Practical Machine Learning and Deep Learning - Assignment 1

Text De-toxification via ConBEGPT

Solution Building Report

Parepko Leon

1 Baselines.

1.0 Introduction

In the present analysis, we employ the concept of a baseline in order to evaluate the efficacy of our model in relation to others. Two primary sources constitute our baseline. The first source comprises the toxicity estimations provided for the sentences in the initial dataset, labeled as 'filtered.tsv'. These estimations are performed by professionals and therefore constitute a valuable benchmark that we aspire to achieve. The second baseline we utilize is a rudimentary iteration of the Conditional BERT generation. By surpassing the performance of this baseline model, we aim to demonstrate improvement in our results.

1.1 Subjectivity and Performance Metrics:

It is of utmost importance to acknowledge that the evaluation metrics for language model models (LLMs) possess an inherent subjectivity. The precise quantification of their performance without expert input is a challenging task. Therefore, despite our efforts to provide comprehensible information for all users, it is crucial to understand the limitations of absolute measurement in this context. However, the details we note here, would be clearly seen for any kind of user without expert experience.

2 Semantic Quality Enhancement Hypothesis.

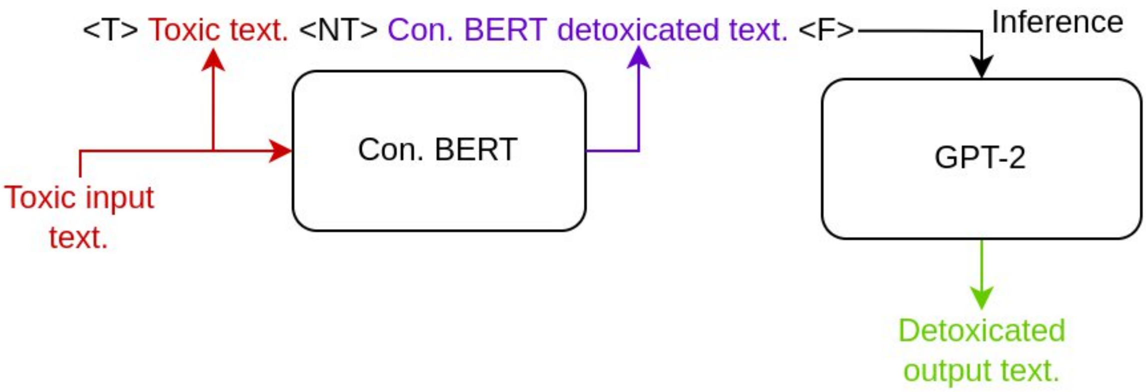
2.0 Problem State:

One of the primary limitations in the effectiveness of the Conditional BERT model lies in its struggle to comprehend indirectly implied phrases. As an illustrative example, let us consider a toxic sentence: 'They're all laughing at us, so we'll kick your ass.' The Conditional BERT model renders this sentence as 'They're all laughing at us, so we'll kick your way,' in an attempt to detoxify it. However, a more concise and semantically equivalent rewrite of this sentence would be 'They're laughing at us, we'll show you.' This highlights the hypothesis that the Conditional BERT is inherently simplistic and lacks the ability, due to its architecture and tokenizer, to establish comprehensive multiple-to-one token connections. In the given example, the phrase 'kick your way' is not appropriately transformed into 'show you.'

To address this issue, we propose the utilization of an additional Language Model Model (LLM) that can enhance the quality of the text generated by the Conditional BERT model and enable indirect transformations. Furthermore, this LLM should incorporate relevant context information obtained from the initial sentence.

2.1 Purposed Solution:

To fulfill these requirements, we have introduced GPT-2 as the novel LLM in our approach. This model offers the advantage of being easily fine-tuned through transfer learning and also exhibits satisfactory inference performance. To facilitate the fine-tuning process, we have designed a new text input data structure and introduced new tokens specifically tailored for this purpose.



2.3 GPT-2 Training Process:

The GPT-2 model underwent a training process encompassing a limited number of 4 epochs. An examination of Figure 1 reveals the utilization of custom tokens, while Figure 2 exemplifies their processing as text. Evidently, the graphical representation in Figure 1 demonstrates a superior level of coherence and aesthetic appeal, thus rendering it a favorable training approach. Nonetheless, it should be noted that an extended training duration of 15 or more epochs may yield enhanced network performance. The allocation of additional time and computational resources to augment the training process could potentially produce more refined output.

