Introduction

Text De-toxification. To solve this task, I created a dataset from raw data and used fine-tuning of the language model. Then, several experiments with training methods were conducted in order to improve results. Finally, I achieved sufficient toxicity reduction performance.

Data analysis

From the raw data, I extracted only toxic and neutral sentence pairs and then distributed them across the train/test/val datasets with ratios of 0.7, 0.2, 0.1. I decided not to use other fields like ref_tox, similarity, length_diff because they seemed useless to me for my type of solution. The data also underwent simple pre-processing.

Model Specification

As a base for my solution I decided to use the pre-trained ceshine/t5-paraphrase-paws-msrp-opinosis model, which I found while exploring the github directory from which I downloaded the raw dataset for this assignment. This model is also part of the T5 family of models, so I was a bit familiar with how to handle it.

Training Process

For the final training, I decided to use only part of the dataset I generated in order to speed up the process: 20 000 training sentence pairs and 2 000 validation pairs. Training took around 45 minutes in the Goggle Collab for 10 epochs:

Epoch	Training Loss	Validation Loss	Bleu	Gen Len
1	1.646000	1.393052	25.087100	13.737000
2	1.446600	1.351189	25.485200	13.633500
3	1.401700	1.337125	25.675900	13.630500
4	1.294900	1.326818	25.923300	13.580500
5	1.249000	1.323992	25.993700	13.565000
6	1.221100	1.321001	25.915100	13.558500
7	1.203500	1.334963	25.123400	13.407000
8	1.237200	1.331684	25.113600	13.457000
9	1.229100	1.332259	25.154800	13.438500
10	1.231000	1.332499	25.148100	13.434500

```
TrainOutput(global_step=6250, training_loss=1.3068408544921875, metrics={'train_runtime': 2769.1287, 'train_samples_per_second': 72.225,
'train_steps_per_second': 2.257, 'total_flos': 1.276588873875456e+16, 'train_loss': 1.3068408544921875, 'epoch': 10.0})
```

10 epochs was enough since validation loss started to increase after 6-th epoch. (See *2.0_final_training.ipynb* notebook for full code)

Evaluation

As an evaluation metric I chose the BLEU metric. It is a common metric to measure paraphrasing accuracy, and it fits my dataset structure. During the training, the highest score I got was approximately 26.

Results

In conclusion, I achieved decent detoxification results using a pre-trained language model after researching it and fine-tuning it with optimal techniques. Also, I gained some knowledge about t5 models and their training processes alongside my work.