Utilising Natural Language Processing (NLP) to identify cherry-picking in environment related ESG disclosures

Chintan Rakeshkumar Patel
Student ID 6066236

Resource Efficiency in Architecture and Planning (REAP MSc.)
Hafencity University, Hamburg

Supervisors:
Prof. Dr.-Ing. Jörg Rainer Noennig
Msc. Maria Alejandra Moleiro Dale

hcu
HafenCity
University
Hamburg

Utilising Natural Language Processing (NLP) to identify cherry-picking in environment related ESG disclosures

by Chintan Rakeshkumar Patel.

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in the Resource Efficiency in Architecture and Planning (REAP) Program at HafenCity University, Hamburg,

under the supervision of

Prof. Dr.-Ing. Jörg Rainer Noennig

Msc. Maria Alejandra Moleiro Dale

27.03.2024

HafenCity
Universität
Hamburg

Acknowledgements

I extend my sincere gratitude to my advisors, **Prof. Dr.-Ing. Jörg Rainer Noennig** and **MSc. Maria Moleiro Dale**, whose profound knowledge and meticulous guidance have been invaluable throughout this endeavour.

Additionally, I am indebted to **Mr. Balázs Cserpes**, who generously shared his expertise which laid foundations for this research.

I would like to extend my heartfelt appreciation to Marlene, Mehul and Aesha for their time, expertise and constructive feedback. Your invaluable insights and thorough evaluation have significantly contributed to the refinement and improvement of this work.

I would like to express my appreciation to HafenCity University and the REAP International Program for providing an academic environment that facilitates cross-cultural educational experiences and fosters personal growth. I am also thankful to my fellow REAPies, whose camaraderie and shared experiences made this Master program truly memorable and enjoyable. Being part of the 11th generation of REAP has been an honour.

To my family, despite the physical distance, your love and support have always been a source of strength and inspiration.

Lastly, I acknowledge with gratitude the support of all those who have been part of my journey thus far, and I appreciate how this research has provided me with a new perspective on sustainability.

Abstract

In response to escalating environmental challenges, the notion of a green economy has emerged as a pivotal approach to address pressing climate-related issues. The green economy framework seeks to harmonise economic growth with environmental preservation and social well-being. In past few decades the environmental, social and governance (ESG) criteria has emerged as a pivotal assessment for private sector institution to address pressing climate-related issues and mitigate environment risks. The significance of ESG in the context of green finance is highlighted by the exponential growth of sustainable investments and ESG oriented funds, reaching €32 trillion globally in 2020. Despite serving as an important assessment measure to guide sustainable investments, the ESG data published by institutions often remains unaudited. The lack of extensive auditing of ESG disclosures poses significant challenges to the accurate evaluation of environmental impact and sustainability efforts undertaken by institutions. This discrepancy between the perceived importance of ESG considerations and the lack of auditing mechanism served as main motivation for this thesis project. Identifying the need for neutral and transparent assessment of ESG data, this thesis proposes a novel approach leveraging Artificial Intelligence (AI) and Natural Language Processing (NLP) to extract information on environmental disclosures from ESG publications.

The thesis research aims to understand the complexity of the ESG data and present NLP methods to adequately address the challenges posed by lack of standardisation and diversity encountered in ESG reports. To address the research objective adequately, the thesis proposes a customised computational approach integrating state-of-art large language models and machine learning to analyse ESG reports published by non-financial institutions hailing from the EU region. Results from the analysis of 87 ESG reports spanning three consecutive years reveal insights into environmental disclosure practices across diverse sectors. Despite challenges and limitations, the thesis sheds light on transformative potential of AI and NLP in enhancing the efficiency and accuracy of ESG analysis.

Keywords : ESG; Natural Language Processing; Environmental disclosures; GHG emissions; Energy Efficiency; Water Consumption; Artificial Intelligence

Legal Disclaimer

This document serves as a legal disclaimer for master thesis research undertaken as part of degree program Resource Efficiency in Architecture and Planning at Hafencity University, Hamburg, Germany. The contents of this thesis project are intended for academic and research purposes only. The information presented herein is not to be construed as legal advice or professional guidance.

The text-mining activities conducted as part of this thesis project adhere to all applicable laws, regulations, and ethical standards governing research practices. The project utilises publicly available material, which have been accessed legally and in accordance with relevant copyright and intellectual property laws. It is important to note that while every effort has been made to ensure the accuracy, reliability, and completeness of the information presented in this thesis project, no guarantee is provided regarding the correctness or suitability of the data and analysis. The findings, interpretations, and conclusions presented in this thesis project are based on the available data and the methodologies employed, and they may be subject to limitations, biases, or inaccuracies inherent in text-mining and research methodologies.

The author of this thesis project asserts that it is a non-commercial project, conducted solely for academic and research purposes. Upon completion of the research project, all copies of the dataset and corpus utilised were deleted. In order to ensure the reference and subsequent verifiability of the results, the body and excerpts of the original material may have been transmitted to a library, museum, or other publicly accessible educational institution for archiving purposes. These actions were undertaken in compliance with applicable copyright laws and regulations, and with due consideration for the privacy and confidentiality of any personal or sensitive information contained within the material.

Any intellectual property rights associated with this thesis project, including any computational algorithms, methodologies, or programming scripts developed or utilised, remain the property of the author and/or respective rights holders. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Author

Chintan Rakeshkumar Patel

Table of Content

List of Figures

List of Tables

INTRODUCTION

Photo by Spiske on Unsplash

1. Introduction

In the wake of environmental challenges and an ever-growing demand for more sustainable solutions, concepts of a green economy have emerged as a pivotal force in addressing pressing climate-related issues and directing a more sustainable development. At its core, the green finance or green economy approach aims to harmonise economic growth with environmental preservation and social well-being. This paradigm shift has gained significant traction in the last few decades, serving as a beacon of change for a more sustainable and equitable future. A report published in 2020 by the Global Sustainable Investment Alliance (GSIA) noted investment in sustainable assets worldwide has reached a staggering valuation of €32 trillion (US$35.8 trillion)* and the momentum has continued to grow in the last three years (Global Sustainable Investment Alliance, 2020). This valuation equates to 36% of all professionally managed assets. Apart from the investments in sustainable assets, sustainability-oriented bonds also experienced a surge in issuance. In the first quarter of 2023 alone, the Climate Bonds Initiative disclosed the issuance of approximately €192 billion (US$ 205 billion)* through various green bonds, social bonds and sustainability bonds (Climate Bonds Initiative, 2023).

As global awareness has grown towards the realisation of the effects of climate change and a dire need for sustainability, humanity stands at a critical juncture where investment in environment-friendly assets is not merely an option but an imperative. However, deceptive or misleading practices adopted by the private sector pose a hurdle in the accurate assessment of the disclosures and impact of the climate-friendly initiatives undertaken by the institutes. In order to address these ongoing and critical issues, Environment, Social and Governance (ESG) criteria are widely accepted and utilised by a large portion of pension scheme trustees, financial regulators, private and institutional investors, sovereign funds, and wealth funds to holistically assess the green finance-related decisions (Krueger et al., 2020). Denoted as ESG, these environmental, social and governance indicators encompass a wide range of information providing an overall profile of activities undertaken by the institutes through various indicators and goals. The Environmental dimension focuses on issues including but not limited to climate change, biodiversity, resource depletion, waste management, pollution control and deforestation. The Social dimension covers information related to human rights, modern slavery, child labour, working conditions and employee relations. The governance dimension encompasses bribery and corruption, executive pay, board diversity and structure, political lobbying and donation as well as tax strategy.

The ESG documentation has grown in popularity and increasingly become integrated into financial and non-financial institutions' corporate activities,

* a constant exhange rate of 1 US$ = 0.94 € (as of 11.2023) is assumed.

planning as well as risk management strategies. It has also emerged as an effective tool to maintain/manage institutional climate-related risks, targets, opportunities and sustainable development of business. The trend of including non-financial measures in investment decisions has been growing steadily thanks to growing awareness towards concerns related to climate change, social inclusion, inequality and ethical investment. Each year institutions publish enormous amounts of intriguing ESG data through various channels such as press releases, reports and blogs. Thanks to a radical societal shift in attitude towards sustainability and regulatory push towards transparency in data disclosure as well as distribution, abundant ESG data is reported or published publicly each year by numerous companies. Although ESG disclosures are widely used to assess company's performance and inclination towards overall sustainable development; the ESG disclosures are not extensively audited or controlled by public authorities in the European Union (EU) demographic. The legislative framework revolving around non-financial disclosures is discussed at length in the second Chapter (EU legislation) of this thesis. The recent developments in the regulations including the implementation of EU taxonomy and the Corporate Sustainability Reporting Directive (CSRD) have strengthened the framework regulations, the lack of standardisation and convergence across institutions operating in various sectors. This leads to malpractice and strategic omitting of certain information from the report as well as cherry-picking the positive information which is advantageous to the firm. Such behaviour could be a barrier to successfully integrating ESG factors in strategic green financing and hinder the transition to carbon neutrality.

The so-called 'E' pillar of ESG disclosure refers to specific information and data related company's environmental impact and performance as well as sheds light on investment strategies implemented towards sustainability. These environmental-related disclosures are an essential part of ESG reporting as they entail details on the range of environmental factors such as greenhouse gas (GHG) emissions, water usage, land use, waste management, energy consumption, biodiversity preservation and overall environmental management practices of the institute. Traditional methods and manual analysis of 'E' pillar disclosures are time-consuming and they struggle to keep up with vast volumes of data in order to extract intricate quantifiable knowledge. Technological advancement has revolutionised the way we digest and disseminate information completely, especially when it comes to automated assessment utilising Artificial Intelligence (AI) and Machine Learning (ML). A sub-branch of AI, Natural Language Processing (NLP) utilises interoperability systems to assess and effectively understand complex human language. NLP showcases the capability to extract essential insights from extensive and voluminous datasets. Its ability to comprehend nuances within the textual and visual data enables efficient categorisation and interpretation of vast arrays of information such as key topics, trends, and sentiment of speech, to name a few.

This discussion about the importance of assessment and standardisation in ESG documentation together with exploring the possibility of automated assessment is the primary motive of this master's thesis which revolves around finding an answer to following question :

Are we able to extract information related to environmental variables in sustainability disclosures of EU-based firms through computational analysis?

Given the magnitude of the main research question, it needs to be broken down into several parts to facilitate a well-directed effort and to understand the intricacy of the 'E' in ESG as well as computational-based analysis, especially Natural Language Processing. A well-formatted and large sample is crucial for successful computational analysis. Financial and non-financial institutes situated in Europe tend to publish more information in their sustainability and annual reports due to numerous reasons including legislative framework and stricter guidelines (Bingler et al., 2021). The research undertaken requires a customised computational approach integrating various aspects of machine learning and NLP. Due to the limitation of available resources and time restrictions, the scope of the dataset is limited to scrutinising disclosure-related reports published by non-financial institutions headquartered in the EU region. The dataset characteristics are discussed at length in the fourth chapter of this thesis.

The second chapter of the thesis sheds light on the importance and emergence of ESG. It holistically covers various aspects included in ESG and further addresses the critical value of the 'E' pillar of ESG in current demographic. The chapter continues with a description of the currently adopted legislative structure for ESG-related disclosures at the EU level as well as prevalent AI based methods in the realm of ESG. The research questions for this chapter can be defined as:

- **What is ESG and why is it important for humankind?**
- **What is the difference between ESG framework and ESG regulation?**
- **What are the existing regulations and guidelines related to ESG disclosures within EU jurisdiction?**
- **What are the quantifiable environment related ESG variables?**
- **What are the prevalent AI based ESG analysis techniques? Which AI analysis techniques are state-of-art?**

The third chapter delves into methodology and precautions implemented to conduct ethical and legally viable datasets containing ESG reports from various institutions across the EU. It also addresses the inclusion of diverse sectors to provide an opportunity to identify underlying trends in environmental ESG disclosures. The second section of the chapter entails a detailed description of document preprocessing. As ESG reports are crafted to convey information in a portable and visually pleasing format, it is inherently not compatible with machine readability. The chapter continues with a systematic review of ESG reports to identify elements of interest for the development of the database and understand document characteristics to gain initial impressions for building a valid NLP pipeline. Furthermore, the fundamentals of document parsing and implementation of a preprocessing pipeline to develop locally stored data directories are explained. Research questions for the third chapter are the following:

» **What precautions should be implemented to adhere to the legal framework of the Copyright Protection Act?**

» **Which elements in the ESG report contain environmental-related disclosures?**

» **Which computational methods are needed to perform information extraction accurately?**

The fourth chapter explores the data analysis methodology for the collected ESG reports, focusing on machine readability and interpretation of text. It discusses the fundamental NLP methodologies required and their significance in extracting insights from ESG reports. As the thesis deals mainly with the extraction of environmental disclosures from ESG documents, this chapter also explains how such data should be processed correctly. The chapter is divided into two sections differentiating between two separate NLP pipelines utilised to extract insights from Textual-based data and Tabular-based data. The textual data analysis section highlights the new developments in AI and the evolution of large language models (LLMs) based technology such as transformers and fine-tuning processes for leveraging the open source NLP resources for text classification tasks. The tabular data analysis section addresses the complexity of extracting information from tables and briefly discusses the experimentation with zero-shot learning methods and LLM-based fine-tuning. The supporting research questions for the fourth chapter are as follows:

» **What are the basics of computational linguistics and NLP?**

» **What kind of attributes do environmental disclosures have?**

» **What is Transformer architecture and how it has impacted information extraction in NLP?**

» **How does LLM-based text classification work?**

» **What are the challenges/pitfalls that need to be addressed?**

» **How can we aggregate and visualise the output from NLP pipelines?**

The results from the two separate NLP pipelines constitute the fifth chapter. The data extracted from 87 ESG reports amounts to almost 190,000 text segment sentences and 682 unique tables. The ESG reports under investigation are collected from three consecutive years (2020, 2021, 2022). The research aimed to identify environmental disclosures from ESG reports and automate this process by leveraging computational tools. An overarching objective was to evaluate the intensity of environmental disclosure inclusion across institutions operating in various sectors. Although concrete correlation between environmental disclosures and influencing factors such as regulatory framework or location of the institution is difficult to establish, various speculations are derived. The chapter continues with discussion dwelling on limitations and challenges of utilising NLP to analyse ESG data. A SWOT analysis is performed to provide critical overview on benefits and future research opportunities derived from the project.

The final chapter (Conclusions) shifts the focus back to research objective surrounding environmental disclosures and NLP techniques recapitulating the research objective. It summarises the most important findings and gives an answer to the main research question.

1.1 Scope

The main objective of this thesis is to investigate the usability of NLP-based techniques to identify environmental-related disclosures from ESG documents. Considering the available resources and challenges in adhering to the legal framework (explained in section 2.7), the scope of the study covers non-financial institutions located in the EU region. This criteria for scope selection is justified further in chapters 2 and 3 of this thesis which essentially entails the scrutiny of possible correlations between environmental disclosure and ESG-related EU legislation changes. As the study involves computational analysis on a local machine, the specifications and capability of computer play a crucial role in the size selection of the dataset. Although the methodology described in Chapter 4 is limited to the study of 87 ESG reports, the pipeline itself is scalable and expected to produce similar or better results if adopted on a larger selection of institutions and ESG report database.

Machine Specifications

```
Processor : 11th Gen Intel(R) Core(TM) i7-11370H @ 3.30GHz
Memory: 16,0 GB (15,8 GB usable)
Storage capacity: 512 GB internal SSD + 5 TB external HDD
Additional remarks: No dedicated GPU processor
```

Implementation in Python

The execution of document preprocessing (as discussed in Chapter 3) and subsequent NLP pipelines (outlined in Chapter 4) necessitates an adaptable and tailored approach. While a wide range of open-source services and software solutions exist for algorithm development, the programming language Python is employed for implementing the computational techniques detailed throughout this thesis. Conceived in the 1990s by Guido van Rossum, Python offers the requisite flexibility and applicability essential for crafting the necessary NLP techniques. Python boasts a syntax that is remarkably intuitive and straightforward to implement. A comprehensive description of the Python syntax utilised throughout this thesis would exceed its confines. However, few resources are outlined in *Appendix 1* that provide thorough explanations.

Python features an extensive array of libraries and modules that can be seamlessly imported and customised to execute a wide variety of tasks. To ensure a smooth reading experience and maintain consistency throughout the thesis, only the concepts and important Python modules are incorporated into the main text framework. Various supplementary Python modules are required to ensure functionality of the computational analysis. These supplementary libraries and its versions are cited, more information and online resources about these libraries are documented in *Appendix 1*.

This approach streamlines the main discussion while still providing comprehensive information on additional tools and modules used in the research.

(E)nvironment (S)ocial (G)overnance

Photo by Geranimo on Unsplash

2. ESG

Since the late 1950s and escalating into the 21st century, the world and humanity have experienced a multitude of significant challenges encompassing environmental risks, the adverse impact of climate change, scarcity of natural resources, unparalleled ecological catastrophes, gender disparity, poverty, inequality, inadequate corporate governance, and economic instability to name a few. Within this array of challenges, several are particularly pivotal for the future and very survival of humankind. In order to address these pressing and enduring challenges, diverse organisations and voices began to express their concerns and advocated the implementation of necessary actions to reverse the current trends. In the early 1980s, this movement initiated a significant dialogue that led to the formulation of Environmental, Social and Governance (ESG) criteria, identifying the risks that institutions encounter in their business conduct. Although discussions on socially responsible investing (SRI) had already been underway for decades, a comprehensive ESG framework was introduced by the United Nations (UN) in their landmark 2004 report "Who Cares Wins: Connecting Financial Markets to a Challenging World" (United Nations, 2004). The core objective of this report is to encourage institutes to integrate ESG concerns into their primary business strategies, promoting the idea that taking responsibility for the impact of business activities on society and the environment can yield long-term benefits.

Along with the idea of responsible investment, the integration of ESG criteria in the strategic operations and management of the institutes are directed to advocate the efforts to mitigate and adapt to climate change and create a more sustainable business environment. Without the intervention of private sector initiatives and global green finance movement, the constant increase in degrading environment could lead to irreversible consequences and devastating natural disasters (Lyon & Maxwell, 2011). A more sustainable economy that takes ESG factors into account is not only meant to lower the risks induced by climate change but in turn induce a positive impact on activities, assets, robust regulations and encourage habitual change in consumers (Moodaley & Telukdarie, 2023). A well-structured ESG framework provides valuable insight into assessing an institute's overall performance and goes beyond mere compliance and targets to drive long-term sustainable value creation (Henisz et al., 2019). By considering ESG factors, institutions can identify sustainability-oriented opportunities for innovation, efficiency and risk management; thereby contributing to a more resilient, conscious and responsible business landscape (Höck et al., 2020).

Environmental factors

- Carbon Emissions
- Raw Material
- Water Stress
- Energy Consumption
- Biodiversity
- Waste Generation

Social factors

- Health and Safety
- Equal Opportunities
- Chemical Safety
- Data Security

Governance factors

- Board Diversity
- Transparency
- Executive Pay
- Business Ethics

Figure 1 : Selected presentation of ESG factors outlined in GRI Standards (GRI, 2021)

This figure has been designed using images from Flaticon.com

2.1 ESG Standards vs. Regulations

The disparity between ESG standards and regulations arises from the enforceability nature of the guidelines. ESG standards and regulations operate as distinct mechanisms, differing drastically in their nature and legal compliance within the ambit of sustainable investing and corporate practices (Henisz et al., 2019).

An ESG standard has a broad scope and its primary purpose is to outline a structured system of principles, methodologies and practices aimed at integrating ESG factors into strategies, investment decisions and reporting. Usually, these standards serve as a voluntary tool providing a structured and elaborated approach for assessing, managing and reporting on sustainability-oriented efforts and ethical considerations. Although these standards are not enforced by government bodies, they offer guidance to institutions opting to evaluate their ESG performance. Fundamentally, ESG standards describe a comprehensive set of criteria and best practices covering various aspects, including but not limited to climate change, labour practice, diversity, corporate governance and community engagement. Although ESG standards and principles are crucial in transition to green economy, they lack harmonisation and convergence. Global Reporting Initiative (GRI), Task Force on Climate-Related Financial Disclosures (TCFD) and Sustainability Accounting Standards Board (SASB) have attempted to create and promote standardisation of the ESG disclosures and framework. Such standardised frameworks are mostly used by institutes on a voluntary basis to report ESG-related disclosures.

Although there are numerous ESG standards available, guidelines from the GRI and TCFD are amongst the most widely used in EU countries. GRI standard provides a comprehensive overview of sustainability reporting with set guidelines and principles for institutes to disclose information on a wide array of topics including its organisational profile, economic performance, environmental impact, social performance, supply chain management, product responsibility as well as stakeholder engagement (GRI, 2022). On the other hand, the TCFD focuses primarily on disclosing climate-related financial risks and opportunities. Recommendations from TCFD provide a foundational approach to disclosing information related to climate-related operations within organisational governance and details on metrics and targets used to assess climate-related risks and opportunities (TCFD, 2017). Even though GRI and TCFD standards are not regulated by the government, an informal survey claims that these standards are used by more than 70% of the publicly traded institutes in the EU country’s annual sustainability report (KPMG, 2022).

However, ESG regulations represent mandatory guidelines or rules established by authorities and stock exchanges. These regulations are legally binding, compelling institutions to adhere to specific reporting standards, disclosures or minimum disclosure requirements. In European Union countries, prioritising ecological and sustainability-focused transition has become paramount (Dai et al., 2023). The EU parliament is improving the current frameworks to cultivate ESG standards and unify the reporting practices within and outside of the European Union (EU). Even though the EU continues to lead in transitioning ESG regulations, this movement is not confined to Europe alone but has become a global trend.

2.2 EU ESG Regulatory Framework

The European Union has been at the forefront of the implementation of regulations in order to facilitate the ecological transition to a low-carbon economy. The European Commission introduced the European Green Deal in December 2019 and set specific targets towards a more sustainable and climate-resilient future (The European Green Deal, 2019). The European Green Deal encompasses the following three principal complements :

- **Objectives**: The primary objective of the European Green Deal is for the EU to become the first climate-neutral continent by 2050. This ambitious goal entails eliminating net GHG emissions and compensating the emissions by creating infrastructure that removes GHG from the atmosphere. Additionally, it aims to implement regulations and solutions that effectively curtail pollution, safeguard biodiversity as well as promote and foster innovations in clean, sustainable technologies while ensuring an equitable and inclusive social transition

- **Strategy**: The strategic foundation of the European Green Deal includes an intermediate target to reduce net greenhouse gas emissions by at least 55% by 2030, in comparison to emission levels from 1990. It calls for concerted action from all sectors to achieve significant reduction in GHG emissions and deployment of renewable resources, energy efficiency improvements, industrial transformation and adoption of sustainable agricultural and land management practices.

