Preprocessing - READMe

This READMe describes how to use the files contained in this folder.

The preprocessing files are the only files that were coded as plain python files and that were executed locally on a home PC. As removing duplicates may need a lot of RAM (or at least in our previous versions it did), we used a gaming PC with 32GB (local) RAM. If you want to execute these files, you have to download them and run them locally (except the minimal\_preprocessing jupyter notebook).

We only used the full dataset files (train\_neg\_full.txt and train\_pos\_full.txt), as the small dataset was a subset of the bigger one anyway.

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

Content: This folder contains

- a python file "preprocess.py" preprocesses every tweet given the options hardcoded in the file

- a python file "preprocessing\_cleaned.py" same file as above, but we cleaned the source code so that it’s nicer to look at.

- a python file "remove\_dups\_and\_split.py" removes duplicates from the (preprocessed) train\_neg\_full.txt and train\_pos\_full.txt and then does the 90/10 training/validation split.

- a google colab file “minimal\_preprocessing.ipynb”

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

Prerequisites (P):

Tools:

Before running the code, please install the following (if not already done) via pip:

pip install num2words

pip install symspellpy

pip install re

pip install nltk

or alternativly using pip3:

pip3 install num2words

pip3 install symspellpy

pip3 install re

pip3 install nltk

Files:

This folder is located in ./

The twitter datasets files have to be in to the folder ./raw/

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

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

"preprocess.py":

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

Prerequisites:

- Prerequisites described above (P)

- User action:

There is a variable "options" that stores a list of lists. Each list encodes if stemming, lemmatizatization, stopword removal and spell correction is to be performed on the tweets to be read in. The user can manually select (via commenting out/not commenting out lists) what the preprocessing step should consists of.

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

How to run "preprocess.py":

python3 preprocess.py

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

Output:

After running this file, and depending on the user's choices of elements for preprocessing step (See Prerequisites, User action), there are folder appropriately named in ./ .

These subfolders contain the user-specified-preprocessed (see Prerequisites, User action) tweets, still seperated by positive and negative labels.

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

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

"remove\_dups\_and\_split.py":

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

Prerequisites:

- Prerequisites described above (P)

- python file "preprocess.py" must have been run beforehand

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

How to run "preprocess.py":

python3 remove\_dups\_and\_split.py

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

Output:

For each of the subfolders, created by running python file "preprocess.py" (except one case \*), the positive and negative tweets are combined in a set, duplicates are removed it, from which training and validation dataset is created.

Exception case (\*):

By raw, no preprocessing is done on the files.

“minimal\_preprocessing.ipynb”:

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

How to run "minimal\_preprocessing.ipynb":

Run all the cells inside the file

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

Output:

Files corresponding to the minimal preprocessing