Introduction

Text Detoxification Task is a process of transforming the text with toxic style into the text with the same meaning but with neutral style. My goal was to create a solution for detoxing text with a high level of toxicity. It can be a model or set of models, or any algorithm that would work.

Data analysis

We were given the filtered ParaNMT-detox corpus to solve the problem. The most important columns from the dataset for me were: reference, ref_tox, translation, trn_tox.
From given dataset I created three datasets in parquet format:

- original.parquet - original dataset in parquet format
- toxicity.parquet - a dataset that correlates texts with their toxicity number
- texts.parquet - a dataset with just texts

Model Specification

Initially I wanted to build a chain of two models: the first one generates text, the second one evaluates its toxicity, the goal being to minimize the toxicity of the generated text (something like a GAN), but I had problems in training: I lacked computational resources in either Google Colab or Kaggle, so I decided to settle for a straightforward solution. The solution is that I trained a BART model for the Seq2Seq task.

Training Process