Final Solution: Text Detoxification Using Fine-Tuned T5-Small Model

Artyom Makarov

Introduction

Text detoxification, the process of converting toxic or offensive text into a non-toxic form, is a challenging problem in natural language processing. In this report, I present a solution that employs a fine-tuned T5-Small Model to address the task of text detoxification. This model has the potential to identify both toxic words and toxic phrases based on contextual information.

Data Analysis

For the fine-tuning process, I used a subset of the ParaNMT corpus, which consists of pairs of toxic sentences with their detoxified translations. The dataset provides a valuable source of toxic language examples and their corresponding non-toxic counterparts, allowing our model to learn the necessary transformations.

Model Specification

The T5-Small Model, developed by Google AI, is a versatile and efficient text-to-text transfer learning model. It is well-suited for text detoxification due to its ability to understand and generate text in a sequence-to-sequence manner. This model is capable of not only identifying toxic words but also toxic phrases, making it a suitable choice for the task. The "small" variant of T5 balances efficiency and performance, making it well-suited for this task.

Training Process

The training process involved the utilization of the Seq2SeqTrainer from Hugging Face. One advantage of using the T5-Small Model is that it requires minimal preprocessing of the data, unlike the previous hypotheses. The only required modification was the addition of a prefix "Detoxify text: " to the input sentences, which guided the model to perform the text detoxification task. The model was trained to understand the context of toxic phrases and generate detoxified translations. It was fine tuned on a GPU, and the process did not consume an extensive amount of time due to the model's relatively small size. However, the use of a smaller model may have drawbacks for the final results.

Evaluation

To assess the effectiveness of the fine-tuned T5-Small Model, it was evaluated using four metrics: semantic similarity, style accuracy, fluency and J score, introduced in [2] (Krishna, Wieting, & Iyyer 2020). Additionally, a human evaluation was conducted to gather valuable insights.

The model excelled in terms of semantic similarity and fluency, indicating its ability to preserve the original meaning and generate coherent text. However, it was far from perfect in terms of style accuracy. This might be a drawback associated with using a small model. During the human assessment, the assessor (me) was impressed by the results, in comparison with previous hypotheses.

Examples of detoxification:

Original: *Hurt you and I want to find out who that was.*
Detoxified: *you and I want to find out who that was.*

Original: *I can't believe we haven't fucked for two years, nine months, three weeks and... 69 hours.*
Detoxified: *I can't believe we haven't slept for two years, nine months, three weeks and... 69 hours.*

Original: *So forgive me for being a little fidgety, but if it's lunch we're talking, I'm gonna eat a fat pork sandwich, and I'm sure as shit not gonna eat it here.*
Detoxified: *so forgive me for being a little bit of a snitch, but if it's lunch we're talking, I'll eat a fat pork sandwich, and I'm sure I'm not gonna eat it here.*

Original: *There is no fucking soy milk!*
Detoxified: *there's no soy milk!*

Original: *What's wrong with people having sex?*
Detoxified: *what's wrong with people having sex?*

Original: *What the fuck are you talking about?*
Detoxified: *what are you talking about?*

Original: *There were fewer horses that needed stealing than men who needed killing, after all.*
Detoxified: *there were fewer horses that needed to steal than men who needed to kill.*

Original: *Hey, Bob, thanks for giving us a chance to grieve together. Shit!*
Detoxified: *hey, Bob, thanks for giving us a chance to grieve together.*

Original: *But my landlord's another fucking story-- Trying to kick me out of my home.*
Detoxified: *but my landlord is another story - trying to get me out of my home.*

Original: *Gryilus, the father of young Phylon. I slashed his face with a whip.*
Detoxified: *Gryilus, the father of Phylon, I slashed his face with a whip.*

Results

In conclusion, the fine-tuned T5-Small Model shows great promise in addressing the text detoxification problem. It demonstrates high performance in semantic similarity and fluency, suggesting that it effectively replaces toxic language while preserving the original text's meaning and readability. However, the model's style accuracy remains an area for improvement, and it may be impacted by the choice of a smaller model.