Project Report 2:

In the second report, ***describe your final solution****.*

Practical Machine Learning and Deep Learning Assignment 1 - Text Detoxification

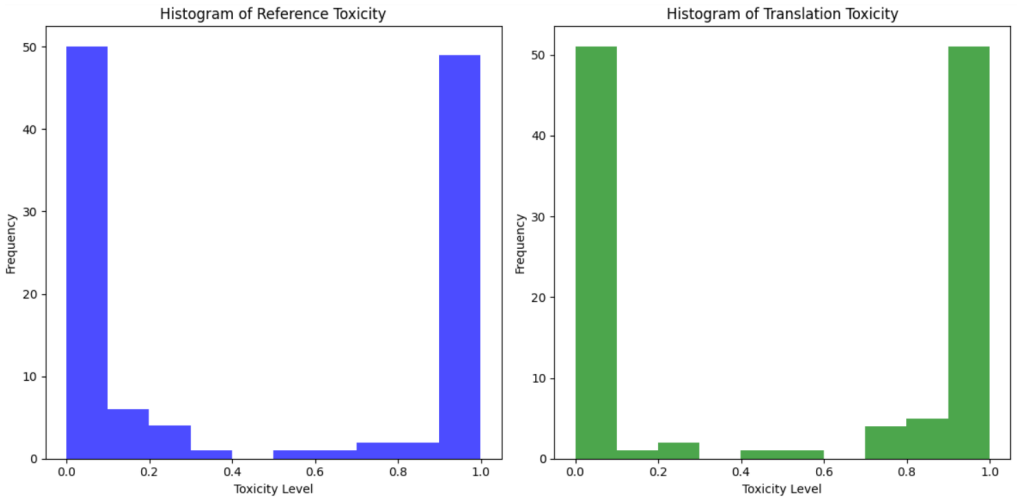
---

Introduction

The main goal of text detoxification is to **reduce rudeness** while keeping the **original meaning** of the sentence. In this report I will describe my solution of the text detoxification task using **pretrained model GPT-2**. I fine tuned GPT-2 on the filtered, preprocessed data, which were provided alongside with the assignment. Moreover, I **tested the toxicity** of my solution using the best toxic classification solution on Kaggle and measured the **semantic similarity** of the translated text using WebBertSimilarity.

