CIL2022 - READMe

This READMe describes how to generate our Kaggle submission file.

To train our models, we used a Pro subscription for Google Colab. We had shared access to the shared google drive folder “CIL 2022”. This had the advantage: we could store model states and embedding vectors in the (shared) cloud and access them from Google Colab.

Each addressed folder contains a READMe file to describe the workflow.

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

Content:

- folder "data": contains data used for preprocessing and training/validation

- folder "Models": contains notebooks of models

- folder "Preprocessing": contains code for preprocessing

- folder "Word embeddings": contains downloaded/stored embeddings

- folder "Saved Model States": contains the best models as saved on the disk

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

Prerequisites (P):

- Logged in with a Google account, with access to Google Colab, Google Colab's CPUs, and GPUs (subscription needed to have prioritized access GPUs for a significantly longer time)

- Mount drive at /content/drive/ You should create a shortcut called “CIL 2022” to this folder in your root directory of google drive. Because there was no other way known to us, we hardcoded all paths to "/content/drive/MyDrive/CIL 2022/". Adding a shortcut to the top directory of your google drive enables all jupyter notebooks to access the correct data.

- Granting permission to run each of our Google Colab notebooks and python files. Once you run a jupyter notebook, it will ask to give permission to access your google drive storage.

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

If you just want to validate our results, head to Models and look for the corresponding readme. As the data is already stored in the Google Drive shortcut you created, the jupyter notebooks can load the models and evaluate the validation dataset. If you want to redo our whole process, read on:

Rough Overview (for the reproduction of results):

- 1. Download tweet dataset and embedding files

- 2. Generate (preprocessed) training and validation dataset

- 3. Train and evaluate models

- 3.1. Train and evaluate baseline models

- 3.2. Train and evaluate advanced models

- 4. Used trained advanced models to compute, via ensemble, the final submission, the Kaggle submission file

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

- 1. Download tweet dataset and embedding files:

-- As for tweet dataset, described in READMe of folder "data"

-- As for embedding files, described in READMe of folder "Word embeddings"

-2. Generate (preprocessed) training and validation dataset:

-- Described in READMe of folder "Preprocessing"

- 3. Train and evaluate models:

-- Described in READMe of folder "Models"

4. Kaggle submission file generation:

-- Described in READMe of folder "Ensemble of top 3 models"; this folder is located in folder “Models”