GRINS FOUNDATION

prometeia

2nd Workshop on Sustainable Finance SPOKE 4
2-3 December 2024

Firm-level data extraction from corporates' Non-Financial Reports

Antonella Fabrizio, **Francesco Giovanardi**, Michele Penza, Lorenzo Prosperi, Lea Zicchino, Sedric Zucchiatti

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Partenariato Esteso Finanziato dal PNRR - Missione 4, Componente 2, Investimento 1.3

Introduction

Research project: **improve on ESG data availability** both within and across firms.

Why:

- **Growing Relevance of ESG data**: investors, regulators, and stakeholders increasingly demand accurate and **comparable** ESG data.
- We have **detailed ESG data just for a limited number of companies** (typically large and/or public).
- SMEs and/or private companies often **disclose ESG information**, through sustainability or Non-Financial Reporting (NFR).
- Is it possible to systematically extract such information?

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
prometeia
GRINS FOUNDATION

Introduction

Research project: **improve on ESG data availability** within and across companies.

How:

- Extracting **precise indicators from NFRs** with Large Language Models (LLMs).
- In NFRs firms disclose targets for sustainability and efforts made to pursue them.
- **Unstructured Data**: data often comes in lengthy (100+ pages) and unstructured non-financial reports.
- We need a pipeline able to **extract consistent and comparable Key Performance Indicators** (KPIs) from NFR of several firms and at different points in time.
- Approaches based on manual extraction are prone to errors and inconsistencies and **will become unfeasible** as soon as number of firms publishing NFRs rises.

The EU Corporate Sustainability Reporting Directive

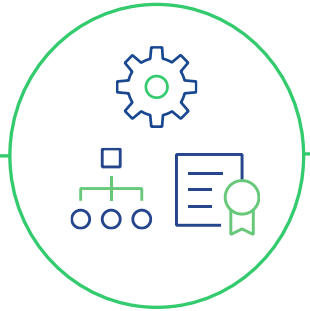
In Europe, some firms are **required to disclose ESG information** through NFRs since 2018 thanks to Non-Financial Reporting Directive (NFRD):

- up to last year only a limited number of firms, such as listed/financial/public interest firms with (a) 500+ employees, (b) €20M+ balance sheet total, or (c) €40M+ net turnover → **approx. 300 Italian and 11k European firms**.
- **Corporate Sustainability Reporting Directive** (CSRD): from 2024 number of firms subject to mandatory reporting will rise and they must disclose comprehensive ESG metrics.
- By 2026, the scope of the CSRD expands to all **large firms, listed SMEs**, and non-EU firm.
- The amount of reporting will increase significantly up to 2029 → **approx. 6k Italian and 50k EU firms.**
- From 2025, NFRs will be available on official portal, the European Single Access Point (ESAP).

Project overview

767 Corporate reports from 20 EU countries
(NFRs, Sustainability Reports, Note to the Financial Statement, Management Reports 📄)

30 KPIs of ESG interest
(Environmental, Social, Governance, Circular Economy, Green Finance 🌍)



Prompting the GPT model and validating the results
We tackle the challenge of extracting information from firms' statements combining AI expertise with the domain knowledge of economic-ESG research

Create a repository/database of the results/end user interface
(allow further documents processing 🔎)

The extraction process at a glance

How we collected ESG indicators

- We hand-collected more than 700 **Non-Financial Reports (NFRs)** for European firms, for both 2022 and 2023.
- We defined a comprehensive collection of **ESG-related KPIs**, consisting in quantitative and qualitative/binary (true-false) variables.
- **Cost-effective pipeline**: we do not want to feed GPT the entire document (too many irrelevant tokens + LLMs tend to underperform when feeded with too much information).
- For each KPI, we defined a list of coherent keywords to **extract relevant paragraphs** from the documents. We used ESG-economic expertise to refine keywords and paragraph extractions up to satisfactory performance.
- We then feed these paragraphs to GPT and refined **prompts several times through human validation** to assess the accuracy of the extraction process on a subsample of documents.

