PMLDL Fall 2023. Assignment 1.

Adelina Kildeeva B21-DS-02

a.kildeeva@innopolis.university

Text Detoxification

Final solution report

Introduction

Text Detoxification Task is a process of transforming the text with toxic style into the text with the same meaning but with neutral style. 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.

The process of determining the solution and the main results of other algorithms can be found in the Building solution report.

Data Analysis

After some data exploration I can draw some conclusions:

1. Some reference-translation pairs of the texts are mixed-up. I swapped out them to keep toxic texts in the reference and their detoxified versions in the translation.
2. The level of toxicity depends on the presence of toxic words in the text.

Final Solution

For the final solution, I decided to use pretrained BERT model that replaces toxic words with their neutral synonyms.

Model

To replace words with their neutral synonyms I used BERT base model (cased). It is a pretrained model on English language that uses a masked language modeling (MLM) objective.

This model has the following configuration:

- 12-layer
- 768 hidden dimensions
- 12 attention heads
- 110M parameters.

More information can be found here: https://huggingface.co/bert-base-cased.

Vocabulary

The algorithm replaces words in the text defined in the toxic word list. The toxic word list consists of 1929 occurrences and is taken from: https://github.com/Orthrus-Lexicon/Toxic

Evaluation

I used the toxicity metric (https://pypi.org/project/detoxify/) and cosine similarity to measure the performance of the algorithm.

Table 1 compares the metrics for the reference, translation, and final replacement with BERT algorithm.

	Reference (toxic text)	Translation (detoxified text)	Replacement with BERT algorithm
Toxicity*	0.73673	0.16165	0.31616
Cosine similarity	1	0.61879	0.91552

* The toxicity metric was calculated on 2000 random elements of the dataset.

Table 1.

Examples of work

Table 2 shows examples of algorithm work.

Input	Output
Doesn't anybody in this town speak in complete fucking sentences anymore?	Doesn't anybody in this town speak in complete coherent sentences anymore?
you even tried to wipe your butt off.	you even tried to wipe your face off.
My eyes are fucked up.	My eyes are tearing up.

Table 2.

Results

I wrote a simple algorithm to replace toxic words with their neutral synonyms using a pretrained BERT model. This algorithm shows good results: it reduces the toxicity of sentences, preserves the meaning of the text, and does not break the grammatical logic of texts. However, it is limited in the list of toxic words and is unable to replace toxic parts of sentences consisting of several words. Transformers (e.g. t5) can be used for more accurate text detoxification.