Learning from Data – Final Project
International Climate Change News

General remarks

Differently than in the other assignments, for this final project we are not giving you a specific task nor telling you what methods you should use, but we're only giving you some data that we deem interesting and that lends itself to a variety of problems/tasks. In other words: the *tasks* you choose to solve, and the *way* you approach the problems is up to you, and it is an important part of the assignment itself. Grading will be determined by your models, your report, and the final presentation you will give for everyone in class.

Please note that you will work in groups. It is important that work is spread as equally as possible, and that both the report and the presentation are prepared and given by all members. The choices you make for the project should also be adequate for the size of your group. The report will need to specify as much as possible the division of labour.

Deadline for submission on Nestor:: October 30th, 2020, end of day.

What you have to hand in by the deadline:

- script with your system. Your script should take a `.json` training file and a corresponding `.json` test file as arguments, as you've done for previous assignments.
- report with system description. See Section 3 for details. Please, make sure to hand in a **pdf** file following the usual report template.

After submission, you will be provided with newer test data which we have withheld, and you can report results on your presentation, together with a description of what you have done and your system.

Presentations:

- November 3rd, 13-15 (Tuesday groups)
- November 4th, 13-15 (Wednesday groups)

1 Data: International Climate Change News

Size and collection You will get data that represents over 35,000 articles about climate change that have been published in the time period around the first 25 *Conference of the Parties*[1] meetings (COP1 until 24).

The articles have been collected from *Lexis Nexis*[2] based on a set of keywords and newspapers. After collection, the articles have been filtered to include at least 3 unique co-occurring keywords.

The articles are from 9 newspapers from 4 different countries and 2 opposing political orientations. For most countries, both political orientations are available. See Table 1 for more details. The orientations are based on analyses by *AllSides*[3] and *Media Bias Fact Check*[4]. Note that the newspapers have a varied availability over time; not every newspaper is represented in each file.

Country	**Newspaper**	**Political orientation**
Australia	The Australian	Right-Center
	Sydney Morning Herald	Left-Center
	The Age	Left-Center
India	The Times of India	Right-Center
	The Hindu	Left-Center
South Africa	The Times	Right-Center
	Mail and Guardian	Left-Center
United States	The Washington Post	Left-Center
	New York Times	Left-Center

Table 1: Countries and political orientations for the newspapers represented in the dataset.

Format The data is provided as a set of `.json` files, one for each COP meeting. The files can be downloaded (as a `.zip`) from `https://teaching.stijneikelboom.nl/lfd2021`.

Each file has the following structure:

- `cop_edition`: Number of the edition of the COP meeting.
- `collection_start`: Start date of collection, always one week before the start of the COP meeting.
- `collection_end`: End date of collection, always one week after the end of the COP meeting.

[1] `https://unfccc.int/process/bodies/supreme-bodies/conference-of-the-parties-cop`
[2] `https://www.nexisuni.com`
[3] `https://www.allsides.com/media-bias/media-bias-ratings`
[4] `https://mediabiasfactcheck.com`

- `articles`: Array of article objects with the following structure:
  - `path`: Path to the original PDF the article was extracted from.
  - `raw_text`: Raw text that was extracted from the PDF.
  - `newspaper`: Cleaned name of the newspaper the article was published in.
  - `headline`: Cleaned headline of the article.
  - `body`: Full and cleaned body text of the article.
  - `classification`: Object of classifications by Lexis Nexis. Note that it is not explicitly disclosed by Lexis Nexis how these classifications came about. They are structured as follows:
    - `subject`: Names of subjects discussed in the article
    - `organization`: Names of organizations discussed in the article.
    - `industry`: Names of industries discussed in the article.
    - `geographic`: Names of geographic entities discussed in the article.

      Each classification type is represented as an object, containing:
      - `name`: Name of the class.
      - `percentage`: Percentage of match with the article.

Note that not all instances are necessarily complete in all fields.

