Solution Building Report

Konovalova Oksana, BS21-DS-02

1. Data exploration

Exploring the data, I made two key observations. Firstly, I was delighted to find that there are no empty values, which eliminates the need to handle missing data during preprocessing. Secondly, it seems that the data is initially mixed up. Upon closer examination, it became apparent that all toxic sentences should be placed in the "reference" column, while the non-toxic ones should be in the "translation" column. This reorganization of the data would certainly facilitate work with it.

2. Baseline solution

While still in the process of researching the data, the idea of creating a simple algorithm that simply removes toxic words from a sentence arose, but such an algorithm requires a list of toxic words. In order to better evaluate the performance of the algorithm, I prepared two such lists. The first one I compiled during the data research phase, the second one I found on the Internet. As a result, I can say that this algorithm is not suitable for this task, because in some situations the sentence becomes incorrect, in other situations the meaning is lost because the algorithm does not look at the context.

```
24. Before the algorithm: 'Shut up, you two, 'said Granny.
After the algorithm: ' up, you two, ' said Granny.
40. Before the algorithm: Fuck! Get out of the fucking way!
After the algorithm: Get out of the way!
41. Before the algorithm: Trying to kill Ethan.
After the algorithm: Trying to Ethan.
```

3. BERT model

After I tried the basic solution, I concluded that it is important to consider the context of the sentence, so while studying the problem, I found a solution such as the BERT model. Since I used a pre-trained model, I again need to have a list of toxic words. I chose a list from the Internet because it works better with the basic algorithm. In the end, I concluded that this model is not so bad for solving this problem, since it looks at the context of the sentence and inserts the appropriate words. Unfortunately, the model depends on a list of toxic words, so the result was not good enough.

```
Cosine similar: 0.6708203932499369
Predicted: let's drink to that.
Target: Let's drink to somethin' else.
Source: let's drink to fuck.

Cosine similar: 0.801783725737273
Predicted: you want to see another doctor first?
Target: You want to annoy another doctor first? Eventually...
Source: you want to molest another doctor first?

Cosine similar: 1.0
Predicted: well, that was good.
Target: well, that was good.
Source: Damn,that was good.
```

4. T5

After working with the BERT model, I came to the conclusion that it is worth either finding a more suitable list of toxic words, or finding a solution that does not depend on the list. When I started looking for a solution, I found the T5 pre-training model, this model does not depend on the list of toxic words, and according to research it gives good results. I managed to make fine-tuning models on a small amount of data, but unfortunately, due to lack of time, I could not do this on large data, so I decided to focus on the BERT model.