TEXT DETOXIFICATION

PMLDL Assignment

Authors
Stepanova Anastasiia
aa.stepanova@innopolis.university
Innopolis University, Nov, 2023

Contents

What is the detoxification problem?

> ”On Definition of Toxicity Toxicity is an umbrella term for almost any undesirable behaviour on the Internet. It ranges from “mild” phenomena like condescending language (Perez Almendros et al., 2020) to grave insults or oppression based on racial or other social-demographic characteristics.” *source*

The task at hand is to enhance the politeness of the text input to my model while minimizing any loss of meaning.

From the given problem statement, I have identified three key components that need to be defined: the model, data preprocessing techniques, and evaluation metrics.

1 Exploring the Data

The ParaNMT-detox corpus dataset consists of several components: ”reference” (the original text), ”translation” (a proposed paraphrasing of the text), ”ref_tox” (the toxicity level of the reference text), and ”trn_tox” (the toxicity level of the translation).

The distribution of text lengths (see fig. 1) shows that the majority of the dataset falls within the range of 50 to 140 characters.

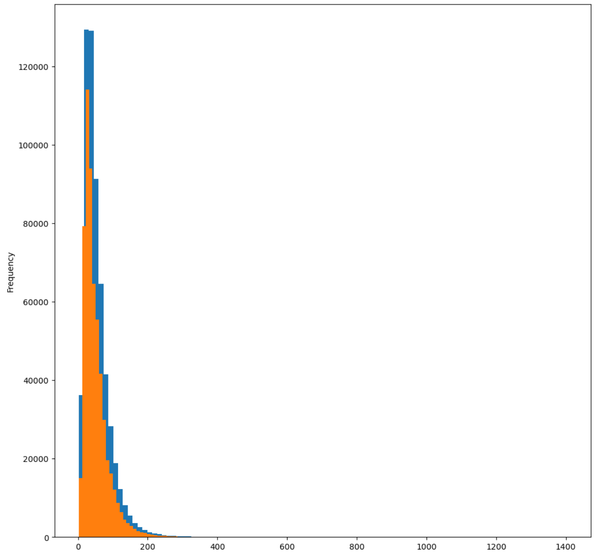


Figure 1: Distribution of the texts length

It is worth noting that in the ParaNMT-detox corpus dataset, there are over 250,000 samples where the toxicity level of the reference text is lower than that of the translation.

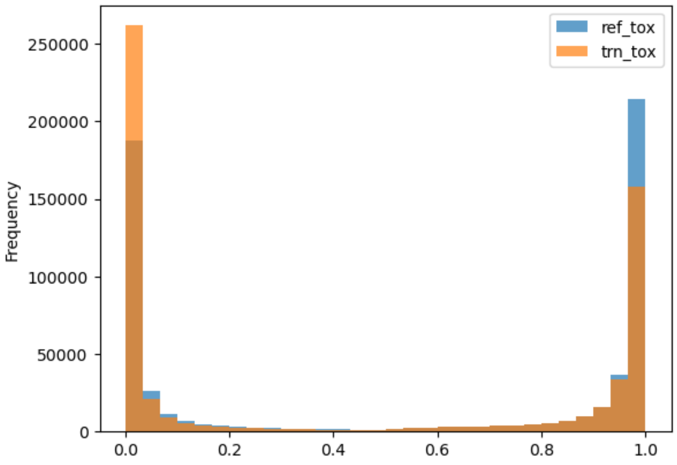


Figure 2: Distribution of text toxicity

Since I intend to use the data in the ”reference” column as input for training and the data in the ”translation” column as the target, I need to swap them (see Fig. 3).

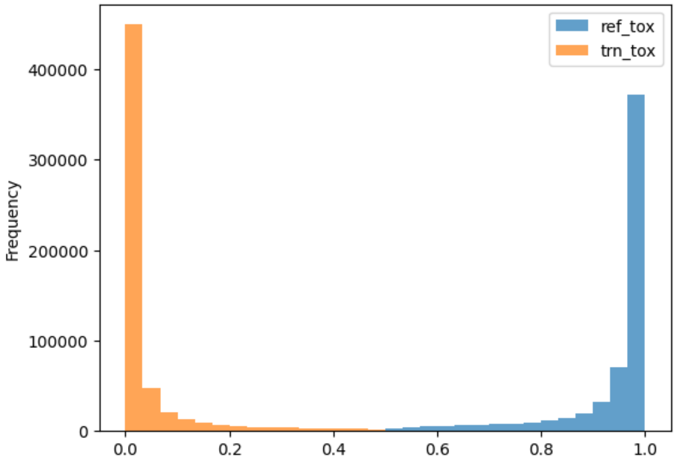


Figure 3: Distribution of text toxicity after swapping

That’s an improvement!

