Final Solution Report

Introduction

The objective of this project is to create a text detoxification solution, which entails eliminating “toxic” vocabulary from the text while attempting to maintain its semantic meaning. This can be framed as a text style transfer job between a hazardous source style and non-toxic target style, as stated officially in the referenced research article.

Data Analysis

After data analysis, it was found that the reference (input) sometimes has less toxicity than the translation (target). That was probably because the dataset was generated as a subset of the general paraphrasing dataset (not specified for detoxification). It was generated in such a way that the difference between reference and translation toxicities is greater than some threshold (probably 0.5). In order to fix this, some pairs were swapped so that the reference toxicity is always higher than the translation toxicity.

Model Specification

The t5-small was used as the base model. It was fine-tuned for detoxification purposes on the given dataset.

Training Process

For training, the Seq2SeqTrainer from the Transformers library was used. The training arguments were as follows:

```
per_device_train_batch_size=8,
per_device_eval_batch_size=8,
max_steps=1000,
save_steps=100,
eval_steps=100,
save_total_limit=3,
fp16=True,
learning_rate=2e-5,
weight_decay=0.01,
```

Evaluation

For evaluation, the similarity score of input and translated texts were used, as well as the toxicity score of the translated text. Toxicity score was measured using a pretrained model. The results of evaluation are shown below.

	Similarity (mean)	Toxicity (mean)
Original data	0.70	0.10
Baseline	0.68	0.76
Our Model	0.64	0.30

Results

The results show that there is still a way to improve the quality of the model in terms of both toxicity reduction and content preservation, but the model is promising to be good at detoxification. It can be improved by longer training and more complex base model (t5-large, for example).