Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
Autumn Call, A.Y. 2023/2024

Last update: September 10, 2024

1 Project dates

Start date: September 9, 2024 at 23:59 (CET)
Due date: September 24, 2024 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Different people express different sentiments using different words. This phenomenon is particularly evident on online social platforms. Further, as psychological studies reveal, sentiment and emotions vary over a broad spectrum and are characterized by taxonomies.

In this project, you are required to predict the sentiment of a given content on Twitter. In a simplified manner, the sentiment provided in this task is either positive or negative.

2.1 Dataset

The dataset consists of a collection of tweets in tabular format. Several attributes characterize each record. The following is a short description of each of them.

- *ids*: a numerical identifier of the tweet;
- *date*: the publication date;
- *flag*: the query used to collect the tweet;
- *user*: the username of the original poster;
- *text*: the text of the tweet.

The sentiment of the tweet is reported on the feature named ***sentiment*** and is equal to **1** for the Positive class and **0** for the Negative one.

The dataset is located at this URL.

Within the archive, you will find the following files:

- **development.csv** (development set): a comma-separated values file containing the records from the development set. This portion does have the `Sentiment` column, which you should use to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the records from the evaluation set. This portion does not have the `Sentiment` column.

2.2 Task

You are required to build a classification pipeline to predict the sentiment of tweet in the Evaluation set.

2.3 Evaluation metric

Your submissions will be evaluated in terms of the F1 macro score. Here you can find the function used to evaluate your submissions.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
0,0
1,0
2,1
3,1
4,0
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set. It corresponds to the column `Id` in the evaluation CSV file.
- the `Predicted` binary outcome. It must be in numerical form, e.g., 0 or 1, as the development set provides.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `lorenzo.vaiani@polito.it`. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_autumn_2024**.

Info: A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing `.zip` extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.

5 Fill in the LLM usage form

As discussed in the exam rules, adoption of Large Language Models (e.g. ChatGPT) is allowed for the production of the report (**not** for the implementation of the solution). Each team **must** provide information about whether they used, and to which extent they did, LLM-based tools.

To do so, please fill in this form by the due date of this project. Failure to do so will result in a void project.

Warning: This is an additional requirement that was not required in past years. Make sure you remember to fill in the form by the due date, or your project will not be considered valid!