1. Data preprocessing.

First, I need to preprocess the dataset to make my system work better. This operation can be divided into the following points:

* remove all special symbols
* remove all user information provided after the last sentence (in this dataset, it comes after the last ones (. ) (! ) (? ), but but not when the last symbol is (.) without space)
* remove all \n
* remove all numbers
* remove all stop words
* remove all duplicate spaces

For tokenization of sentences I used tiktoken (byte pair encoding) tokenizer.

1. Creating a model.

For text classification, I decided to make my model look like an encoder from a Transformer.

* First I created a layer to embedding the view for the input information.
* Created Positional Encoding.
* Created encoder similar to the original article.
* Replaced block of fully connected layers with block of convolutional layers.
* Created class for each EncoderLayer and Classifier (Tr) which combines all the functions of the model and makes predictions.

Schematic representation of model

Зображення, що містить текст, знімок екрана, монітор, програмне забезпечення

Автоматично згенерований опис

1. Model training and testing.

In order to quickly execute the model, I created a small-sized model.

With an additional dataset, a higher accuracy can be achieved.



Зображення, що містить знімок екрана, Графік, ряд, схема

Автоматично згенерований опис

Pink – Validation, Blue – Train