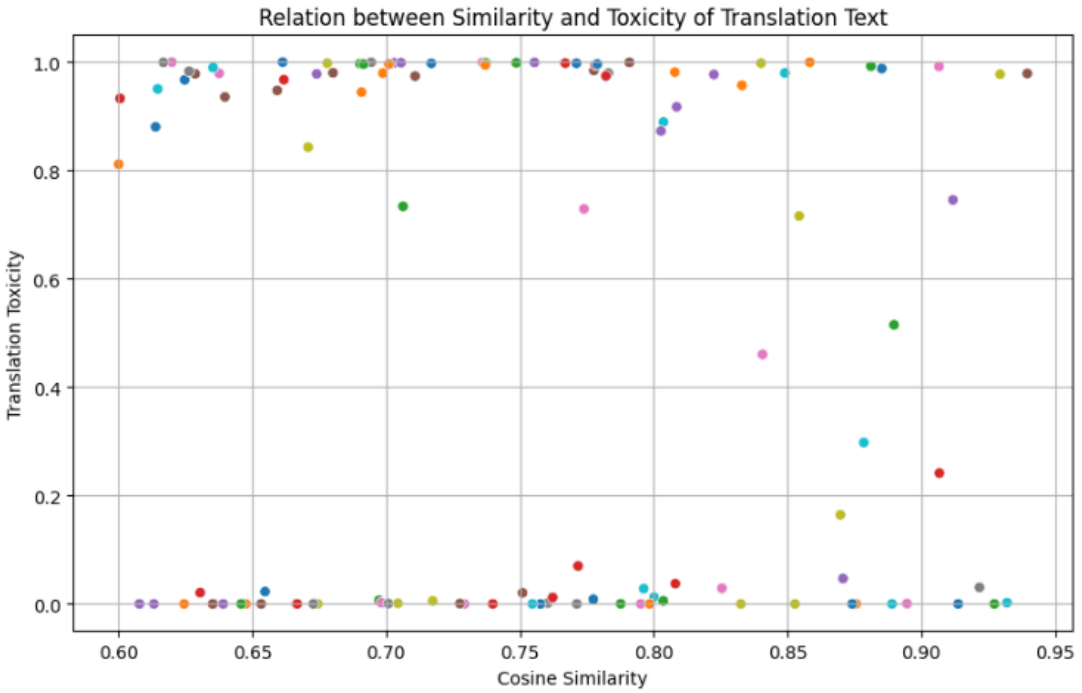
Analysis

The provided dataset consists of **577777 rows**, each row contains the **original text, its translation, their similarity, difference in their length, and the toxicity** of both of them. I explored the data, readed some samples and realised that with high probability the **translated text was created by a machine** not a human. On the histograms below it is visible that **roughly half of the translated data remains to be toxic**.



With a closer look at the data it turns out that **some reference text which was not initially toxic becomes toxic** with the translation. Though, the dataset contains also rows which were corrected well.

Also, **I explored how the similarity is related to the toxicity** of the translated text. On the graph below it is visible that the **similarity does matter much**. There are sentences which are very far from the original text and still toxic as well as very close to the original but improving its toxicity.



Training Data

Based on this data analysis I decided not to use all the provided dataset but to **filter** the data and use only these rows which maintain low toxicity after the translation. Also, I decided to **drop data containing the similarity or the difference in length**. I use only **100,000 sentences** which have the **best translation toxicity score**.

Each sentence must be preprocessed before training. For the GPT-2 pretrained model the training data have to follow the **pattern**:

*<s>Reference</s>>>>>> Translation *

For example:

*<s>You just fucking filmed it, yes?</s>>>>>> you just made your movie, didn't you? *

The model is then fed with data in such a format. This method was described in **Paraphrasing with Large Language Models** by Sam Witteveen and Martin Andrews (https://arxiv.org/pdf/1911.09661.pdf).

Model Specification and training

In the paper mentioned above authors use **GPT-2** for paraphrasing. Since GPT-2 showed quite good results in the paraphrasing and the detoxification is also a more specific version of the paraphrasing task I decided to use GPT-2.

Surprisingly, GPT-2 fine-tuned on a very small amount of data (only 100 filtered lines) performed quite well. However, for the final training I used **80000 rows of filtered data** (20000 I keep for testing).

Regarding the specifics of the training: I use **3 epochs** since it is enough to lower the **loss to 1.6**. I tried more epochs but it turns out that after the 3rd epoch the model does learn so fast and only waste the time.

Evaluation

The best way to evaluate the results of the detoxification is human based. However, such an evaluation is quite difficult, so I tried **3 ways** of automatic evaluation.

I use part of the given dataset, which was not in the training part.

First, I **measure the toxicity** of the translation. I do so with the use of **unitary/toxic-bert**. This solution provides a value ranging from **0 to 1**, where 1 means highly toxic and 0 not toxic at all.

Second, I use **Sacrebleu** metrics for evaluating the difference between the reference text and its translation. However, the metric is quite sensitive to changes and the translation requires replacing words or even removing them, causing the metrics to give very low marks. I tried to compare mine translations with the given translation, but the original translation sometimes are of a very low quality causing biases in the evaluation.

Third, I use **semantic similarity** to evaluate if the translation keeps the original meaning of the sentence. For that I use **WebBertSimilarity**. This metric returns a value from **0 to 5**, where 5 means very similar and 0 different.

Overall, these metrics altogether evaluate both - the **quality of the detoxification** and **similarity** of the meaning to the reference text.

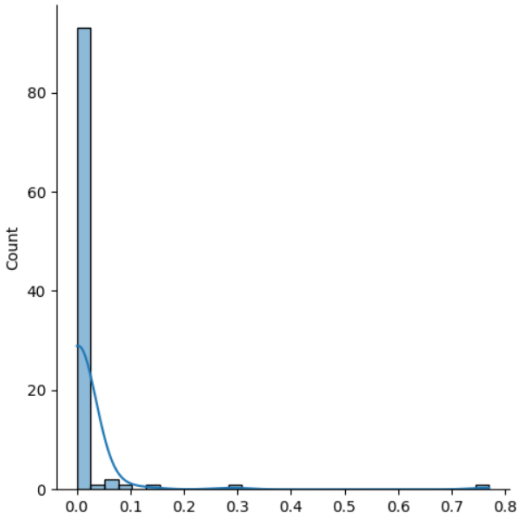
Results

1. **Toxicity of the translated text:**

   The evaluation of the toxicity of the translated sentences proved to almost fully detoxify the text. See the results of the evaluation bellow (the scale is form 0 to 1, where 0 dedicates low toxicity).