- **Funding**: To realise the transformative objectives and milestones outlined in the European Green Deal, EU estimates a need for private investments estimated around € 1 trillion (US$1.06 trillion) within the next ten years.

This European Green Deal paves an elaborated pathway for the EU and European Commission to address societal challenges. It leads to ESG regulatory framwork with the key objectives of eliminating greenwashing practices, enhancing transparency and disclosure as well as ensure comparability of ESG data. A recent report published by European Securities and Market Authority (ESMA) defines greenwashing as :

"practice where sustainability-related statements, declarations, actions, or communications do not clearly and fairly reflect the underlying sustainability profile of an entity, a

financial product, or financial services. This practice may be misleading to consumers, investors, or other market participants". (Progress Report on Greenwashing, 2023)

The aim of the ESG regulations is to neutralise greenwashing practices and encourage sustainable investments by providing clear and transparent ESG information to investors on their investment choice. This regulatory framework includes multiple directives: namely the Sustainable Finance Disclosure Regulation (SFDR), EU Taxonomy and the Corporate Sustainability Reporting Directive (CSRD).

First introduced in June 2020, the EU Taxonomy serves as a standardised classification system that focuses on the 'E' pillar of the ESG. It identifies economic activities substantially contributing to six environmental objectives: climate change adaptation, climate change mitigation, biodiversity and ecosystems, circular economy, pollution and water (EU Taxonomy, 2020). It acts as a tool for institutes to disclose to what extent their activities align with sustainable practices. Although the EU Taxonomy provides a framework to recognise sustainable investments, it does not mandate investments in the ecological transition outlined in the European Green Deal. The classification system outlined in the EU Taxonomy is dynamic and is meant to evolve over time. Currently it only covers two environmental objectives related to climate while the other four objectives are under scrutinisation. However, the scope of activities covered by the EU Taxonomy is anticipated to broaden. The EU parliament is also considering to propose a social Taxonomy in addition to existing environmental focused Taxonomy. Under Taxonomy regulations both financial and non-financial institutions are mandated to disclose:

- **Taxonomy eligibility**: Effective from 2022, reporting of specific activities must be aligned with the existing screening criteria outline in the EU Taxonomy.
- **Taxonomy alignment**: Economic activities must contribute substantially to at least one of the six environmental objectives without causing any significant harm to the other objectives. Taxonomy also requires implementation of a minimum safeguard to limit the exposure to damage. Non-financial and financial institutions are respectively mandated to report their activity's environmental alignment from January 2023 and January 2024. The mandated reporting extents to disclosure of proportions of turnover, capital expenditures and operational expenditures for the activities that align with the EU Taxonomy. Furthermore, banks in particular will be required to disclose their Green Asset Ratio (GAR), revealing the proportion of EU Taxonomy-aligned assets and investments in relation to total assets and investments under management. To accurately report GAR, a transparent and exact data from the clients will be crucial for banks.

In addition to EU Taxonomy, the European Parliament has implemented Sustainable Finance Disclosure Regulation (SFDR) and Corporate Sustainability Reporting Directive (CSRD) as part of its action plan to strengthen the existing ESG regulatory framework and encourage transparent disclosure of information.

Proposed in November 2019, SFDR compliments the efforts outlined in the Paris agreement and subsequent adaptation to Sustainable Development Goals (SDGs) (SFDR, 2019; United Nations, 2015). EU financial institutions as well as financial institutions with presence and/or operations in EU are covered under the SFDR framework. It mandates institutions to disclose the ESG related information at entity and product level. Mainly, SFRD aims to increase transparency related to the ESG profile of investment portfolio through Article 6 (Integration of sustainability risks into investment decisions and investment advice), Article 8 (Pre-contractual disclosures for products promoting environment or social characteristics) and Article 9 (Pre-contractual disclosures for products with sustainable investment objectives). This legislative frame also mandates disclosure of Principle Adverse Impact (PAI) indicators of financial products to increase awareness amongst investors regarding potential sustainability risks associated with these investments. It allows investors access to more transparent portfolios and encourages environment conscious investment (Scheitza & Busch, 2023).

In December 2022, EU Parliament initiated CSRD framework that replaces 2014's Non-Financial Reporting Directive (NFRD) (CSRD, 2022). Applicable since January 2023, CSRD covers a larger scope of institutes head-quartered in the EU as well as other non-EU institutes with presence in the EU. CSRD will be applicable to more than 50,000 large, medium and small-scale non-financial institutes from January 2024 onwards. Any institute meeting two out of following three criteria will be subjected to abide by the CSRD regulations:

» Entities with more than €40 million (US$42.55 million) turnover yearly.

» More than 250 employees.

» Assets valued at more than €20 million (US$21.27 million).

CSRD paves way for the radical shift from voluntary disclosures of ESG related factors to mandatory disclosures. It encompasses a large set of ESG information and provides a standardised way to report. It will include information institutes business model and strategy, social matters, sustainable business opportunity, governance diversity, respect of human rights, etc. (Badenhoop et al., 2023).

Apart from the regulations discussed above, EU has also implemented several regulations to promote and achieve transition outlined in the European Green Deal and EU's Action Plan. In relation to this master thesis, EU Taxonomy, SFDR and CSRD are important and provide an outlook on future disclosure trends within EU. Apart from the policy and regulations presented so far, EU has approved several additional regulations, certifications and frameworks such as Renewed Sustainable Finance Strategy and European Green Bond

Standards (EuGBS) supplementing the endeavours outlined in the EU Taxonomy (Renewed sustainable finance strategy, 2021; European Commission, 2023).

2.3 E in ESG

The 'E' environmental pillar within the ESG framework has garnered significant attention due to its focus on the impact on the natural environment by private sector activities. This pivotal component of the ESG primarily addresses the institute's influence on the environment and is often considered the most complex pillar from a reporting perspective. The institutes provide information and disclosures related to environmental-related risks, natural resource management as well as environmental undertakings and activities. The information also includes but is not limited to institutes reliance on fossil fuels, the adept management of water and other natural resources, pollution levels, climate change mitigation strategies, hazardous waste generation and disposal practices, and the measurement of its carbon footprint. This faction is integral to assessing the potential risks to an institute's long-term financial well-being and survival as investors actively weigh environmental opportunities in financial decisions.

Traditionally the information reported under the environment pillar could be classified into following three categories :

- **Emissions**: Emissions refer to the release of greenhouse gases, pollutants or other harmful substances into the atmosphere as a byproduct of various industrial activities or energy consumption. The emissions are typically measured into three different scopes; Scope 1, Scope 2, and Scope 3. These scopes are further defined and elaborated in the upcoming section of this chapter. Measuring, managing and ultimately reducing emissions are essential while demonstrating sustainable practices and mitigating climate change induced challenges.

- **Resource Use**: This classification evaluates information related to the consumption of natural resources within institutional operations comprising water, energy, raw materials, and land. Efficient resource use is essential for minimising environmental degradation, conservation of finite resources and reducing the ecological footprint. It also sheds light on supply-chain management and eco-efficient solutions adopted by the institutes.

- **Innovation**: Innovation refers to the development and implementation of novel ideas, technologies, processes, or business models that drive positive environmental outcomes. This could involve investing in cleaner technologies, exploration of alternative materials or creative solutions to reduce the environmental cost of operations.

2.4 Measuring Environmental Disclosures

Human activities or 'footprints' are often held accountable for the current degradation of the environment and high pollution levels. While there are complex resources available to measure and accurately predict the human footprint, a fair assumption of the human footprint could be gauged by the rate at which humans consume resources and generate waste versus how fast the Earth can assimilate the waste and replenish resources (Wackernagel & Galli, 2007). Following the engagements to reduce the human footprint, institutions across the globe started reporting their emissions and resource usage. From the perspective of corporations and the private sector, measurement and assessment of such emissions and resource use are valuable not only for regulatory compliance or accounting but also for the development and deployment of climate-friendly strategies to reduce emissions and promote sustainable activities. On the other hand, an accurate baseline of GHG emissions is required to assess the alignment with frameworks such as carbon pricing policies and track processes towards the envisioned net-zero emission targets (Silva Lokuwaduge & Silva, 2022).

This imperative is further strengthened by the forthcoming Corporate Sustainability Reporting Directive (CSRD), set to be enforced in 2024 for large companies and in 2026 for Small and Medium-sized Enterprises (SMEs) with operations inside the EU. Under the CSRD, institutes will be obligated to disclose audited GHG emissions alongside a quantitative trajectory and mitigation strategy aimed at achieving net zero emissions (CSRD, 2022). In light of recent developments in regulation and framework, environment disclosures and the inclusion of information in an institute's ESG reports have evolved and grown drastically. These disclosures typically cover information on the institute-specific impact on various environment-related topics such as waste generation, climate risks, energy transition, soil degradation, natural resources consumption, biodiversity, land degradation, supply chain, water consumption, air pollution, land resettlement, and surface water pollution. Although institutional impact could be quantified and measured in any of these topics, this research project focuses on the three most discussed topics in non-financial institutes' ESG disclosures: Greenhouse Gas emissions, Energy Efficiency and Renewables, and Water Consumption.

2.4.1 GHG Emissions

It is evident that the mean temperature of our planet has been increasing. A 2022 Intergovernmental Panel on Climate Change (IPCC) report describes the impact of GHG emissions on climate change as a direct result of GHG emissions in the atmosphere. Various studies insist on the exceeding in the emissions and resulting consequences (Pörtner et al., 2022). United Nations host an annual Climate Change Conference at the World Conference of the Parties (COP) to review the goals and objectives of the global efforts targeted to fight climate change. Along with the global policy frameworks, the United Nations has also included ‘climate action’ as the 13th sustainable development goal (SDG) emphasising rapid action in accurate monitoring of GHG emissions.

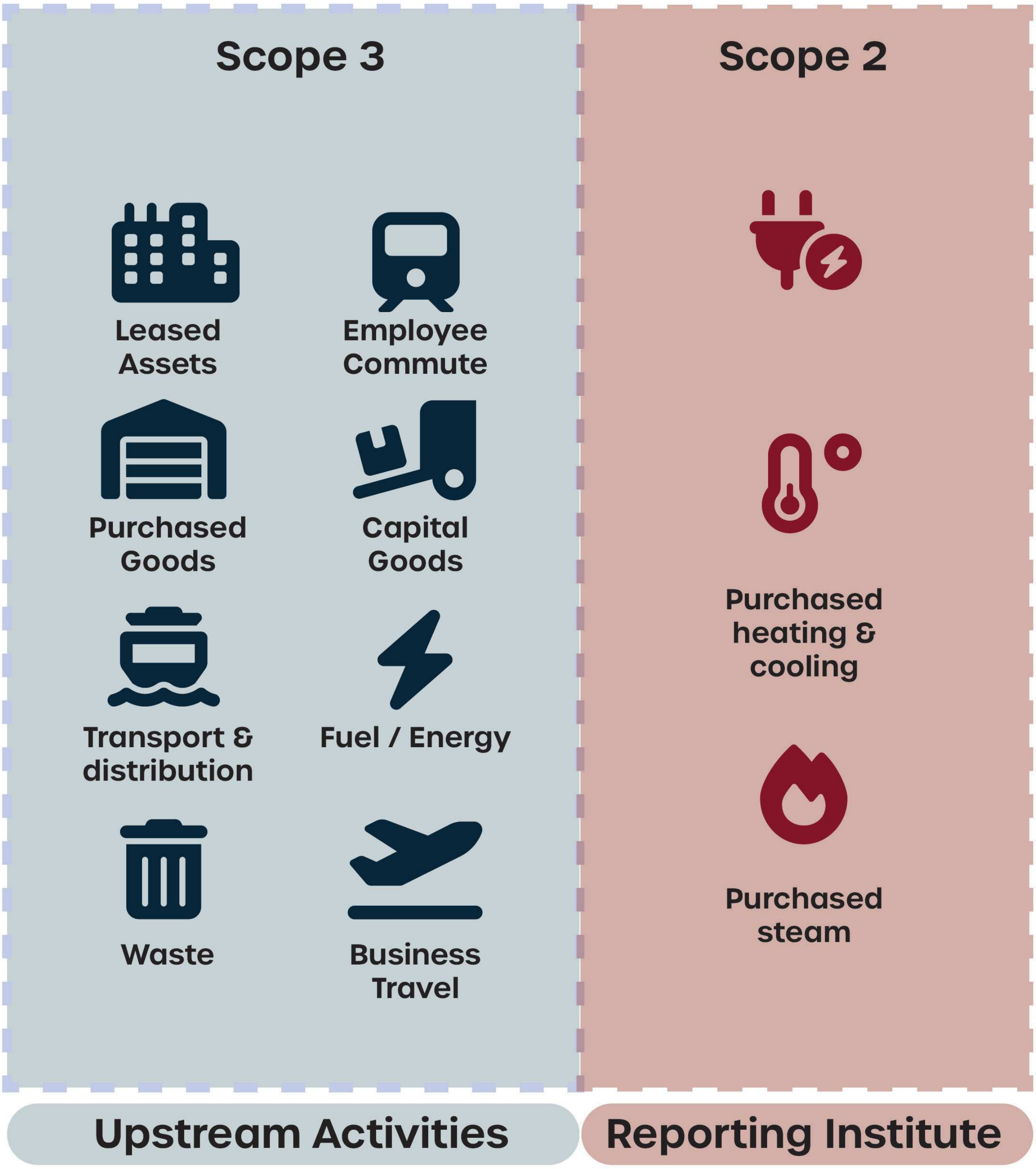
From a scientific viewpoint, GHG emissions serve as a means to quantify the Earth’s global warming, enabling the calculation of radiative forcing: a positive value indicates that the Earth’s system absorbs and retains more energy than it radiates back to space. The assessment of GHG emissions typically comprises all emissions attributed to an individual, event, organisation, service, location or product (Garvey et al., 2018). It is expressed in the unit ‘carbon dioxide equivalent (CO_2-eq.)’, offering a standardised scale for measuring the effect on climate and global warming potential of various sources such as water vapour, carbon dioxide, methane, nitrous oxide, hydro fluorocarbons, etc. According to the Global Warming Potential (GWP) framework, the CO_2-eq value for any gas is determined by the amount of carbon dioxide (CO_2) that would produce an equivalent warming effect on the Earth. In most cases, the amount of CO_2 equivalency is measured in metric tonnes.

At this stage, the location and size of the private-sector institute and its operations determine whether it is mandatory to report the GHG emissions or voluntary. The calculation methodology for GHG emissions is non-uniform and varies along with the regulations and is often sector-specific. The variation in these methodologies can at times impede comparisons between institutes operating across diverse countries or sectors, potentially introducing biases in the information. Additionally, ESG reports often do not sufficiently document calculation methods and assumptions, resulting in more bias and potentially misleading claims. In practice, comprehensive data on operations from a large set of stakeholders is necessary to achieve accurate measurement of GHG emissions. To avoid biases and ensure consistency in calculation methods the GHG Protocol is adopted as standard by major institutes, the World Business Council for Sustainable Development (WBCSD) and the World Resource Institute (WRI). First introduced in 2001, the GHG Protocol provides a comprehensive set of guidelines and tools to measure GHG emissions from cross-sector and country-specific operations. Along with the GHG Protocol, ISO 14064 standard or France’s carbon-balance tool are also utilised by many institutes. To differentiate and manage accurate measurements of direct and indirect emissions, the GHG inventory is divided into three scopes.

- **Scope 1** refers to emissions stemming directly from sources owned or controlled by the institutes. It encompasses emissions generated by energy sources located on the institutes premises, such as natural gas, fuels, and refrigerants, as well as emissions from the operation of boilers and furnaces. Additionally, it also includes GHG emissions from institute-owned fleets, such as cars, vans, trucks, helicopters and aeroplanes.

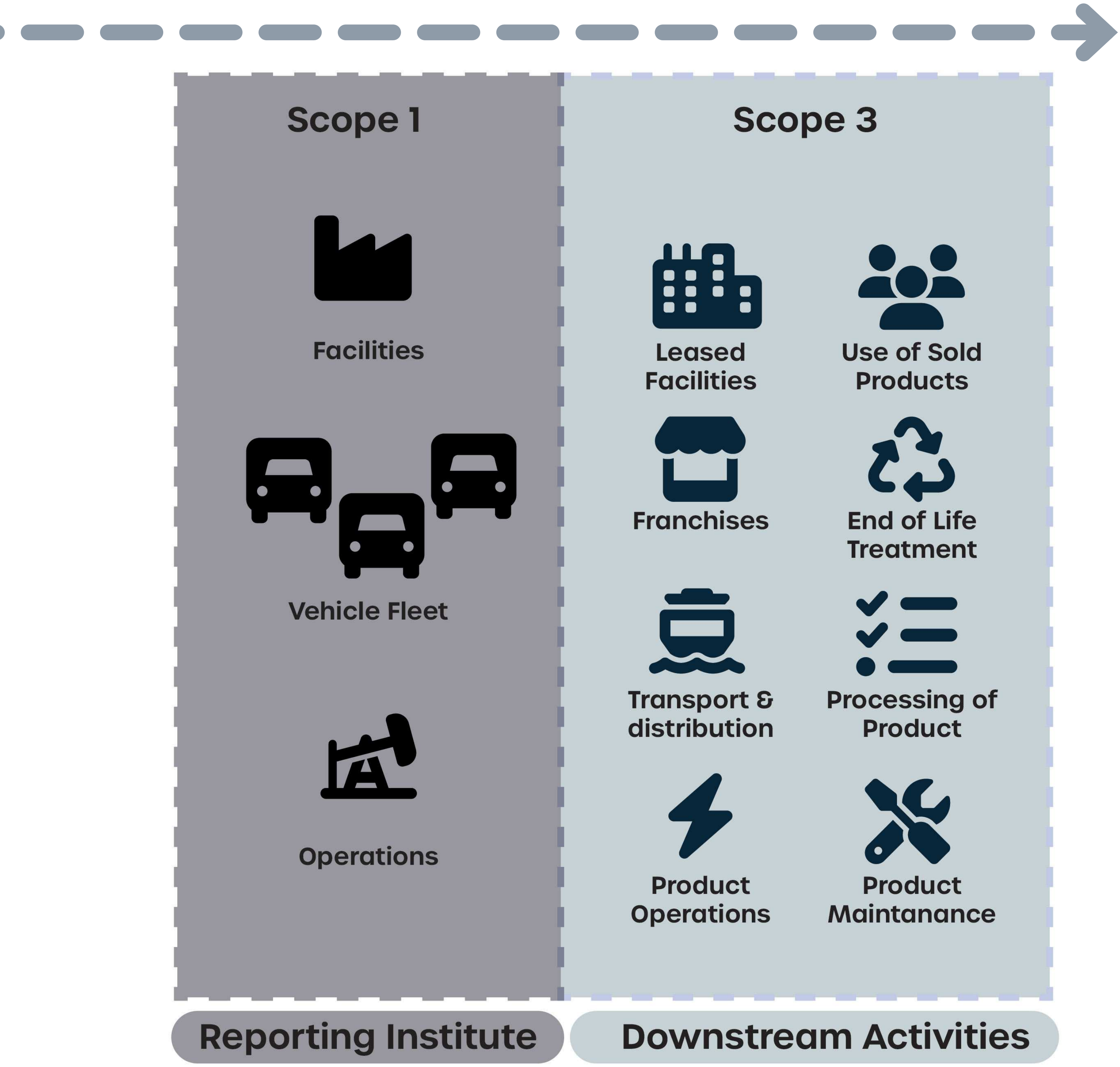
- **Scope 2** accounts for GHG emissions associated with the generation of purchased electricity, power, heating or cooling energy consumed by

Material Supply Chain Flow



Upstream activities refer to processes involved in acquiring raw materials or inputs used in a product or service. In case of energy supplier upstream activities might involve oil drilling, construction of wind farms, solar panel installations, etc.

Figure 2 : GHG emission scopes (Adapted from GHG protocol, 2001)



Scope 1

Facilities

Vehicle Fleet

Operations

Scope 3

Leased Facilities

Use of Sold Products

Franchises

End of Life Treatment

Transport & distribution

Processing of Product

Product Operations

Product Maintanance

Reporting Institute

Downstream Activities

Downstream activities pertain to processes involved in distribution, use and disposal of the final product or service. In case of energy supplier downstream activities encompass distribution of energy to end-users, customer energy consumption, etc.

institutes.

- **Scope 3** refers to all other indirect emissions that are consequences of the activities or operations of the institutes but occur from sources not owned or controlled by it. Some examples of Scope 3 activities are extraction and production of purchased materials, construction of the facilities, and financed emissions via block-chain or investments.

While current regulatory standards enforce reporting on Scope 1 and 2 emissions for large institutes, reporting on Scope 3 remains optional. Often denoted as value chain emissions, Scope 3 emissions are usually the largest component of institutes' total GHG emissions, especially for financial institutions, insurance providers, and the automotive industry as their upstream and downstream activities involve long-term and high emission-prone operations. *Figure 2* illustrates the various activities included in the scope of GHG emissions. The activities and boundaries of the scope depend heavily on the nature of the operations and services offered by the institutes. For example, the petrochemical industry's Scope 1 and 2 emissions would be significantly higher than similar scale institutes operating in the banking sector as emissions occurring from various activities related to fossil fuel such as extraction, production and distribution are relatively very high. A uniform industry-specific emission calculation approach streamlines not just the calculation process but also provides an opportunity to compare the emissions across institutes. The ability to measure GHG emissions properly not only has a positive impact on the development of sound GHG reduction strategies but also provides a better outlook on transition risk and avoids exposure to potential non-compliance fines.

2.4.2 Energy Efficiency and Renewables

Despite the innovation and technological advancements, fossil-fuel dependent operations in the supply chain remain the largest contributor to cumulative GHG emissions. Mitigating the energy intensity of operations or service provision is crucial for achieving a balance between sustained economic growth and environmental preservation. Concerns across the world are growing and carbon-neutrality targets continue to gain momentum. Undoubtedly, energy efficiency improvements and increased use of renewable energy have a significant impact on not only emissions but also on the treatment of pollutants. Because of the long-term benefits and carbon credit initiatives, private sector entities in the EU have shown a great inclination to adopt energy-efficient operations and utilise renewable energy sources. Although the current regulation does not include specifics on disclosures related to energy consumption and usage of renewables in energy-mix, non-financial institutes tend to disclose such information as it has been shown to increase economic stability (Widianingsih et al., 2024). In addition to generating financial benefits to stakeholders, energy efficiency and utilisation of renewable energy contribute directly or indirectly to six sustainability development goals (SDGs 7,9,11,13,16 and 17) (Chen et al., 2024).

