Information extraction on sustainability reports using Transformer-based models

Enora Petry

Aggregate Confusion Project
Massachusetts Institute of Technology

April - July 2024

Table of Contents

Context

Why sustainable finance ?

Sustainability metrics

Three main pillars, divided in various subtopics :

Environmental	**Social**	**Governance**
Climate Change	Workplace safety	Board composition
Resource Use	Fair wages	Lobbying
Toxic Emissions	Product Safety	Corruption and Fraud

Figure: Examples of ESG factors

Sustainability metrics

Three main pillars, divided in various subtopics :

Environmental	**Social**	**Governance**
Climate Change	Workplace safety	Board composition
Resource Use	Fair wages	Lobbying
Toxic Emissions	Product Safety	Corruption and Fraud

Figure: Examples of ESG factors

Quantify how sustainability affects the companies' decisions

Sustainability reports

Sustainability reporting is the disclosure of non-financial information to stakeholders.

They provide information concerning environmental, social, economic and governance issues.

PDF format, ranges from 20 to 100 pages long.

Objective

The objective of the research project is to:

Investigate on the correlations between ESG disclosure and ESG Scores

Methodology

Sustainability reports are difficult to parse:

- Unstructured text, no common framework
- Large variety of topics
- Long textual information

Methodology

Sustainability reports are difficult to parse:

- Unstructured text, no common framework
- Large variety of topics
- Long textual information

A need to work at the paragraph scale,
cluster according to topic and extract key sentences.

Methodology

To work with sustainability reports:

1. Divide into paragraphs $\rightarrow$ **PDF-to-text pre-processing**
2. Cluster according to topic
3. Extract key sentences

Methodology

To work with sustainability reports:

1. Divide into paragraphs → PDF-to-text pre-processing
2. Cluster according to topic → **Topic analysis for each paragraph**
3. Extract key sentences

Topic selection

MSCI ESG Ratings has specific key issues scores per industry.

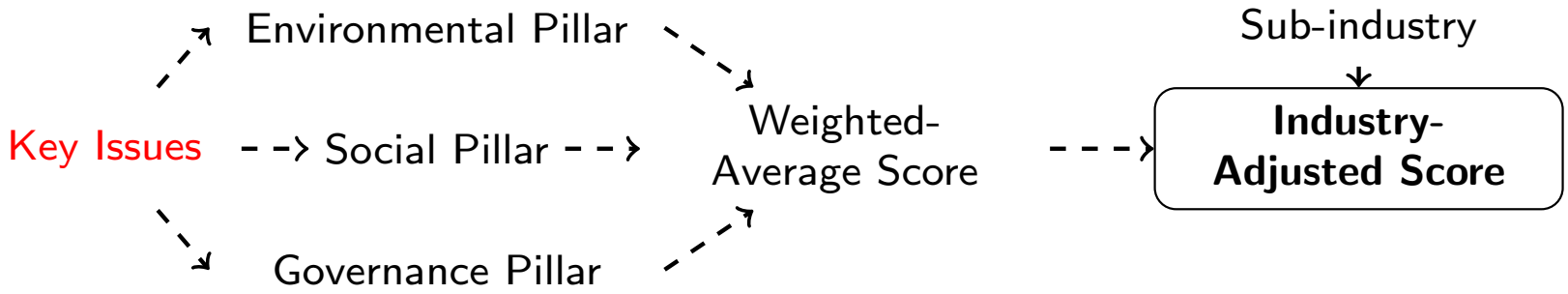


Figure: Overview of MSCI ESG Methodology

Topic selection

MSCI ESG Ratings has specific key issues scores per industry.

