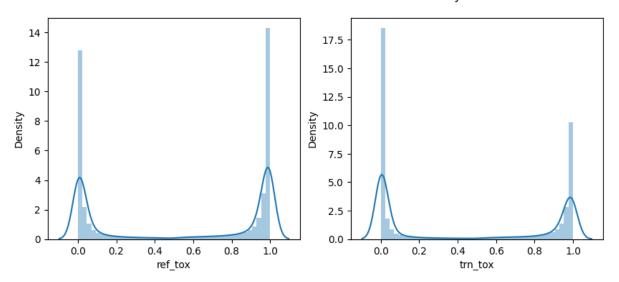
First Report

1. Data analyzing:

The path in solution began with analyzing the dataset. After studying "reference" and "translation" sentences I have got the following distribution:



Distribution of Reference and Translation toxisity

I discovered that toxic text and neutral ones were mixed. Additionally, there are some sentences with toxicity scores around 0.5. Such sentences will not affect training results.

2. Data pre-processing:

- By analyzing dataset, I dropped the rows with toxicity level around 0.5
- Next, I changed non-toxic reference sentences with corresponding toxic translation sentences. Here I visualized the words that are the most frequently used in toxic style text:



• I removed the punctuation, separated the dataset into train and validation sets and tokenized all sentences.

3. Models:

- First attempt was to fine tune pretrained
 Vamsi/T5_Paraphrase_Paws · Hugging Face model. However, due to computational resource limitations, we couldn't train this model on the whole dataset.
- Next, I attempted to fine tuned lighter models: <u>t5-small · Hugging Face</u> and <u>mrm8488/t5-small-finetuned-quora-for-paraphrasing · Hugging Face</u> on a smaller data sample. They both provide bad results due to lack of the data.

Before	After	Before	After
0.908439	0.6947	0.971542	0.5256
0.986022	0.1955	0.984656	0.0423
0.957481	0.1968	0.998577	0.0686
0.998181	0.6947	0.963460	0.2158
0.715417	0.1395	0.986255	0.9131
0.999102	0.6947	0.978353	0.1601
0.957807	0.0263	0.996818	0.0491
0.973515	0.8552 y	0.991037	0.0404
0.991525	0.0821	0.971545	0.3608
0.999473	0.2306	0.980846	0.6521

Fig 1. Scores in t5_small (Left) and t5_paraphraser (Right)

4. References:

- https://arxiv.org/abs/2109.08914
- https://github.com/skoltech-nlp/detox/releases/download/em nlp2021/