Building solution report

Hypothesis 1: A simple dictionary

The first idea that came to my mind was simply using a dictionary to replace toxic words with ordinary ones.
To do this, I searched for bad words in each sentence and what they are replaced with in the translation.

Input: Help me, you cunt!
Output: Help me, you forgot!

This was an example of how this dictionary will work.
The disadvantages of this method is that we do not check whether the replacement was toxic and because of this, the effectiveness of the model decreases
This algorithm reduces the level of text toxicity by ~61%

Hypothesis 2: BERT model and smart dictionary

Then I decided that I needed to choose words more grammatically, for this I started looking for more information about detoxification and from sources I realized that the BERT model would cope well with this task, created a list of bad and good words and already with confidence that the toxic word is replaced by a less toxic one, I used the model

Input: Oh, crap!
Output: oh , what !

This was an example of how this model will work.
The disadvantages of this method is that sometimes toxic words do not seem to the model as such.
This algorithm reduces the level of text toxicity by ~62%

Hypothesis 2: Final solution

Studying the results of how previous algorithms detoxified the text, I noticed that often when 1 solution has a toxic translation, 2 is normal, and vice versa. so, I decided that we could compare the results of each algorithm and insert into the resulting dataset the translation where the level of toxicity would be less.
This algorithm reduces the level of text toxicity by ~80%