1. Train a BART model as text generation task

This is the most straightforward solution, just train the model to generate non-toxic text on the input text.

2. Train two sequential models

Train two models: the first one gives an estimate of the toxicity of the input text, the second one uses the first model as a loss function and minimizes its value so that the toxicity of the output text is minimal