Project Report 1:

*In the **first report**, describe your **path in the solution creation** process. List any **architectures**, **ideas**, **problems** and **data** that leads to your final solution.*

Practical Machine Learning and Deep Learning Assignment 1 - Text Detoxification

---

Hypothesis 1: **Using 2 way translation using BERT**

I tried to use a double way translation (without any fine-tuning). First, I translated the reference text into Russian and then back from Russian to English, hoping the translation model has some embedded polit-correctness sense. However, it turns out that the translation module has a solid knowledge of toxic words both in Russian and English language and it has no problem to translate them without changing its toxicity.

Hypothesis 2: **Searching for possible toxic words**

I have found a dataset, from which it would be possible to extract a small database of the most often toxic words. I wanted to use it and replace or remove these words from the text. However, it needs some handmade substitution for every single toxic word. Though such a result is stable, it does not provide enough paraphrasing power, since it generically substitutes dictionary based values.

Hypothesis 3: **Low toxicity causes low similarity**

Based on the data analysis I concluded that it is possible to translate a highly toxic sentence to one with low toxicity level while keeping them very similar. There is little to no relation between the similarity and toxicity.

Hypothesis 4: **Using a random sample of the data**

First, I wanted to use a randomly chosen portion of the given data, so that the different topics, vocabulary and style is used. However, the provided dataset seems to be a hot human translated, but rather a machine translated. Moreover, approximately half of the translations were unsuccessful.

Hypothesis 5: **The smaller loss, the better**

Actually, in this task, there is one important tradeoff - the similarity and the toxicity. One of the biggest problems is the fact that the similarity might be syntactic (the sentences look similar) or semantics (the sentences mean the same, but might use different words). Most of the loss functions can take into the account only the similarity (most often just the syntactic one). Therefore, the standard (default) loss functions do not bring a trustworthy message about the real power of the model.

Results:

Overall, I learnt from the previously mentioned tries, these important things:

- ➔ The data for fine-tuning have to be carefully chosen - I decided to sort them according to the toxicity of the translated version and choose the best translated version (those with the lowest translation toxicity).
- ➔ I did not use any additional dataset, since I found enough data in the provided one after sorting it and choosing only correctly detoxified samples.
- ➔ I decided to fine-tune a gpt-2 model, which performed quite well.
- ➔ For the evaluation I decided to measure the toxicity of the generated outputs by the best solution on Kaggle, which classifies the comments to be or not to be toxic (based on BERT)