Final Solution Report

Data Analysis

By selecting samples from the provided dataset with reference sentence toxicity levels greater than 0.9 and their corresponding reference sentence toxicity levels less than 0.2, we have obtained approximately 100,000 samples. These samples have been meticulously cleaned to remove any redundant punctuation and organized into a JSON file with the following structure:

```
{
  "instructions": "Rephrase with lower toxicity level.",
  "inputs": "Reference text",
  "outputs": "Translation text"
}
```

You can access the "subset.tsv" file and the generated JSON file in the "data" folder within the repository. However, due to my use of Colab and Hugging Face, it is more convenient to load the dataset from the JSON file and upload it to Hugging Face for easier accessibility and usability.

You can access the dataset by following this link: Detoxify Dataset on Hugging Face.

Model Specification

Model Introduction:

ALPACA is a language model designed to follow instructions, conceived and developed by scientists at Stanford University's Center for Research on Foundation Models. This compact yet powerful 7B language model was refined from the LLaMA 7B model of Meta AI and was educated on 52K demonstrations of instruction-following, which were created in the self-instruct style using Open AI's text-davinci-003.

It is a rather recent model, the authors of “Alpaca: A Strong, Replicable Instruction-Following Model” paper has not published its architecture

Model Parameters:

Parameter	Value
Trainable parameters	4,194,304
All parameters	6,742,609,920
Trainable %	0.06220594176090199

Training Process

Efforts have been made to optimize resource utilization in Colab. Low-Rank Adaptation (LoRA) accelerates large language model training while saving memory by introducing update matrices, reducing trainable parameters for cost-effective and efficient processing.

This has been done with the usage of PEFT and bitsandbytes. PEFT is a library that specializes in the efficient adaptation of pre-trained language models (PLMs) to various downstream applications, without necessitating fine-tuning of the entire model's parameters. Instead, PEFT methods focus on fine-tuning only a limited number of additional model parameters, substantially reducing both computational and storage costs. This is crucial, as fine-tuning large-scale PLMs can be prohibitively expensive. While bitsandbytes offers lightweight wrappers around CUDA custom functions, including 8-bit optimizer.

Thus the model was loaded 8-bit quantization.

Loss for training and validation :

Step	Training Loss	Validation Loss
10	2.917700	2.086030
20	1.527200	1.096352
30	1.011600	0.955284

Then the finetuned model was pushed to huggingface for easier usage.

Evaluation

As explained in the previous report, two metrics were used to evaluate the model, toxicity level of model’s output and the cosine similarity between the generated output and the target sentence.

Results