Energy-related ESG disclosures encompass a range of metrics including targets, total energy consumption over a fiscal period, utilisation of renewable energy, and investments in renewable infrastructure projects. GRI standards and TCFD standards recommend institutions to disclose total energy consumption across two distinct categories, delineating between energy sourced from non-renewable and renewable sources (GRI, 2020; TCFD, 2017). The disclosure of the energy matrix includes various fuel types such as electricity, heating, cooling and steam. It is measured in unit joules or watt-hours.

2.4.2 Water Consumption

In the domain of non-financial institutions, water consumption and reuse represent prominent focal points following GHG emissions and energy efficiency in discussions surrounding quantifiable environmental measures. ESG reporting concerning water typically encompasses a concise portrayal of the institutes interaction with water as a natural resource, delineating the methods and locations of water withdrawal, consumption, and discharge. The water consumption parameter encompasses not only the quantity withdrawn from sources but also the amount of water recycled and utilised during workplace activities, operational and manufacturing activities. Given the criticality of water as an indispensable and limited resource, entities operating in regions characterised by water scarcity or susceptibility to droughts undertake various water reuse initiatives and invest in replenishing water reservoirs to ensure steady and sustainable operations.

Although the current EU legislation does not mandate disclosure of water-related quantitative measurements, widely adopted ESG standards advocate for the inclusion of comprehensive water-related disclosures to mitigate investment vulnerabilities associated with water-related risks. By delineating water consumption, institutions can provide insights into their overall water utilisation, spanning various operational facets. This transparency extends to water reuse practices, reflecting the institution's commitment to sustainable water management practices. Despite it not being a particularly important disclosure in traditional financial institutes' ESG reporting, in the scope of this research scrutinising water-related information in ESG documentation could be insightful as it includes utilities and service provider institutes (Simionescu et al., 2020). The water consumption is typically disclosed in cubic meter as a unit.

2.5 Cherry-picking in ESG

Cherry-picking in the realm of natural science is primarily defined as the selective disclosure of information or emphasis on certain favourable metrics or initiatives while downplaying or ignoring less favourable aspects. In the context of ESG, cherry-picking is associated with malpractice which only highlights positive aspects of an institutes ESG performance while disregarding or downplaying any negative impacts or practices. For example, an institute might heavily promote and include its environmental sustainability efforts such as investment in renewable energy projects while failing to adequately address issues like the construction of fossil-fuel-dependent assets as its core operations or its water usage and waste management systems. This kind of selective focus misdirects investors and stakeholders through false impressions of the institutes overall ESG performance. Cherry-picking undermines the integrity and transparency of ESG reporting, potentially misleading investors who rely heavily on ESG data to make informed decisions (Bowen, 2014; Yu et al., 2020). An underlying aim of this thesis is to exploit the recent technological advances in the field of artificial intelligence computational methods to automate environment related disclosure.

Recognising the importance of addressing this issue in ESG disclosures, the European Securities and Market Authority (ESMA) has been actively providing clear definitions and possible remediation actions some of which have already been amended in corporate sustainability reporting (CSRD) (Progress Report on Greenwashing, 2023).

2.6 AI in ESG

Artificial intelligence (AI) is a growing area of research that utilises numerical methods to provide task-oriented solutions to a wide range of problems. Essentially AI is a simulation of human intelligence in machines, enabling machines to understand natural language, recognise patterns and many more tasks. The term AI is an umbrella term used for various techniques and technologies, including machine learning, deep learning, natural language processing, computer vision, robotics and expert systems. The advantages of AI in the context of ESG disclosures originate from its inherent capability to handle extensive and diverse datasets, its implementation through state-of-the-art computational methods, and its rigorous scientific and research-led underpinning. There has been an exponential growth in ESG data generation over the last few years and this growth is foreseen to continue at a rapid pace, thanks to the mandates in regulatory framework and growing awareness. In the current AI demographic, two subcategories of AI are most promising for ESG data: natural language processing (NLP) and computer vision in combination with deep learning.

Underlying ESG data necessary for the computational analysis can come in various forms such as reports, messages, news articles or transcripts. As abundant ESG-related data is published each year, it enables a wide spectrum of possibilities and scrutiny through available AI tools. Scholars have grasped the opportunity to analyse the ESG data with a financial as well as non-financial perspective. Perazzoli et al. (2022) utilised basic NLP analysis holistically and evaluated the ESG data from a General Systems Theory perspective. The researchers acquired approximately 55,000 ESG-related texts through web crawling which were then tested to gain insights into environmental risks and occurrence of greenwashing-related terms. Moreover, the developed model was able to digest and forecast the upcoming trends in high-value disclosures from the investors' perspective. Taking a different approach Kang & Kim (2022) collected 60 ESG reports from six institutes situated in developed countries and performed sentiment analysis and thematic analysis. Although this approach was able to identify the thematic structure as well as the balance of positive and negative sentiments in the text, the analysis could not identify the underlying tone of the text. Keyword matching and rudimentary sentiment analysis tools lack ability to extract insights and provide comparable results. On the positive side, these basic NLP algorithms can process huge amounts of data fairly quickly and require less intricate software architecture.

Somewhat advanced NLP algorithms are structured based on non-linear models, such as neural networks that can learn relationships from billions of data points. These neural networks can handle complex language tasks and extract insights from textual data corpus. Moreover, non-linear NLP models demonstrate robustness and scalability, enabling them to handle large-scale text datasets efficiently. With advancements in deep learning techniques and parallel processing architectures, these models can effectively process and analyse massive volumes of textual data, making them well-suited for ESG data analysis tasks. To train and develop such models high-quality, labelled training datasets are required. Unfortunately in the financial and climate sectors, there are several recurring issues with the availability of labelled as well as non-biased datasets which hinders the machine learning applicability. To highlight this research gap, Kölbel et al. (2020) undertook a study to identify transition and physical climate risks using a non-linear large language model. Although the techniques employed were able to identify the various causes that impacted the risk factors, a direct correlation between the effect and cause is difficult to establish due to the lack of labelled data. Scholars Yang et al. (2020) adopted a novel approach to utilise and train a contextual large language model to perform the classification of finance-related ESG texts into primary topics namely corporate fraud, stock returns and volatilities. Going one step further into the topic classification task, Varini et al. (2020) trained a large-language model to perform climate-related topic detection from an unclassified textual-based ESG dataset. Due to the complexity of climate-related text, performance of the trained model was sub par and was outperformed by a more basic keyword-based model. A different approach taken by Bingler et al. (2021) to perform similar task on a pre-classified dataset was able to extract and classify the texts into a desired output. These studies signal that there is a keen interest in analysing ESG data with AI tools. The projects have gained some interesting insights and highlighted inherent challenges in working with ESG data. Apart from researchers, commercial service providers and third-party ESG auditors are also invested in finding viable techniques to automate traditional ESG data auditing practices. In one way or another, all well-known ESG rating consultants employ AI tools to various degrees.

From initial experimentation and literature study, a customised data pipeline and utilisation of a diverse set of AI tools are necessary in order to sufficiently address the research question(s) undertaken in this thesis. The specifics of these are discussed at length in the next chapters.

2.7 Legal Framework

Working with ESG data might lead to some serious legal concerns based on a variety of factors. This is especially true in the case of critique-prone research that might open doors to negative publicity and harm the image of institutions. The legal framework around text-mining projects such as this thesis, is intricate and requires strict adherence to a diverse set of conditions.

Since the research scope involves the formulation of a database of ESG reports collected from publication channels. Before the data is analysed automatically or the data collection task is performed, it should be clear whether usage of the material is permitted for the user case and does not violate any restriction imposed by copyright law. As this thesis falls under German jurisdiction, specific sections of 'Copyright and Related Rights' (Urheberrechtsgesetz - UrhG) are relevant. The following paragraphs highlight the relevant texts from an official English version of UrhG:

> …. "Section 44b - Text and data mining
>
> (1) 'Text and data mining' means the automated analysis of individual or several digital or digitised works for the purpose of gathering information, in particular regarding patterns, trends and correlations.
>
> (2) It is permitted to reproduce lawfully accessible works in order to carry out text and data mining. Copies are to be deleted when they are no longer needed to carry out text and data mining.
>
> (3) Uses in accordance with subsection (2) sentence 1 are permitted only if they have not been reserved by the rightholder. A reservation of use in the case of works which are available online is effective only if it is made in a machine-readable format." …

As the thesis falls under a scientific research purpose following section 60d of UrhG is also relevant.

> …"Section 60d - Text and data mining for scientific research purposes
>
> (1) It is permitted to make reproductions to carry out text and data mining (section 44b (1) and (2) sentence 1) for scientific research purposes in accordance with the following provisions.
>
> (2) Research organisations are authorised to make reproductions. 'Research organisations' means universities, research institutes and other establishments conducting scientific research if they
>
> 1. pursue non-commercial purposes,
> 2. reinvest all their profits in scientific research or

3. act in the public interest based on a state-approved mandate.

The authorisation under sentence 1 does not extend to research organisations cooperating with a private enterprise which exerts a certain degree of influence on the research organisation and has preferential access to the findings of its scientific research.

(3) The following are, further, authorised to make reproductions:

1. libraries and museums, insofar as they are accessible to the public, and archives or institutions in the field of cinematic or audio heritage (cultural heritage institutions),

2. individual researchers, insofar as they pursue non-commercial purposes.

(4) Those authorised in accordance with subsections (2) and (3) and pursuing non-commercial purposes may make reproductions made pursuant to subsection (1) available to the following persons:

1. a specifically delimited circle of persons for their joint scientific research and

2. individual third persons for the purpose of monitoring the quality of the scientific research.

The making available to the public must be terminated as soon as the joint scientific research or the monitoring of the quality of the scientific research has been concluded,.

(5) Those authorised under subsections (2) and (3) no. 1 may retain reproductions made pursuant to subsection (1), thereby taking appropriate security measures to prevent unauthorised use, for as long as they are needed for the purposes of the scientific research or the monitoring of the quality of the scientific findings.

(6) Rightholders are authorised to take necessary measures to prevent the security and integrity of their networks and databases being put at risk on account of reproductions made in accordance with subsection (1).".... (Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG), 2022)

To validate the legality of the work being conducted under the copyright protection law, it is paramount to identify whether the data being analysed is protected by copyright at all or not. For the research being conducted, the data corpus (ESG reports) for the computational analysis would be gathered from publicly available sources in portable document format (PDFs). The nature and purpose of this kind of report is to inform the stakeholders, potential investors or curious beings about the operations and initiatives performed during the previous fiscal year as well as forward-looking statements and future goals. Furthermore, such documents also contain information that is mandatory to be disclosed under various regulations from governing authorities. According to section 5 (§5 UrhG), the information made available under statutory compliance is not considered copyrighted work and hence could be reproduced without any repercussions. In the case of the ESG reports, it is interpreted as the parts or sections of the report that are made available to secure statutory compliance do not enjoy copyright. Although it is hard to determine whether the whole document should be considered copyrighted material or not, as the reports include a variety of elements which could be protected and mentioned explicitly in machine-readable format; reproduction of such reports would violate §44b UrhG.

Assuming the material being reproduced is copyrighted, moving on to the next stage of the decision-making process, §60d UrhG is considered. §60d UrhG applies only to non-commercial scientific research, and this thesis research can invoke the clause if certain conditions are met. As this thesis is not being commissioned under any research funds and not geared toward commercializing the findings, it qualifies for using material for text and data mining purposes under §60d UrhG.

§60d UrhG enables the replication of the copyrighted material in order to

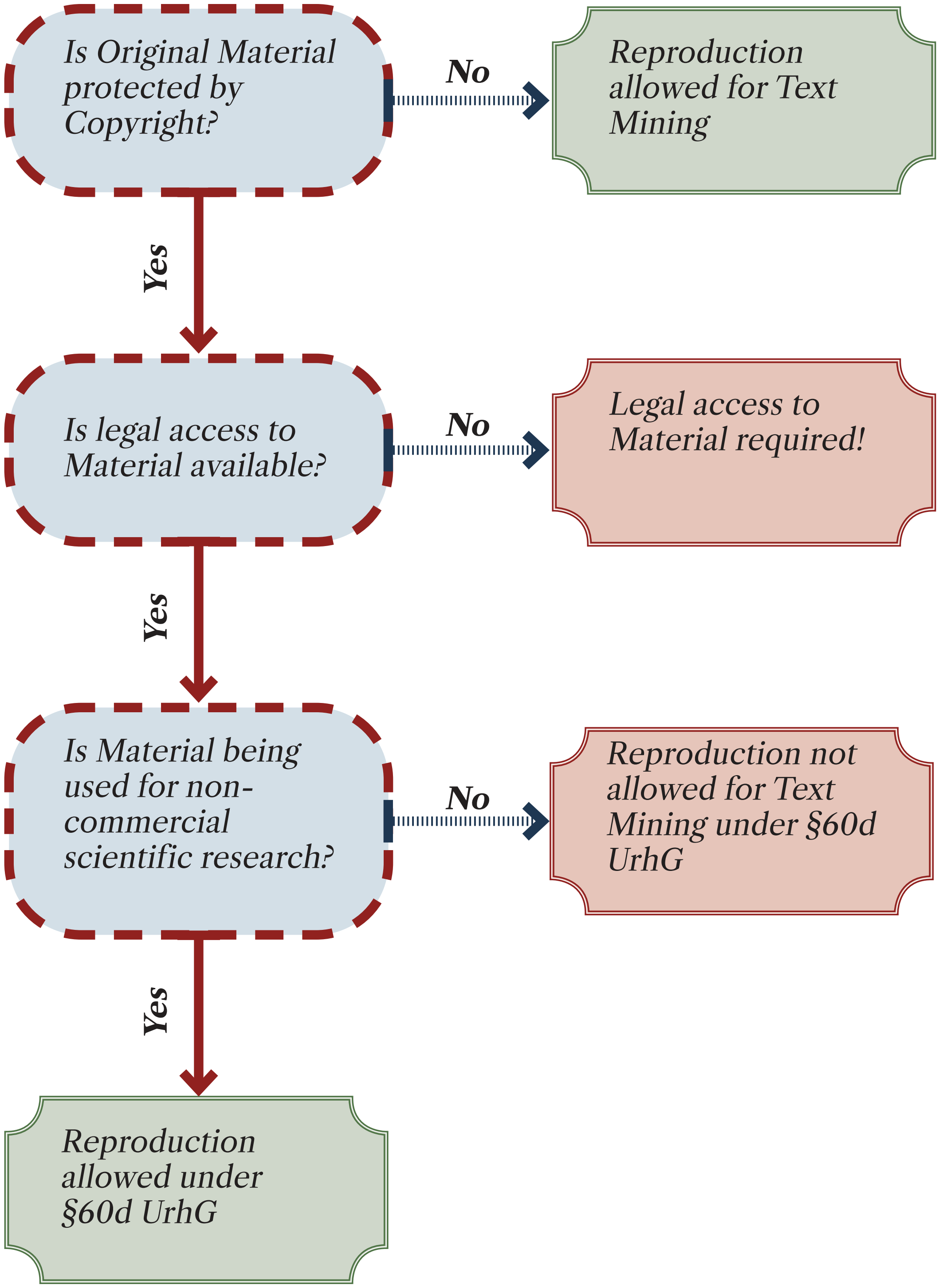

Is Original Material protected by Copyright?
No
Reproduction allowed for Text Mining
Yes
Is legal access to Material available?
No
Legal access to Material required!
Yes
Is Material being used for non-commercial scientific research?
No
Reproduction not allowed for Text Mining under §60d UrhG
Yes
Reproduction allowed under §60d UrhG

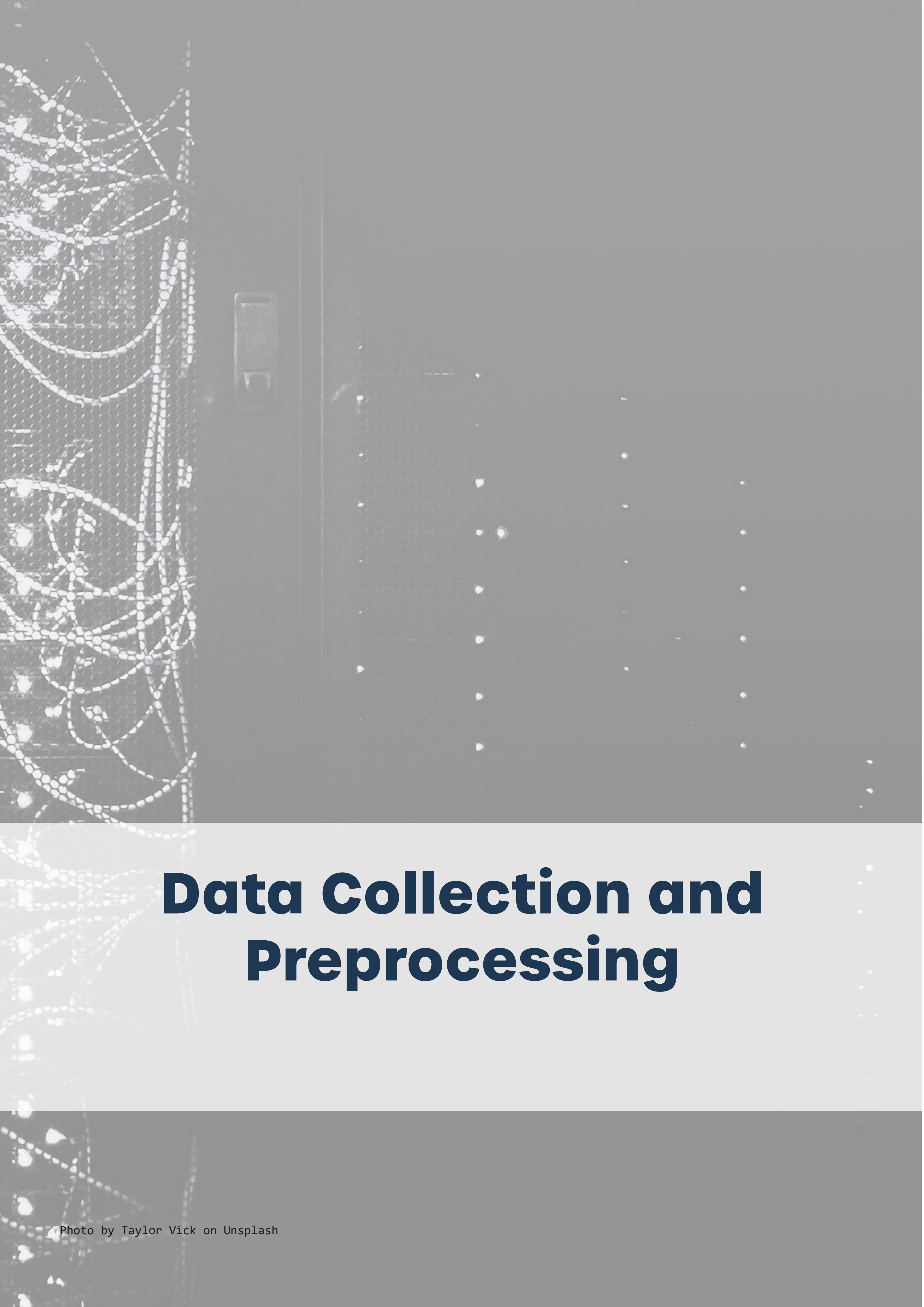
Figure 3 : Decision tree for viability of material reproduction under copyright law (UrhG)

create a corpus (database) through normalisation, structuring, categorisation or other processing methods. These processing methods also include the conversion of original material into another format as part of text and data mining. It should be noted that §60d UrhG does not grant a right to access the material to be analysed or included in the corpus, but rather assumes that the material is being accessed and reproduced legally. For this thesis, the corpus contains material from ESG reports that are made publicly available for anyone to access; either through the official website, regulatory filings or specialised databases. It is generally legal to access ESG reports that are made publicly available unless explicitly stated otherwise (see Legal Disclaimer).

Although under the §60d UrhG it is viable to utilise the selective ESG reports for text and data mining purposes provided following prerequisites are met:

» The created corpus may be shared electronically with a definable group of people, however, it is not permitted to publish the original material utilised for the corpus on own website(s) or similar services.

» Upon completion of the thesis project, the original material must be deleted along with the corpus directories. However, for the purpose of ensuring the reference and subsequent verifiability of the results, the body and excerpts of the original material may be transmitted to a library, archive, or other publicly accessible educational institution. (Brettschneider, 2021)



Data Collection and Preprocessing

Photo by Taylor Vick on Unsplash

3. Data Collection and Preprocessing

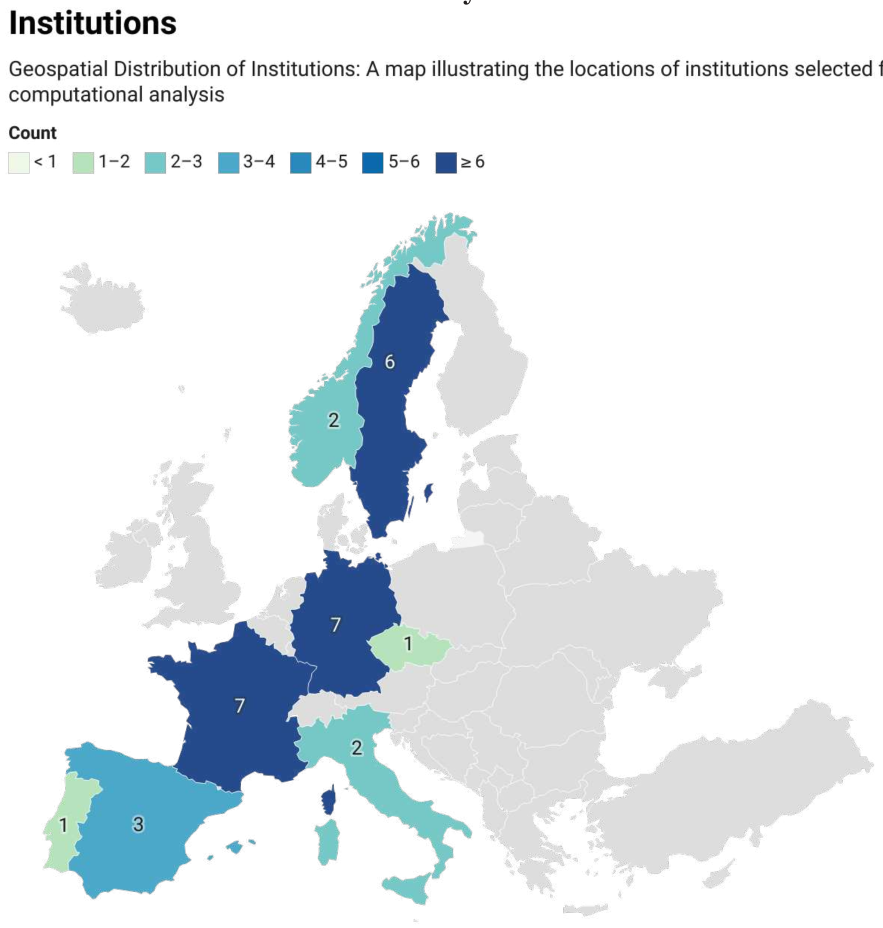
The first stage in any artificial intelligence based knowledge extraction system is to acquire the data from which the algorithm pipeline can be trained to extract the data. For this thesis research, non-financial activity reports published by the institutes serve as a database. Essentially classified as ESG reports, these documents are published under a variety of titles including but not limited to sustainability report, non-financial annual report, corporate social responsibility report, impact report, etc. The institutions typically provide the ESG reports in PDF format. This format is widely regarded as an optimal digital substitute for traditional paper-based documents due to its seamless compatibility across various devices and operating systems. One of the primary advantages of PDFs is their portability, platform independence, and human-readable nature. However, it lacks structure, posing diverse challenges for in-depth analysis and machine readability.

