Solution Building Report

Data Exploration Results

During the initial phase of our project, we conducted a thorough exploration of the dataset to better understand the nature of the dataset:

Text Length: The majority of text samples in the dataset consist of fewer than 50 words. This insight indicates that the challenge lies in detoxifying short text snippets.

Toxicity Distribution: We observed that both the reference and translation parts of the text can have high toxicity. However, it's important to note that they cannot be simultaneously toxic in a single sample. This constraint ensures that we have valid training data for our models. With these insights in mind, we proceeded to develop and evaluate two hypotheses for text detoxification.

Hypothesis 1: BART-based Paraphrasing Methodology:

Our first hypothesis focuses on using a BART (Bidirectional and Auto-Regressive Transformers) model for text detoxification. We employed a two-step approach:

Fine-tuning BART: We fine-tuned a pre-trained BART model on our dataset. This step is crucial as it enables the model to understand the specific nuances of toxicity present in the data. The model was initially trained using paraphrasing datasets called ‘Quora’ and ‘Paws’.

Paraphrasing with Fine-Tuned BART: After fine-tuning, we used the BART model to paraphrase toxic text, generating non-toxic versions of the input text.

Justification:

- BART's bidirectional architecture allows it to capture contextual information effectively, which is vital for understanding the context of toxicity.
- Fine-tuning the model on our dataset makes it specialized for text detoxification, enhancing its performance on this task.
- The model was initially trained on a dataset for paraphrasing which is closely related to our problem because both of them preserves meaning.

Hypothesis 2: Toxicity Classifier and BART Paraphrasing Methodology:

Our second hypothesis incorporates a two-step process involving a toxicity classifier and the BART model:

Training a Toxicity Classifier: We trained a BERT-based text classifier to determine whether a given text snippet is toxic or non-toxic.

Paraphrasing with Fine-Tuned BART and Toxicity Classifier: We used the fine-tuned BART model to generate multiple paraphrased versions of the toxic text and then leveraged the toxicity classifier to select the paraphrase with the least toxicity score.

Justification:

- The toxicity classifier serves as a valuable tool for assessing the degree of toxicity in text, helping us identify the most suitable paraphrased version.
- By combining the strengths of both models, we aim to strike a balance between preserving the original meaning and detoxifying the text effectively.

Results:

The BERT classifier was trained to classify text samples as toxic or non-toxic. Here are the evaluation metrics for the BERT classifier:

- Accuracy: 0.8649
- F1 Score (Macro): 0.8649

I couldn’t evaluate the BART model because the training kept crashing on google colab….