# Practical Machine Learning and Deep Learning - Assignment 1

# Text De-toxification via ConBEGPT model.

Final Solution Report

Parepko Leon

## Abstract.

The topic of automatically rewriting offensive content has received relatively limited attention thus far, despite its potential for various practical applications, such as promoting a more respectful and inclusive online environment by encouraging users to post more neutral versions of emotionally charged comments.

Previous studies have proposed a Conditional BERT architecture as a solution to this problem. While this approach has demonstrated notable efficiency in achieving its goals, it is not without drawbacks. It is hypothesized that by carefully considering the limitations and negative phenomena observed in the behavior of ConBERT approach, it is possible to enhance the results. By addressing these shortcomings, we can develop a more robust and effective method for automatically rewriting offensive content. So we define 2 key observations to fix:

- **Semantic quality:** One limitation of the Conditional BERT model is its struggle to accurately comprehend indirectly implied phrases, which hampers its overall effectiveness. The model exhibits difficulty in establishing comprehensive multiple-to-one token connections, thereby limiting its ability to capture the nuanced meaning behind certain phrases.

- **Insufficient detoxification:** An additional phenomenon observed in the behavior of the Conditional BERT model is the occurrence of suboptimal levels of detoxification. This arises from the simplistic nature of certain sentences, wherein only a few words are present. Consequently, the model lacks the necessary information to effectively analyze and neutralize potentially offensive language within such sentences. Moreover, in cases where a sentence is excessively long or complex, the Conditional BERT model may intermittently overlook or miss certain offensive words or phrases, thereby compromising its detoxification capabilities.

This research project proposes a novel approach to address the challenge of text de-toxification by leveraging the capabilities of Conditional BERT (Con. BERT) and GPT-2 models. The ConBEGPT network architecture is employed to enhance the quality of text generated by Con. BERT by incorporating the GPT-2 block. Through experimentation, it was observed that this integration resulted in improved overall network performance. However, the introduction of the Estimator concept introduced certain ambiguities. Both the inclusion and exclusion of the Estimator produced data, in conjunction with the GPT-2 model, yielded minimal differences in results. Despite this, observations suggest that the Estimator model effectively compels the GPT-2 model to learn specific scenarios, which we define as special cases of improvements sought through the Estimator approach.

## Data Analysis and Preprocessing.

The initial dataset underwent preprocessing and simplification to facilitate its interpretation. This involved selecting specific columns that contain relevant information, namely the text and its toxicity estimation. Consequently, the resulting dataset comprised four columns:

- **Toxic text:** This column contains the string of the original text with a high toxicity level.
- **De-toxificated text:** This column contains the transformed or paraphrased version of the "Toxic text" intended to sound less toxic.
- **Toxic text estimation:** This column captures the manually estimated toxicity levels of the corresponding text entries, as evaluated by professionals.
- **De-toxificated text estimation:** This column mirrors the "Toxic text estimation" column but corresponds to the "De-toxificated text" entries.

From this preprocessed dataset, two separate training subsets were constructed. The first subset, known as the GPT-2 training corpus, was intended for training the GPT-2 model using the given text samples. The second subset, referred to as the Estimator training dataset, served to train the Estimator model, during the subsequent training processes.

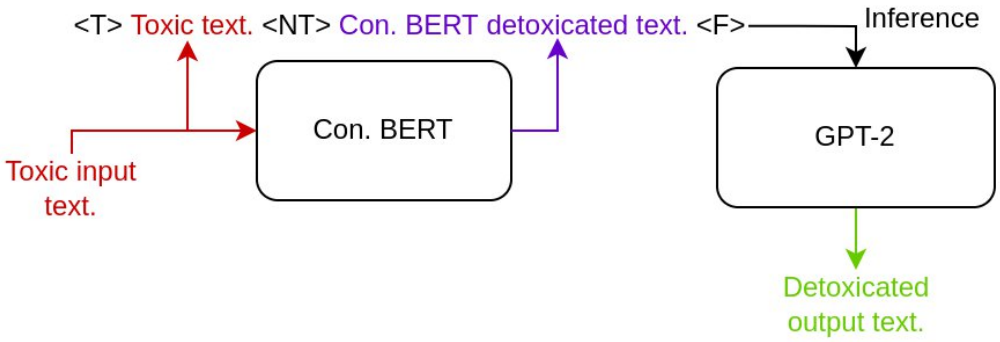
The estimator dataset involves shuffling toxic and de-toxified texts and undergoes several preprocessing stages required for the Estimator model. Initially, all the text is converted to lowercase and any numbers, punctuation, and other semantically useless information are removed. The text is then split into an array of words and tokenized using the standard 'nltk' tokenizer. This methodology is derived from the lectures and lab materials.