As the scope of the thesis is limited to non-financial institutions head quartered in the EU, a rather rudimentary manual data collection approach is chosen instead of employing more advanced automated web crawling and web scraping. A primary reason to collect and reproduce local copies of the report in the pipeline is to avoid discrepancies in the database as well as unintentional violations of the statutory regulations. Several open-source websites provide hosting services for ESG reports, yet it's worth noting that they may operate under different statutory regulations than those in Germany. Therefore, maintaining local copies manually instead of utilising web-scrapping algorithms ensures compliance with regulations and mitigates the risk of inadvertently breaching legal requirements. Additionally, web-crawling activities involve accessing and extracting data from diverse sources, which has a higher potential of inconsistencies or discrepancies in the formulation of the database if not managed carefully. A manual approach to reproduce local copies of the report can ensure data integrity, and accuracy while adhering to the applicable statutory regulations.

3.1 Data Collection

3.1.2 Data Directory Structuring

A well-formulated directory structure enhances the integrity and reliability of computational-based analysis. As the reports are being sorted and stored manually, it decreases the chances of data duplication and misclassification. Additionally, it ensures the accuracy and trustworthiness of the machine-learning process undertaken during the document preprocessing stage. The selection of the institutions is a crucial stage in ensuring the validity and reliability of the research findings. Since the primary objective of the research is to investigate the environmental-related disclosures in non-financial institutions located within the EU, the selection of institutions is based on the dataset obtained from Statista to avoid selective biases (Statista, 2023). The dataset comprises a curated list of top financial performers in terms of revenue generation from the EU region. The dataset obtained from Statista classifies the institutions into three sectors namely: Construction, Real Estate and Utilities (Statista, 2023). Cumulatively, a total of 36 institutions are selected which match the criteria for both non-financial activities and EU presence. In the manual iteration, seven of the initially selected 36 institutions were eliminated due to the unavailability of ESG-related disclosures in report format. ESG reports and disclosure documents from the last three fiscal years (2020, 2021 and 2022) for the remaining 29 institutions are manually downloaded from verified sources and organised in a local directory. In total 87 reports are included in data directory.



Institutions

Geospatial Distribution of Institutions: A map illustrating the locations of institutions selected for computational analysis

This map has been created using Datawrapper

Figure 4 : Geospatial Distribution of selected Institutions

Figure 4 illustrates geospatial locations of the headquarters from 29 institutions whose ESG data is being analysed. These 29 institutions are further classified into three sectors, geospatial distribution of which is shown in *Figure 5*.

Construction Sector

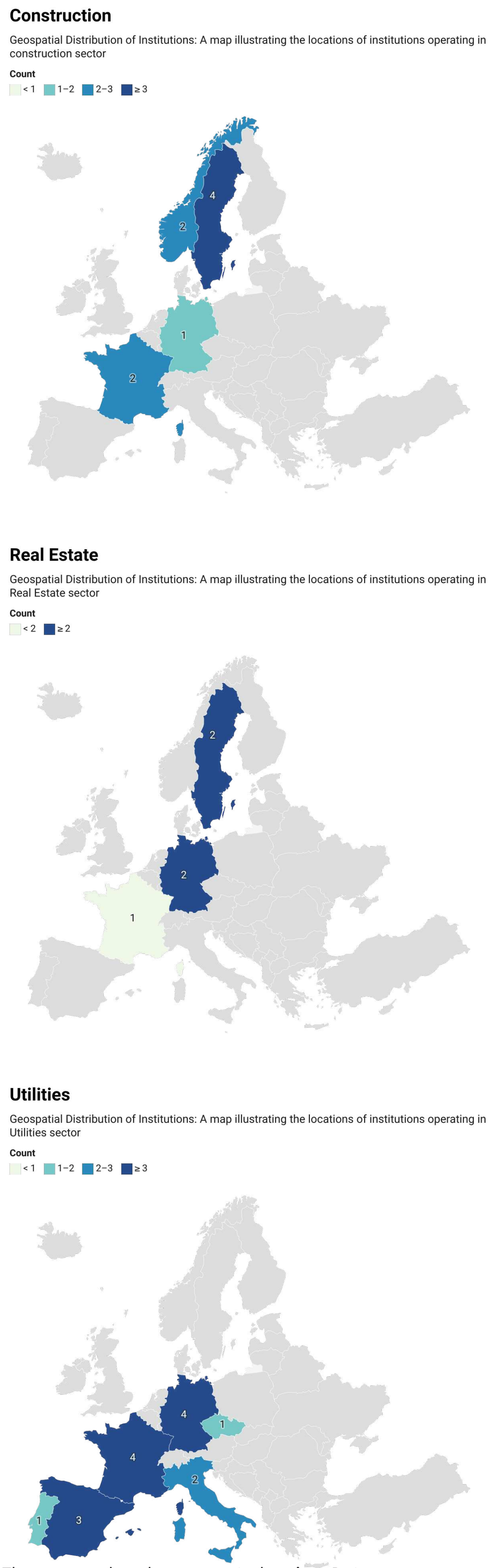
The set of institutions classified as the construction sector encompasses firms engaged in various facets of the built environment such as infrastructure development, and engineering services. In the context of the analysis and environmental impact, institutions operating in this sector hold particular significance due to their substantial environmental impact. From material sourcing to project completion and maintenance, the construction sector typically has a large footprint in resource consumption, energy usage and waste generation.

Real Estate Sector

The real estate sector comprises a diverse array of institutions involved in property development, management, investment, and brokerage services. These institutions are pivotal in shaping urban landscapes, providing residential, commercial, and industrial spaces. Throughout the lifecycle of real estate assets, from construction and operation to renovation and decommissioning, these institutions influence various environmental aspects such as resource consumption, energy efficiency, and greenhouse gas emissions.

Utilities Sector

The set of institutions classified under the utilities sector comprises a wide range of service providers involved in the generation, transmission, distribution, and provision of essential services such as electricity, water, and natural gas. Utility firms have a substantial environmental impact and are major contributors to greenhouse gas emissions and resource consumption. Within the EU, utility providers are also major players in directing the transition to a low-carbon economy



Construction

Geospatial Distribution of Institutions: A map illustrating the locations of institutions operating in construction sector

Real Estate

Geospatial Distribution of Institutions: A map illustrating the locations of institutions operating in Real Estate sector

Utilities

Geospatial Distribution of Institutions: A map illustrating the locations of institutions operating in Utilities sector

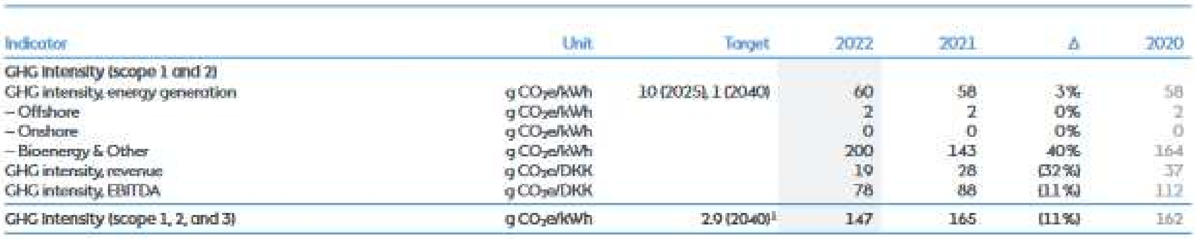
These maps has been created using Datawrapper

Figure 5 : Sector-wise Geospatial Distribution of selected Institutions

3.1.2 Data Characteristics

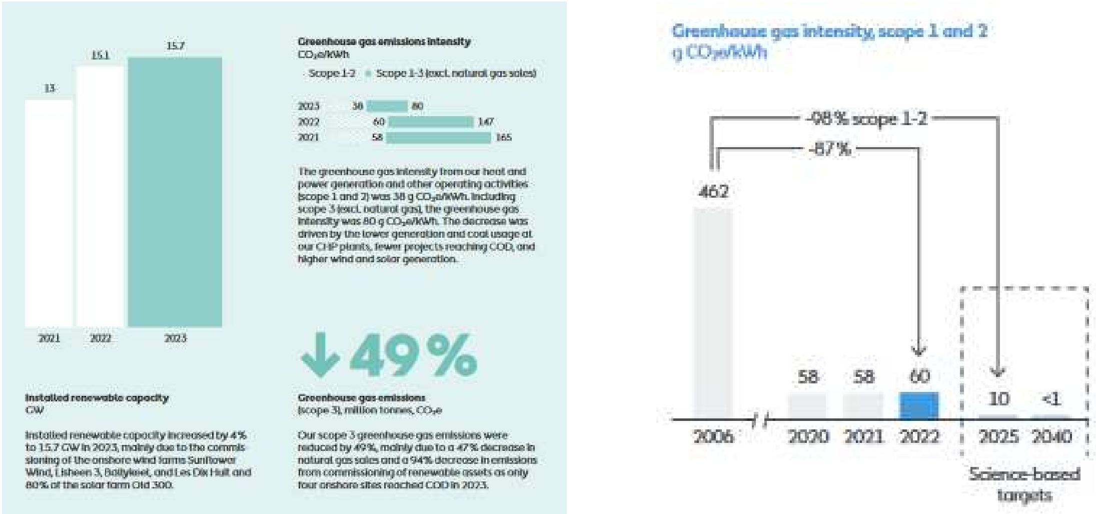
As stated earlier, the ESG reports are published in PDF format and targeted towards a broad audience including shareholders, stakeholders, potential investors, government entities and society. Although PDF format allows for graphically pleasing visualisations, it hurdles machine readability and raises challenges in employing computational based knowledge extraction techniques. This limitation hinders the ability to leverage advanced analytical techniques to extract meaningful information from the reports. Consequently, the lack of standardisation in reporting formats further complicates data aggregation and heightens the inability to incorporate meta-data.

Understanding the document structure and focusing on analytically important sections of the report is crucial to increasing the accuracy of correct text annotation and text classification task performance. Typically environmental disclosures related to greenhouse gas emissions, energy efficiency, and water consumption are published in the form of a small text paragraph, table, info-graphic or combined with indicative visuals. Such examples are shown in *Figure 6* as snippets taken from a sample ESG report.



Indicator	Unit	Target	2022	2021	Δ	2020
GHG intensity (scope 1 and 2)						
GHG intensity, energy generation	g CO_2e/kWh	10 (2025), 1 (2040)	60	58	3%	58
– Offshore	g CO_2e/kWh		2	2	0%	2
– Onshore	g CO_2e/kWh		0	0	0%	0
– Bioenergy & Other	g CO_2e/kWh		200	143	40%	164
GHG intensity, revenue	g CO_2e/DKK		19	28	(32%)	37
GHG intensity, EBITDA	g CO_2e/DKK		78	88	(11%)	112
GHG intensity (scope 1, 2, and 3)	g CO_2e/kWh	2.9 (2040)[1]	147	165	(11%)	162

GHG data shown in Tabular format



The greenhouse gas intensity from our heat and power generation and other operating activities (scope 1 and 2) was 38 g CO_2e/kWh. Including scope 3 (excl. natural gas), the greenhouse gas intensity was 80 g CO_2e/kWh. The decrease was driven by the lower generation and coal usage at our CHP plants, fewer projects reaching COD, and higher wind and solar generation.

Installed renewable capacity increased by 4% to 15.7 GW in 2023, mainly due to the commissioning of the onshore wind farms Sunflower Wind, Lisheen 3, Ballykeel, and Les Dix Huit and 80% of the solar farm Old 300.

Our scope 3 greenhouse gas emissions were reduced by 49%, mainly due to a 47% decrease in natural gas sales and a 94% decrease in emissions from commissioning of renewable assets as only four onshore sites reached COD in 2023.

GHG data illustrated in Info-graphic & in combination with text paragraphs

Figure 6 : Snippets from sample ESG report (adapted from KPMG (2022))

3.2 Document Preprocessing

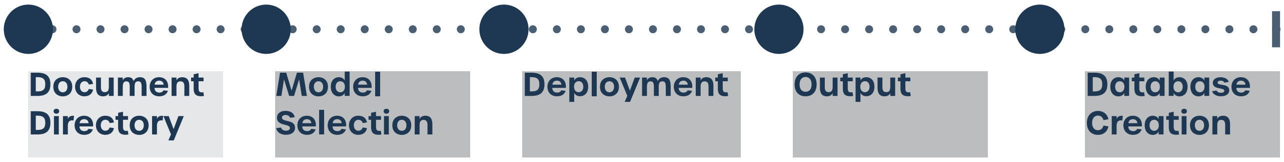


Figure 7 : Document preprocessing pipeline (author)

Preliminary investigations into ESG report documentation indicated that the data crucial for computational analysis is commonly presented in structured formats such as tables or textual content embedded within infographics. Following this observation, the implementation of document layout analysis techniques becomes imperative for the extraction and processing of the data and information from the unstructured ESG reports. Thus application of computational algorithms to analyse the visual layout and structure of documents is utilised to enable the identification and segmentation of different components within the PDFs such as text blocks, tables, figures, and captions. A data pre-processing pipeline (*Figure* 7) is created to identify the various elements from the document and process the text blocks and tables into a more manageable JSON (JavaScript Object Notation) array. Following sections describes the pipeline and computational concepts utilised during each stages.

3.2.1 Model Selection

In recent years, number of document layout analysis algorithms have emerged. Essentially all of these algorithms follow a basic concept denoted as computer vision. Computer vision refers to the interdisciplinary field of artificial intelligence and computer science concerned with enabling programming languages to interpret and understand visual information from the real world. Text mining techniques enable the processing of unstructured textual input; however, they often overlook the visual cues that humans rely on for correct document comprehension. Computer vision techniques, specifically document layout aware analysis, enables identification and segmentation of different components such as text blocks, tables, figures, captions, etc.

Selecting a computer vision based model or toolkit is an iterative process, for the use case undertaken LayoutParser toolkit is chosen as it is light-weight and features wide range of customisation in implementation.

LayoutParser

LayoutParser is a Python-based toolkit that utilises deep learning (DL) models to identify various elements in a document and perform wide array of document image analysis (DIA) tasks including layout detection, table detection and scene text detection (Shen et al., 2021). In essence, deep learning is a subset of machine learning (ML) that focuses on developing algorithms inspired by the structure and function of the human brain's neural networks. At its core, deep learning seeks to enable programming languages to learn from data representations in a hierarchical manner, allowing the models to automatically discover patterns and features within the data. This is achieved through the use of deep neural networks which are composed of multiple layers of interconnected artificial neurons, or units. To understand the basic components utilised in the LayoutParser toolkit, a basic understanding of artificial neural networks (ANN) and Mask Region-based Convolutional Neural Network (R-CNN) is paramount.

Artificial neural networks (ANN) are statistical models inspired by biological neural networks in the human brain. In layman's terms, biological neural networks process the input signals such as breathing, drinking and eating and process the subsequent instructions given to our body. Similarly, ANNs are statistical models that are capable of modelling and processing complex, non-linear inputs and provide output actions based on the given inputs. At their core, ANNs consist of interconnected nodes, or neurons, organised into layers. These neurons process and transmit inputs through a network of weighted connections, with each connection representing the strength of influence between neurons. ANNs are designed to learn from data through a process called training, where the network adjusts its internal parameters or weights, in response to input-output pairs provided during the training processes. Since ANNs are capable of learning complex patterns and relationships within data, they can perform complex tasks such as classification, regression, and pattern recognition. In LayoutParser library, an advanced ANNs system architecture is employed to create a multi-stage pipeline for object detection, known as a Region-based Convolutional Neural Network (R-CNN). A simplified model architecture of R-CNN is shown in *Figure 9*, in essence, R-CNN models identify regions within the input nodes (images, documents) that are likely to contain objects of interest. Each proposed region is then cropped and warped to a fixed size, forming a region of interest (ROI). Once the ROI is formed, the R-CNN framework utilises a neural network to extract features from each ROI. These features capture the visual characteristics of the objects within the proposed regions and are fed into a classifier. These features capture the visual characteristics of the objects within the proposed regions and are fed into a classifier with metadata including a bounding box and similarity to trained dataset.

The LayoutParser library offers various models pre-trained on an extensive set of image-based documents as well as textual data with the objective of transforming the unstructured dataset into a structured database. It also supports exporting the layout details and associated metadata into various formats like JSON and CSV. A sample layout analysis and interpretation by LayoutParser algorithm is shown in *Figure 10* with its bounding boxes.

After a few iterations through a trial-and-error approach, pre-trained models PubLayNet is selected for extraction of the textual elements and YOLOx is selected for extraction of tabular data. (Shen et al., 2021).

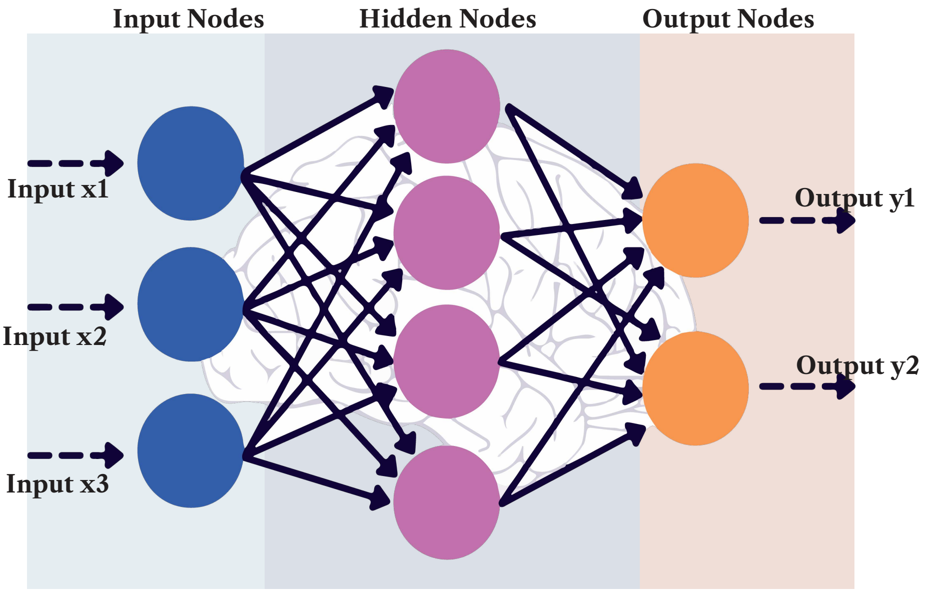


Figure 8 : Architecture of Artificial Neural Network (ANNs) (adapted from Blanchy et al., (2023)

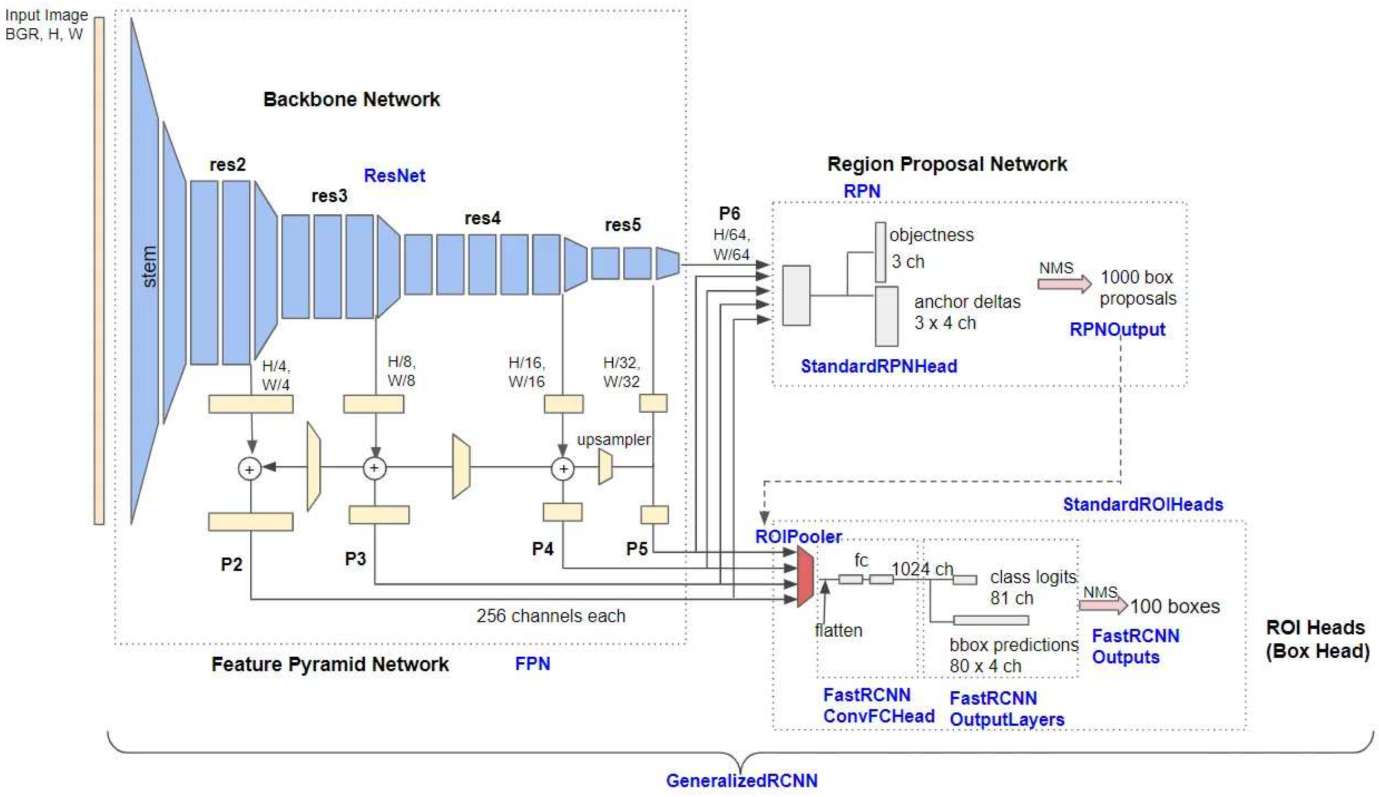


Figure 9 : Detailed architecture of R-CNN (adapted from Pham et al., (2020))

The blue labels represent class names. The process represented under Backbone Network extracts feature maps from the input image which are then fed into regional proposal network (RPN). RPN detects object regions and 1000 box proposals with respective confidence scores are obtained. The confidence scores are then fed into ROI pipeline together with feature maps.

