Data Science and Machine Learning Lab
Politecnico di Torino

Project Assignment
Winter Call, A.Y. 2025/2026

Last update: January 10, 2026

1 Project dates

Start date: January 09, 2026 at 23:59 (CET)
Due date: February 1, 2026 at 11:59 (CET)

Due date is a **strict deadline**.

2 Problem description

The proliferation of digital news platforms has fundamentally transformed the way information is produced, distributed, and consumed, leading to an unprecedented volume of online news content. As a result, the automatic organization and categorization of news articles have become critical challenges for both media organizations and information retrieval systems. Accurately identifying the thematic category of a news article – such as international affairs, business, or technology – is essential for content recommendation, targeted advertising, and efficient news aggregation. The central objective of this project is to develop and evaluate a machine learning model capable of automatically classifying news articles into predefined categories based on their textual content and associated metadata.

2.1 Dataset

Warning: For this project, you are not allowed to use external datasets other than the one provided.

The dataset contains approximately 100,000 instances describing online news articles collected from multiple international news sources. Each record corresponds to a single news article and includes both metadata and textual content. Articles originate from various publishers and span different publication dates. A categorical label is associated with each instance.

Several attributes characterize each record. The following is a brief description of each of them:

- id: Unique identifier of the article.
- source: News outlet or publisher of the article.
- title: Title of the news article.
- article: Full textual content of the article.
- page_rank: Page rank associated with the article source.
- timestamp: Date and time of publication.

- label: Target label associated with the article (used for the classification tasks).

The label attribute corresponds to a predefined set of news categories. The mapping between category names and their numerical labels is defined as follows:

- International News: 0
- Business: 1
- Technology: 2
- Entertainment: 3
- Sports: 4
- General News: 5
- Health: 6

The dataset is located at:
https://drive.google.com/file/d/1dWNUwC47MfPs7mxsP0_e2rmUY4_AUDCk/view?usp=sharing

Within the archive, you will find the following elements:

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

2.2 Task

You are required to build a classification pipeline to predict the actual topic of each news in the Evaluation Set.

2.3 Evaluation metric

Your submissions will be evaluated through Macro F1.

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
1,2
2,5
3,4
...
```

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` label for the corresponding record.

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_winter_2026**.

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!