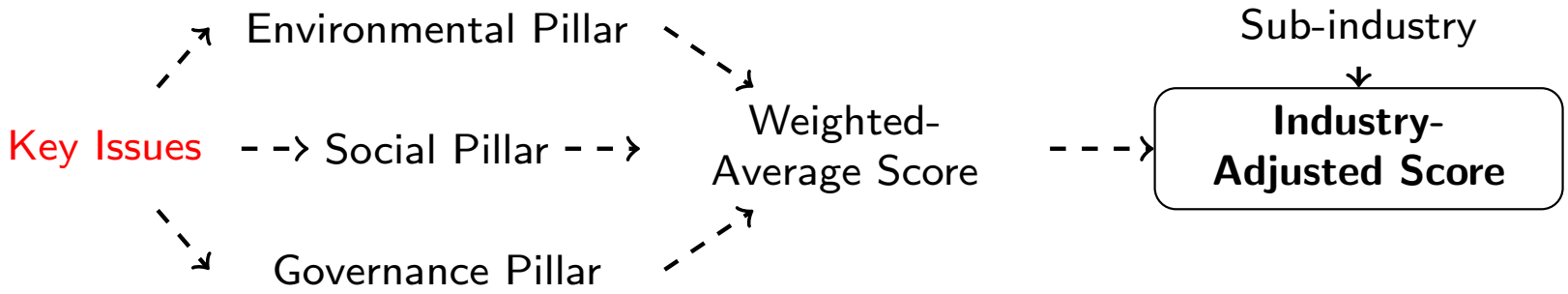


Figure: Overview of MSCI ESG Methodology

Use six industry-specific key issues for clustering

Topic classication

Zero-shot classification

Natural language processing task which aims to correctly label a fragment of text with unseen categories.

Topic classication

Zero-shot classification

Natural language processing task which aims to correctly label a fragment of text with unseen categories.

Several advantages:

- No additional training required
- Captures semantic meaning
- Flexible

Zero-shot classification

Yin, Hay and Roth reframed text classification as a natural language inference problem:

Probability of text being in topic $\longrightarrow$ Given the text, is the following hypothesis true ?
text : text
hypothesis : "This example is topic"

Classification problem **Entailment problem**

Zero-shot classification

Yin, Hay and Roth reframed text classification as a natural language inference problem:

Probability of text being in topic $\longrightarrow$ Given the text, is the following hypothesis true ?
text : text
hypothesis : "This example is topic"

Classification problem **Entailment problem**

Remarkable results for zero-shot classification.

Topic classification

Zero-shot classification with Bart-MNLI

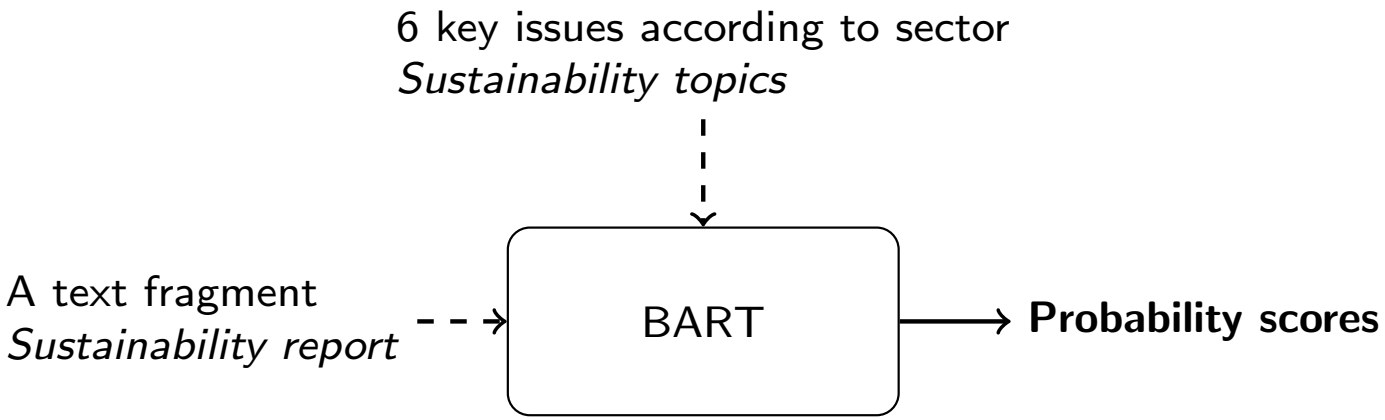


Figure: Overview of Methodology

Methodology

To work with sustainability reports:

1. Divide into paragraphs → PDF-to-text pre-processing
2. Cluster according to topic → Topic analysis for each paragraph
3. Extract key sentences → **Use probability score to select sentences**