Selection of the KPIs

Selection of 30 KPIs

E

- Scope 1 Emissions
- Scope 2 Emissions
- Scope 3 Emissions
- Environmental Monitoring & Management Systems
- Circular Economy initiatives
- Water Use
- Environmental Certifications
- Recycling
- Hazardous Waste Production
- Non-Hazardous Waste Production
- Consumption and Production of Renewable Energy
- Taxonomy-aligned Revenues
- Product Certifications
- Company Certifications
- Emission of Polluting Agents in the Atmosphere
- Emission of Polluting Agents into the Water
- GHG Emission Reduction Policies
- Product Life Cycle Extension
- Life Cycle Assessment (LCA)
- Eco-design

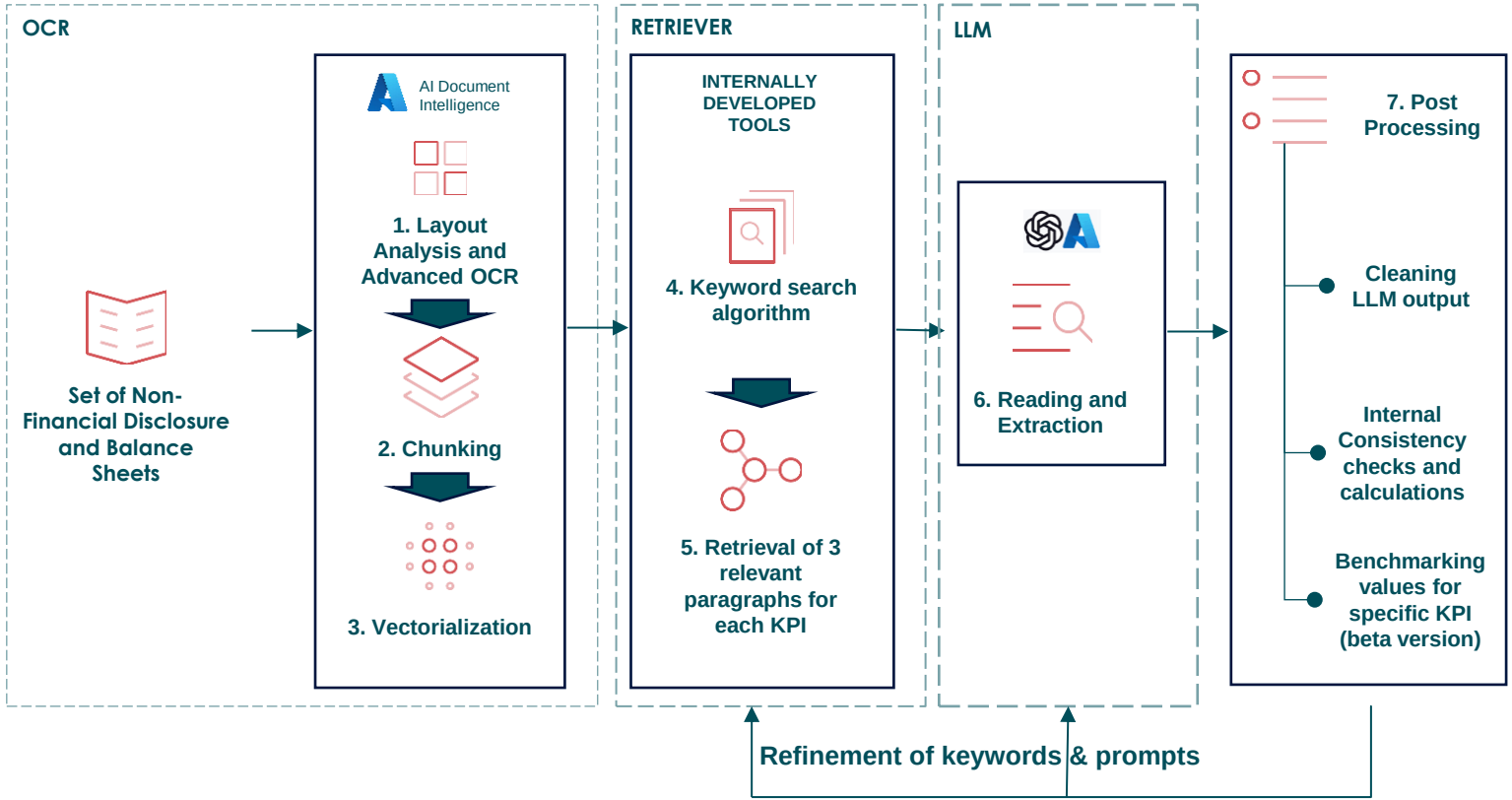
S

- Percentage of Female Employees
- Percentage of Female Employees in Managerial Positions
- Employee Training Hours
- Percentage of Fixed-Term Employees
- Number of Injuries
- Benefit/B Corp Corporations
- Quality Management Certifications

G

- CSR Committee
- Percentage of Female in Board of Directors
- CEO = President of the Board

The pipeline



The pipeline

Once the KPIs of interest were defined, we followed a 3-step pipeline to extract the information:

1. **Optical Character Recognition (OCR)**: we converted the visual content of the document into digital text through the **vectorization of the paragraphs**.
2. **Relevant paragraph retriever**: we selected a set of **keywords specific to each KPI**. A keyword should be a word that is more likely to appear in the paragraphs/table containing the information. This involved a **trial-and-error process to identify the set of keywords** that better performed in the extraction of the correct paragraph.
