Solution Building Report

Baseline

I explored the dataset and found out that there are many words that are present in the toxic dataset, but not in the detoxified dataset. This gave me an idea for the baseline model. The paper [1] suggested in the assignment statement also described the method. I filtered the words that are “bad” (that occur in toxic dataset much more frequently than in detoxified). Then, I searched through the sentences, masked them one by one and let the pretrained BERT [2] model guess which word was there. I repeated the procedure until there were no “bad” words in the sentence.

Example:
And what the fuck is up with your white knight? -> And what the [MASK] is up with your white knight? -> And what the heck is up with your white knight?

Advantages:

- No need for training
- The model used is trained on a larger dataset and is therefore good for human-like text generation
- Simplicity

Limitations:

- The meaning of the sentence might be corrupted because the model does not see the “bad” word (I feel like someone just fucked up.-> I feel like someone just woke up)
- If the “toxicity” of a sentence is formed with something but the “bad” words, then the model won’t change anything

Results on a random sample of 1000 sentences:

BLEU (zeros excluded)	ROUGE1 F1	ROUGE2 F1	Toxicity	Semantic Similarity with original sentences

0.3615	0.5722	0.3318	0.3154	0.8724

Hypothesis I: Pretrained T5

For this hypothesis I have found a T5 model [3] that was pretrained on the dataset. This approach allows us to save time and resources. Additionally, the model achieves good results in the task.

Example:
And what the fuck is up with your white knight? -> What's up with your white knight?

Advantages:
- No need for training
- Trained for this task
- State-of-the-art architecture
- Meaning of the sentence is preserved

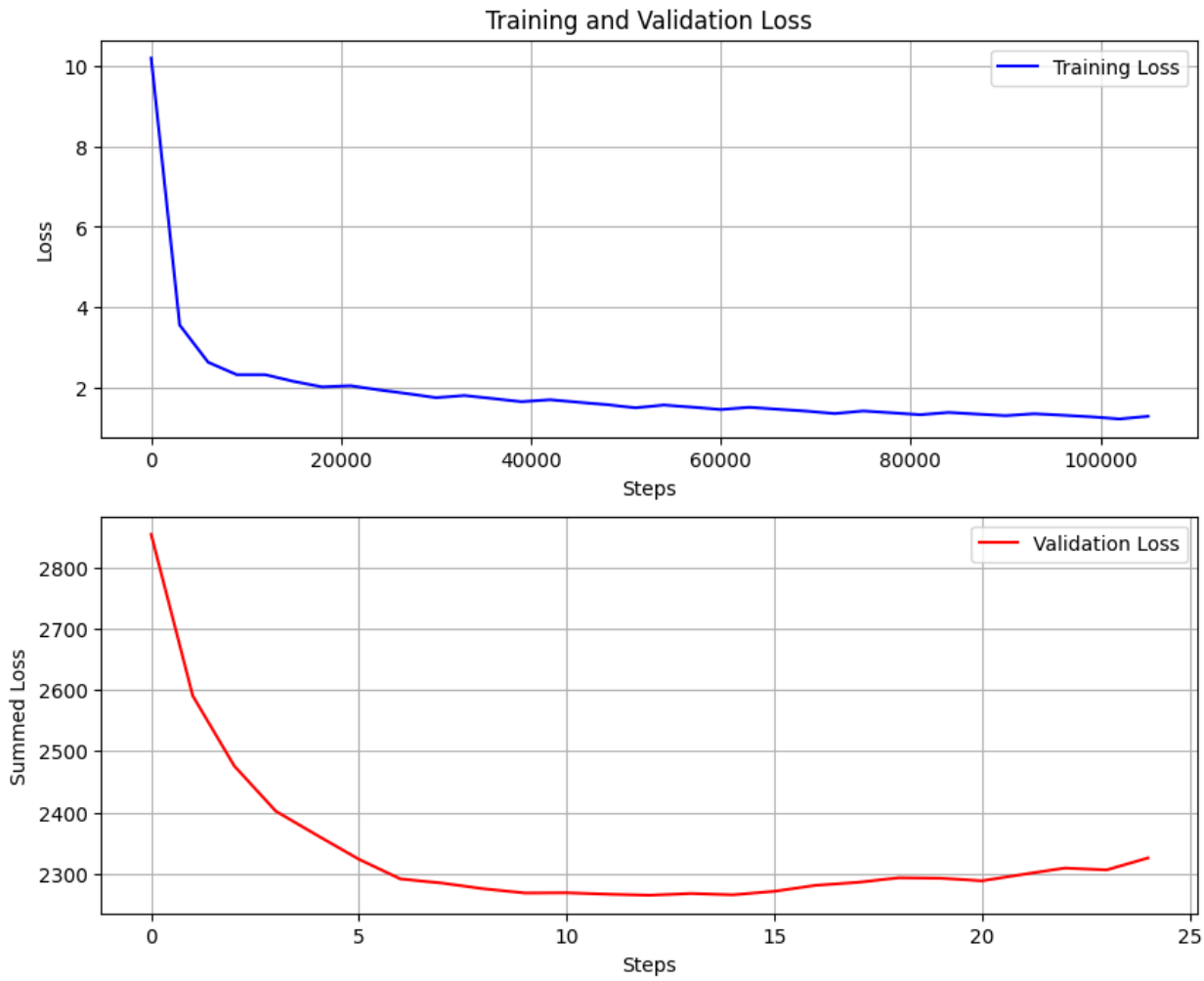
Limitations:
- The model learnt to produce text from the dataset which includes many bad quality translations

