Project: Towards Sustainability or Greenwashing – Analyzing Companies Sustainability Reporting

# General instructions

0. Pick your team members (3 people per team). Decide beforehand who will contribute what. Declare it clearly in your project git page, which files each person contributed. Pick carefully your team members so that each person has complementary skills.

1. Find first which [UN sustainability goals](https://sdgs.un.org/fr/goals) this project addresses.

2. Do some

a) [literature search](https://scholar.google.com) to see what has been done in this area in the last 5 years (up to 2 pages).

b) Are there any commercial solutions out there? Is there room for improvement in your solution/prototype?

3. Decide how you will model or solve the problem.

4. Collect data, clean data, annotate data.

5. Quantify your solution (eg accuracy, precision, recall, etc). Here, you also need to provide some baselines and explain in what aspects your technique is better. It may be faster, simpler, more accurate, more “visual”, more intuitive, etc, or maybe it is the first method to address and solve this problem.

6. **Deliverables:**

* A github page with a **readme.md** that is your report and your results (*not* a doc, pdf etc). The git should contain the **related code** and experiments and also a youtube link to a **video** showcasing your solution (no more than 15mins).
* *[optional] Write a simplified 1-2 paragraph article about the problem and your solution in LinkedIn. Put the link to the LinkedIn article in your git page.*

# Problem description

The new European *Corporate Sustainability Reporting Directive* (CSRD) will make mandatory for companies to report on their environmental and social impacts. Already today, most large companies publish an annual sustainability report to detail their environmental and social performance. But are companies implementing sustainable practices or are these reports simply greenwashing? In this project, you will implement text classification algorithms to detect environmental claims. You will then use your classifier to develop an “environmental claims” score for companies’ sustainability reports. Finally, you will compare this score to companies’ environmental metrics e.g., greenhouse gases emissions, air pollutants, and water intensity.

# Tasks

1. Download the “environmental claims” dataset (training, test, and validation set), available [here](https://huggingface.co/datasets/climatebert/environmental_claims/tree/main), and explore your dataset. Data analyses EDA
2. Implement text classification algorithms to detect environmental claims
   1. Compare the performance of various models
   2. Try to improve the accuracy by using various texts embeddings and any other techniques you can think of.

**Note: Your grade will largely depend on the results of this step.** The idea is not to only implement the best model possible. Rather, it is to compare the performance of various models and how the accuracy improves (or not) when you add more complexity.

1. Using your “best” classifier, compute the “environmental claims” score of a few company sustainability report (for instance, you can download 5-10 reports of the companies of your choice; alternatively, you could extract texts from the companies’ websites using web scraping). The “environmental claims” score could be, e.g., the share of sentences that are environmental claims, but feel free to design other metrics.
2. For the chosen companies, compare the “environmental claims” score to their actual environmental performance, using the provided Trucost dataset “Companies environmental performance”. You could for instance look at the evolution of the company’s greenhouse gas emissions or compare your company to others from the same sector   
   *Note: you should not store Trucost data on GitHub. Only extract the information you need for your analysis, and delete the dataset afterwards.*
3. [optional] Create a simple interface (eg using [streamlit](https://streamlit.io)) that, given a company and/or company sustainability report, returns the environmental claims score and some environmental performance metrics

Try to be creative, think out-of-the-box and remember to have fun!