Introduction

First of all, I want to remind that I use small version of the initial dataset because of capacity constraints. I rearranged translated sentence and reference sentence (in cases when translated sentence much more toxic than reference sentence). I found some rows with inappropriate data. I deleted such columns because it negatively affects on the training stage.

Baseline: Dictionary based

One of the easiest and understandable solutions is to identify toxic words in each sentence and then just delete them. Let's create toxic words identifier. I used RobertaForSequenceClassification, RobertaTokenizer from transformers for evaluating toxicity of each word in a sentence. All in all, model works well, the main goal was achieved, but meaning of many sentences was lost.

	bleu	rouge1	rouge2	TER
0	84.32542	0.95095	0.895144	7.657658

Default BERT

The second idea is using pretrained BERT for masking language modeling. However, I do not BERT to work as usual. BERT usually replace 15% of words with "[MASK]" token. Instead, I want to replace toxic words with "[MASK]" token. I use the following algorithm:

- evaluate toxicity of each word (using RobertaForSequenceClassification)
- replace each toxic word with "[MASK]" token
- use BERT to predict "[MASK]" token
- replace toxic word with predicted word

	bleu	rouge1	rouge2	TER
0	67.597228	0.957931	0.906898	12.672701

Fine-tuned BERT

The third idea is fine-tuning BERT from the second idea. I want to fine-tune BERT on translated part of the dataset. My idea is the following: usually reference text and translated text are similar. I believe that fine-tuning BERT on translated dataset will increase performance of the

model because BERT will learn dependencies between words and will predict the most appropriate word instead of "[MASK]" token. For fine-tuning I will replace 15% of translated text with "[MASK]" token.

	bleu	rouge1	rouge2	TER
0	74.007204	0.956069	0.909206	11.009174

T5

The 4-th idea is using T5 model. Firstly, I want to train T5-small model and then find model with similar task and try to fine-tune it. The idea is pretty obvious. Just translate reference sentence (in my case it would be toxic-sentence) to the non-toxic sentence.

	bleu	rouge1	rouge2	TER
0	0.184257	0.547436	0.311757	67.929089

T5 fine-tuned

As I said, 5-th idea is using T5 model that was already trained on the similar task. Model s-nlp/t5-paranmt-detox is suitable in my case. I want to fine-tune this model for 2 epochs

Bleu	Rouge1	Rouge2	Ter	Gen Len
24.661200	0.579100	0.356100	63.112700	12.968000

Bart

The last idea is using Bart model. The idea is the same as with T5 fine-tuned model. I choose s-nlp/bart-base-detox.

	bleu	rouge1	rouge2	TER
0	0.23752	0.590495	0.369169	62.610798