Methodology

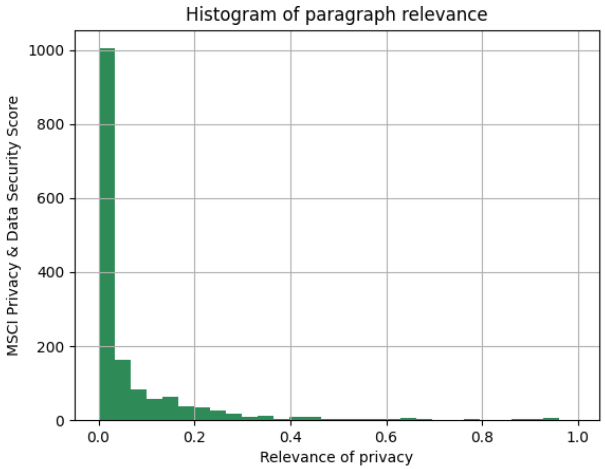
To work with sustainability reports:

1. Divide into paragraphs → PDF-to-text pre-processing
2. Cluster according to topic → Topic analysis for each paragraph
3. Extract key sentences → **Use probability score to select sentences**

Select the 20% most relevant sentences.

Compare ESG disclosure with ESG scores

Aggregation function : necessity to calculate a **disclosure score** for every topic at the document level.



→

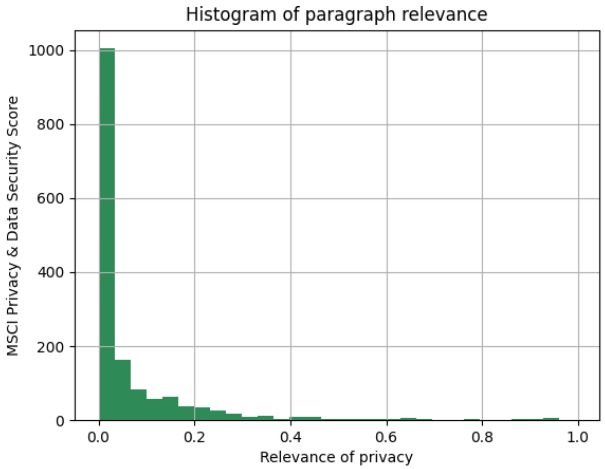
Score from ESG report:

0.28

Average probability score
of the top 20% of text paragraphs.

Compare ESG disclosure with ESG scores

Aggregation function : necessity to calculate a **disclosure score** for every topic at the document level.



Score from ESG report:

0.28

Average probability score
of the top 20% of text paragraphs.

Captures the quantity and relevancy of the sentences.

Data

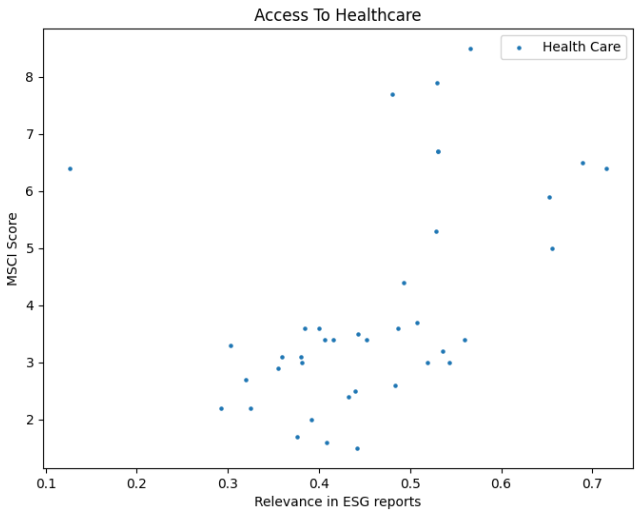
For this project, 1863 ESG reports were analysed.

