Text Detoxification Report

Timur Aizatvafin

November 2023

1 Final Part

1.1 Goal

The primary objective of this project is to implement and evaluate a text detoxification system. The system aims to transform text with toxic content into a neutral counterpart without changing the original meaning. This contributes to a more positive discourse in digital communication platforms.

1.2 Data Understanding

I began by examining a dataset consisting of sentence pairs, where each pair includes a ”toxic” reference sentence and its ”detoxified” translation. Key features of the dataset include toxicity levels, cosine similarity scores, and the relative length difference between the paired sentences. Graphs will be included to illustrate:

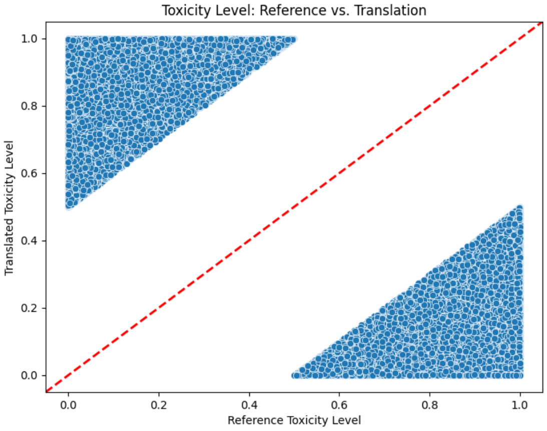

Figure 1: In some cases ref with trn need to be swapped

- The distribution of toxicity levels in the reference and translated texts.
- A scatter plot comparing the toxicity levels of reference vs. translated texts.
- The length distribution of the reference and translated texts.
- The distribution of cosine similarity scores between the text pairs.

1.3 Data Preparation

Data preparation involved:

- Tokenizing the sentences and calculating their lengths to create new features for text length analysis.
- Cleaning and normalizing the text data to ensure consistent tokenization and comparison.
- Splitting the dataset into training and validation sets to evaluate model performance.

1.4 Models

The first model was a Sequence-to-Sequence (Seq2Seq) model with an attention mechanism. It leverages an encoder-decoder architecture to understand the context of the toxic sentences and generate neutral equivalents.

The ReplacerModel effectively sanitizes input text by obscuring toxic words, thereby allowing the text to be displayed or used in contexts where such content is undesirable or harmful. This approach is straightforward and relies on a predefined list of toxic words, which can be customized for different applications or levels of content moderation.

1.5 Predictions

To evaluate the models, new toxic sentences were processed to generate detoxified outputs. A qualitative review was conducted to verify that the detoxified versions maintained the original intent of the sentences.

1.6 Results

The results showed that:

- The Seq2Seq model demonstrated a promising ability to lower toxicity in the content, generally preserving the intended meaning.
- Observations of the toxicity scores post-prediction suggested a trend toward less toxic language.

- The ReplacerModel showed good performance on given input of toxic words.

The models demonstrate promising applications in content moderation and support for creating healthier online interactions.