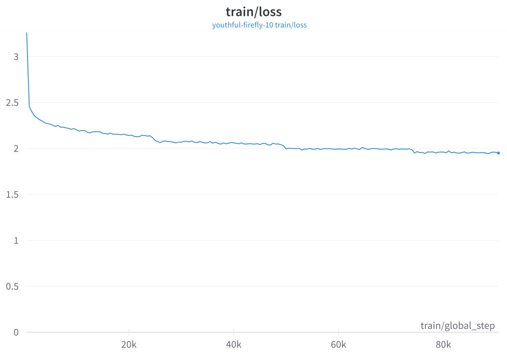


Figure 1

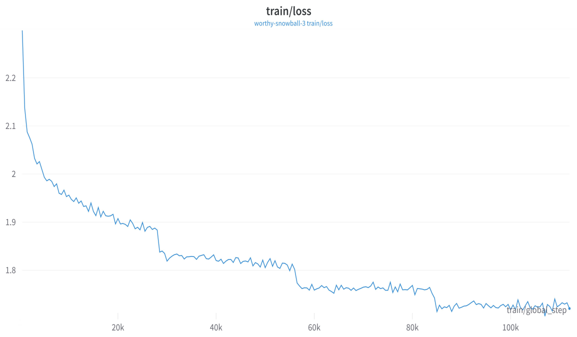


Figure 2

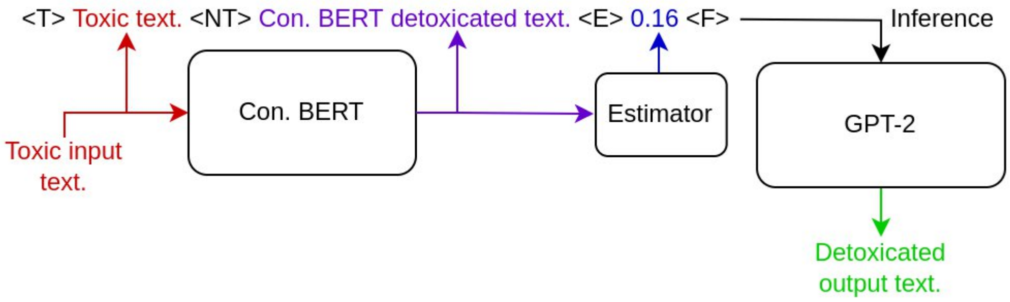
3 Estimator Hypothesis.

3.0 Problem State:

Another phenomenon observed is the suboptimal level of detoxification due to the simplistic nature of certain sentences. For instance, the initial phrase 'Does anal...' was transformed into 'Anal...', which still retains its offensive connotation. In order to address this issue, we propose the implementation of an estimator neural network. This network would provide an estimation of the toxicity level of the text generated by the Conditional BERT model. It is evident that the previously discussed problem related to semantic quality is closely intertwined with this issue.

3.1 Purposed solution:

To further improve the performance of the GPT-2 model, we sought to integrate the predicted toxicity estimations into its functioning. This would enable GPT-2 to simultaneously enhance the quality of the text and facilitate the detoxification process. Our approach is rooted in the understanding that LLMs, such as GPT-2, tend to exhibit superior performance in such tasks after undergoing sequential data distillation, which involves repetitive input from their own output.



3.2 Technical Details:

For the implementation of the estimator, we selected the simple Embedding with Fully-Connected layer network due to its simplicity and ability of fast inferencing full text without any sequence processing such as RNN’s. Additionally, we introduced a new token '<E>' to incorporate the estimated toxicity into the context of GPT-2. This necessitated the compilation of a restructured training corpus, specifically tailored to accommodate these modifications.

4 Results.

The challenge of enhancing the semantic quality has been effectively resolved, as evidenced by the GPT-2 model exhibiting superior performance on subjective evaluations when compared to the rudimentary Conditional BERT model. Notably, the achieved outcome represents a notable advancement from the initial baseline architecture. It is also noteworthy that the GPT-2 model entails an efficient inference process, rendering it feasible for deployment even on less powerful computing devices such as laptops. It is plausible to presume that with further training iterations, the model's capacity for generalizing complex phrases would significantly improve, thus warranting exploration of extended training procedures for heightened performance.

The second concept pertaining to the estimator encountered a number of challenges. The performance of the estimator neural network proved to be notably unsatisfactory. Upon closer examination, it was discovered that the issue resided in the initial dataset. Remarkably, the dataset exhibited a trend where sentences were predominantly classified as either 0.999 (indicating a high level of toxicity) or 0.001 (pointing to a significant lack of toxicity), with minimal representation of intermediate values (imbalanced dataset). It is conceivable that this classification imbalance arose due to the subjective judgments of the specialists responsible for curating and evaluating the dataset, as human assessors often display a tendency towards extreme viewpoints. This binary classification phenomenon presents a hindrance to our objective of accurately estimating toxicity levels, necessitating a dataset that encompasses a more uniformly distributed range of estimations. Regrettably, after applying normalization techniques to address this issue, the dataset becomes considerably reduced in size, rendering it unfeasible for training the estimator. In summary, our observations indicate minimal disparity when employing or excluding the estimator data in conjunction with GPT-2. Nonetheless, estimator model forces GPT-2 to learn scenarios similar to those previously described in Section 3.0.