Introduction

The aim of the assignment is to create a solution for the text detoxification problem. For this, I tried different approaches: guessing masked “bad” words, training a transformer model, or using a pretrained one.

Data Exploration

I explored the data and found out that the dataset itself is of low quality, some “non-toxic” texts are actually quite toxic, sometimes the meaning is not preserved. Other features were examined in the first notebook.

Best model

For the task I tried three different models: model masking “bad” words, and trying to guess, what should be instead, the transformer model, and the pretrained t5 model. The best model happened to be the pretrained model. I think this is because the authors of the model had better recources.

Evaluation

I used the following metrics to evaluate the final result:
BLEU score - measures the similarity between the machine-generated text and one or more reference texts.
Rouge1 F1 score - measures the overlap of unigrams (single words)
Rouge2 F2 score - measures the overlap of bigrams (couples of words)
Pretrained toxicity classifier: gives a toxicity score from 0 to 1, see the reference section.

The results:

The results for the best model 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