**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.**

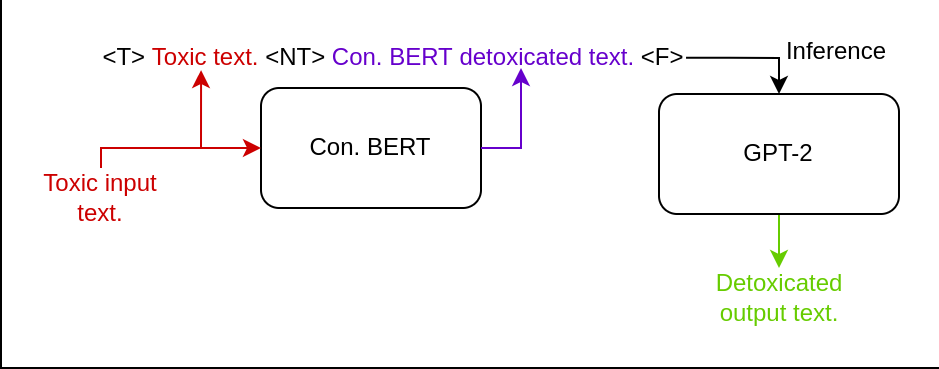
**2.0 Problem State:**

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.



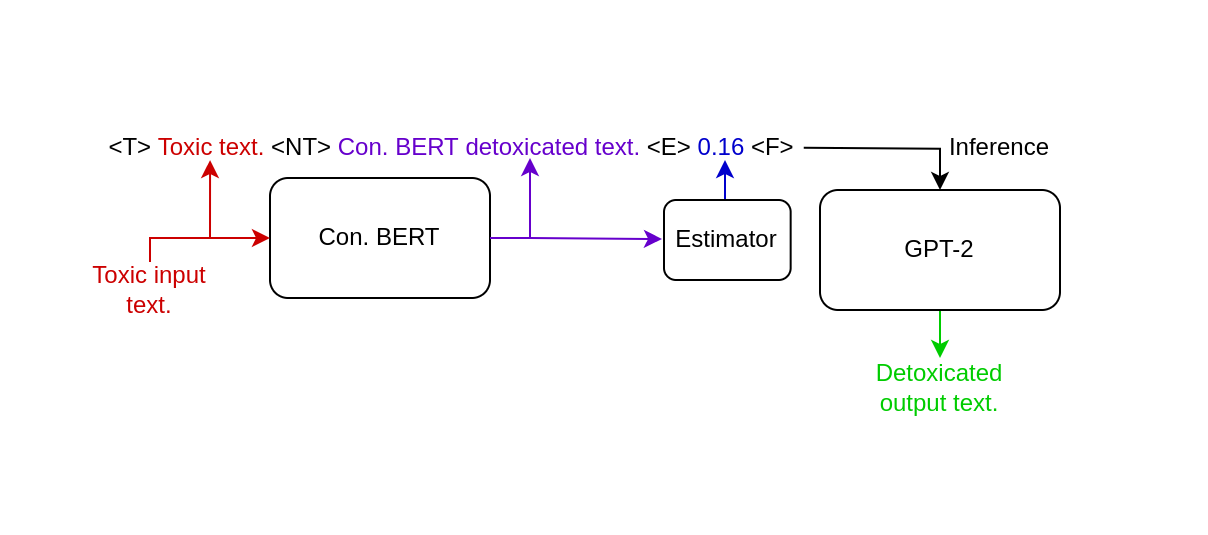
**3 Estimator Hypothesis.**

**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 Gated Recurrent Unit (GRU) due to its capacity to provide satisfactory single neuron output. 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.**