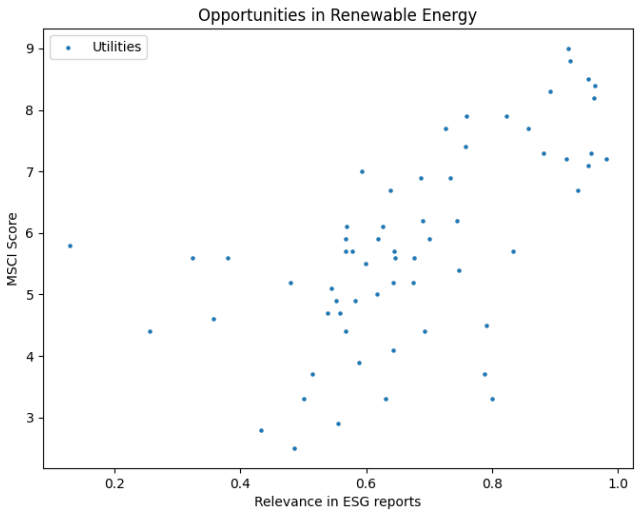
Year	ESG reports
2020	621
2021	804
2022	438
Total	1863

Smaller sample due to time constraints.

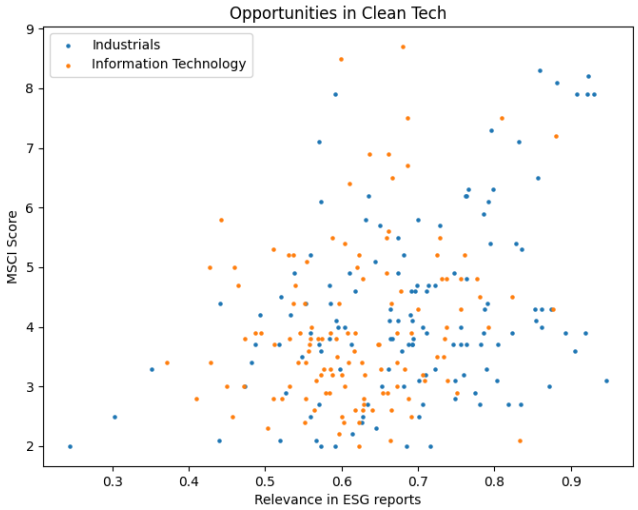
Results

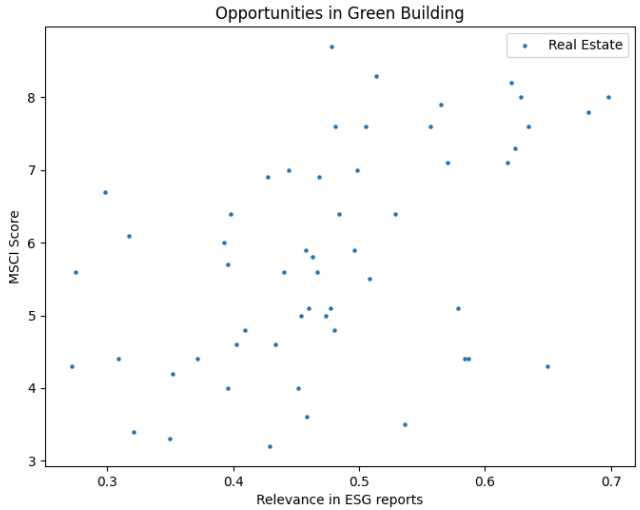
Correlation between calculated ESG disclosure score and the ESG Score for six key issues:

