Text Detoxification Model Development Report

Timur Aizatvafin

November 2023

0.1 Introduction

This project embarked on the development of a text detoxification model, aiming to neutralize toxic sentences while preserving their original meaning. The endeavor contributes to enhancing the quality of digital communication by mitigating the impact of harmful language.

0.2 Challenge and Approach

The primary challenge was to accurately identify and transform toxic text without altering the core message. To address this, I utilized Seq2Seq models known for their efficacy in tasks that require understanding of context and generation of corresponding outputs.

0.3 Model Training and Evaluation

Initial training was conducted on a subset of a comprehensive dataset, focusing on the model's ability to discern and mitigate toxicity. Despite the promising direction, the results were suboptimal, likely due to the limited training data and the model's nascent stage of learning.

0.4 Technical Hurdles

During the process, I encountered typical machine learning challenges, such as model loading errors and discrepancies in data dimensions. Solutions involved ensuring consistency in the model's architecture during both training and inference phases and rectifying dimension mismatches by aligning the training environment with the prediction context.

0.5 Preliminary Outcomes

Preliminary tests on new sentences indicated a notable shift towards reduced toxicity, demonstrating the model's potential. Qualitative analysis further suggested that while the model reduced toxicity, extensive training on a larger corpus could yield better results.

0.6 Conclusion and Future Work

The work concluded with a model that showcases a good approach towards text detoxification. Future improvements include expanding the training dataset and refining the model architecture to enhance performance and reliability.