Results on a random sample of 1000 sentences:

BLEU (zeros excluded)	ROUGE1 F1	ROUGE2 F1	Toxicity	Semantic Similarity with original sentences
0.4175	0.5558	0.3023	0.3346	0.8535

Hypothesis II: Training Transformer

Another idea is to create a custom transformer [4] model trained especially for this task. Transformer architecture has shown itself outstanding for seq-to-seq tasks, so I decided to implement my own Transformer model. As an architecture I have taken the one shown in the tutorial [5] (8 attention heads, 3 encoder and decoder layers, 4 as a generator dimension). I also used some code parts shown in the video. Due to limited computational resources, I limited the size of the vocabulary to 25 000 and the sentence length to 128. The number of epochs used was 25, however the model seemed to overfit after approximately 10-15 epochs, so I saved the model with the best validation loss. The loss graph is the following:



Example:
And what the fuck is up with your white knight? -> And what is up with your white knight ?

Advantages:

- The architecture good for the task
- The model trained on the dataset
- Meaning of the sentence is preserved

Limitations:

- The model learnt to produce text from the dataset which includes many bad quality translations
- Computationally expensive, which implied cut vocabulary and simple architecture

Results on a random sample of 1000 sentences:

BLEU (zeros excluded)	ROUGE1 F1	ROUGE2 F1	Toxicity	Semantic Similarity with original sentences
0.5269	0.6175	0.4051	0.1603	0.7668

Results

Model name	BLEU (zeros excluded)	ROUGE1 F1	ROUGE2 F1	Toxicity	Semantic Similarity with original sentences
Baseline	0.3615	0.5722	0.3318	0.3154	0.8724
Pretrained T5	0.4175	0.5558	0.3023	0.3346	0.8535
Transformer	0.5269	0.6175	0.4051	0.1603	0.7668

Baseline model has shown acceptable results, however, the metrics BLEU, ROUGE and Toxicity metrics are not perfect. It is unable to capture complex toxicity and sometimes preserve the meaning of the text. Pretrained T5 and custom Transformer have shown similar results. The preservation of the meaning is worse forTransformer, but T5 loses in terms of other metrics. Even though both models have shown somehow good results, I would like to take as a final solution my own model, because the main task was to reduce toxicity, and the model showed the best result in this, preserving the meaning good enough. Also, I can better describe model specification and training process for my own model