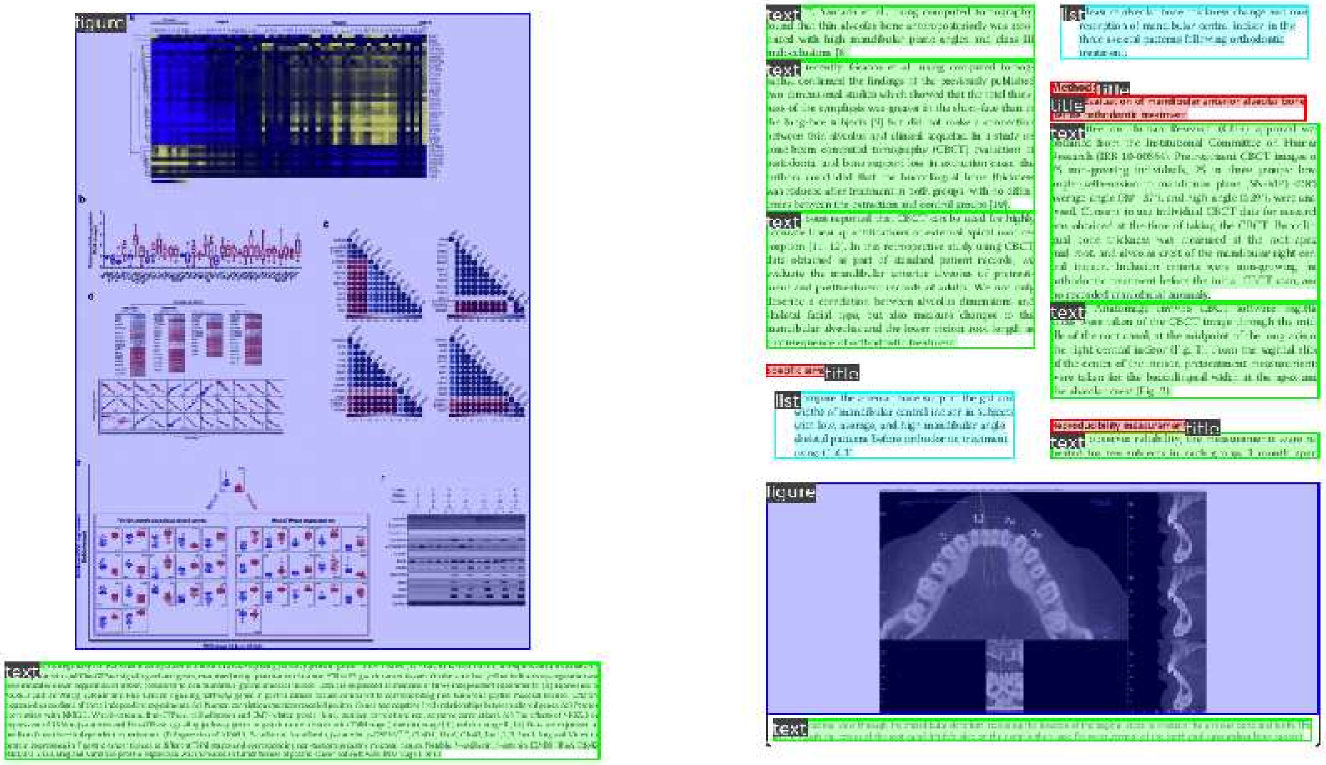


Figure 10 : Sample annotations using LayoutParser library (adapted from Shen et al. (2021))

3.2.2 Deployment

Document Directory

Model Selection

Deployment

Output

Database Creation

Python Libraries for Deployment

opencv-python==4.9.0.80
tqdm==4.66.1
detectron2==0.5
pycocotools==2.0.7
pillow==10.2.0
omegaconf==2.3.0
pyyaml==6.0.1
fvcore==0.1.5.post20221221
tabulate==0.9.0
torchvision==0.16.2

psutil==5.9.7
matplotlib==3.8.2
datasets==2.16.1
scipy==1.11.4
termcolor==2.4.0
cloudpickle==3.0.0
onnx==1.15.0
pandas==2.1.4
beautifulsoup4==4.12.2
unstructured==0.11.6

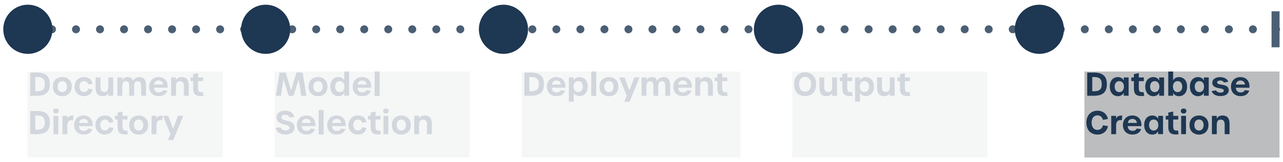
Once the deep learning models have been chosen to parse the documents, the deployment stage involves series of configuration to prepare the PDFs for element -based extraction. An open-source Python library ‘Unstructured‘ is employed in deployment phase.

Unstructured library offers capabilities in extracting elements and metadata from the document database into a more manageable JSON array format.

Furthermore, it supports the configuration of pre-trained models based on LayoutParser library as well as offers extensive configuration capabilities to combine multiple models for better data extraction (Unstructured Technologies, 2023). The initial attempts of data extraction through various LayoutParser models showed that rather than using one single model for the PDF partitioning and element extraction task, a combination of different models for text extraction and tabular extraction provides better results.

Unstructured library also supports data extraction in HTML format which is necessary for the information extraction tasks performed in the analysis. The partitioning functions in Unstructured library break a document down into elements such as ‘Title’, ‘NarrativeText’, ‘ListItem’, ‘Table’, ‘Image’, etc.; enabling customisation for further analysis. Output in JSON format is desired for this thesis research as it provides a higher degree of compatibility than textual file formats as well as custom configuration capabilities. A sample output in JSON array from a Python script based on an Unstructured library entails the structure described in *Table 1* including rich metadata.

3.2.3 Database Creation



A well-formulated data directory is essential for ease in machine readability and it also streamlines the NLP pipeline flow. JSON array provides a standardised and easily interpretable format for storing the data from documents, making it more manageable and accessible for subsequent stages of NLP pipelines. Moreover, JSON arrays incorporate additional layers of metadata, enriching the parsed data with contextual information extracted through computer vision.

Although the JSON array offers document partitioning in various elements, for the purpose of this thesis research, metadata extracted as “type” and “text“ are at utmost importance. The extracted text under the element types “NarrativeText“, ”Image”, “ListIteam“, “Title“ are further complied into txt files for text classification and other NLP based analyses while the data under element type “Table“ is extracted into HTML file through compiling “text_as_HTML“ array.

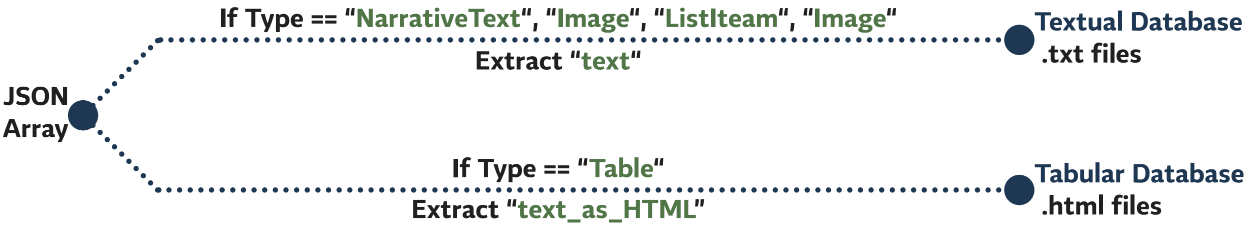


Figure 11 : Database development flow

Table 1: JSON array characteristics

Field Name	Format	Description	Example
element_id	String	A unique identifier for the element	1871b254b4461baa9d9a41930b1874a2
metadata	Hybrid	A list of string and numericals divided in sub-arrays comprising various information associated with element such as coordinates, layout, class probability, filename, file type, languages, page number.	"coordinates": { "layout_height": 2339, "layout_width": 1654, "points": [[47.0, 514.7], [47.0, 1064.9], [1656.9, 1064.9], [1656.9, 514.7]], "system": "PixelSpace" }, "detection_class_prob": 0.54058, "file_directory": "C:/../../Dataset", "filename": "Institute.pdf", "filetype": "application/pdf", "languages": ["eng"], "last_modified": "YYYY-MM-DD T15:41:42", "page_number": 1
text	string	Parsed and extracted text from the element under lens. This field contains the data which is sorted to create '.txt' file.	"text" : The number of housing units sold to consumers was in line with the preceding year: 1,263 (1,273). The sales rate for ongoing pro- duction was 57 per cent (62). The number of housing units sold to investors was 146 (332), with the transactions being conducted in Hamburg...
type	string	Identifier for type of the element. Tags include : • **ListItem** • **NarrativeText** • **Table** • **Title** • **Image** • **PageBreak** • **Header** • **Footer**	"type": "NarrativeText" "type": "Title"
text_as_HTML	HTML	Field unique to elements associated with type 'Table'. Includes the tabular structure in HTML which is sorted to '.html' file.	<table><thead><th>Net sales</th><th>7,276</th><th>7,466</th></thead><tr><td>perating profit</td><td>914</td><td>752</td></tr><tr><td>.....

Methodology

Photo by Florian Olivo on Unsplash

4. NLP Methodology

Humans can normally identify and comprehend the information presented in ESG reports easily, even if the information is presented in the form of an info-graphic or complex tables. We tend to attribute specific characteristics to particular keywords, people and base our ability to recognise and interpret patterns on our prior expectations, knowledge, and experiences. The way we convey information and present the content has evolved greatly in recent years thanks to developments in data visualisation tactics and computers. In the realm of ESG, an experienced person could effortlessly associate various information or text snippets from ESG reports with a particular aspect of ESG in mere seconds. However, at times it could be a daunting task as more and more data on ESG is being published in the form of blogs, news articles and reports. In contrast, computational programs can progress large magnitude of data efficiently. Still, it lacks the understanding and expertise to understand the content in the context of various ESG aspects. Thanks to the advancements in computational linguistics and ANNs, it is possible to define rules and conditions to extract desired outputs from large quantities of data. Without diving into too much detail about the background and inner workings of the computational algorithms, this chapter describes the analysis pipeline utilised to gain insights into the environmental disclosures from collected reports. The first section of the chapter holistically defines and describes the important terms that come up while working with the analysis pipeline.

NLP is a sub-branch of artificial intelligence connected to linguistics, and computer science that leverage various technology to understand and interpret human languages. This branch covers a wide range of methodologies that are used to bridge the gap between human language and machine understanding for instance computer vision, machine translation, text processing and text generation. A drastic development has happened in recent years in the NLP domain thanks to advancements in machine learning algorithms, availability of large-scale datasets and improvements in computational power. In the context of scientific research, NLP can be leveraged in various ways to revolutionise traditional research methodologies by automating tedious tasks as well as extracting valuable insights from textual data.

From the initial experimentation, it is observed that to properly address the research problem at hand, separate evaluation and extraction of environmental disclosures from textual and tabular dataset performs better. Two separate pipelines are developed to assess these datasets and address challenges raised during analysis. Although currently available hybrid modules for NLP can integrate multiple modalities, a dedicated pipeline for each data type performs better and requires considerably lower computing power.

4.1 Textual Data Analysis

More than often, environmental-related indicators, vision and goals are conveyed in a text-based format such as paragraphs, statements or lists. Extracting actionable insights and classifying the text into suitable categories requires a structured NLP pipeline. Traditionally this pipeline covers a broad range of NLP tasks such as text preprocessing, tokenization, part-of-speech tagging and text classification. For the objective at hand, an NLP pipeline that can ultimately perform text classification is deemed necessary. Text classification or categorisation process assigns parts of texts to predefined categories. Simply put, a text classification task allocates a discrete label from a set of possible tags to a given text. To illustrate the concept further by a simple example, consider the importance of understanding sentiments on reviews for e-commerce platforms. It is crucial for businesses or service providers to classify customer feedback as positive, negative, or neutral to make informed decisions in product modifications. This task, known as sentiment analysis, falls under the umbrella of text classification in NLP, where the goal is to categorise the text into sentiments such as good, bad, or neutral, based on text from customer reviews. For the case study undertaken, the goal is to categorise the text into various environmental disclosures namely GHG emission, energy efficiency and renewables, and water consumption.

Figure 12 illustrates the stages of the text classifier NLP pipeline. Each stage is broken down and explained in detail in upcoming sections of this chapter.

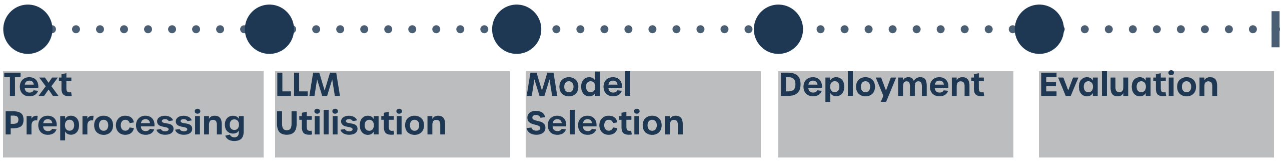


Figure 12 : NLP pipeline for textual data analysis

4.1.1 Text Preprocessing

Organised text data from the document preprocessing pipeline contains noise such as spelling mistakes, hyphens and white spaces. Although computer vision techniques are used during the document preprocessing stage to ensure accurate text extraction, they are also prone to including text kerning and white spaces (Shen et al., 2021). This additional noise into the dataset could raise challenges in obtaining meaningful insights. The text was cleaned up by removing special characters such as bullet points, line breaks and additional white spaces. For the subsequent stages of analysis, the text needed to be broken down in a way that the models utilised during the NLP pipeline could understand and process it. This text preprocessing step is called tokenization.

Text Tokenization

While humans can effortlessly comprehend the structure and semantics of the textual content, a machine perceives text merely as a sequence of characters or ‘tokens’. Thus for the machine readability of input textual data, a systematic breakdown of text segments into discrete linguistic units is required. In many instances, identifying tokens is relatively straightforward, particularly while working with a segmented language like English. In segmented languages words are separated by blank spaces, allowing tokens to be recognised as the segments between these spaces. However, the tokenization task becomes considerably more intricate when dealing with abbreviations, or numerical inputs. For instance, a full stop at the end of a sentence constitutes a single token, and in the case of a decimal number (10.25), it is treated as part of the token without distinct separation (Blanchy et al., 2023). Tokenizing ESG texts presents unique challenges as various units, symbols and extensive use of numerical. To address these complexities, tokenization algorithms often rely on predefined lexicons to compare text entries. Some algorithms possess the capability to analyse text both retrospectively and prospectively, applying multiple sets of rules simultaneously and selecting the most appropriate algorithmic approach to represent the text in tokenized form. These algorithms are based on a deep-learning model which possesses the ability to capture contextual information and dependencies between words and numbers more effectively than traditional tokenization methods. Unlike traditional tokenization techniques, which treat each word as an independent entity, deep-learning-based models leverage attention mechanisms to consider the context of each token within the entire text inputs. This enables it to generate more meaningful text representations, allowing for a more accurate and nuanced understanding of the input data. Additionally, these models can handle varying lengths of text inputs more efficiently, making them more suitable for processing varying text lengths. Deep-learning-based model, transformers is utilised for tokenization and further text classification tasks in the pipeline.

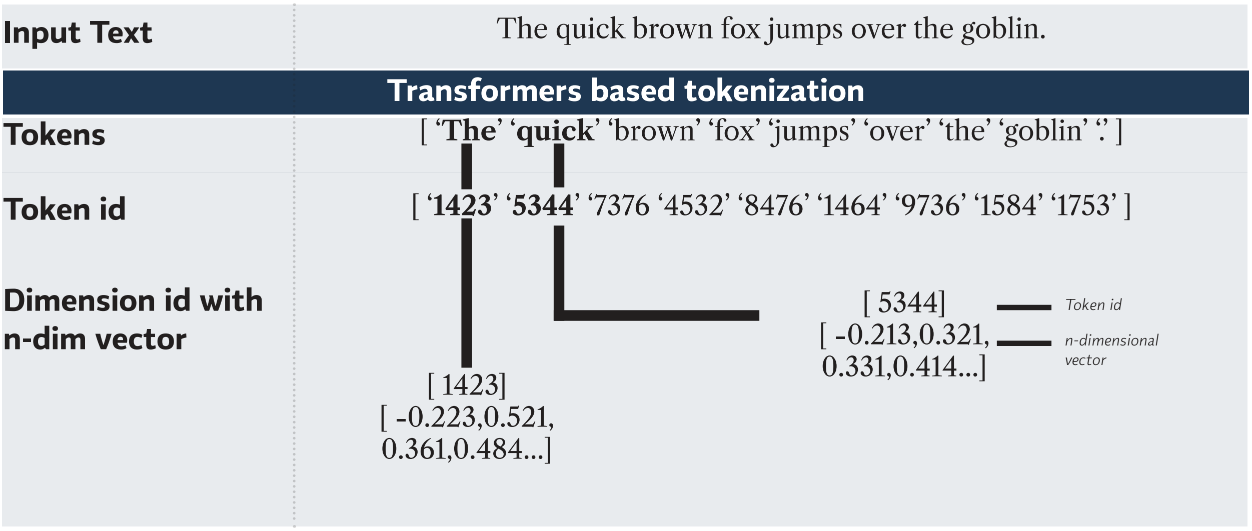


Figure 13 : Text tokenization through transformer based models

Transformer

Introduced by Vaswani et al. (2017) in "Attention is All You Need" research publication, transformer is widely used in the current NLP demographic. The sequential architecture of Transformer relies on self-attention mechanisms, which allow the model to weigh the importance of different words in a sentence while tokenizing. This attention mechanism enables the model to capture the contextual dependencies between the words more flexibly and reliably compared to traditional tokenization models. As illustrated in *Figure 14* transformer architecture is a very complex process structure. Only the essentials and basic mechanics are covered in this section to establish a fundamental understanding and relevance to this thesis research.

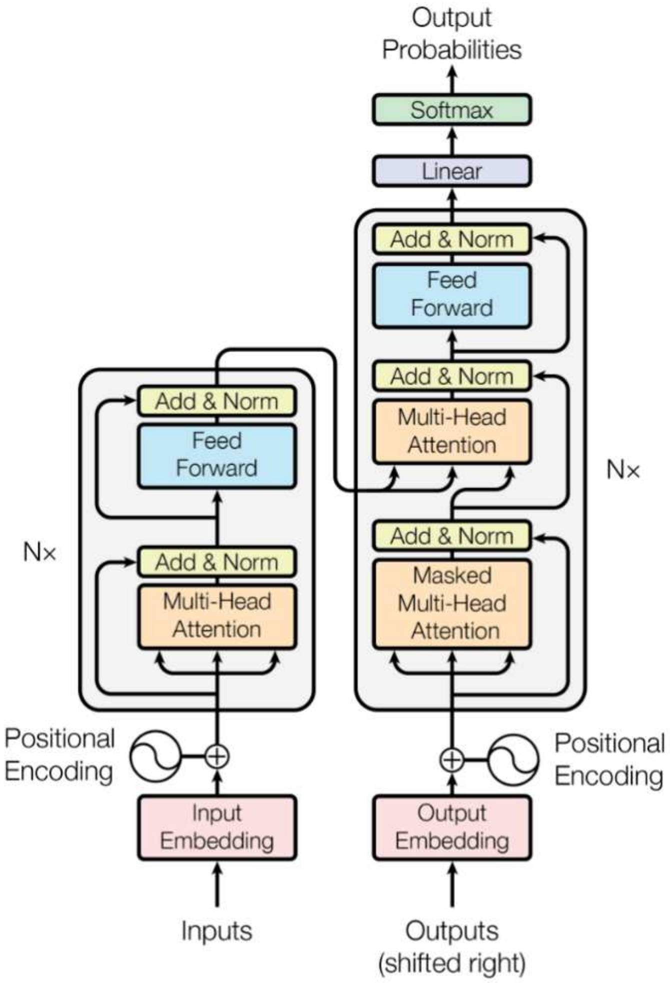


Figure 14 : Transformer model architecture (Vaswani et al, 2017)

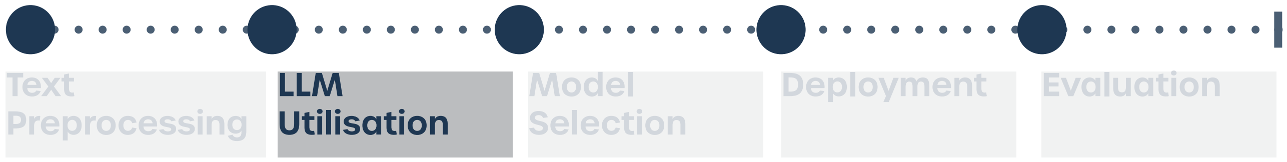
Simply put, the Transformer architecture consists of multiple layers of self attention mechanisms and feed-forward neural networks. Each of these layers performs two main operations: self-attention and feed-forward transformation.

In the self-attention step, the model computes attention scores for each word in the input sequence, determining how much focus should be given to each word when processing the entire sequence. These attention scores are computed based on the similarity between the word representations in the sequence. Once the attention scores are computed for each word, they are used to weight the word representations, allowing the model to generate context-aware representations for each word. These context-aware representations capture both the local and global context of each word in the input sequence.

