Practical Machine Learning and Deep Learning

Assignment 1

Andrey Starodumov

B20-DS-01

1 Thoughts behind the solution

After the data analyzing, I decided to implement the solution for just translation of toxic sentences. So The goal is knowing that the sentence is toxic put it into the model and get less toxic sentence.

I wanted to try some simple approaches that are described below.

1.1 Deleting the toxic words based on a built dictionary

As an example of the potential result of this method, the sentence *"We fucked up"* should become *"We up"*. It is obvious that this approach is bad since we can lose the sentence semantic.

1.2 Pre-trained models

After all tries with Bert models and problems with it, I decided to try the Pegasus model.

I discovered that it can already exclude some toxic words (even if it wasn't implemented for that). For example:

"At least one of you Dunham cunts are gonna pay for my fucking boy." $\Longrightarrow$ *"One of you will pay for my boy."*

"It is a shit day." $\Longrightarrow$ *"It is a bad day."*

Of course, the model is needed to be trained more for our task because sometimes it gives sentences which we do not expect. For example:

"Fuck you." $\Longrightarrow$ *"Fuck you."*.

So it has no changes, but it is actually very toxic sentence.

2 Final decision

Since I saw that the Pegasus paraphrasing model applies both changing the words and just loosing them, I decided to try to implement the additional training process for the model.

the loss function is embedded in model, so we will put the token input, token output, and token expected output. Let us see, what it will give us.