PMLDL Fall 2023. Assignment 1.

Adelina Kildeeva B21-DS-02

a.kildeeva@innopolis.university

Text Detoxification

Solution building report

The main goals of detoxification are to reduce the toxicity of the sentence and preserve meaning. I decided to use these metrics to measure the performance of algorithms.

Baseline

Remove the “toxic” words from the text. I found a list of inappropriate words on the internet and filtered the sentences by simply deleting occurrences of bad words. This method removes toxicity from the text, but is limited in the word list and loses the logic and meaning of some sentences.

With this method I obtained the improvement of toxicity from 0.737 to 0.495.

Hypothesis 1

Use pretrained BERT model that will replace toxic words with their neutral synonyms. We start by replacing all inappropriate words with a mask, and then the BERT model inserts the most contextually appropriate words in place of the mask. So, we remove the toxic words while leaving the meaningfulness of the text. However, this method is limited to replacing only certain words, whereas sentences may contain larger toxic parts.

The toxicity reduced from 0.737 to 0.55.

Hypothesis 2

Utilize a more comprehensive list of toxic words. It would make the texts in the baseline solution even less understandable, so I didn't apply the new vocabulary to it. New vocabulary improved the performance of the BERT solution as the model is able to fill in the gaps.

The toxicity improved from 0.737 to 0.314.

Results

I settled on the option of replacing toxic words using BERT. This is a simple but quite effective way to somehow solve the problem.