BaseLine:

I explored the dataset and found out that there are many words that are present in the toxic dataset, but not in the detoxified dataset. This gave me an idea for the baseline model. I found the words that are “bad” (that are present mostly in toxic dataset), and masked them. Then, using pretrained BERT model, I tried to guess, which words should be there.

The results are the following:
Average toxicity in the results of the model: 0.3153660014080524
Average bleu score of the model (excluding zeros): 0.3615299800038338
Average rouge1 score of the model: 0.5722149046957493
Average rouge2 score of the model: 0.33180146803334354

Even though the results are not bad, the actual text produced by the model usually differs a lot in terms of meaning. So, the model is good enough as a baseline, but not sufficient for the task.

Hypothesis I: Pretrained T5

For this hypothesis I have found a T5 model that was pretrained on the dataset. This approach allows us to save time and resources, receiving good results. Additionally, the model achieves good results in the task. The model link can be found in the references section.

The results are the following:
Average toxicity in the results of the model: 0.3381002720839897
Average bleu score of the model (excluding zeros): 0.4175015958159734
Average rouge1 score of the model: 0.5573721431158483
Average rouge2 score of the model: 0.30413524706289174

The model shows significantly better results in bleu score, and approximately the same results in other metrics as the baseline solution. However, the text is much more meaningful, so the model is better for this task.

Hypothesis II: Transformer

Another idea is to create a custom transformer model trained especially for this task. I created and trained the model for 20 epochs and achieved the result of validation accuracy equal to 2.14, which is much better than in the beginning. It took approximately 5 hours. Unfortunately, even though there is the checkpoint for this model, it is not currently available for inference, so there are no metrics.

Results:

Taking into account the good metrics and sufficient meaningful result of the pretrained model and unavailability of the transformer model, we can conclude, that it is the best of the models.