Final Solution Report

Arina Yartseva

November 5, 2023

1 Introduction

The objective of this project was to develop a model capable of converting toxic texts into non-toxic language. This report documents the final solution after exploring various architectures and approaches.

2 Data Analysis

An in-depth analysis of the dataset was conducted to understand the distribution and nature of toxicity within the texts.

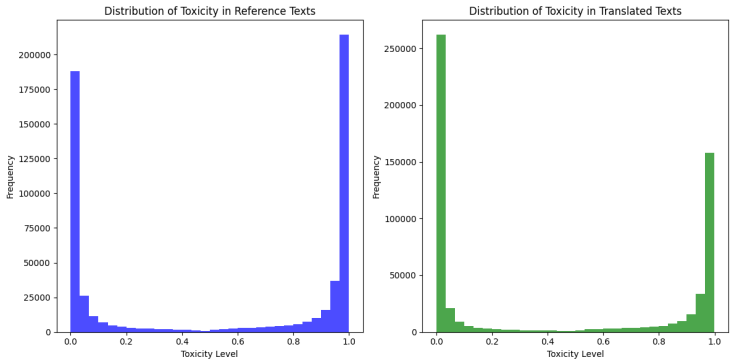

Figure 1: data analysis

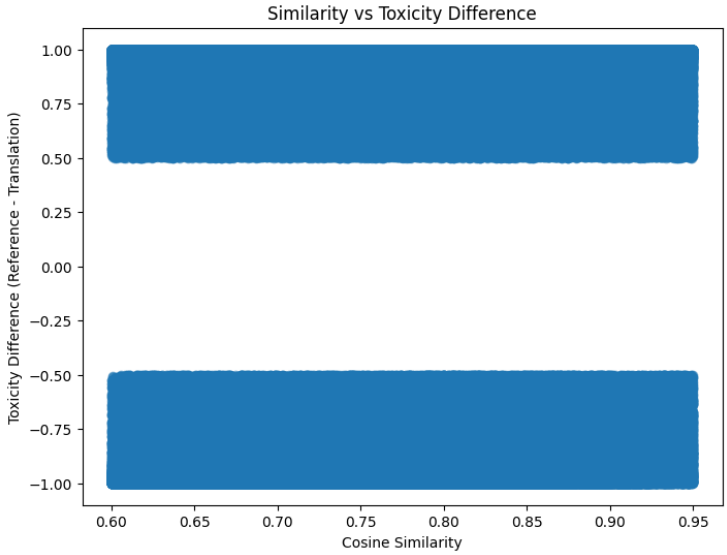

Figure 2: data analysis

The data analysis suggests that most translations either maintain the same level of toxicity or reduce it, while retaining high content similarity to the original text. There's also an observation that a portion of texts has high toxicity both before and after translation.

3 Model Specification

The final model is a fine-tuned version of the T5 base model, chosen for its ability to understand and generate natural language.

4 Training Process

The T5 model was fine-tuned on a curated dataset of toxic and non-toxic text pairs. We employed a custom learning rate with warm-up and used a batch size of 16 for efficient training.

5 Evaluation

The model's performance was evaluated on a validation set, showing significant improvements over the baseline and earlier hypotheses. A combination of BLEU score, level of toxicity and semantic similarity was used as the evaluation metric.

$$\text{Score} = \text{BLEU Score} \times (1 - \text{Toxicity Level}) \times \text{Semantic Similarity}$$

6 Results

The final model demonstrated proficiency in context comprehension and the generation of non-toxic alternatives, maintaining the semantic integrity of the original expression. The evaluation was based on combination of BLEU score, level of toxicity and semantic similarit.