After the self-attention step, the model applies a feed-forward neural network to each word representation independently. This feed-forward transformation further refines the representations, capturing higher-level features and interactions between words. By stacking multiple layers of self-attention and feed-forward transformations, the model can learn increasingly complex patterns and dependencies in the input data proving a highly efficient and scalable framework for processing text data.

As shown in *Figure 13*, the transformer based tokenization assigns dimensional vector id to each token processed from the input text. These vector ids are based on linguistic structure, contextual importance of the token within the processed text and self-attention.

4.1.2 Large Language Model



Language models are computational algorithms designed to understand and generate human language utilised in various downstream NLP tasks such as sentence predications, text classification, machine translation, text summarisation, and sentiment analysis. Classically the LMs were defined as the task of predicting subsequent words from the query text. The discovery of Transformer architecture led to a revolutionary change in the field of Language Modelling as it enables machines to understand and integrate the contextual meaning of the text inputs. In comparison to traditional statistical and neural language models, the stake architecture of the Transformer can estimate conditional probabilities of the next words simultaneously using gradient-based supervised learning. This opened the door to even more complex and effective LMs and optimisation of the Large Language Model (LLM). The emergence of these transformative models redefined the benchmarks across various NLP tasks. Prominent examples of such models based on transformer architectures include the Bidirectional Encoder Representational Transformer (BERT), the Generative Pre-trained Transformer (GPT) series, MPNet, and the character-level language model T5. Unlike the earlier approaches that relied on training individual models on specific labelled datasets for particular tasks, these sophisticated models introduced a paradigm shift by leveraging self-supervised pre-raining on vast amounts of unlabelled text data. For instance, BERT's training corpus consists of the Book Corpus (800B words) and English Wikipedia (2500M words), while GPT-3 is trained on a staggering 500B words sourced from diverse datasets, including books and content from the internet. These models employ innovative techniques like masked language modelling (MLM), next-sentence prediction (NSP), and generative pre-training during their self-supervised learning phase. Importantly, these techniques eliminate the need for manual labelling of data to understand the language structure, as the models learn to predict the inherent structure and semantics of the text without explicit supervision. (Devlin et al., 2018)

Fine-Tuning

Pre-training and Fine-tuning are two crucial stages in the development and deployment of large language models in an NLP pipeline. These stages are fundamental in leveraging the power of deep-learning models to adapt them to specific tasks or domains. The primary objective of pre-training is to capture general language patterns and features from diverse text resources, enabling the model to acquire a broad understanding of language.

Fine-tuning occurs after pre-training and involves adapting the pre-trained model to specific downstream tasks or domains. During fine-tuning, the model is further trained on task-specific labelled data to enhance its performance and tailor it to the target application. The main research question for the thesis emphasises finding the environmental-related disclosures from ESG documentation, a model fine-tuned on considerably large ESG data is necessary to perform the text classification on the prepared datasets.

Hugging Face

Hugging Face is an open-source deep-learning platform that provides APIs and resources for accessing and fine-tuning state-of-the-art pre-trained models (Hugging Face, 2024). These models support a wide range of tasks across various domains, including text classification. Hugging Face provides a comprehensive repository of over 120,000 fine-tuned models, 20,000 datasets and 50,000 demo applications on their Model Hub. These resources are open-source and available to the public to utilise for a wide range of use cases including research. The Hugging Face library facilitates the utilisation of available models for downstream NLP applications, these resources can be accessed through online wrapper services or can be downloaded for local processing.

4.1.3 Model Selection

Fine-tuning a large language model from scratch is a computationally expensive and memory-intensive method. To address the research objective adequately, an open-source ESG domain fine-tune model is needed which could be processed on a local machine independently without any external dependencies. Furthermore, it should be fine-tuned with a considerably large annotated dataset including environmental disclosure labels. As a wide range of fine-tuned models are available on Hugging Face, it could be a daunting task to choose a valid fine-tuned model which qualifies through in required parameters to be integrated into an NLP pipeline. To choose an appropriate fine-tuned model various experimentation and iteration were deemed necessary. The list of models tested including their pros/cons and model description is jotted down in *Appendix 2*.

After careful consideration and weighing the benefits of tested models, ESGify was chosen to be used in the text classification task. ESGify is a fine-tuned model made available to perform multilabel text classification on ESG texts (si-ai-lab, 2023). It is able to classify the text input into 46 various ESG classes including the ones which are relevant to this thesis research: Greenhouse Gas Emissions, Water Consumption and Energy Efficiency and Renewables.

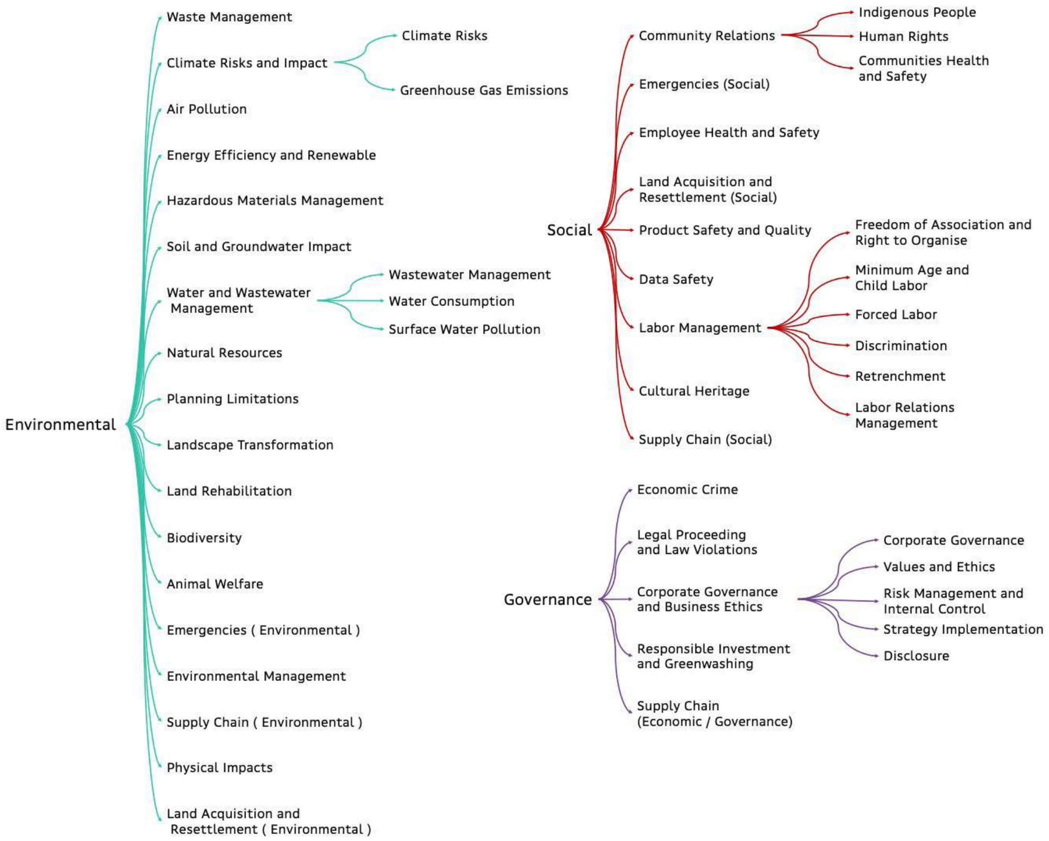

Environmental
Waste Management
Climate Risks and Impact
Climate Risks
Greenhouse Gas Emissions
Air Pollution
Energy Efficiency and Renewable
Hazardous Materials Management
Soil and Groundwater Impact
Water and Wastewater Management
Wastewater Management
Water Consumption
Surface Water Pollution
Natural Resources
Planning Limitations
Landscape Transformation
Land Rehabilitation
Biodiversity
Animal Welfare
Emergencies (Environmental)
Environmental Management
Supply Chain (Environmental)
Physical Impacts
Land Acquisition and Resettlement (Environmental)
Social
Community Relations
Indigenous People
Human Rights
Communities Health and Safety
Emergencies (Social)
Employee Health and Safety
Land Acquisition and Resettlement (Social)
Product Safety and Quality
Data Safety
Labor Management
Freedom of Association and Right to Organise
Minimum Age and Child Labor
Forced Labor
Discrimination
Retrenchment
Labor Relations Management
Cultural Heritage
Supply Chain (Social)
Governance
Economic Crime
Legal Proceeding and Law Violations
Corporate Governance and Business Ethics
Corporate Governance
Values and Ethics
Risk Management and Internal Control
Strategy Implementation
Disclosure
Responsible Investment and Greenwashing
Supply Chain (Economic / Governance)

Figure 15 : Text classification labels for ESGify Model (si-ai-lab, 2023)

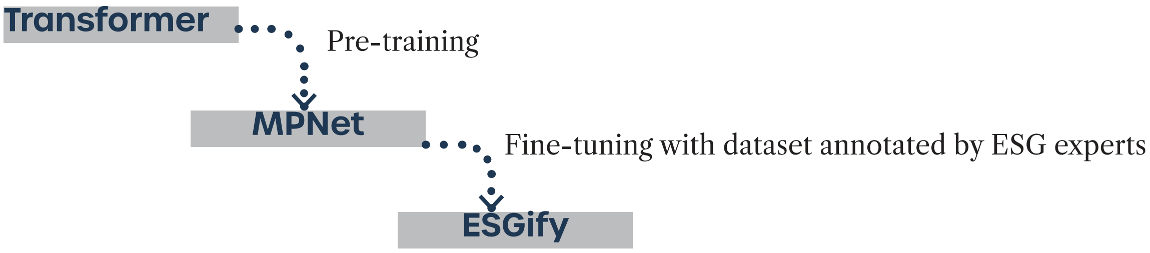
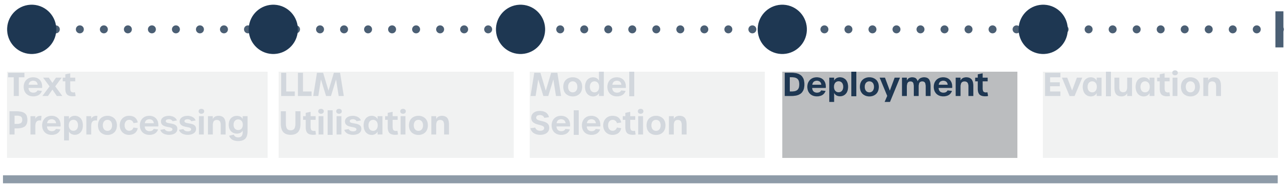


Figure 16 : ESGify model base and Fine-tuning

Integration of ESGify on a local machine is fairly simple as it is based on MPNet. MPNet is a lightweight large language model trained on 160 GB of text corpora and provides relatively better results in text classification tasks than other pre-trained models BERT and XLNet (Song et al., 2020). Before fine-tuning ESGify model was domain-adopted using Mask Language Modelling (MLM) from texts of ESG reports. Simply put, MLM randomly masks 10 to 30% of the words from the input text and then the pre-trained model (in this case MPNet) predicts the masked words. The precision-recall of the predictions determines the viability and accuracy of the model. Once adequate accuracy results from the MLM task were achieved (>80), the ESGify model was fine-tuned on approximately 2000 texts manually annotated by ESG specialists (sb-ai-lab, 2023). The manual annotation task is time-consuming and requires the utmost precision as it serves as a baseline for the evolvement of the model. Fine-tuning performed on a high-quality dataset was one of the major decision drivers in utilising ESGify as a text classifier.

4.1.4 Deployment



Python Libraries for Deployment

```
iopath==0.1.8
torch==2.1.2+cu118
portalocker==2.8.2
setuptools==68.2.2
transformers==4.36.2
nltk==3.8.1
flair==0.13.1
numpy==1.26.2
pytorch==4.12.2
pandas==2.1.4
transformers==2.1.4
scipy==1.11.4
matplotlib==3.8.2
fvcore==0.1.5.post20221221
psutil==5.9.7scipy==1.11.4
```

Once the model selection choice is concurred for the extraction of the desired output, it needs to be deployed to serve its intended purpose. This stage typically involves packaging the model along with any necessary preprocessing steps or dependencies into a working environment where it can receive input data, process it using the trained model, and produce the desired outcome. Before the prepared textual dataset is fed into the model for inference, a word masking step is performed to avoid biases in output.

Named-Entity Recognition (NER) masking

The base model (MPNet) used for fine-tuning to achieve desired text classification is trained on large corpora comprising text-based data sources such as books, articles, websites, social media posts, etc (Song et al., 2020). The diversity of the training dataset increases the chances of false associations between the words that are obvious indicators. To illustrate with an instance, consider a text classification model trained to classify discrimination in the working environment and the input text we wish to classify reads as “Sophie promotes diversity and inclusion in the workplace culture at ABC incorporation”. Although the true classification of the input text would be ‘Values and Ethics’, ample references in training data to the words ‘Sophie’ and ‘ABC incorporation’ could lead to unexpected associations in classification results. To minimise such ambiguity, false association and biases between environmental disclosures and institutions, named-entity masking is performed.

Named Entity Recognition (NER) is an NLP task aimed at identifying and classifying named entities with a body of text into predefined categories such as names of persons, institutions, locations, etc. Typically NER is performed to identify these entities and provide a meaningful output, but for the case study undertaken NER is employed to mask these entities and avoid inaccuracies. The *Figure 16* shows sample input text and masked version of the input text.

Input Text	“**John Smith**, the CEO of **Grn Innovations** based in **San Francisco**, pledges to enhance the company’s environmental, social, and governance practices to drive positive change within the local community and beyond.”
	Flair based NER Masking
Masked Text	“<**Person**>, the CEO of <**Organisation**> based in <**Location**>, pledges to enhance the company’s environmental, social, and governance practices to drive positive change within the local community and beyond.”

Figure 16 : NER masking

Pipeline Architecture

The computational architecture of the textual NLP pipeline, as depicted in *Figure 17*, illustrates the flow of information and iteration stages.

Initially, a textual database containing .txt files is accessed. These files undergo preprocessing via NER libraries to create masks specifically identifying entities such as institutions and individuals. Subsequently, a method known as large language-based tokenization is applied to segment the text into individual sentences, facilitating subsequent text classification tasks. These sentences are then subjected to the ESGify model, which employs classification techniques to identify and extract text segments pertaining to predetermined labels. The model exhibits the capability to classify text segments into 47 distinct categories. However, to mitigate computational resource consumption, the focus is directed towards three specific labels that align closely with the research objectives. These labels encompass Greenhouse Gas Emissions, Energy Efficiency and Renewables, and Water Consumption.

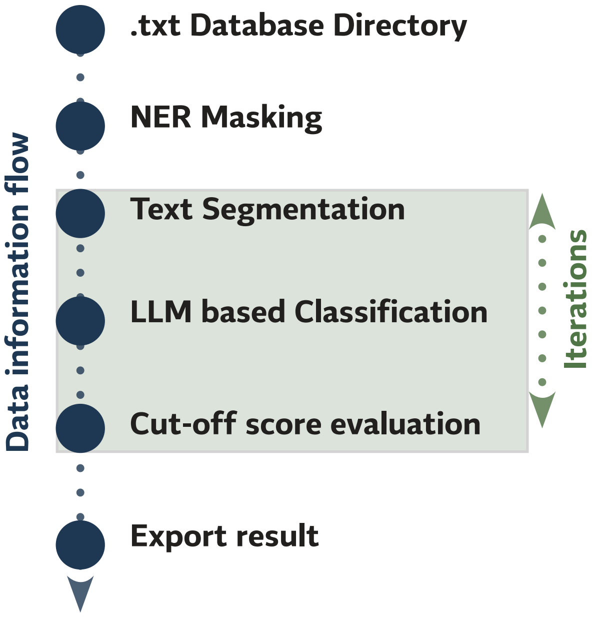
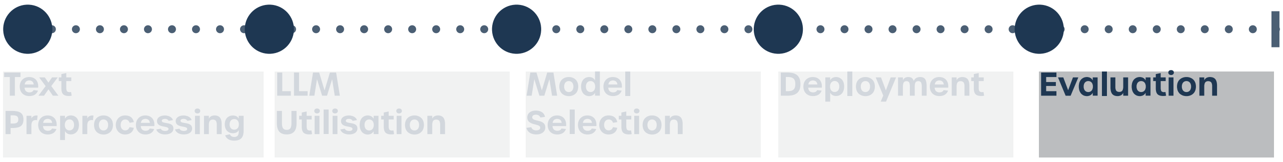


Figure 17 : NLP pipeline flow

4.1.5 Evaluation



Evaluation of the obtained text classification labels involves an iterative process that requires manual oversight. The model analyses the content of each sentence and assigns a score ranging from 0 to 1. This scoring mechanism allows for the assessment of the relevance of each sentence to the designated labels. After manually analysing the quality of the labels assigned by the model, a decision is made to establish a cut-off threshold at a score of 0.85. This cut-off point indicates that the text classification label is considered accurate to an extent of 85%, signifying that the labelled text segments with scores equal to or higher than 0.85 are deemed correct and highly relevant to the designated classification labels. Conversely, text segments with scores below this threshold are somewhat relevant to the classification labels and are avoided from consideration to decrease ambiguity in result interpretation.

Following the determination of relevant text segments based on the established cut-off threshold, the output data is structured into data frames for further analysis. These data frames organise the labelled text segments along with their respective scores, facilitating a comprehensive overview of the classification results. Subsequently, the structured data is exported into an Excel file format. In the subsequent chapter 5, a detailed discussion is made regarding the outcomes derived from the textual NLP pipeline. This discussion delves into the implications of the classification results, highlighting key insights and significance within the context of the research objectives.

4.2 Tabular Data Analysis

Quantitative environmental disclosures are often published in tabular format, as it is easier to understand and relate the disclosures to its scope, measurement unit and activities. These tables encountered in ESG reports are not structured and do not follow the traditional tabular structure found in typical data frames. Instead, the tabular data presented in ESG reports exhibit non-standardised formats and greater complexity, posing numerous challenges in their interpretation through computational methods.

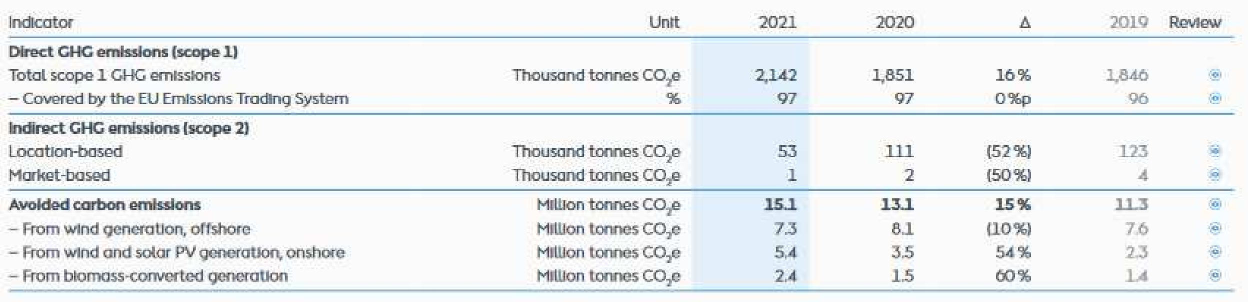
4.2.1 Challenges

In a traditional tabular structure, data is organised into rows and columns, where each row represents an observation or data point, and each column represents a variable or attribute. Additionally, the data within each column is typically of the same data type (e.g. integer, float, string) and the first row usually contains column names or labels, providing context to the data stored in each column. This structure allows for easy indexing and slicing of data; simplifying the process of computational algorithms.

In contrast, the tabular structure used in ESG reports exhibits a mix of unstructured and nested data frames lacking uniformity. Furthermore, these tables often contain multidimensional data representing diverse metrics across different time periods, geographical regions and activity units. As a result, the tabular structure in ESG reports features hierarchical arrangements, nested columns, merged cells, etc. to represent diverse dimensions of the disclosure metrics. While current computational algorithms can parse the boundaries of tables accurately, parsing complex tabular structures through various machine learning and deep learning algorithms is a challenging task for multiple reasons.

Since the ESG reports are created to be visually appealing, the cell structure of the tables is not always marked with clear boundaries hindering the machine readability. The absence of boundaries also makes it difficult to parse the data organised in multiple levels of categories or subcategories. Consider the tabular representation illustrated in *Figure 18* delineating GHG emissions quantities. This table encompasses data related to Scope 1 and 2 emissions, in addition to other GHG emission metrics. While the hierarchical organisation and categorisation of information are readily discernible to human observers, the absence of clear cell boundaries and the multilevel information structure pose daunting challenges for computational algorithms tasked with interpreting and extracting such data. The intricate nesting of categories and the lack of clear boundaries between data elements hinder the parsing of information by AI algorithms. Despite the recent strides in machine learning and deep learning techniques, algorithms for the seamless extraction of such information preserving its original structure are still evolving.

Another primary challenge lies in the inconsistencies across institutions in



Indicator	Unit	2021	2020	Δ	2019	Review
Direct GHG emissions (scope 1)						
Total scope 1 GHG emissions	Thousand tonnes CO_2e	2,142	1,851	16 %	1,846	◉
– Covered by the EU Emissions Trading System	%	97	97	0 %p	96	◉
Indirect GHG emissions (scope 2)						
Location-based	Thousand tonnes CO_2e	53	111	(52 %)	123	◉
Market-based	Thousand tonnes CO_2e	1	2	(50 %)	4	◉
Avoided carbon emissions	Million tonnes CO_2e	**15.1**	**13.1**	**15 %**	11.3	◉
– From wind generation, offshore	Million tonnes CO_2e	7.3	8.1	(10 %)	7.6	◉
– From wind and solar PV generation, onshore	Million tonnes CO_2e	5.4	3.5	54 %	2.3	◉
– From biomass-converted generation	Million tonnes CO_2e	2.4	1.5	60 %	1.4	◉

Figure 18 : Sample GHG emission disclosure table (KPMG (2022))

terms of table formatting and data presentation. Each institute may adopt its unique schema to represent the data resulting in the absence of a standardised format for training deep learning algorithms. The non uniformity in the layout and structure of tables makes it difficult to devise a one-model-fits-all solution. To develop such a model, the algorithms must contend with varying levels of complexity, nested hierarchies and irregularities in cell organisation. An NLP model addressing these challenges requires a nuanced approach that incorporates flexibility, adaptability, and continuous refinement to navigate the complexities inherent in ESG reporting.

Considering the intricate computational requirements and limitations, after a few iterations of experimentation (Appendix 2), a simplified pattern-matching pipeline is proposed to retrieve information about environmental disclosures.

