Practice IV

Clickbait detection
Second stage

Objective

- Train a LLM for detecting whether a text is a clickbait or not

Specification

In teams of 3-4 members do the following:

1) Load the corpus *TA1C_dataset_detection_train.csv*

   - *Teaser Text* column will be used as feature

   - *Tag Value* colum will be used as target (class)

2) Split the corpus in *train and dev* sets using 75% for training and 25% for development

3) Set metrics for evaluation

4) Tokenize the input

5) Train a LLM using the train set and the dev set as evaluation dataset

6) Make the necesary adjustemnts to the model to improve perfomance over the development *set*

7) Use the best adjusted model to predict instances of *test* set available in the *TA1C_dataset_detection_dev.csv* file

Corpus split

- Use *train_test_split* function of scikit-learn
- Instance of corpus should be shuffled (*shuffled* = *True*)
- Set the the fixed random seed equals to 0 (*random_state=0*)
- Activate the stratify function (*stratify=y*)

Evaluation metrics

- A classification report (*classification_report*) should be applied to the predictions on the dev set for each experiment
- The confusion matrix should be generated to verify how the instances are classified
- The main metric is f1_macro, this one will be used for selecting the best model

Tokenization and LLM training

- A model based on Bert adjusted to Spanish (*dccuchile/bert-base-spanish-wwm-cased*) is recommended as a baseline
- The same model should be used for tokenization an training
- The evaluation of the model is performed using the *dev* set (25% of the original *train* set)

LLM adjustment

- The parameters of the LLM shoud be adjusted to improve the performance
- Some relevant parameters are:
  - eval_steps
  - num_train_epochs
  - learning_rate
- You can also try different LLM models

Predictions on the test set

- The best model with the selected parameters must be used to predict the *Tag Value* of instances in the *TA1C_dataset_detection_dev.csv* file

Predictions on the test set

- The output CSV file must have the following features:
  - Name: *detection.csv*
  - Columns: *Tweet ID* and *Tag Value*
  - Separator character: Comma “,”

Evidence

- Source code
- A report in PDF format describing the following:
  - Task to be solved
  - Selected machine learning methods
  - Adjusted hyperparameters
  - Classification report of each experiment

Evidence

A table describing the experiments performed showing the best configuration of each LLM on the *dev* set

LLM	LLM hyperparameters	Average f-score macro
dccuchile/bert-base-spanish-wwm-cased	eval_strategy = ="steps",eval_steps=100	0.85
dccuchile/bert-base-spanish-wwm-uncased	default	0.88
...	...	...
FacebookAI/xlm-roberta-large	num_train_epochs=3, learning_rate=0.001	0.9

Evidence

- The file *detection.csv*