In the proposed *article*, the authors used an architecture that replaces a flag (a mask for toxic words) with predictions. To use this architecture, I need to mask the toxic words in the ”reference” column of the dataset. However, this method will only work if the data passed to the model contains words exclusively from the toxic set. The approach is not universal. In the article, the authors explain that not only ”bad words” determine the toxicity of a sentence, but humans can also use seemingly neutral words to convey a toxic meaning.

For future work, I performed the masking process in the ”data-filtering.ipynb” notebook. I utilized the toxic dictionary and replaced all the toxic words with the flag [”MASK”].

2 NLP models exploration

To select a model, let's define the task in NLP terms: the process of making a sentence less toxic involves paraphrasing, which can be classified as a seq2seq or text-to-text task. Additionally, I would like to utilize pre-trained models to avoid the need for training a transformer from scratch.

However, with the abundance of transformers available since 2017, it can be overwhelming to choose just one. Therefore, I am particularly interested in multitasking models that can be provided with a task description to give them a basic understanding of what I want to achieve.

In the past, researchers commonly solved NLP problems using RNN models. However, with the invention of the attention mechanism, NLP solutions were greatly enhanced. Subsequently, the publication "Attention Is All You Need" revolutionized the field, leading to the creation of the transformer model by the Google team. "'

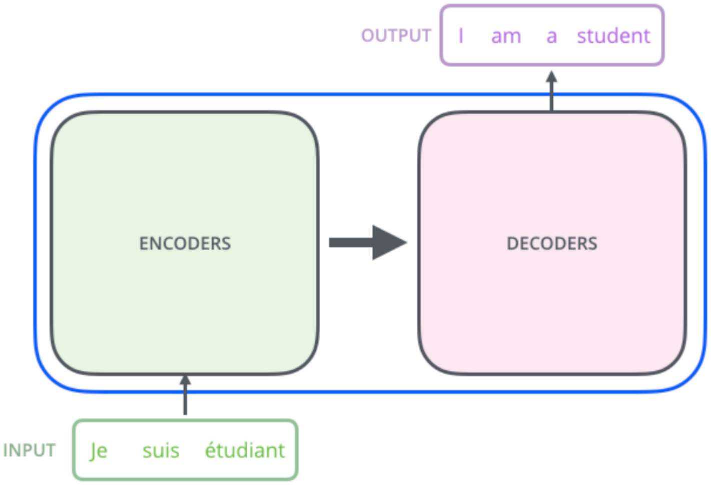
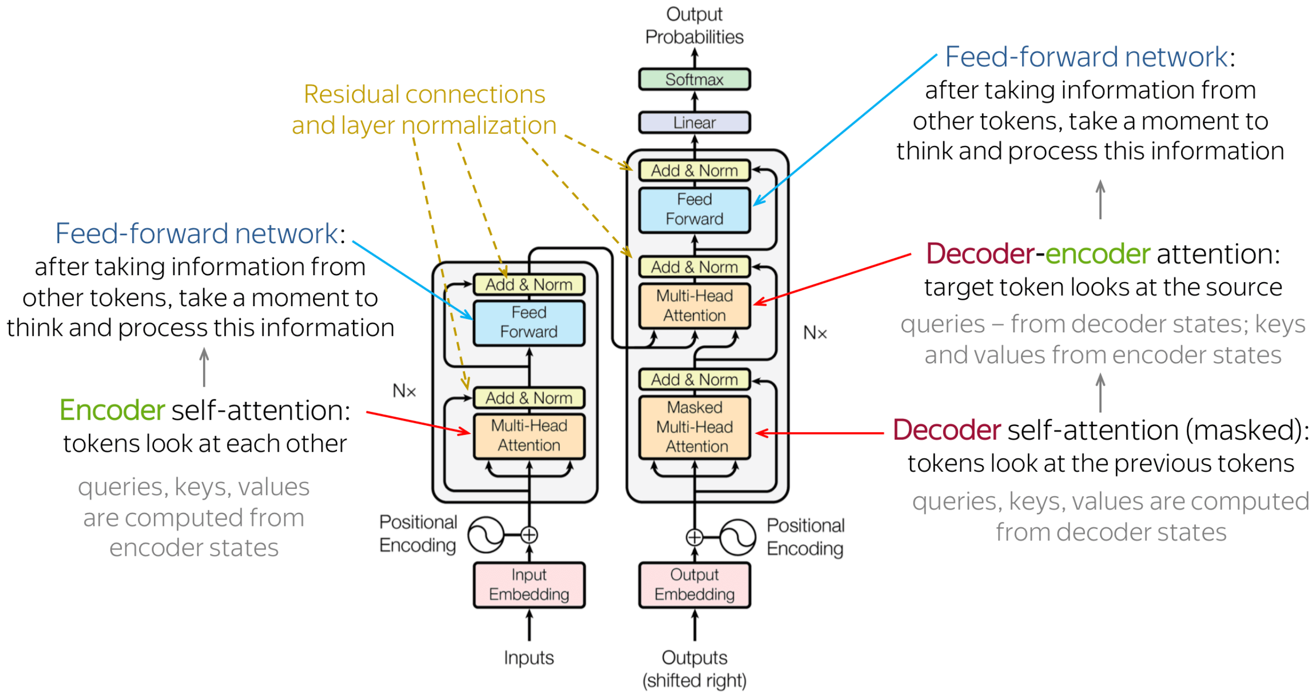


