Text Detoxification

Solution Process Report
Roman Makarov

Baseline + Hypothesis 0

As a baseline, we had a main dataset composed of approximately 500k entries of reference and translation sentences.

Upon a thorough examination of the dataset, I discovered that I could significantly reduce its size by filtering out irrelevant entries, specifically those with a reference toxicity score of less than 0.95 or a translation toxicity score greater than 0.01. A detailed explanation of this process can be found in the first section of the Main Solution Report.

Hypothesis 1

During the process of reviewing the primary dataset, I also noticed that some of the translations appeared to be somewhat unusual or awkward. To address this concern, I made the decision to seek out an alternative dataset, ultimately selecting one from Hugging Face datasets.

This approach allowed me to incorporate high-quality, human-translated sentences, hence enhancing the overall quality of the dataset.

Hypothesis 2

Throughout the training phase, I noted a substantial drop in the model's loss during the first epoch. However, the reduction in loss occurred more gradually, with little to no further changes afterwards. This led me to consider an alternative approach: utilizing more data and reducing the number of training epochs. This strategy allows to create a more generalized model, one that could adapt more effectively across various inputs.

Hypothesis 3

The default metric for sequence-to-sequence training in Hugging Face is the exponential of the cross-entropy loss.
Initially, I considered adding a toxicity metric, but then I noticed that the model removes all toxic words and constructions with ease, and its only challenge is preserving semantic meaning and sentence completeness.
Consequently, I wanted to track my performance in a slightly different way. I wanted

to check how similar it is to a reference translation. Additionally, I did not want to use BLUE score, as it focuses on n-gram overlaps and lacks semantic similarity.. Instead, I chose to utilize a BERT similarity score to evaluate and track the model's performance.

Result

As a result, my project involves optimizing a dataset by filtering out low-toxicity reference and high-toxicity translation entries. Additionally, high-quality human-translated sentences were added to enhance dataset quality. Training was set up with a focus on utilizing more data and fewer epochs to achieve a generalistic model, and a BERT similarity score was used for semantic performance tracking.