4.2.2 Tabular Analysis Pipeline

A computer vision-powered document preprocessing is used to identify and extract the tables from the ESG reports into HTML (Hypertext Markup Language) format. As previously mentioned, this extraction process is intricate and necessitates manual validation to assess the inaccuracies in the table's data structure. Although computer vision techniques generally produce precise outcomes when extracting text or numerical data from tables, discrepancies are noted in cell formatting and column headings, among other aspects.

As the research objective revolves around automation in environmental disclosure extraction, a manual keyword-based pattern-matching algorithm is implemented to address the objective. The step-by-step processing of the tabular data directory through algorithm is illustrated in *Figure 19*.

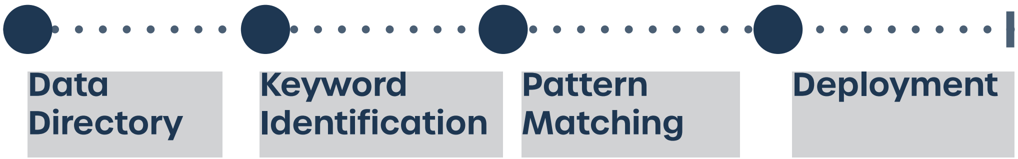


Figure 19 : Tabular Analysis Pipeline

Dataset Overview

As discussed in the section 3.2 , Tables parsed from the collected ESG reports are stored in a local data directory in HTML format. HTML is a standard markup language that is structured and organised in a grid format. Furthermore, it also incorporates various formatting elements to provide context to the data. The sample HTML syntax of a parsed tabular data :

```
<table><thead><th>INDICATOR</th><th>UNIT</th><th>1Q22</th><th>1Q21</th><th>A%</th><th>2025</
th><th>2030</th></thead><tr><td>Revenues aligned with EU taxonomy</td><td>%</td><td>53%</td><td>66%</
td><td>-13p.p.</td><td>70%</td><td>&gt;80%</td></tr><tr><td>Scope 1 &amp; 2 Emissions Intensity</td><t-
d>gCO,/kWh</td><td>152</td><td>112</td><td>35%</td><td>˜100</td><td>0</td></tr><tr><td>Renewables Gen-
eration</td><td>%</td><td>1.40</td><td>1.35</td><td>4%</td><td>1.55</td><td>&lt;1</td></tr><tr><td>Fe-
male on Leadership</td><td>%</td><td>26.9%</td><td>24.6%</td><td>+2p.p.</td><td>30%</td><td>35%</td></
tr><tr><td>ESG &amp; equity linked compensation for Top Management?</td><td>.</td><td>Va</td><td>4</
td><td></td><td>Vav</td><td>Va</td></tr><tr><td>Cybersecurity</td><td>bitsight rating</td><td>800</
td><td>800</td><td>0%</td><td>Keep</td><td>advanced*</td></tr></table>
```

Keyword Identification & Pattern Matching

Identification of the keywords requires manual iteration and validation as these keywords needs to be aligned with the information extraction objectives. To extract the information related to GHG emission, energy efficiency and renewables and water consumption; several observations are deducted through studying characteristics of the tables and content included in parsed dataset. During the initial impressions, it was established that the tabular syntax is not accurately preserve original structure of the table. Nonetheless, textual content, numerical values and symbols are accurately parsed and included in the syntax.

Pattern-matching tasks involve defining specific patterns or structures that the keywords are expected to follow within the table or table cells. It could also include sequences of characters, words, or symbols, within the body of the tabular data. For the objective at hand, multiple patterns are discovered through observation and deducting the most common expressions, symbols associated with target environmental variables. *Table 2* explicates various keywords and pattern combinations utilised to identify occurrences related to GHG emission, energy efficiency and renewables, and water consumption in tabular data.

Table 2: Keywords and Pattern

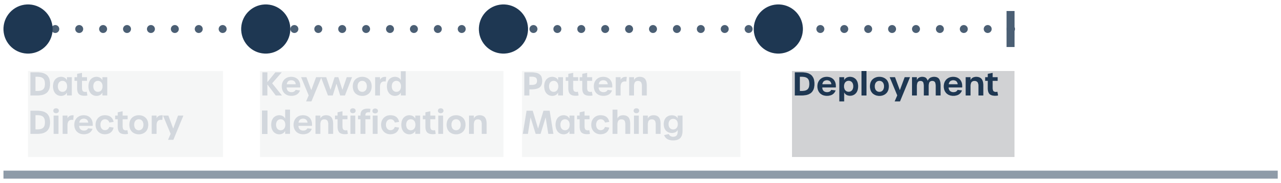
Target	Keywords	Pattern
GHG emission	- GHG - Greenhouse Gas - Emission - Scope 1, 2, 3 - Scope 1 - CO_2 eq. …….	- Keyword 'and' Unit (CO_2 eq.) - Numerical value 'and' Keyword - Numerical value 'and' Unit - Keyword 'or' Keyword

Table 2 (continue): Keywords and Pattern

Energy Efficiency and Renewables	- Energy Consumption
	- Renewable
	- Renewable energy
	- Watt hour (Wh)
	-KWh, MWh
	- Energy Infrastructure
	....
		`- Keyword 'and' Unit (KWh, MWh..)`
		`- Numerical value 'and' Keyword`
		`- Numerical value 'and' Unit`
		`- Keyword 'or' Keyword`

Water Consumption	- Water Consumption
	- Ground-water
	- Water regeneration
	- Cubic Meter (cuM)
	....
		`- Keyword 'and' Unit (KWh, MWh..)`
		`- Numerical value 'and' Keyword`
		`- Numerical value 'and' Unit`
		`- Keyword 'or' Keyword`

4.2.3 Deployment



Python Libraries for Deployment

beautifulsoup4==4.12.2
nltk==3.8.1
flair==0.13.1
numpy==1.26.2
pandas==2.1.4
scipy==1.11.4

Compared to the textual data analysis pipeline, deployment and Python implementation of the tabular data analysis is fairly simple. Instead of employing advanced algorithms like Boyer-Moore or Knuth-Morris-Pratt algorithms, a conventional sequential scan approach is utilised for pattern matching across the dataset (Crowston et. al, 2012). While these sophisticated algorithms demonstrate superior performance in the context of Big Data analyses, the relatively small size of the tabular dataset ensures optimal efficiency with the sequential scan method.

During sequential scanning the algorithm essentially highlights and extracts whether the pre-loaded keywords through pattern matching and validates its presence in the html syntax. It provides insight into whether the processed ESG report contain or mention target environmental disclosure in parsed tabular syntax or not. This binary result is then exported into simple Excel data frames for further scrutiny.

Results & Discussion

Photo by Fateme Alaie on Unsplash

5: Results & Discussion

With the aim of advancing knowledge in AI and NLP within the realm of environmental disclosures, the question ‘Are we able to extract information related to environmental variables in sustainability disclosures of EU-based firms through computational analysis?’, has served as a guiding principle throughout this thesis research. Collating a corpus of findings on the efficacy of various NLP-based algorithms and sources of sustainability disclosure, this chapter sheds light on the answer to the question. By discussing collected results from textual data analysis and tabular data analysis pipelines in conjunction with the reviewed literature, an assessment is drawn on the general accuracy of text classification and other methodologies employed throughout the data analysis.

From 87 ESG reports published during the last three years (2020, 2021 and 2022), approximately 190,000 text segments and 682 tables were analysed. For ease of understanding and to maintain compatibility, the results from both pipelines described in previous chapters are discussed separately to conclude.

5.1 Environment in ESG

5.1.1 Text Classification

Following the methodology developed in section 4.1, text classification results are obtained from a model called ESGify. The results contain text classification label scores concerning the text segments analysed. The cut-off score of 0.85 for each three classification labels GHG emissions, Energy efficiency and renewables, and water consumption is observed. The text segments corresponding to these classification labels are accumulated to facilitate trend analysis and to gain insights.

Figure 20 represents the total sentences parsed from ESG reports collected from each 27 institutes and *Figure 21* shows classification labels distribution. Initial visual insights from these area graphs suggest that the length of the documents certainly affects the distribution of labels. The institutions whose ESG reports tend to be lengthy and contain more sentences in general also have a higher number of sentences classified under selected environmental disclosures. Observing the distribution of label classifications, it appears that greenhouse gas (GHG) emissions are the most frequently mentioned environmental factor, followed by energy consumption. On the other hand, only a handful of institutions seem to address water consumption in their sustainability reports.

Establishing a comparative scale for classification label extracted from documents with a varying length is challenging without utilising statistical

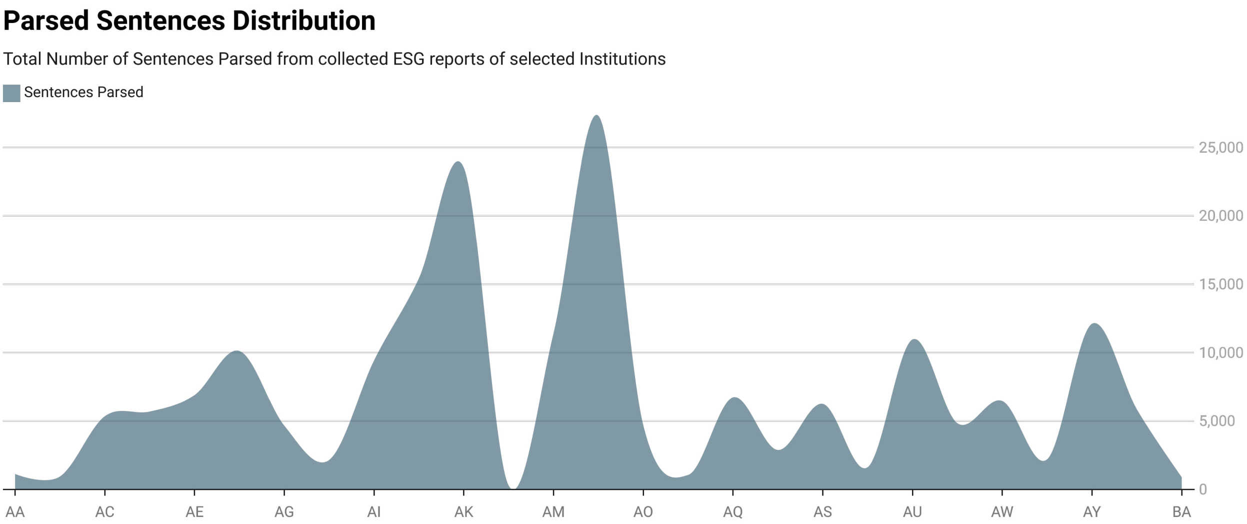


Figure 20 : No. of Parsed Sentences/ Text Segments from each institution

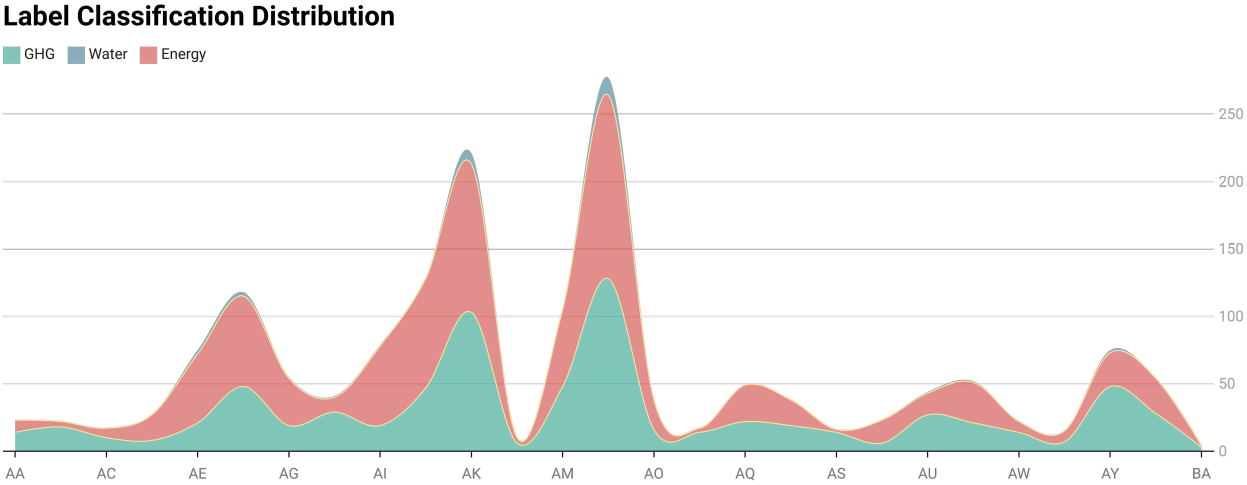


Figure 21 : Classification labels distribution

methodologies. A simple relative weighing method is employed to establish contextual relevance and further scrutiny. Relative weights represent the significance or importance of different classification labels relative to each other within a dataset (Antoncic 2020). It is calculated by following formula :

$$\varepsilon_d = \frac{L_d}{T_d}$$

with

- E_d being Relative Weightage for label L in document d
- L_d : number of sentences classified as L
- T_d : total number of sentences in document d

Subsequently normalised weights are applied to standardise the relative importance of each label across the dataset for equitable comparisons and refined analyses. Normalised weights ensure that the values assigned to each label are scaled to a consistent range (between 0 and 1) which in turn removes any biases introduced by differences in scale or magnitude. Normalised weightage is expressed by following formula :

$$\mathrm{N}_d = \frac{\varepsilon_d}{\max(\varepsilon_d)}$$

```
with
- N_d being Normalised Weightage for label L in document d
- E_d : Relative Weightage of for label L in document d
- max(E_d): maximun Relative Weightage of label L
```

The selected environmental factors under research scope exhibit varying degrees of presence within analysed ESG reports, with disclosures and statements related to GHG emissions emerging as prominent focus. To gauge relative presence of each factor, overall score for each label is calculated. The overall score of label is mathematically expressed as :

$$L = \sum_{i=o}^{n} N_d$$

```
with
- L being overall score for label
- n being total number of documents
```

The overall score presented in *Figure 22* seek to explicate the relationship between the classification labels within the dataset. Although the statements classified under GHG emission label are highest in the dataset, relative comparison of classification labels tells a different story. Despite the substantial representation of GHG emission related statements, the elevated relative prominence of Energy efficiency and renewables classification suggests a disparity between the absolute frequency and the relative importance of classification categories within the dataset.

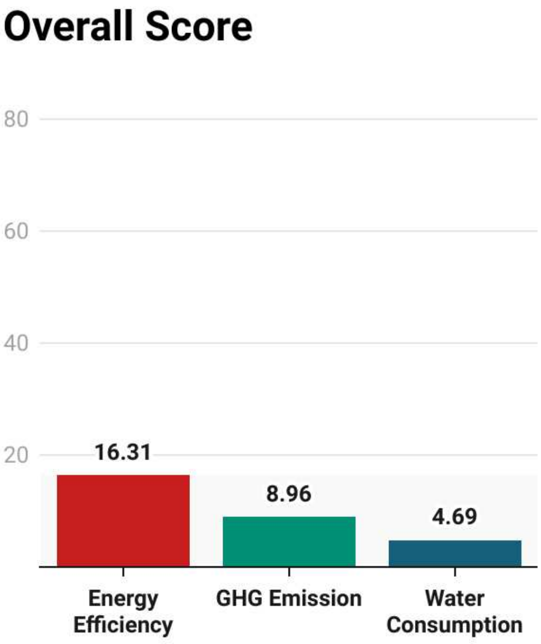


Figure 22 : Overall Score

This relative comparison prompts further inquiry into potential underlying factors. In order to identify the potential driving factor, a year wise and sector wise analysis of overall score trend is carried out (*Figure 23*). The year-wise comparison suggest growing trend in disclosure of environmental variables in ESG

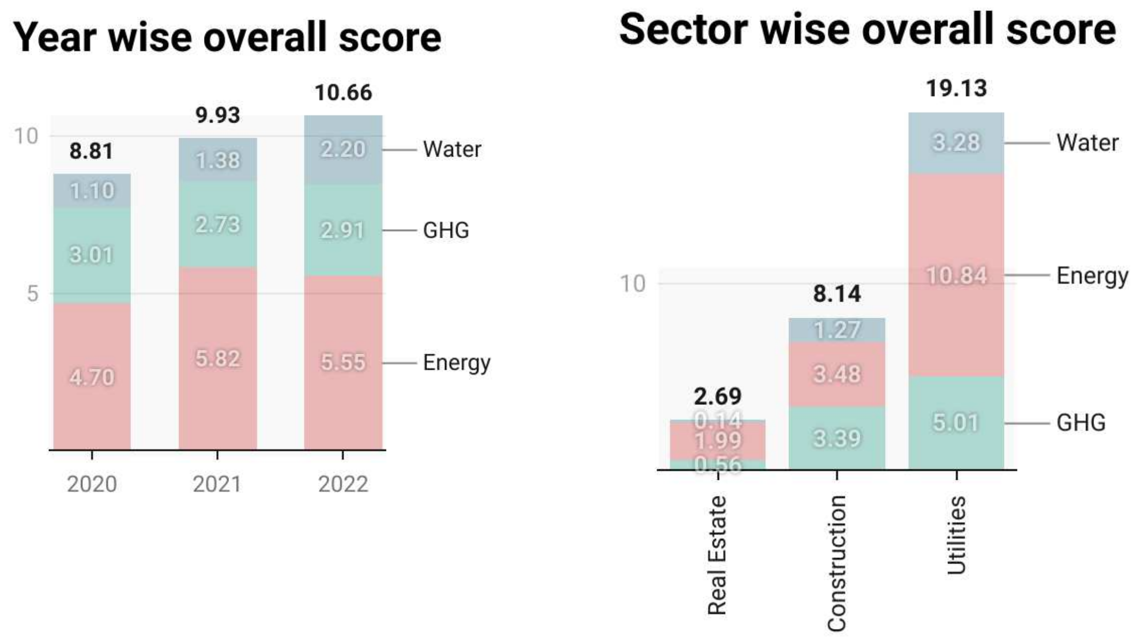


Figure 23 : Year wise and Sector wise comparison of overall label score

reports. This trend aligns with a growing emphasis on sustainability and corporate responsibility discussed in Chapter 2 along with reinforcement of regulatory frameworks and heightened awareness regarding environmental impacts.

The sector-wise comparison of the overall label score suggests that institutions operating in Utilities sector disclose more information on environmental factors than institutions operating in Real Estate or Construction sector. This trend highlights the utilities sector's heightened emphasis on environmental considerations in ESG. This trend could be attributed several factors. Firstly, the institution operating in utilities sector inherently has a significant environmental footprint due to its operations involving energy production, distribution, and infrastructure. Consequently, these institutions may face greater stakeholder pressure to disclose environmental information to address risks associated with their operations.

While the trend assessment based on the overall score of text classification labels may lack clarity and does not adhere to conventional research methodologies, speculations regarding the potential trends and correlations can still be examined. Despite the inherent ambiguity in deriving concrete causes, such analysis offer valuable indications of emerging trends or patterns within the dataset. In the subsequent sections, we delve into a country-wise analysis of all three labels and explore potential correlations.

Country Wise Comparison

Greenhouse Gas Emission

An examination of the country-wise distribution of GHG emissions classification is presented in *Figure 24*. The dataset subjected to classification analysis encompasses data from eight countries. The normalised label score pertaining to GHG emissions and its temporal evolution over the course of three years are shown. This comparative analysis indicates a progressive increase in disclosures related to GHG emissions over the past three years. However, Portugal stands out as an outlier. This ambiguity may stem from the inclusion of only one institution from Portugal in the dataset. France, Germany and Norway emerge as leaders in GHG disclosure practice.

This analysis of country-wise distribution sheds light on the variations in GHG emissions reporting practices across different geographical regions. A steady growth is observed in GHG disclosures across the EU region. This strengthens the speculation of heightened awareness on environmental accountability within the EU and evolving regulatory framework. However, the anomaly observed in Portugal emphasises the importance of considering contextual factors and dataset composition when interpreting results.

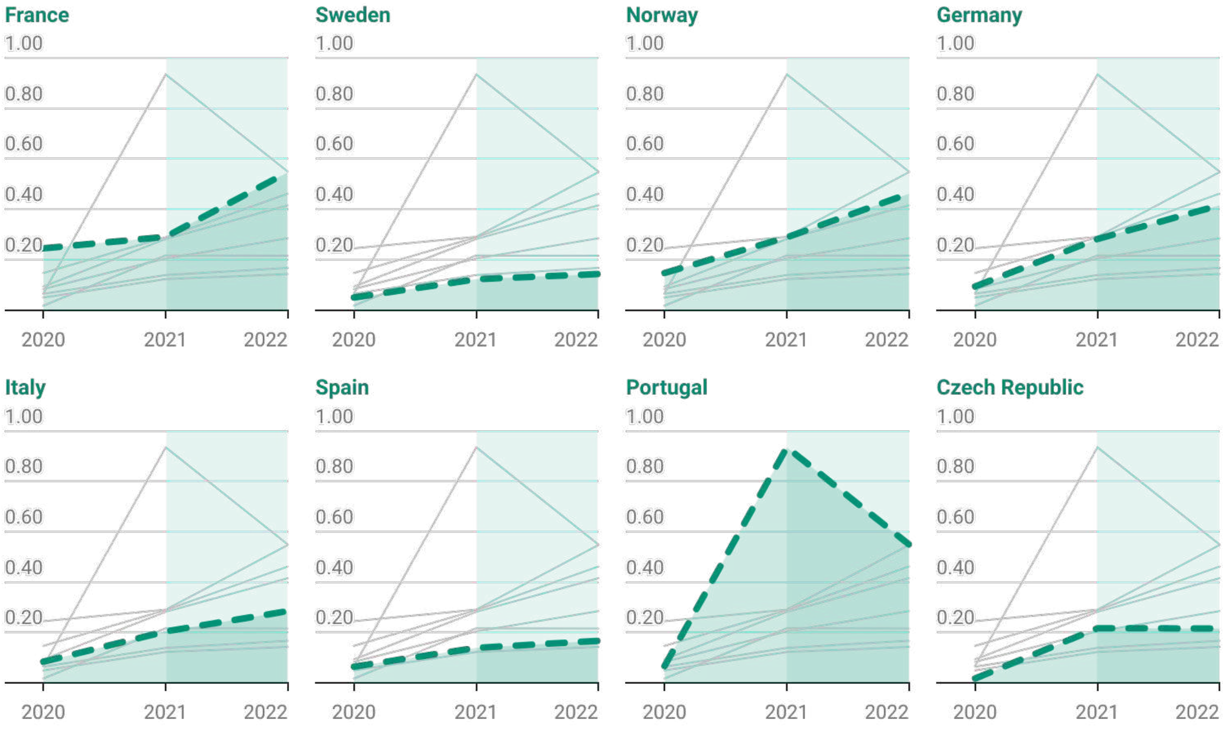


Figure 24 : GHG emission classification : Country wise distribution

Energy Efficiency and Renewables

Similar to GHG emission classification, the normalised weightage of energy efficiency and renewables classification is illustrated in *Figure 25*. As opposed to GHG emission country-wise analysis, the temporal evolution over the course of study period shows slight decrease in relative significance of energy disclosures. The similar ambiguity is observed in case of Portugal. This parallel ambiguity strengthens the necessity for dataset characteristics considerations.