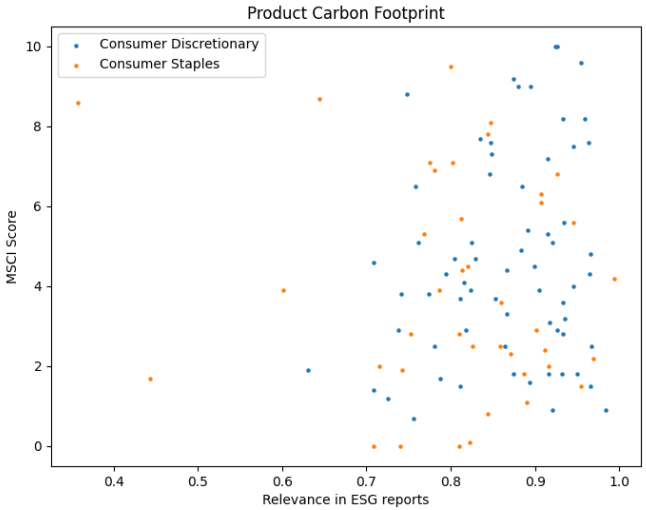


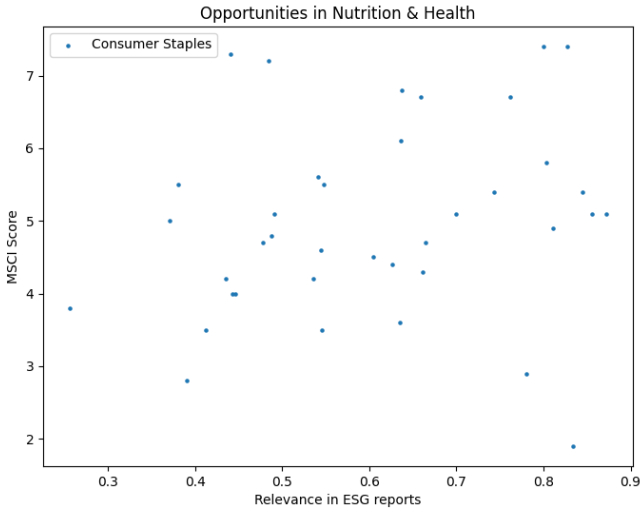


Results


Opportunities in Clean Tech
Industrials
Information Technology
MSCI Score
Relevance in ESG reports


Opportunities in Green Building
Real Estate
MSCI Score
Relevance in ESG reports


Product Carbon Footprint
Consumer Discretionary
Consumer Staples
MSCI Score
Relevance in ESG reports


Opportunities in Nutrition & Health
Consumer Staples
MSCI Score
Relevance in ESG reports

Discussion

Key takeaways:

- Positive, but noisy correlation
- One-dimensional key issues
- Large focus on opportunities

Discussion

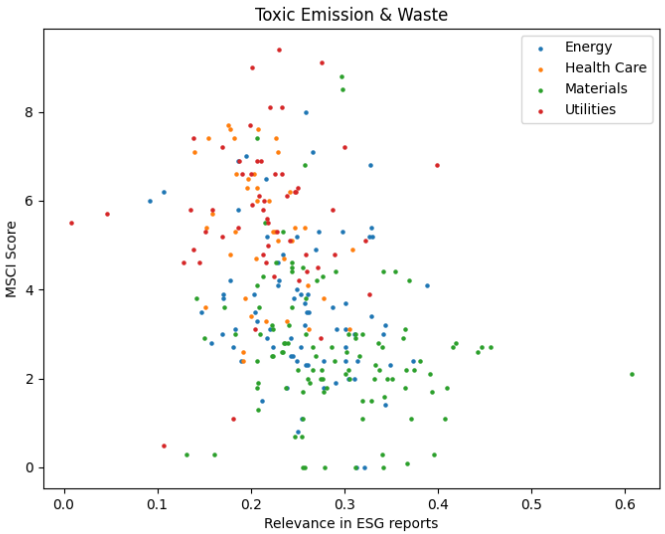
Key takeaways:

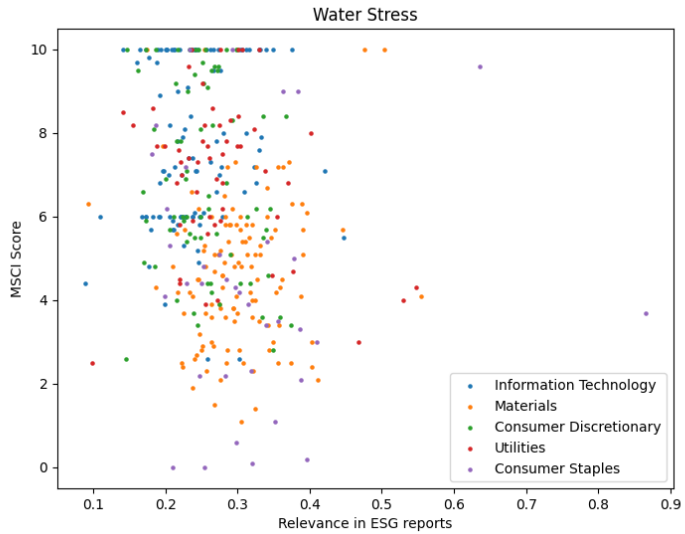
- Positive, but noisy correlation
- One-dimensional key issues
- Large focus on opportunities

