# Solution Building Report

Karina Denisova, BS21-DS-01

November 5, 2023

## 1 Baseline: Delete

The study used a basic text removal approach. This approach is based on a simple idea borrowed from the media sphere: you will often notice that profanity is replaced with asterisks or removed from television shows or articles. Based on this principle, I decided to "hide" profanity by removing it from the text. For this purpose, a dictionary with profanity was taken. After tokenizing and lemmatizing the sentence, we remove the tokens found in the dictionary from the input toxic sentence.

Using this approach allows for initial filtering of profanity and making the text more acceptable to the audience. However, it should be noted that this approach has its limitations and is not an ideal solution to completely solve the problem of text toxicity.

| **Toxic Sentence** | **Detoxified Sentence** |
|---|---|
| I like that shit. | I like that. |
| Does anal... | Does ... |

Table 1: Toxic Sentence and Detoxified Sentence Examples

As we can see Baseline successfully delete all "bad" words so sentence become non-toxic. In most of the cases we can follow the referenced text contest, but in some cases not...

## 2 Hypothesis 1: Using T5 model

As Baseline not always can follow the content, I decided to use some model that can do paraphrasing tasks and pay attention to the referenced text. I used pre-trained small T5 model checkpoint from HuggingFace. To make this model not simply paraphrased, but detoxified text, I fine-tuned it in parallel dataset (toxic/detoxified)

| **Toxic Sentence** | **Detoxified Sentence** |
|---|---|
| He could actually die here. | he could have died. |
| Everybody nice and fucking comfy now? | are you all comfortable? |

Table 2: Toxic Sentence and Detoxified Sentence Examples

So as in shown in a table model starts to perform way better results in compare to baseline, because t5 can follow the content way better then just deleting words

## 3 Results

I made some experiments about detoxification sentences. As it shown from examples,there is still a need to improve methods. Sometimes it is enough to remove obscene words from the text, and in others, detoxification of texts is possible if they are completely reformulated. The most promising direction for development may be considered to be a combination of all the presented strategies and their application depending on the nature of toxicity in specific proposals.