Solution building report

Baseline: Dictionary-based Approach

The standard baseline for text detoxification usually entails taking a pre-compiled list of harmful words or phrases and substituting them with non-toxic alternatives. This approach is straightforward but frequently insufficient to handle complicated instances of toxic language, and it could miss harmful information that is peculiar to a certain situation.

Hypothesis 1: T5-Small Model

My first hypothesis aimed to improve the Text Detoxification task by leveraging the T5-Small model, a variant of the Transformer model, which is a state-of-the-art architecture for sequence-to-sequence tasks. We can preprocessed the dataset, split it into training and validation sets, and fine-tuned the T5-Small model to learn to detoxify toxic text while maintaining the original meaning.

The T5-Small model offers several advantages:

- It can handle context-dependent toxicity, making it more suitable for nuanced cases.
- It generates paraphrases that are often more coherent and contextually accurate than simple dictionary-based substitutions.
- Fine-tuning allows the model to adapt to the specific task and data.

Hypothesis 2: CondBERT

Second hypothesis was to explore the use of CondBERT, a conditional variant of the BERT model, for Text Detoxification. CondBERT is designed to generate text with specific attributes while conditioning on input text. We hypothesized that by conditioning the model on the input text and instructing it to produce non-toxic text, we could achieve accurate detoxification. CondBERT offers the following potential advantages:

- It can condition the generation process on the context of the input, allowing for context-aware detoxification.
- Fine-tuning can help adapt the model to the detoxification task and improve its performance.

Results

Following testing using the T5-Small and CondBERT models, we discovered the following outcomes:

When it came to the quality of detoxification, the T5-Small model fared better than the dictionary-based baseline. It successfully maintained the original meaning while using more neutral language that was more contextually suitable for the hazardous parts. It also produced paraphrases that were more logical.

Although the CondBERT model needed a lot of conditioning and fine-tuning to operate on par with the T5-Small model, it showed potential. A model's performance was contingent upon both the calibre and volume of training data.

Generally speaking, the T5-Small model performed better on the Text Detoxification challenge and needed less data and fine-tuning.

To sum up, the Transformer architecture's T5-Small model emerged as a formidable contender for the Text Detoxification challenge. It fared better than the dictionary-based method and shown great promise for managing text's context-dependent toxicity. Although the CondBERT model has potential, additional fine-tuning and conditioning work is needed to make it work. When selecting between the two models, one should take into account the particular demands and limitations of the Text Detoxification job, in addition to the data and computing resources that are at hand.