

Project – ISIS 4221  
Stage II: Challenge Selection

**HW04 report:** 17-05-2021

**Partial results presentation date:** {18-19}-05-2021

**Final project presentation:** {01-05}-06-2021

At this stage you are going to select the challenges that you are going to take on in your project. There are a total of 5 challenges of which 3 are mandatory, and 2 are optional. The purpose of the challenges is that you face an NLP project in a similar way to an industrial setting. The raw datasets are not ready for the task, which means that you must do the appropriate processing and labelling steps.

Challenges (blue->mandatory):

1. HW04: Automatic identification of “trending/viral” topics around COVID in time in the World using deep learning architectures. Type: Different approaches.
2. Automatic identification of COVID vaccines/vaccination/mental-health documents. Type: Sequence classification.
3. Automatic identification of COVID school-reopening/household-violence documents. Type: Sequence classification.
4. Context Q&A system about COVID in a general setting. Type: Transformer.
5. Fake News identification in COVID public discussion. Type: Different approaches.

You must decide if you take more than 3 challenges based on the benefits specified below.

Note: A perfect solution or a high degree of precision is not expected in some of the challenges. What is expected is that you explore an approach, be able to justify and validate it with a solid data set.

For each challenge you must follow the “data-centric” methodology:

1. Search for "recent" approaches in the literature that solve the problem or a similar problem. You must document your findings.
2. Based on the findings and the available knowledge, propose a multilanguage dataset for the construction and evaluation of the models for the challenge. You should evaluate your dataset in light of: a) completeness, b) labeling methodology, and c) verification. You must document in detail the dataset construction process and the tools used. Note: all datasets must be submitted to validate results.
3. Definition and training of NLP models. You must justify the selection of each model, its hyperparameters, and architecture if applicable.
4. Validation and comparison of results at the level:
  - a. Models.
  - b. Languages.

- c. Dataset.
  - d. Features.
2. You should also compare your work with works previously found in the literature.

Important note: The teacher/monitor must be able to replicate the results by executing the python scripts that each group attaches in the submission.

#### Exchange of information between groups

You carry out collaborative processes with other groups at the data level exclusively. You must describe in detail how you carried out the integration processes. **No code or models can be exchanged.**

#### Final project submission dates

With the purpose that you advance in the development of the final project, HW04 will be the realization of challenge number 1.

**HW04:** {17}-05-2021. You must send the report, datasets, and code, like any other previous HW. (10%)

**Partial results presentation date:** {18-19}-05-2021 Prepare a presentation with your progress on each of the challenges at the data, model, and results level. On the presentation day, you must choose whether to apply for the bonuses or not. (5%)

**Final project presentation:** {01-05}-06-2021 You must deliver the project before June 1, 2021 (code + datasets + report). From June 2 to 5, group presentations will be made. Each challenge will have an equivalent percentage in the grade. (15%- or more based on the bonus selection).

#### Bonus

1. Completing an additional challenge replace (or add) to the worst grade in the course assignments (you will not be able to exceed the maximum grade).
2. Completing all the challenges will exempt you from taking the final exam. The grade you get in the additional challenges will be your final exam grade.