Polina Bazhenova: p.bazhenova@innopolis.university

Final Report

1. **Introduction:**
   This report presents the solution of the text detoxification problem. The main task was to paraphrase toxic style texts to non-toxic ones. The solution contains two T5 models and one simple Seq2Seq LSTM model. All models were trained on the filtered ParaNMT-detox dataset which you can download from here.

2. **Data Preprocessing:**
   - 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.
   - I transformed all sentences to lowercase, removed the punctuation, separated the dataset into train and validation sets and tokenized all sentences.

3. **Models:**
   - T5 models:
     For this problem I tried to evaluate pretrained t5-small · Hugging Face and mrm8488/t5-small-finetuned-quora-for-paraphrasing · Hugging Face models. However, due to computational resource limitations, we couldn't train this model on the whole dataset.
     That is why, they were fine tuned on the smaller sample. To estimate them I trained the model to establish the toxicity score on the same filtered ParaNMT-detox dataset. It provided good validation results. However, results of t5 models were not good due to lack of the data.

   - Seq2Seq LSTM model:
     This model has two main components: an encoder and decoder. The encoder is responsible for processing the input sequence and capturing its meaningful representations. The decoder generates the output sequence based on the encoded representations. Both of them use a single LSTM layer in themselves. Again this model could not be trained on the whole dataset, but it takes a larger sample than previous ones. Unfortunately, this model was not estimated for now.