3. **Large Language Models**: we defined and **refined the query prompted to the LLM** for the extraction and the storage of the information. We addressed the questions in such a way that mathematical computations were limited (or avoided) and hallucinations were prevented.

Testing on a restricted sample of documents

Page retriever

As the algorithm needs human validation, for each we defined a sub-sample to test the extraction of the indicators.

- We then extracted and validated the output for **30 KPIs on 65 documents** (35 purposefully selected for each KPI + 30 randomly selected), mostly in Italian.

To extract and analyze the KPIs:

- We manually reviewed the documents to identify relevant information (**Ground Truth**) and **recorded the pages** where it was located;
- We ran the pipeline to extract data;
- We measured performance by verifying that the **extracted paragraph pages matched those we had identified during the manual review**.

Testing «out-of-sample»

KPI value extraction reader

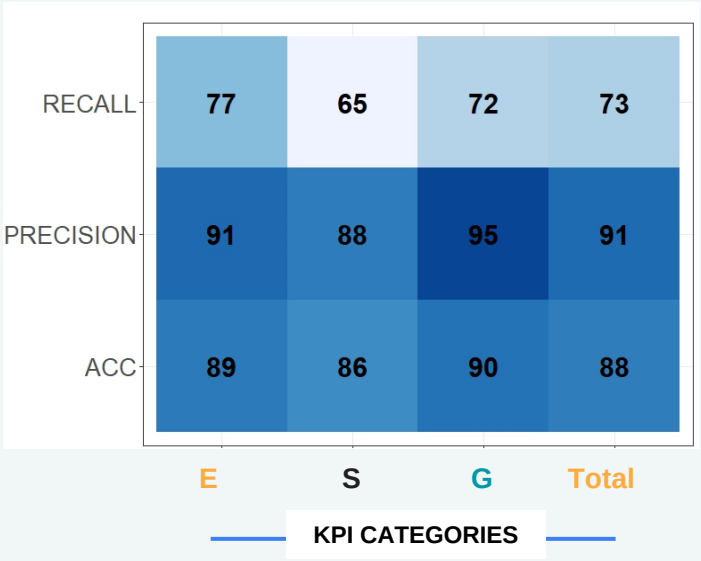
For 10 new documents, we applied our pipeline and extracted the 30 KPIs. Here, we performed a more detailed manual review to ensure accuracy. For this smaller sample:

- **Performance evaluation**: we verified that **the extracted KPI value matched the value reported in the document**.
- **Post-processing**: we worked with the extracted data to retrieve the final output (e.g., calculating percentages to prevent LLM hallucinations, standardizing units of measurement, etc.).
- **Error analysis**: we noticed that most errors (57%) are due to the optical character recognition infrastructure (OCR), responsible for the conversion in digital text.

