Assignment 1 - Text De-toxification - Report 1

Sergey Golubev, B20-AI-01 student

s.golubev@innopolis.university

Solution link

Hypothesis 1: N-gram model

Firstly, I was planning to build an N-gram model as a baseline solution. The intent to implement this kind of model was to simply mark some words as toxic ones, and replace them with non-toxic N-gram endings. However, I left this idea after I discovered the main assignment dataset is too complicated for this approach.

In particular, some entries could be corrected well with the N-grams, e.g., `I didn't fuck him` -> `I didn't screw him`. But others are unlikely to be de-toxified well, because they don't just contain several toxic words, but utilize complex syntactic structures that themselves represent toxicity, or sarcasm. For instance, `I'm going to hit you in all directions, civil and criminal, on all counts` - the N-gram model wouldn't even recognize any toxic smell here.

Hypothesis 2: Paraphrasing

notebooks/3.0-t5-finetuning.ipynb
notebooks/3.1-t5-inference.ipynb

I decided to look for a more robust approach, and discovered that the problem could be solved with paraphrasing. The text intent will be preserved after high quality paraphrasing. And the high quality paraphrasing could be achieved with some large seq2seq model. So, the idea is to fine-tune a powerful text2text paraphrasing model on the given dataset.

There is a problem with de-toxification via paraphrasing. A good paraphraser model would change the text toxic segments, however there is a probability for the output to

preserve toxicity. Indeed, it is often easier to paraphrase a toxic sample into another one that would appear toxic, too. There are two points that helped me to overcome this:

1. In general, there are more non-toxic words than toxic ones. A powerful paraphraser would come up with several correct options for each input. So, we only have to choose the best paraphrase, i.e. the least toxic one.
2. I decided to evaluate the toxicity level with a separate helper model, and choose the paraphrase option that has the minimum toxicity.

Hypothesis 3: CondBERT Inference

notebooks/4.0-condbert-inference.ipynb

As the second solution, I decided to perform an inference of some powerful text de-toxification model and compare the results to the Hypothesis 2 implementation.

For this purpose, I used the CondBERT model for text de-toxification proposed in the paper. You can find more details on it in the Second Report.

Results

Please refer to the Second Report.