Mathematics for Big Data
Deep Learning

The project assignment involves classifying various texts. You must work in teams of two or three people to complete this task.

1 Datasets

The training set contains 216,974 sentences in Catalan, each assigned to one of 21 possible classes, numbered from 0 to 20.

Here are a few examples from the training set:

Class 0	**Class 16**	**Class 17**	**Class 19**
Licitació de la concessió administrativa de l'ús privatiu del domini públic municipal d'una peça de terreny.	Aprovació definitiva de l'avanç del Pla de millora de la Franja Nord.	Notificació de l'acord de deixar sense efecte l'aprovació definitiva d'un projecte de reparcel · lació.	Notificació als titulars d'uns vehicles abandonats.

Each sentence in the training set is associated with a unique numerical identifier. The trainset can be found in the `train.csv` file, which contains the following columns:

- `id`: A unique identifier for each sentence.
- `text`: The text of the sentence.
- `class`: The class label (ranging from 0 to 20).
- `year`: The year in which the text was created (from 2013 to 2020).

In addition to the training set, there is a separate test set consisting of 56,997 sentences from the years 2021 and 2022. This test set will be made publicly available once all teams have completed training their models. The test set will contain only the `id` and `text` columns.

2 Models

To classify the texts, you must train one or more models using Python and libraries such as TensorFlow, Pytorch or other text-processing frameworks in a Jupyter Notebook.

You must consider the following restrictions:

- The models you train must be based on neural networks.
- You are allowed to apply preprocessing techniques to the text data (e.g., tokenization, stopword removal, stemming, etc.) to improve model performance.

- You may also create ensemble models, but all models in the ensemble must be based on neural networks.

3 Deliverables

The following list contains all the items you must deliver for this project.

- A presentation with 8 to 14 slides in PDF format, containing:
  - Name, surnames and NIU of every team member.
  - If you do a descriptive analysis of the dataset, the results and conclusions you obtained.
  - If you apply preprocessing techniques to the texts, the detailed description of such process.
  - Description of the experiments you have carried out, together with the results and conclusions obtained.
  - Description of the final model you will use to classify the texts in the testset.
  - Description of the hyperparameters you used during the training of the final model.
  - Estimation of the accuracy (in percent) you expect to get when executing the final model on the testset with $56,997$ texts.
  - Justification for your accuracy estimation.
  - Comments on what your needs would be to improve the accuracy. For example, would you like a larger data set? Or do you need a more powerful CPU? More RAM? More run time? Etc.
  - Any other comments on your development that you deem appropriate.
- Jupyter Notebook showing the construction and training of the final model, from the dataset reading process until the writing of the predictions file, detailing also the training of the model and all the processing you do on the texts.
- CSV file with two columns, `id` and `class`, with the predictions of your final model for the texts in the testset.

The deliverables must be uploaded to the Virtual Campus as follows:

- Before June 12th at 11:59pm you must submit your presentation and Jupyter notebook.
- On June 13th at 8:00am the testset will be made public.
- You must submit the CSV file with your predictions for the texts in the testset the same day (June 13th) before 11:59pm.

4 Grading

The grading of the project will be done on 10 points, and will consist of three items:

- N_1 (5 points) is the grading of the written presentation. The grade will be based on the clarity of the explanations and the justification of the decisions taken during the development of the project.
- N_2 (2 points) measures the the estimation of the accuracy. In particular, if p_j denotes the

real accuracy obtained by one of the teams in the testset and $\hat{p}_j$ denotes the estimation of the accuracy written in the presentation, then,

$$N_2 = \begin{cases} 2 & \text{if} \quad |p_j - \hat{p}_j| \leq 2 \\ 1 & \text{if} \quad 2 < |p_j - \hat{p}_j| \leq 4 \\ 0 & \text{if} \quad 4 < |p_j - \hat{p}_j| \end{cases}$$

- N_3 (3 points) takes into consideration the actual accuracies obtained by all teams. In particular, assume we have G teams and we sort the accuracies from highest to lowest as follows:

$$p_1 \geq p_2 \geq \cdots \geq p_i \geq 25 > p_{i+1} \geq \cdots \geq p_G$$

Then, the grade for the team with accuracy p_j is

$$N_3 = \begin{cases} \frac{2}{p_1 - p_i}(p_j - p_i) + 1 & \text{if } p_j \in \{p_1, \ldots, p_i\} \\ 0 & \text{if } p_j \in \{p_{i+1}, \ldots, p_G\} \end{cases}$$

The final grade will be then computed as

$$N = N_1 + N_2 + N_3$$