Solution Building Report

Arina Yartseva

November 5, 2023

1 Introduction

This report outlines the journey taken to develop a solution for transforming toxic texts into non-toxic equivalents. The task was approached with a series of hypotheses, each leading to iterative development and refinement of the final model.

2 Baseline: Dictionary-Based Approach

I started with a simple dictionary-based approach where toxic words were replaced with non-toxic synonyms. This method provided a quick and straightforward way to address the problem but lacked contextual understanding.

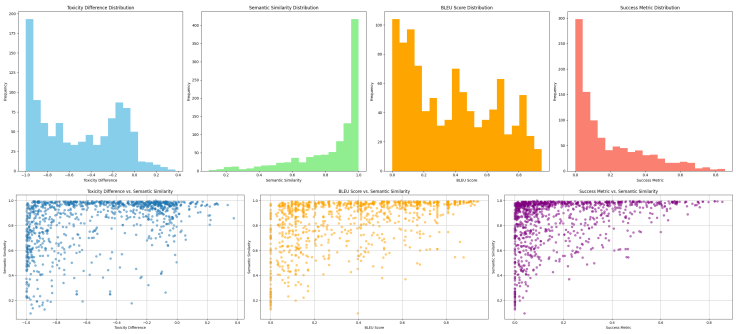

Figure 1: Result of Dictionary-Based Approach

The histograms represent the distributions of four different metrics:

1. **Toxicity Difference**: Most values are negative, suggesting a reduction in toxicity after an intervention.

2. **Semantic Similarity**: There is a strong tendency towards higher similarity, with most values clustering near 1.0, indicating very similar textual content.

3. **BLEU Score**: The scores are spread across the spectrum but with a lean towards lower scores, suggesting varying quality in translation or text generation relative to a reference.

4. **Success Metric**: This metric is skewed towards lower values, indicating a lower success rate for the measured activity or performance.

Overall, these histograms provide a quick visual summary of the effectiveness of interventions on toxicity, the degree of similarity in content, the quality of translations, and a general measure of success across different activities.

3 Hypothesis 1: Custom Embeddings

To capture context, we experimented with custom embeddings trained from scratch on our dataset. This allowed the model to develop a nuanced understanding of the text, but it required extensive data and training time.

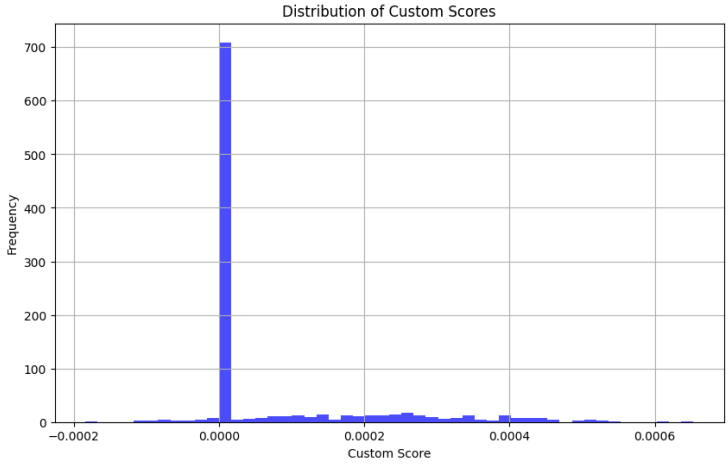

Figure 2: Results of Custom Embeddings

4 Hypothesis 2: More RNN Layers

We hypothesized that increasing the depth of the RNN layers would enhance the model's ability to understand longer dependencies in the text. We tested various configurations, eventually settling on a three-layer LSTM structure.

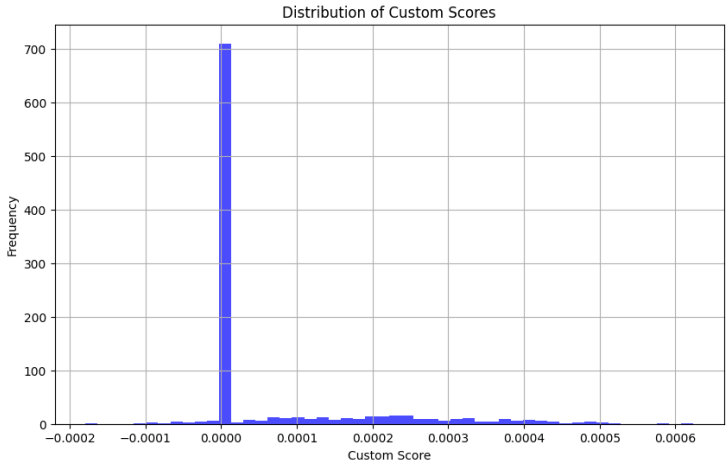

Figure 3: Results of More RNN Layers

5 Hypothesis 3: Pretrained Embeddings

The final hypothesis involved leveraging pretrained embeddings from models like T5, which have been trained on a vast corpus of data. This allowed us to benefit from transfer learning, significantly improving the model's performance on our task.

6 Results

Through iterative testing and refinement, we found that the combination of pretrained embeddings with fine-tuning on our dataset yielded the best results. The model was able to effectively transform toxic texts into non-toxic ones with a high degree of success rate.