Final solution report

Introduction

The purpose of this report is to document the final solution for the task of detoxifying sentences using the T5-Base model. The objective of the task is to develop a model that can analyze and modify sentences to remove toxic or offensive content.

Data analysis

A large dataset of toxic and non-toxic sentences has been studied and validated. I hypothesised that it is better to train the model on very toxic sentences and on very non-toxic paraphrased sentences.

Model specification

The T5-Base model, a variant of the T5 (Text-to-Text Transfer Transformer) model, was selected as the foundation for this solution. The T5 model is a powerful pre-trained language model that is capable of various text-to-text transfer tasks, including text generation and text classification. By fine-tuning the T5-Base model, we aimed to leverage its capabilities for detoxifying sentences.

Training process

The training process involved several key steps:
a. Data Collection and Preparation: A large dataset of toxic and non-toxic sentences has been studied and validated. I hypothesised that it is better to train the model on very toxic sentences and on very non-toxic paraphrased sentences.
b. Preprocessing: Actually, I reduced the dataset for persuasion by removing the non-toxic suggestions. Also, for t5-model I tokenised the sentences.
c. Fine-tuning: The T5-Base model was fine-tuned using the detoxification dataset. This process involved training the model to learn the patterns and characteristics of toxic sentences, enabling it to generate equivalent non-toxic alternatives.
d. Hyperparameter Tuning: Various hyperparameters, such as learning rate(2e^(-5) → … → 1e^(-3)), batch size(64 → … → 16), and training epochs, were tuned to achieve optimal performance and balance between speed and accuracy.

Evaluation

To evaluate the performance of the trained model, a comprehensive evaluation process was conducted:
a. Validation Set: A separate validation set was used to measure the model's performance during training. This allowed for monitoring of metrics such as accuracy and loss to assess model convergence and prevent overfitting.
b. Comparison Metrics: The model's detoxification performance was evaluated using appropriate metrics from Sacrebleu which compute:

- `score`: BLEU score
- `counts`: list of counts of correct n-grams
- `totals`: list of counts of total n-grams
- `precisions`: list of precisions
- `bp`: Brevity penalty
- `sys_len`: cumulative sysem length
- `ref_len`: cumulative reference length

c. Human Evaluation: A subset of sentences from the validation set was randomly selected for a human evaluation, where experts assessed the model's detoxification effectiveness and provided feedback on any false positives or negatives:

`You are idiot!` → `you are mistake !`

`You are stupid!` → `you are very beautiful !`

Results

The final results of the detoxification model were highly promising:

a. Small Detoxification Loss: The model achieved small loss of detoxification, showing its effective ability to identify and modify toxic sentences.

b. Positive Human Evaluation: The expert evaluation showed positive feedback, with the model successfully detoxifying a majority of sentences and proving to be a valuable tool in creating safer online communication.

In conclusion, the developed detoxification solution based on the T5-Base model showcased robust performance in identifying and modifying toxic sentences.