RoBERTa - All - READMe

-----------------------

Description: This folder contains various Google Colab notebook dedicated to the XLNet model

-----------------------

Content:

- RoBERTa - Training.ipynb: Train the RoBERTa model

- RoBERTa Evaluation.ipynb: Evaluate the RoBERTa model on the validation set and generate the .csv files for the ensemble

-----------------------

Prerequisites:

- Logged in with a Google account, with access to Google Colab, Google Colab's CPUs and GPUs (subscription needed to have prioritised access GPU's for a significant longer time) (for these notebooks, it is recommended to have access to GPUs)

- Granting permission to run each of our Google Colab notebooks and python files.

- Mount drive at /content/drive/ [This is done when running the notebook, see below]

- The folders/files regarding Preprocessed training and validation datasets are at the described location (see READMe for folder "data").

- Install transformers via:

command: pip install transformers

or alternativly via command: pip3 install transformers

- Install transformers via:

command: pip install datasets

or alternativly via command: pip3 install datasets

-----------------------

Which notebook has to run first:

RoBERTa - Training.ipynb

-----------------------

-----------------------

RoBERTa - Training.ipynb

-----------------------

How to run: RoBERTa - Training.ipynb

Before running, the user can specify

- which preprocessed datasets they want to work with

In the second cell section, there is a variable

PREPROCESSING\_CHOICE defined by PREPROCESSING\_OPTIONS[INDEX], where INDEX is an integer [0;8]. The user can choose INDEX from that range, which corresponds to the preprocessed dataset.

After selecting INDEX:

It is highly recommended to have access to GPUs. In tab "Runtime", click on "change runtime type", choose GPU as hardware accelerator and save this setting.

Run all cells, top to bottom.

-----------------------

Output:

In the directory /content/drive/CIL22022, the trained models are stored as file; the stats are found as print statement.

-----------------------

-----------------------

XLNet Evaluation on Validationset for RoBERTa Evaluation.ipynb

-----------------------

How to run: XLNet Evaluation on Validationset for Ensemble.ipyn

Before running, the user can specify

- which preprocessed datasets they want to work with

In the second cell section, there is a variable

PREPROCESSING\_CHOICE defined by PREPROCESSING\_OPTIONS[INDEX], where INDEX is an integer [0;8]. The user can choose INDEX from that range, which corresponds to the preprocessed dataset.

In the 4th last cell, there is a variable called "model" that is defined by model.from\_pretrained(PATH\_TO\_MODEL), where user has to indicate PATH\_TO\_MODEL (similar as how it's done in the code)

It is highly recommended to have access to GPUs. In tab "Runtime", click on "change runtime type", choose GPU as hardware accelerator and save this setting.

Run all cells, top to bottom.

-----------------------

Output:

In the directory /content/drive/CIL22022/data/test data/, the intermediary prediction is stored as csv file.