In this report, I will provide an overview of the entire project, highlighting key stages and decisions made along the way.

Data Analysis

Upon downloading the dataset, I noticed that some translations of toxic text were even more toxic than the reference. Additionally, it became evident that toxic text tended to be longer than non-toxic text. The maximum length of toxic text reached 1401 characters, while its translation was consistently shorter, usually less than 1000 characters. This suggests that making text non-toxic often involves removing offensive language. Consequently, I focused on models that could generate entirely new, non-toxic text, as opposed to those that merely replaced bad words with better alternatives. You can read more about it in the first notebook.

Model Selection

My initial choice was the GPT model, as it showed promise in addressing the problem. However, I encountered limitations due to its non-free nature. I also experimented with the T5 model, which I had prior experience with, but its performance proved to be slower than anticipated. My search eventually led me to the bart-base model, specifically trained for detoxification tasks. This model demonstrated superior results compared to the T5 model. After further exploration, no better alternative surfaced, leading to my decision to adopt the bart-base model as the final choice. You can read more about this part in another report and in the second notebook.

Model Training

For model training, I leveraged the transformers library and its built-in classes. This involved creating an object to store training arguments and configuring the model for the training process. You can read more about this in the third notebook.

Model Evaluation

To assess the model's performance, I opted to use the BLEU score, which was the primary metric discussed at my university. This metric provided a reliable measure of the model's effectiveness.

Results

The model did not finish the training by time issue, but the outcome of this project is a highly successful model that effectively detoxifies text, achieving an BLEU score of 89. This outcome marks a significant accomplishment and demonstrates the model's excellence in its task.