Final Report

Konovalova Oksana, BS21-DS-02

Introduction

Text detoxification is a process designed to identify, filter, or cleanse digital text content to remove or replace elements that may be considered offensive, harmful, or inappropriate. It aims to create a safer and more inclusive digital environment by reducing the presence of toxic language, hate speech, profanity, or any other content that violates community guidelines or legal regulations.

Data analysis

As I examined the data, I made two key observations. First, there are no null values, eliminating the need to handle missing data during preprocessing. Secondly, the data is mixed up. Upon closer inspection, it became apparent that all toxic sentences should be placed in the “source” column, and non-toxic ones in the “translation” column. This reorganization of data will certainly make working with it easier.

Model Specification

I'm using BERT base model (uncased)pretrained model on English language using a masked language modeling (MLM) objective. BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts.

Evaluation

Comparison a predicted sentence with a target sentence using cosine similarity.

```
Cosine similar: 0.6708203932499369
Predicted: let's drink to that.
Target: Let's drink to somethin' else.
Source: let's drink to fuck.

Cosine similar: 0.801783725737273
Predicted: you want to see another doctor first?
Target: You want to annoy another doctor first? Eventually...
Source: you want to molest another doctor first?

Cosine similar: 1.0
Predicted: well, that was good.
Target: well, that was good.
Source: Damn,that was good.
```

```
Cosine similar: 0.4999999999999999
Predicted: I m a good cook.
Target: I'm a terrible cook.
Source: I’m a pathetic cook.

Cosine similar: 0.5477225575051662
Predicted: you hit your own head.
Target: You banged your head real bad.
Source: you hit your fucking head.

Cosine similar: 0.8117077033708014
Predicted: first my potatoes, then my tomatoes, then my salad, and now my salad and now the green beans.
Target: First my potatoes, then my tomatoes, then my lettuces, now my goddam beans.
Source: first my potatoes, then my tomatoes, then my salad, and now my salad and now the damn beans.
```

Results

I made a text detox solution using the BERT model. This model depends on a list of toxic words, so it is necessary to have a quality list. The model shows good results, it does not change the grammar of the sentence, but in some cases it slightly changes the meaning, but this is not very critical, because it copes with its intended purpose.