Figure 4: Basic transformer architecture

The basic transformer architecture consists of encoder and decoder.



Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Figure 5: Basic transformer architecture with explanation

2.1 Hypothesis with GPT-2

The OpenAI has developed GPT-2, which is a decoder model commonly used for NLP tasks. As a decoder, GPT-2 has right-shifted attention. To train the model, it is necessary to separate the request or ”reference” text and indicate the point where the model begins its prediction.

To achieve this, I will utilize the keywords ”User:” and ”Assistant:” to split the results. If you want to learn more about GPT-2 and T5 tokenization, you can refer to this informative article: A good article about GPT-2 and T5 tokenization. It highlights that both GPT-2 and T5 tokenizers operate using subwords.

2.2 Hypothesis with T5

Text-to-Text Transfer Transformer -by the architerture is a classical transformer.

In the research paper https://arxiv.org/pdf/1910.10683v4.pdf the authors of T5 model stated one crucial thing about the dataset used for T5 training:

> ”We removed any page that contained any word on the “List of Dirty, Naughty, Obscene or Otherwise Bad Words”.” *https://arxiv.org/pdf/1910.10683v4.pdf*

That might mean that T5 not have tokens for toxic words, this have to be checked. To answer the question lets look at the T5 tokenizer: the tokenizer works pretty similar to GPT one: it does subword tokenization, therefore surely, T5 have tokens for toxic words.

> ”To specify which task the model should perform, we add a task-specific (text) prefix to the original input sequence before feeding it to the model.” https://arxiv.org/pdf/1910.10683v4.pdf

So, the T5 model gets task-prefix at the beginning of the sentence. I decided to use it because the T5 is trained for multitasking, therefore I can create a new task and the model itself will have a minimal understanding of this at the beginning of the fine tuning process. That’s great!

2.2.1 How to create a new task?

Just add a prefix with brief task description.

3 Results

Both models, GPT-2 and T5, demonstrate impressive learning capabilities, achieving excellent performance after just one epoch.

To showcase the results, let’s establish an evaluation technique. Recent research has highlighted the need to replace classical evaluation metrics traditionally used for NLP models with more comprehensive approaches that involve tasks such as classification, similarity evaluation, and others specifically designed for transformer training. In line with this, I have chosen to utilize the well-performing toxification classifier model provided by https://github.com/unitaryai/detoxify. This model will aid in evaluating the performance of GPT-2 and T5.

Surely, I will not perform the detoxification evaluation based on the training dataset, so I decided to use the one from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

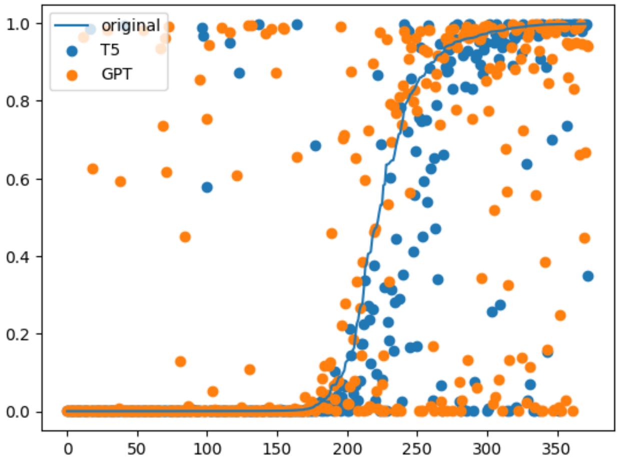


Figure 6: Toxicity level of the T5 and GPT-2 results compared to original text

4 References

Exploring the Limits of Transfer Learning with a UnifiedText-to-Text Transformer https://arxiv.org/pdf/1910.10683v
- paper about T5 from its fathers)

A good article about GPT and T5 tokenizers https://towardsdatascience.com/comparing-transformer-tokenizers-686307856955

References