Test data The test data that is withheld consists of newer data, from COP25. The structure of this data is identical to that of the training data. The test data will be provided to you after submission of your report. Running your model on this data is a good way to validate how future-proof your system is.

```
{
  "cop_edition": "24",
  "collection_start": "26/11/2018",
  "collection_end": "22/12/2018",
  "articles": [
    {
      "path": "data/raw/nexis_20200821_1130/Files(100)/Glob(...) catastrophe(2).pdf",
      "raw_text": "Global (...) climate catastrophe\n     The Australian\n (...)",
      "newspaper": "The Australian",
      "date": "November 29, 2018",
      "headline": "Global inaction puts world on track for climate catastrophe",
      "body": "WASHINGTON: Promises from nations to reduce greenhouse gas (...)",
      "classification": {
        "subject": [
          {
            "name": "GREENHOUSE GASES",
            "percentage": "99"
          }
        ],
        "organization": null,
        "industry": [
          {
            "name": "GLOBAL WARMING",
            "percentage": "91"
          }
        ],
        "geographic": [
          {
            "name": "AUSTRALIA",
            "percentage": "93"
          }
        ]
      }
    }
  ]
}
```


Figure 1: Example of a data file. Sequences marked by `(...)` were shortened.

2 Tasks and Evaluation

Labels We challenge you to come up with one or more sets of labels that divide the data. These can be inferred from the data that has been provided in a more or less direct way.

You can think of, but are not limited to:

- Country of publication
- Political orientation of newspaper
- Presidential administration of publication (for US data)
- Edition of *Conference of the Parties*

Tasks Of course, the labels that you generate for the data influence the task that you will work on. You can also decide to work on multiple tasks.

Possible tasks are, but are certainly not limited to:

- Can you predict in which **newspaper** each article was published? What can you learn about the differences between the newspapers?
- Can you predict in which **country** each article was published? What can you learn about the differences between the countries?
- Can you predict under which **presidential administration** each article was published? What can you learn about the differences between the administrations?
- Can you predict around which **COP edition** each article was published? What can you learn about the differences between the editions? (For compatibility with the newer test data, you can (1) generate labels for ranges of COP editions or (2) determine what old edition the new data is most like.)

For all tasks, it is interesting to optimize the model, but also to look at what the most informative features can tell you about the social and medial impact of the articles.

Based on the task, you can also decide to use only part of the available data. It is all up to you. However, keep in mind that the task should be adequate for the course and the size of the group you are working in.

Evaluation Since there is no predetermined task, evaluation will be done independently by each group, and it won't be possible (apart from very special cases) to run any comparisons across groups.

Nevertheless, you are asked to explain very well which labels you are using for your tasks, and you will have to make sure that your system is evaluated according to the following measures in each task that you run: precision, recall, and F-score per class, and overall macro F-score and accuracy.

Please note that **these measures must be implemented in your script** and printed on output. If you also print a confusion matrix it's a good idea, and you can discuss what you observe accordingly. Some error analysis is also welcome.

3 What you have to do

1. You have to **design one or more classification tasks** using the data you're given. See above for suggestions and potential labels to predict.

2. You have to **develop a system** to address your tasks. You have complete freedom here on what methods and architectures you want to use, which features, which algorithms, the model development is entirely up to you. Also the evaluation settings are your choice (cross-validation, separate dev set) You can obviously use all the support from scikit-learn, Keras, and NLTK for this, and any other library you find useful. You can use existing pre-trained models, if you wish, and any embeddings representations, including training your own, if you wish. You are encouraged to experiment with different features, and please, report on what you observe. Overall, anything you use will have to be mentioned in your report.

3. You are also asked to produce a **report**. The report should contain the explanation of how you tackled this problem, a description of the features you used, any feature selection method you applied, the algorithm(s) you chose to learn your model, including parameter tuning and setting, any additional data/resources you incorporated, and how well you do when developing (either via a separate dev set or via cross-validation) in terms of accuracy, precision, recall, f-score. You should also justify your choices explaining why you selected a certain approach, certain features, the learning algorithm, and so on. One important aspect of the report will be also specifying who did what in the team, how the labour was split, and whether there was any imbalance due to any reason you would like to mention.

4. Finally, you are asked to produce a **presentation** in which you will explain to the others what you have done, and why. You will have 15 minutes for this, including questions (think in terms of 10+5). Please, bear in mind that the presentation will contribute to the final grading as well, so all team members will have to contribute.

Before the presentation you will be able to run your results on the test set (we will make it available after the submission deadline), so that you can refer to those as well.