercise 5 - Sequence and Anger Regression using Transformers

> Contextualized and Transformed We chose German to work with, since there exist pretrained Hugging Face BERT-base models for this languages and it contains 548k sentences in the dataset.

Part 1(0.75) Named Entity Recognition using BERT:

1. When initializing the BertForTokenClassification-class with BERT-base you should get a warning message. Explain why you get this message.

Warning message: “Some weights of BertForTokenClassification were not initialized from the model checkpoint at bert-base-german-cased and are newly initialized: ['classifier.bias', 'classifier.weight'] You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.”

Explain: Specifically, the weights associated with the classifier (the final layer for token classification) were not found in the pre-trained model and are newly initialized. We should train the model on downstream task before relying on it for predictions or inference in order to make the most out of the pre-trained knowledge from bert-base-german-cased and adapt it to the specific use case.

2. Which model performed best on the evaluation set?

Results:

Metrics	Fine-tuned with 1,000 sentences	Fine-tuned with 3,000 sentences	Fine-tuned with 3,000 sentences and frozen embeddings
f1_micro	0.9093318454456689	0.9201004324173525	0.9203236155670247
f1_macro	0.2746443834543513	0.41321798818924504	0.4123807323168641
accuracy	0.9093318454456688	0.9201004324173525	0.9203236155670247
evaluation loss	0.050377	0.036406	0.036059

The model fine-tuned with 3000 sentences and model fine-tuned with 3000 sentences with frozen embeddings have got very similar metrics results. However, they are both better then the

1000 sentences version, probably due to the data amount.

3. Are there differences between f1-micro and f1-macro score? If so, why?

Yes, the f1_micro scores are better than f1_macro scores. f1_micro Score considers the total number of true positives, false positives, and false negatives across all classes and computes the F1 score globally. It weighs each sample or prediction equally, focusing more on the overall performance of the classifier. It is more influenced by the larger classes due to its global nature, while f1_macro Score calculates the F1 score independently for each class and then takes the unweighted mean of these individual F1 scores. It treats each class equally and is not influenced by class imbalance. The model may be good at predicting the most numerous tags but bad at predicting others.

4. How large is the performance gap between 1’000 and 3’000 sentences for finetuning?

The f1_micro and accuracy scores have little difference (only about 0.01), but 3000-sentences model performs better than 1000-sentences model in terms of f1_macro scores (0.14 higher). The evaluation loss is also better (a gap 0.014).

5. Is it better to freeze or not to freeze the embeddings?

Whether we choose freeze or not depends on the specific case, on the size of training data and the similarity between the pre-training data and downstream task. As the results show, there is little difference between before and after freezing the embeddings. The differences in the scores are minimal, suggesting that, in this particular scenario, freezing or not freezing the embeddings doesn't have a substantial impact on the model's performance. In this case, we should consider other factors like training time, computational resources, and interpretability of the model may influence the decision on whether to freeze embeddings or not. From the result, the train_runtime of fine-tuned with 3,000 sentences is 683.1194, while with frozen embeddings the train_runtime is 658.0318. We could choose to freeze the embeddings to save the time.

Part 2 (0.25) Emotion Regression: How angry are you?

Please have a statement in your report as to why and what hypothesis led you to choose this architecture. Defend why you chose/chose not to do any preprocessing? Did your results support the hypothesis? Why/Why not?

We choose three models: roberta, bert-base, distilbert. We compare the three models by computing Pearson-R score. Our hypothesis is that the performance on this task would be roberta > bert-base > distilbert. DistilBERT is a smaller and more light weight version of BERT, which generally performs at a slightly lower level than BERT on certain downstream tasks, from V. Sanh et al. 2019. In comparison, RoBERTa is a larger model compared to BERT, and has demonstrated improved performance on various natural language understanding benchmarks compared to BERT, according to Liu et al. 2019.

We choose to do pre-processing for better training as the model doesn't deal with noisy data, which may influence its results. Because the data is from Twitter, which often contains noise, such as special characters, emojis, URLs, hashtags, and mentions, which might not contribute to the learning task. Removing these elements helps focus the model on relevant content.

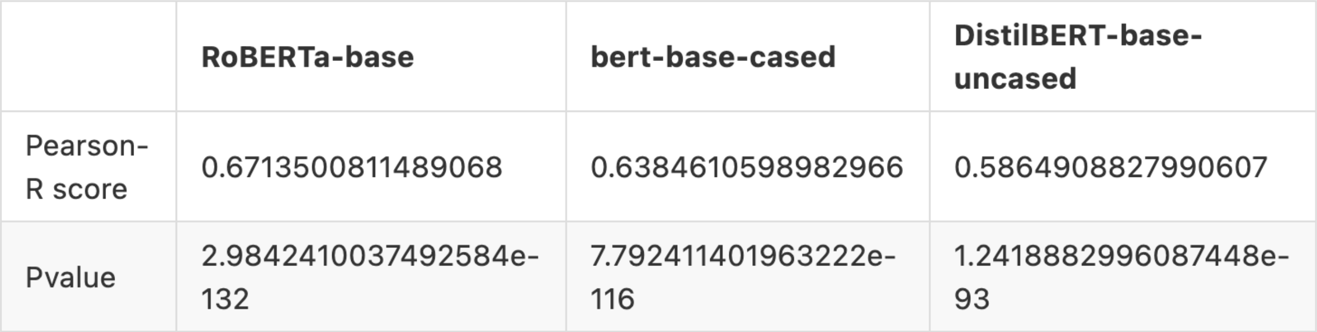
Our procedures:

- Remove usernames start with "@" symbols.
- Remove hashtags start with "#" symbols.
- Remove URLs
- Replace emojis

1. Take the best performing model and evaluate it on the test set. Briefly explain what Pearson-R evaluates and what it tells about the performance of our model and report your test set results (Pearson-r score).

Pearson-R is often used to evaluate the correlation between two datasets, which are predicted values and true values. Specifically, it quantifies the strength and direction of a linear relationship between predicted and actual values. The null hypothesis is that the distributions of the data are uncorrelated and normally distributed. If the p value is lower than the significance level, the null hypothesis could be rejected.

Results:



	RoBERTa-base	bert-base-cased	DistilBERT-base-uncased
Pearson-R score	0.6713500811489068	0.6384610598982966	0.5864908827990607
Pvalue	2.9842410037492584e-132	7.792411401963222e-116	1.2418882996087448e-93

Our reports support the hypothesis: roberta > bert-base > distilbert (in terms of Pearson-R score), where positive correlations occur.The larger Pearson-R score is, the more similar the true values are to the prediction values.