Solution Building Report

Introduction

The main goal of the project is to develop the detoxification model, that is, a model that can paraphrase “toxic” text in neutral form. The assignment is based on the Text Detoxification using Large Pre-trained Neural Models by Dale et al. paper, where the definition of the task is formally stated. We are provided with the training data - a subset of ParaNMT corpus filtered for text detoxification purposes. As a result, we should train and provide a model that can detoxify texts.

Baseline solution

As a baseline solution, the model fine-tuned on t5 for paraphrasing was used. Initially, it was planned to fine-tune this model to get the final solution, but the model was too big for existing computational power, so this idea was rejected.

Final solution

The t5-small model was used as the base. It was fine-tuned on the provided detoxification dataset. To enhance the performance of the model, several outputs from the model were taken and the output with least toxicity score was taken as the final output. Toxicity score was measured using a pretrained model.