Task 12 TAG_it

Paolo Renzi
Sapienza Università di Roma

1 The description of the dataset

The dataset comprises two tasks aimed at detecting hate speech. Task 1 focuses on Hate Speech Detection, where the goal is to classify whether a message contains hate speech or not. The training set consists of 5,600 tweets from PolicyCorpusXL, while the test set includes 1,400 tweets from PolicyCorpusXL and 3,000 tweets from ReligiousHate. In Task 2, termed Contextual Hate Speech Detection, both the content of tweets and their metadata are considered for classification. This task includes two sub-tasks: Political Hate Speech Detection, which utilizes data from both development and test sets from PolicyCorpusXL, and Religious Hate Speech Detection, where only the test data from ReligiousHate is provided, adapted from an original cross-domain task.

2 Methodology to reframe the dataset

I exploited the fact that each row of data ends with the name of the dataset and used that to separate each line. In the test set the data is taken from two datasets so the name changes accordingly but i still used the same approch by first dividing for political_test and then I iterated in each line to divide it again by Religious_test and put it all in a list. I then iterated over the resulting list and created a dictionary where the keys are the id of the tweet and the values are themselfs a dictionary with as key the type of data (Ex. "text", "label", "created at") and the values is the data. I differentiate between test and training and textual data and conte data from the filepath string.

3 Prompts

I created 4 prompts for each task, and i added in task 2 the following metadata, created at and user_created_at beacause the age of the profile at the time of writing might have a correlation with the task. For the same reason I added status, friend e followers count.

4 How to run

To run the parser you should set the folder contating task_12.py as CWD and then run

```
python task_27.py
```

You will not need to download the dataset manually, beacause the script will do it for you and will also unzip it.