Encouraging results

Results

Some graphics seem to indicate a negative correlation between disclosure and scores :

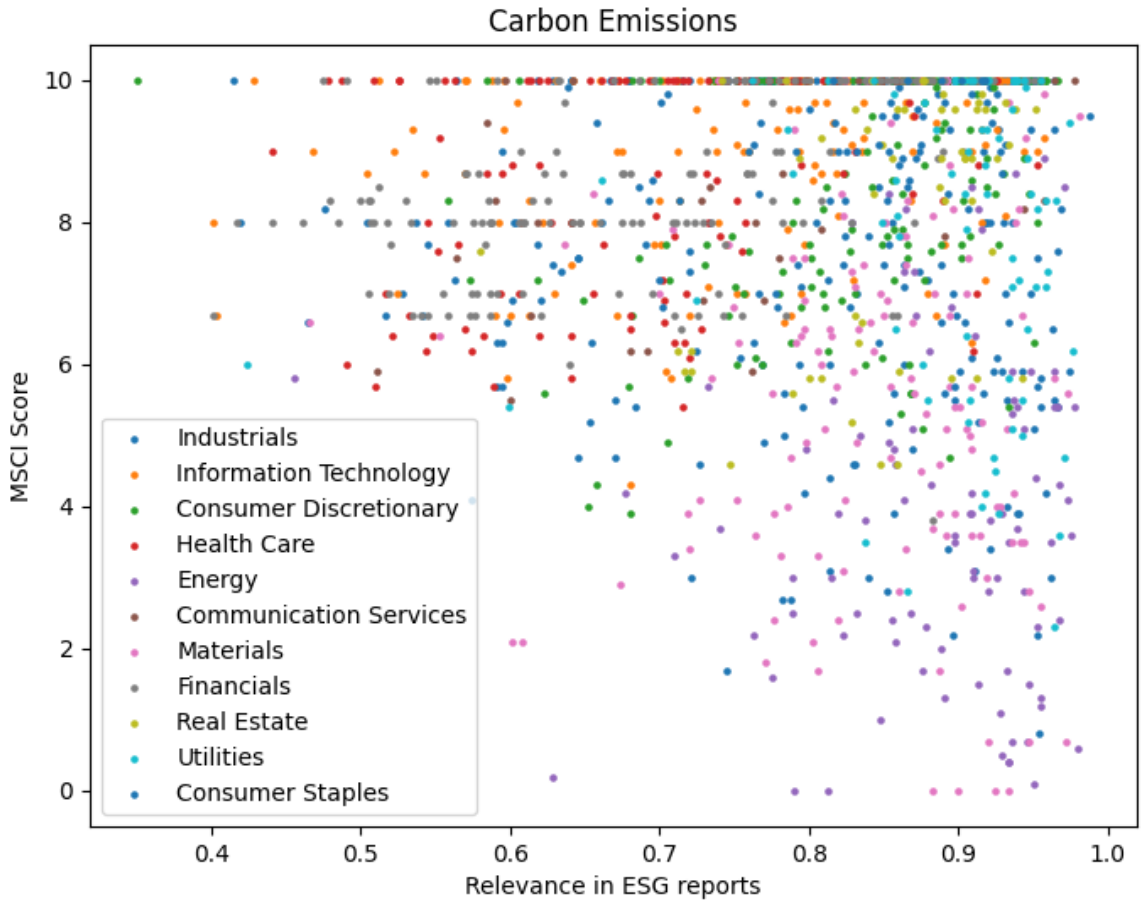




For certains issues, companies that disclose more extensively are the ones with lower scores.