In the case of the GPT-2 training corpus, the preprocessor generates a string with custom tokens. It uses the toxic text from the new dataset as the '<T>' token, then generates data for the '<NT>' token using the pretrained Con. BERT. The toxicity of the Con. BERT generated text is further estimated to obtain the '<E>' token. Finally, the '<F>' token represents the non-toxic or less toxic sentence from the new dataset, serving as the final result.

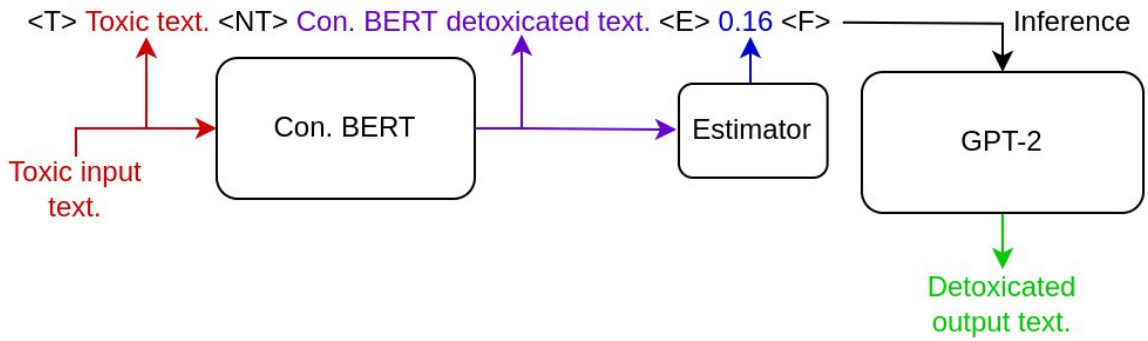
It is important to note that, upon closer examination, it was discovered an issue in the initial dataset. Remarkably, the dataset exhibited a trend where sentences were predominantly classified as either 0.999 (indicating a high level of toxicity) or 0.001 (pointing to a significant lack of toxicity), 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.

## Model Specification.

To fulfill the semantic quality problem, 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. The scheme showed below illustrates the architecture of the overall network with GPT-2 incorporate only.




To further improve the performance of the GPT-2 model and solve the insufficient detoxification problem, 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.




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.

# 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.

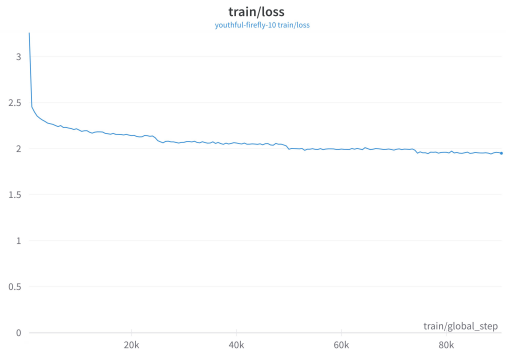



Figure 1


Figure 2

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.

The training process of the estimator model is relatively simple and can be done locally, which was done in this case. The preprocessed data, as described in the "Data Analysis and Preprocessing" section, was used in the same manner as the basic training techniques. The model was trained for 20 epochs using the entire dataset, as it showed active dynamics during training. However, the model did not yield satisfactory results and lacked robustness, mainly due to issues with the initial dataset as mentioned in the "Data Analysis and Preprocessing" section. Increasing the number of epochs or restructuring the dataset did not solve the problem. It appears that additional data with intermediate toxicity estimation would help balance the dataset and potentially improve the stability of the estimator model during training.

## Evaluation.

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 key issues described above might be named solved because manual tests on them showed different an much less toxic transformation produced by a network.

Notably, the Estimator component can be directly evaluated by comparing its predicted toxicity level with the professionally labeled toxicity level from the new dataset. To measure the performance, we used the mean squared error (MSE) as a metric, given that there is a single feature in the model's output. The MSE quantifies the average squared difference between the predicted and actual toxicity levels, providing an assessment of the Estimator.

## Results.

The integration of Conditional BERT and GPT-2 models using the ConBEGPT network architecture proved effective in improving the overall performance of text de-toxification and solve two issues we purposed as a main target. Through experimental analysis, it was observed that GPT-2 integration resulted in enhanced quality of generated text. However, the introduction of the Estimator concept brought about some uncertainties. Both including and excluding the Estimator in conjunction with the GPT-2 model yielded minimal differences in results. Nevertheless, it was noted that the Estimator model effectively compelled the GPT-2 model to learn specific scenarios. These scenarios can be defined as special cases of improvements sought through the Estimator approach.

The findings from this study underscore the potential for enhancing text de-toxification through the ConBEGPT network architecture. Further investigations are required to explore the precise impacts and implications of incorporating the Estimator concept in conjunction with the GPT-2 model. By refining and optimizing these techniques, it may be possible to achieve even more significant improvements in text de-toxification processes.