# Introduction

For my final solution, I utilized pretrained models from Hugging Face for the task of generating detoxified translations. I fine-tuned and tested several models such as Bert, T5, Blenderbot, and MVP, and compared their performance using a radar plot.

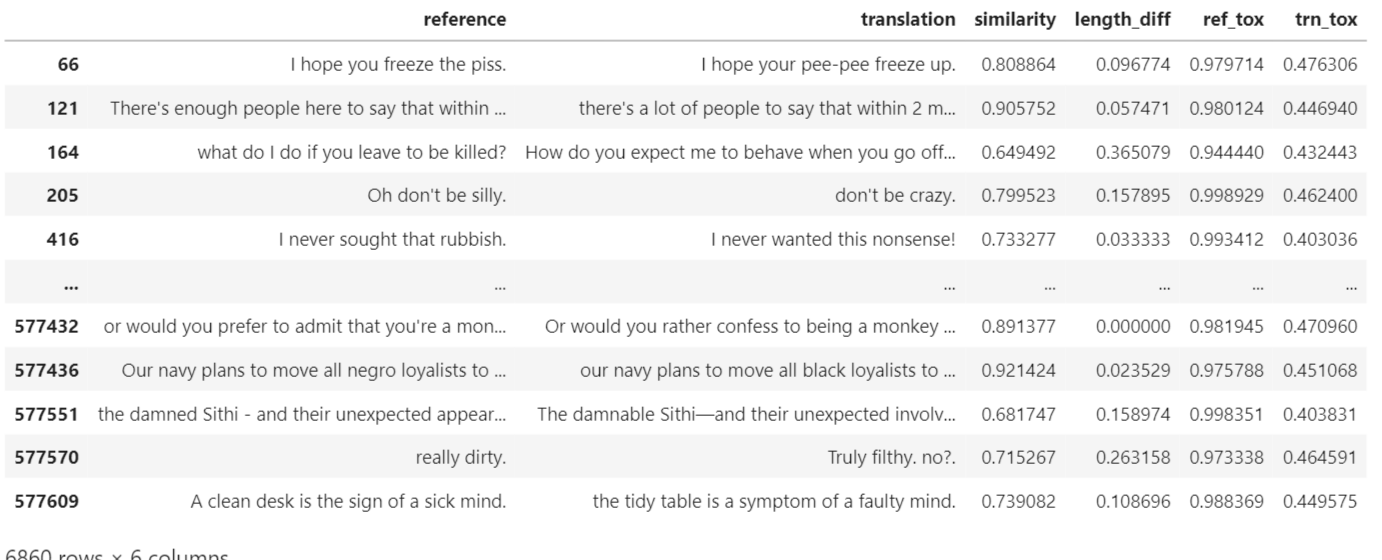
# Data Analysis

I used the original dataset from para-NMT-detox and also incorporated toxic comments from Wikipedia for testing the network.

<https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data>

To preprocess para-NMT dataset I swapped values where trn\_tox > ref\_tox, so that it didn't confuse the model.

Next, I removed the rows, where toxicity was greater than 0.4. I looked up through such translations and made a conclusion that most of them were bad. Examples:



After that, I observed the length difference, which was not more than 0.4 throughout data, so it will not impact drastically our translations.

Finally, I have removed all rows with similarity values less than 0.75 to increase model preference to make the sentences almost similar without paraphrasing them drastically.

# Model Specification

I fine-tuned and tested models from Hugging Face's Seq2SeqTextGeneration class, including Bert, T5, and MVP. These models have shown similar performance for each metric, with a promising BLEU score of around 30.

# Training Process

Due to limitations, I was not able to train all models simultaneously. Instead, I saved the test results and visualized the metrics for each model in order to choose the best one.

Also, I have trained all models on cropped dataset because each time I've ran out of CUDA memory:

1. 5000-10000 train
2. 500-1000 validation
3. 500-1000 test

# Evaluation

I used a radar plot to compare the performance of each model based on three metrics: BLEU score, brevity penalty, and average inference time. The test results were obtained from a test dataset.

# Results

Based on the evaluation, the Bart-base model performed the best with the highest BLEU score, while also having a reasonable average inference time. This led me to choose Bart as the final model for my solution. Finally, I've tried to test it on toxic comments with length of sentence < max input length to model.