   ```
   Mean:   0.01655
   Median: 0.00126
   Mode:   0.00060
   Worst:  0.76984
   Best:   0.00060
   ```

   Here is a histogram depicting the distribution of toxicity of the translated sentences:



2. Sacrebleu

I would say that the main drawback for the Sacrebleu score in the detoxification is that it is too sensitive to even minor changes.

It turns out that the similarity of the reference text and translation is hard to measure, since often the correct translation requires removing words or replacing them. Hence, the length of the translated text differs, however the meaning remains unchanged. Therefore, the score is quite low, though the meaning of the sentence was not changed. See below the score.


```
{'score': 18.810553319365745,
 'counts': [461, 209, 112, 58],
 'totals': [998, 898, 798, 699],
 'precisions': [46.192384769539075,
  23.2739420935412,
  14.035087719298245,
  8.297567954220314],
 'bp': 1.0,
 'sys_len': 998,
 'ref_len': 920}
```

3. Semantic similarity

The model I use is very sensitive to changes in the sentences and also towards sentiment, hence even removing a single toxic word causes a drop in the score by 0.5 on the scale from 1 to 5, where 5 means very similar. Hence I consider it to be similar enough every sentence with the score above the 2.5 threshold. See below the measured values of semantic similarity.

```
Mean:   2.67735
Median: 2.75023
Mode:   0.83485
Best:  4.52251
Worst:  0.83485
```

4. Examples of translated sentences

Here are some examples of well translated sentences with high semantic similarity:

```
Original:  He thinks he's part of a frickin' platoon now.

Translated:     he thinks he's part of a platoon now.

-----------------------------------------

Original:  Can I please have a bit of peace and bastard quiet!

Translated:     I want to have a bit of peace and quiet!

-----------------------------------------

Original:  He did not call you the freakin' Golden Boy of BMS.

Translated:     he didn't call you the Golden Boy of BMS.

-----------------------------------------

Original:  40 kliks, no fucking water.

Translated:     40 kliks, no water.

-----------------------------------------

Original:  It's not wine. It's vinegar you fool
```

```
Translated:     it's not wine, it's vinegar

----------------------------------------

Original:  Damn it, where's Conklin?

Translated:     where's Conklin?
```

Here are some examples of not correctly translated texts:

```
Original:  Like fuck! thought Darcy.

Translated:      like it.

----------------------------------------

Original:  He's as imprudent as his father was at his age.

Translated:      he's quite an adversary, just like his father.

----------------------------------------

Original:  Destroyer!

Translated:      explosion!

----------------------------------------

Original:  "What the hell goes on?"

Translated:      what's going on here?

----------------------------------------

Original:  You know how to play your role damn well.

Translated:      you know how to play your music.

----------------------------------------

Original:  The diagnosis is much simpler, he's a jerk.

Translated:      the diagnosis is a miracle.
```

Limitations and possible improvements:

I would say that the biggest problem is to find some optimal metric, as I showed in the evaluation, the task is important not only to keep low toxicity but also to keep the same meaning. I think it is necessary to combine the semantic similarity with toxicity metrics to achieve better results.

The solution I proposed has sometimes mistakes in paraphrasing long and complicated sentences with high level of toxicity. Also there are mistakes in translating very short sentences including highly toxic words, in such examples the model usually removes the toxic word and only a little piece of the original sentences remains without any meaning. This causes outliers in the scores.

Regarding the false/positive score it can be expressed as the ratio of the high toxic to high similarity and low toxic but also low similarity. The evaluation showed that the second is more likely to occur, having high similarity but with high toxicity is quite rare in the translation.

Conclusion

Overall, I would say the solution was successfully implemented and the results fulfilled the expectation. The translated sentences are of very low toxicity and they keep the original meaning most of the time.