Solution Bulding Report

Data Analysis and Preprocessing Enhancements

Upon initial inspection, I observed disparities in the toxicity levels between some references and their translations—with the references often presenting lower toxicity levels. To ensure consistency and reliability in our data, I have implemented preprocessing steps that adjust the dataset so references consistently exhibit higher toxicity levels than their translations.

Furthermore, the dataset will undergo a random partitioning process to create training and test subsets. The test subset will comprise a 0.2 fraction of the entire dataset, allowing for a robust evaluation of the model's performance.

Refining the Task Definition

The undertaking at hand can be conceived as a **sequence-to-sequence** task, paralleling text translation processes. Consequently, it is amenable to resolution through the application of specialized models that have been expressly devised for such tasks.

To enhance performance, I plan to utilize a pre-trained Large Language Model (LLM). In my quest for a suitable approach, I encountered a resource on the Huggingface platform which delineates the process of fine-tuning a T5 model on a specific dataset. I intend to tailor this methodology to meet the requirements of our task.

In alignment with the recommendations provided in the tutorial, additional data processing is mandated. Each reference will be prefixed with “Perform Text-Detoxification:” to capitalize on the model’s capacity for In-Context Learning, which was honed during pretraining. This strategic fine-tuning is anticipated to bolster the efficiency of the learning process.

Results and Demonstrations

The adapted model has adeptly grasped the task parameters, excelling in the excision of toxic verbiage and expressions while concurrently preserving the original intent and meaning of the text. Displayed below are select instances where the model's proficiency is particularly evident:

```
--------------------------------------------------------------------------------------------------------
REFERENCE: I thought you said you knew this fucking guy.
TRANSLATION: didn't you say you knew him?
MODEL TRANSLATION: [{'translation_text': 'I thought you said you knew this guy.'}]
--------------------------------------------------------------------------------------------------------
REFERENCE: one of my guys is squirming 10 to 15 years old for some kind of electronic voodoo shit from the Feds.
TRANSLATION: My boy tiny's doing 10 to 15 because of some electronic voodoo the Feds pulled.
MODEL TRANSLATION: [{'translation_text': 'one of my guys is squirming from 10 to 15 years old for some kind of electronic voodoo nonsense f
--------------------------------------------------------------------------------------------------------
REFERENCE: You know how I know you fucked him?
TRANSLATION: you know how I know you drove him?
MODEL TRANSLATION: [{'translation_text': 'You know how I know you screwed him up?'}]
--------------------------------------------------------------------------------------------------------
REFERENCE: Paris was throwing up.
TRANSLATION: Paris and I took turns throwing up.
MODEL TRANSLATION: [{'translation_text': 'Paris messed up.'}]
--------------------------------------------------------------------------------------------------------
REFERENCE: Grigio, shut up! I'm trying to sleep!
TRANSLATION: Grigio, quiet, I want to sleep!
MODEL TRANSLATION: [{'translation_text': "Grigio, I'm trying to sleep!"}]
--------------------------------------------------------------------------------------------------------
REFERENCE: step two: They scare them to shit.
TRANSLATION: Step 2: it scares the crap out of them.
MODEL TRANSLATION: [{'translation_text': 'step two: They scare them.'}]
```