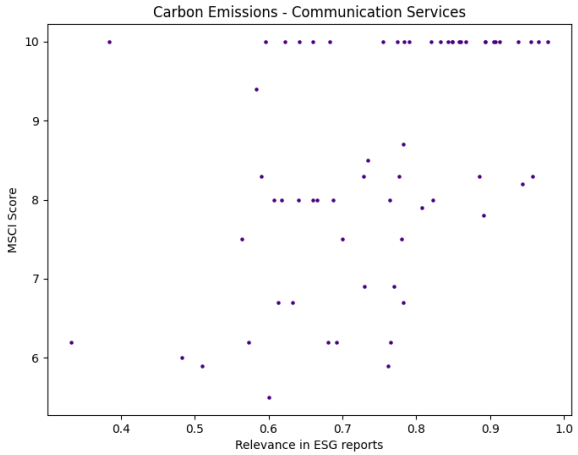
Results

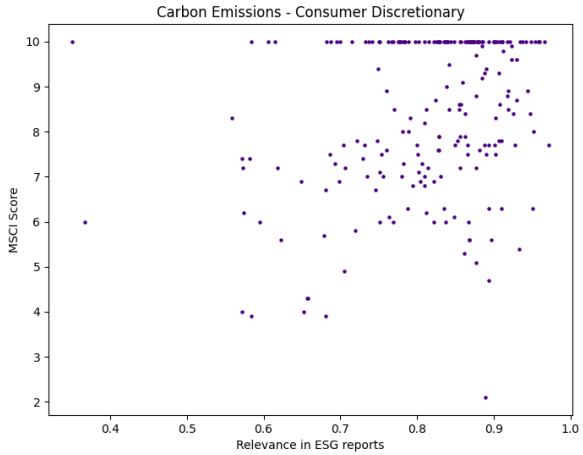
An interesting pattern occurs with the Carbon Emissions key issue:

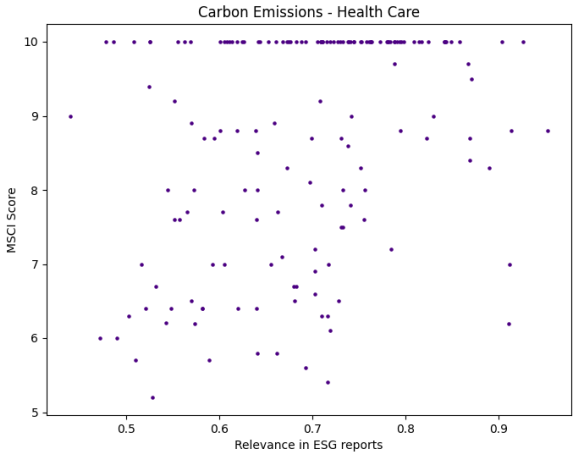


Results - Carbon Emissions

However, positive correlations can be observed for a few sectors:







Next steps

During the remainder of the internship:

- Increase sample size
- Add *Risk*, *Opportunities*, *Management* categories
- Explore additional datasets

Conclusion

This research project has allowed to:

- Explore sustainability datasets
- Develop a method to extract information from sustainability reports
- Analysed patterns between disclosure and scores

Conclusion

This research project has allowed to:

- Explore sustainability datasets
- Develop a method to extract information from sustainability reports
- Analysed patterns between disclosure and scores

Better transparency and understanding of how disclosure affects sustainable metrics.

Acknowledgements

I would like to thanks:

- Aggregate Confusion Project
- Finance Durable et Investissement Responsable

Acknowledgements

I would like to thanks:

- Aggregate Confusion Project
- Finance Durable et Investissement Responsable

Thank you for listening!

Keywords

MSCI Key Issue	Keywords
Access To Finance	expand financial services
Access To Healthcare	improve health access developing countries
Anti-competitive practices	anti-competitive practices
Biodiversity and Land Use	impact operations biodiversity land
Business Ethics Fraud	fraud conflict of interest
Carbon Emissions	manage carbon related risks and opportunities
Controversial Sourcing	efforts sourcing traceability and certification
Corruption	corruption risks bribery
Financing Env Impact	capitalize on opportunities green finance
Finance Product Safety	regulations financial products
Health and Safety	workplace safety standards
Human Capital Development	employee training leadership productivity
Opps in Clean Tech	strategy and revenue clean technology
Opps in Green Building	building regulations and performance real estate
Opps in Nutrition and Health	improve nutritional health profile
Opps in Renewable Energy	renewable power development
Privacy and Data Security	privacy regulations information security
Product Carbon Footprint	reduce carbon footprint
Product Safety and Quality	product safety quality management
Raw Material Sourcing	materials traceability certification
Toxic Emission and Waste	toxic contamination management
Water Stress	water stress risks opportunities