Performance Evaluation

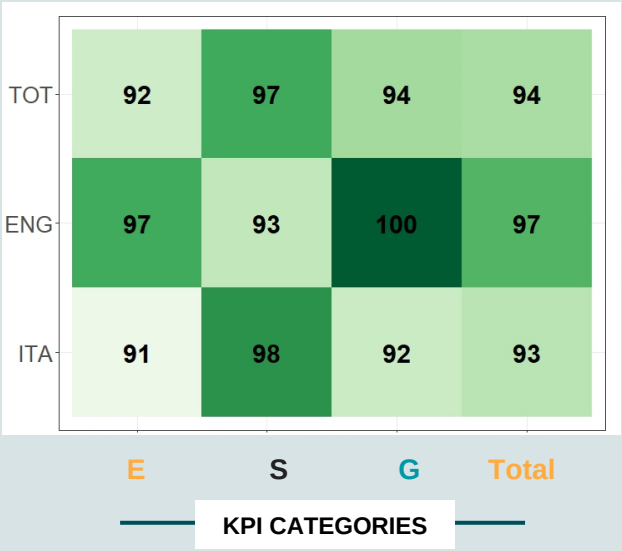
PAGE RETRIEVER PERFORMANCE

Accuracy, precision and recall over the sample of 65 documents; computations over **page numbers**.



KPI DATA VALUE PERFORMANCE

Accuracy over the restricted sample (10 documents), by document language; computations on **KPI's actual value**



ERROR ANALYSIS

There were 14 wrong predictions in total, caused by:

OCR	57%
RETRIEVER	14%
LLM	29%

OCR is the most error-prone text.
Prompting can be improved further.

Recall = true positives / (true positives + false negatives) , Precision = true positives / (true positives + false positives) , Accuracy = correct predictions / all predictions

State of the art and next steps

Task	Main activities	Status
Documents collection	Collection of NFRs in Italian	**Completed**
	Collection of NFRs in English	**Completed**
	Collection of Optical Balance Sheets (OBSs) in Italian	**Completed**
Ground Truth	Collection of the correct page for the restricted sample of NFRs in Italian	**Completed**
	Collection of the correct page for the restricted sample of NFRs in English	**In progress**
	Collection of the correct page for the restricted sample of OBSs in Italian	**In progress**
Keywords	Definition of the keywords for documents in Italian	**Completed**
	Definition of the keywords for documents in English	**Completed**
Prompts	Definition of the prompts to LLM in Italian	**Completed**
	Definition of the prompts to LLM in English	**In progress**
Algorithm training	Extraction of the KPIs for the restricted sample of NFRs in Italian	**Completed**
	Extraction of the KPIs for the restricted sample of NFRs in English	**In progress**
	Extraction of the KPIs for the restricted sample of OBSs in Italian	**To be done**
KPIs total extraction	Extraction of the KPIs for the total set of NFRs in Italian	**To be done**
	Extraction of the KPIs for the total set of NFRs in English	**To be done**
	Extraction of the KPIs for the total set of OBSs in Italian	**To be done**
New KPIs	Definition of new ESG-related KPIs	**To be done**

Conclusions and Applications

- Non-financial reports contain **large ESG information in the form of unstructured data**.
- We developed a **cost-effective pipeline** to extract 30 consistent ESG KPIs from NFRs in 3 steps:
  i) OCR to convert **PDFs into plain text**;
  ii) Retriever to identify **KPI-relevant portion of text**;
  iii) **LLM prompting** to extract the KPI.
- After the initial **fine-tuning necessary for each KPI**, the algorithm is extremely fast and **relatively cheap**.
- Research applications:
  i) in the near future, this pipeline allows to **retrieve ESG KPIs for thousands of European firms** → we can leverage them to **predict ESG ratings for SMEs** (Ozkan, Romagnoli and Rossi, 2023).

  ii) Battiston, Monasterolo and Montone (2024). It's not about ESG, but about **tech greenness and disclosure** → we can **build an index of green technological investment** out of extracted KPIs and further test this hypothesis.

Thank you for the attention