This trend indicating a marginal decline in the relative importance of energy disclosures over time across EU countries, contradicts the comparison of overall scores conducted earlier in *Figure 23*. This discrepancy underscores the need for comprehensive analyses and careful consideration of contextual factors to accurately assess trends in ESG reporting practices. Further scrutiny into the factors influencing these divergent trends such as local regulations, incentives, policies, etc. may provide valuable insights into the dynamics of sustainability reporting across different regions.

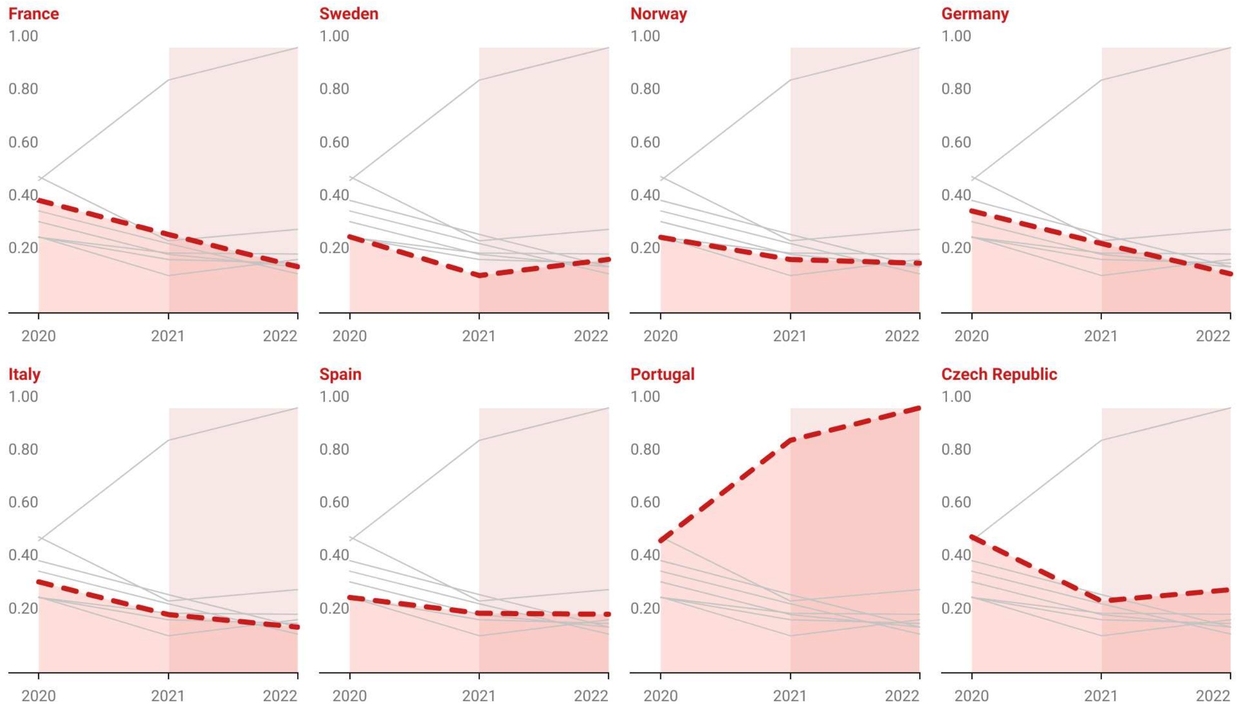


Figure 25 : Energy Efficiency classification : Country wise distribution

Water Consumption

The normalised weightage of water consumption classification is shown in *Figure 26*. Contrary to GHG emission and Energy efficiency classification, country wise trend in water consumption is inconclusive. This could be due to multiple reasons. As previously discussed in the classification distribution (*Figure 21*), the total number of sentences classified under water consumption is the lowest. Only a limited number of analysed ESG reports include information on water consumption. As an outlier, the text classification analysis identified consistent water consumption disclosures for institutions located in Italy.

The inconclusive country-wise trend highlights the need for further research and inclusion of additional ESG reports in data collection to better understand the extent of water consumption disclosures. Additionally, the identification of Italy as an outlier in disclosures related to water consumption raises curiosity about the geo-specific factors influencing reporting practices and regulatory framework.

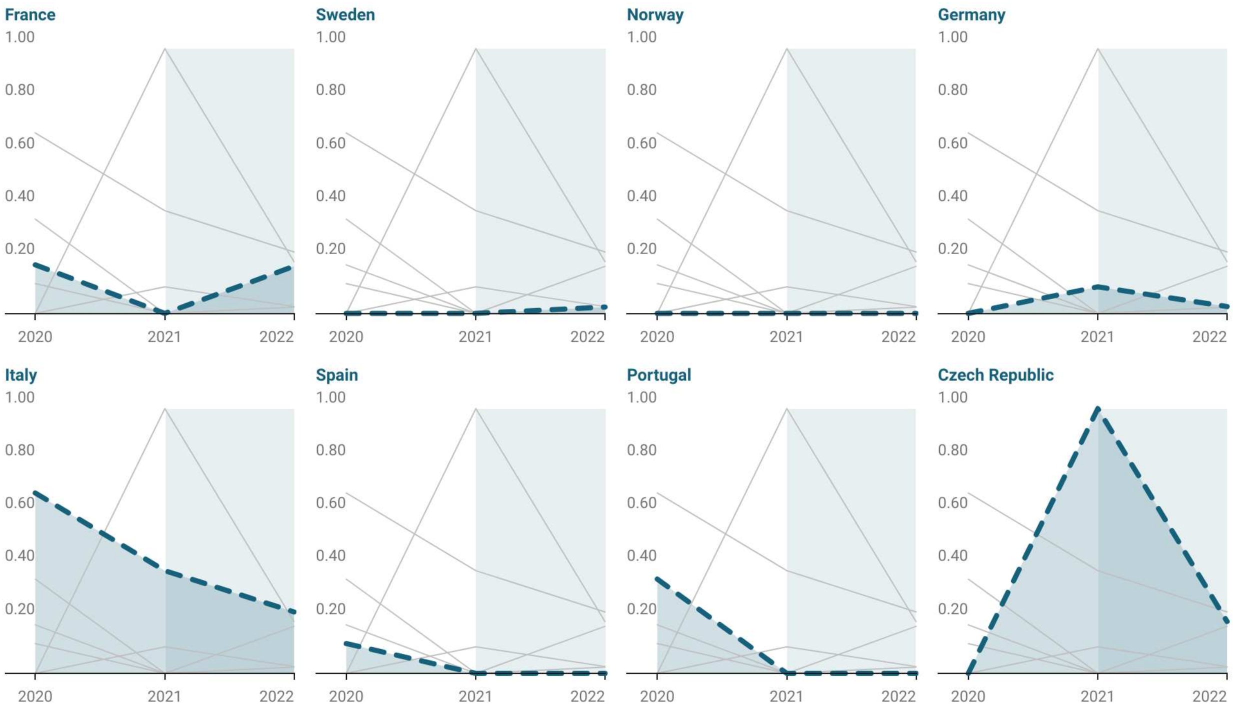


Figure 26 : Energy Efficiency classification : Country wise distribution

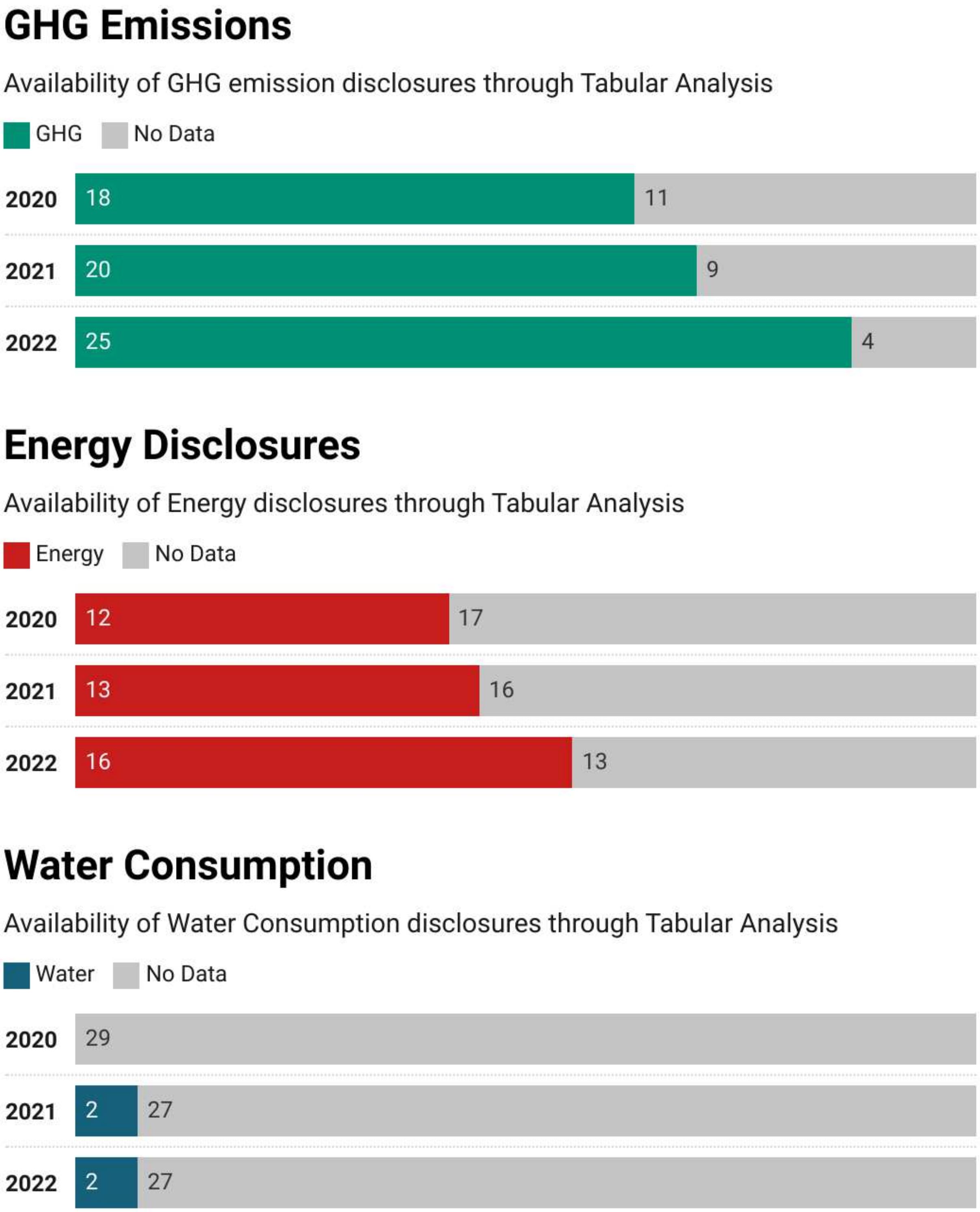
5.1.1 Tabular analysis

In contrast to the NLP pipeline utilised for text classification, the computational pipeline adopted for tabular analysis operates by scanning for predetermined keywords and patterns within the dataset. This pipeline yields a straightforward binary assessment based on the presence or absence of keywords or patterns.

Figure 27 illustrates the yearly variation in environmental disclosures of selected 29 institutions. Similar to results of textual pipeline, disclosures related to GHG emissions are featured prominently within the tabulated data. The energy disclosures are also found in considerable number of institutes, while only handful of references to water consumption are discovered in ESG reports.

The comparative analysis of selected environmental factors indicates a growing trend in their frequency over time. Additionally, a sector-wise examination was conducted to shed light on reporting dynamics within various sectors. Figure 28 presents intensity of environmental disclosure relative to operating sectors. The sector wise analysis reveals that energy-related disclosures are more pronounced than GHG emission disclosures in the real estate sector, whereas institutions operating in the construction and utilities sectors tend to focus more on GHG emissions disclosures.

This sector-specific variation in environmental disclosures reflects the diverse priorities and focus within different industries. The emphasis on energy-related disclosures in real estate sector may stem from a focus on energy


GHG Emissions

Availability of GHG emission disclosures through Tabular Analysis

Energy Disclosures

Availability of Energy disclosures through Tabular Analysis

Water Consumption

Availability of Water Consumption disclosures through Tabular Analysis

Figure 27 : Year-wise distribution of environmental disclosures

efficiency measures in living environment. Furthermore, the heightened focus on GHG emissions disclosures in utilities sector aligns with the significant environmental footprint associated with energy production and distribution.

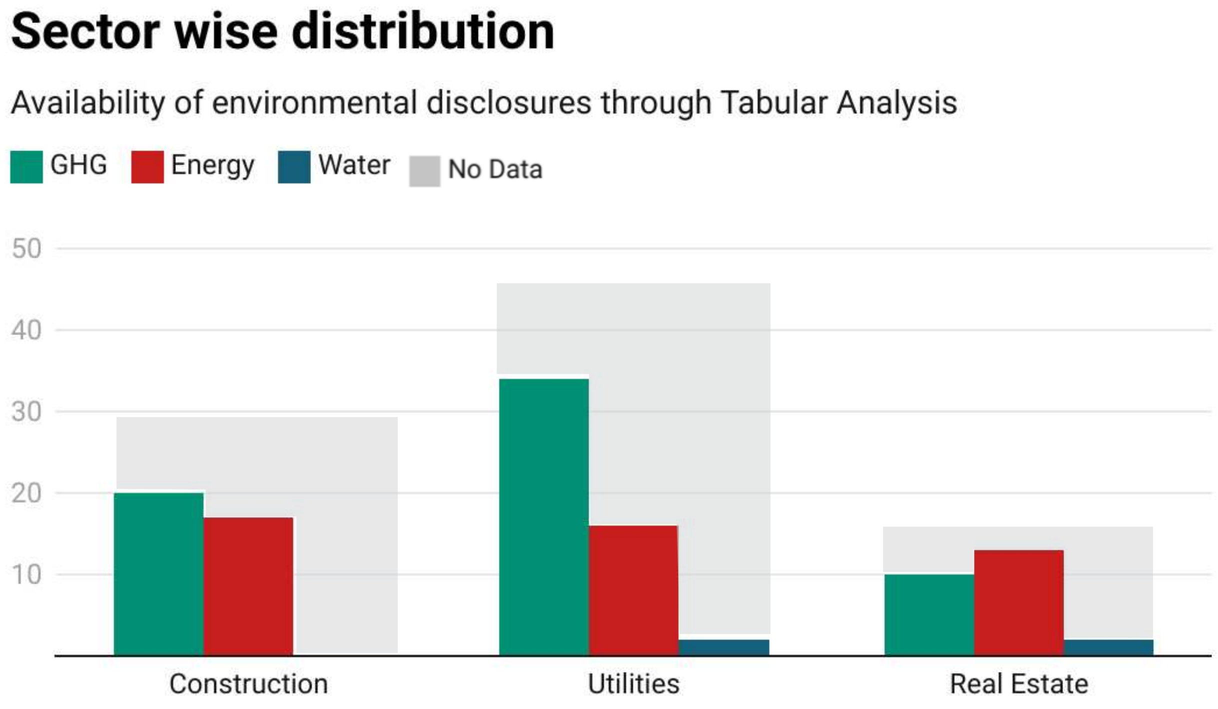


Figure 28 : Sector-wise distribution of environmental disclosures

In general, the reporting practices across different sectors and geographical location varies greatly in ESG reporting. Institutions operate within unique contexts, facing different environmental risks and stakeholder influences. To understand the underlying factors and interpret the extracted data accurately, an approach that acknowledge sector-specific nuances and regional disparities is necessary. The computational based analyses undertaken during this thesis serve as a foundational tool in this endeavour, elucidating the heterogeneous landscape of ESG reporting practices and stimulating speculations on influencing factors. The comprehensive comparison of results from text classification and tabular analysis with various perspective reveals intriguing patterns and complexities. However, substantiating these speculations would require targeted research objectives and tailored dataset.

The results and correlation stated so far suggests that there is a prominent focus on disclosure related to GHG emission and energy efficiency. From all 29 institutions included in analysis, the institutions operating in sectors of construction and utilities tent to disclose more information.

5.2 Evaluation

When evaluating the results found through computational analysis, some limitations need to be considered. The majority of limitations for this study lies in the accuracy and creditability of the algorithms at every stage of document preprocessing and NLP pipelines.

5.2.1 Limitations

The computer vision techniques utilised during the document preprocessing stage are not flawless. The intricate structure and visually diverse content of ESG reports present challenges in accurately extracting and parsing the content. Albeit the recent strides in accuracy of computer vision techniques, it still struggles to interpolate data presented in visual or tabular form to machine readable format accurately.

During the preprocessing of the collected data, several choices were made that could have influenced the outcomes of the study. Specifically, the methodology adapted entailed the removal of special characters and stopwords along with text tokenization. However, alternative preprocessing techniques such as text lemmatization and text stemming could have been implemented to further refine the NLP pipelines.

In the context of ESG reporting, LLM based text classification models struggle to capture the nuanced meanings and context embedded within textual data due to complex language and industry-specific terminology (Varini et al., 2020). Additionally the training data used for fine-tuning can impose biases. Biases in training data could lead to skewed classifications and inaccurate representations of certain topics. Additionally, the size of the training dataset could always be larger in order to improve the models and the classification results.

Keyword identification and pattern matching computational analysis employed for tabular analysis are also limited by their reliance on predefined criteria. It is also important to note that the results and possible underlying correlation explored through comparison are unique to the database. ESG reports collected from institutions situated in disparate geographical regions or engaged in diverse sectors of operation may exhibit distinct patterns and relationships.

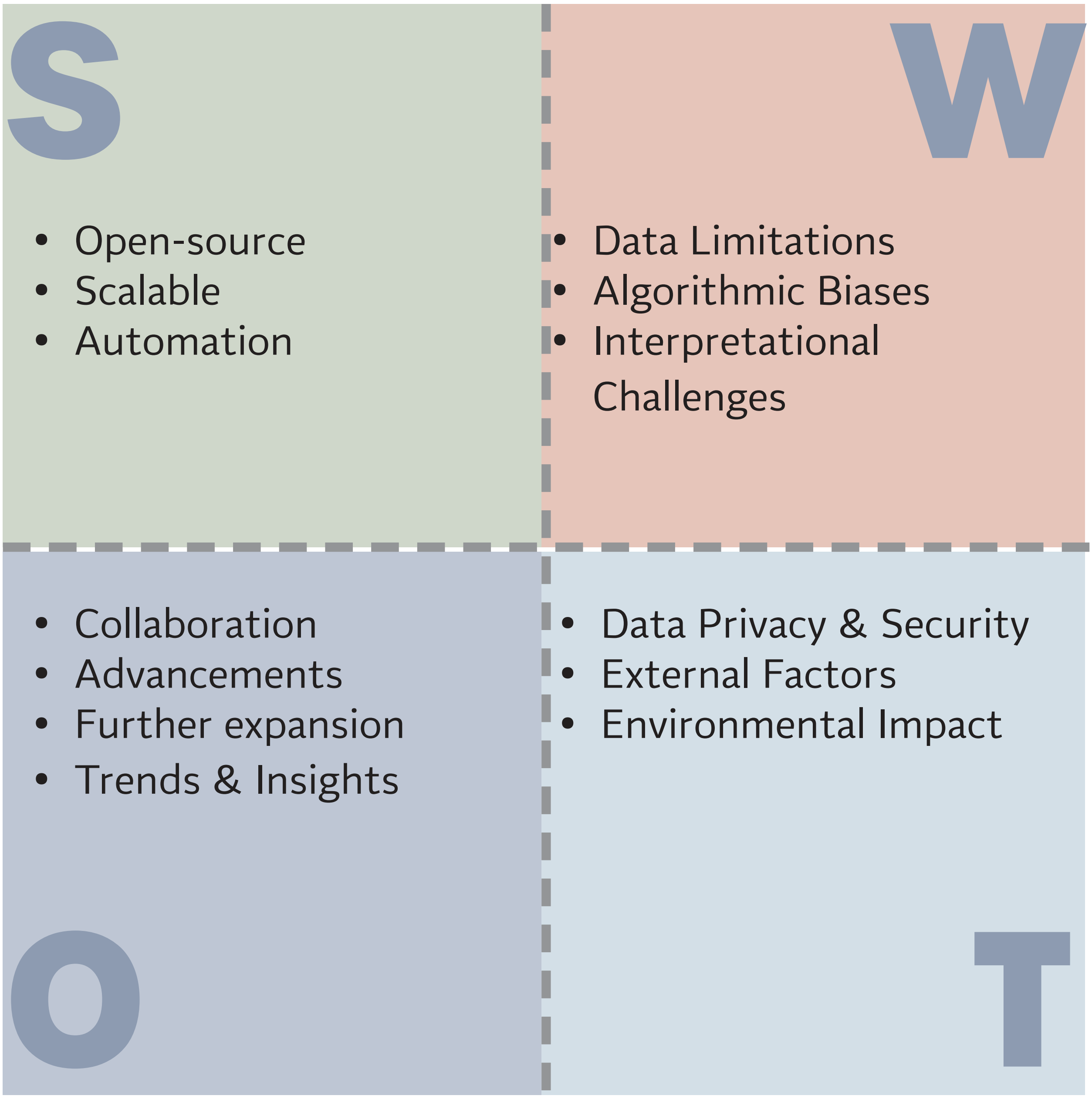
The proposed NLP pipeline addresses the main research question adequately, demonstrating the ability to extract insights from ESG reports through computational measures. However, to address the cherry-picking in environmental disclosures a more sophisticated and complex NLP pipeline is necessary. Identifying cherry-picking is a difficult task that requires computational algorithms to perform multi-level text classification and identification of disclosed quantity of the environmental factors including Scope 1 and Scope 2 GHG emissions, among other variables. Additionally, the NLP pipeline should include external data sources to gauge quality of environmental disclosures impartially.

Albeit the ambiguities and limitations, the computational based approach serve as a stepping stone for gaining deeper insights into ESG reporting data. To elaborate further on the potential of NLP in ESG along with associated opportunities and pitfalls, a SWOT (strengths, weaknesses, opportunities and threats) analysis is conducted.

5.2.2 SWOT

The SWOT analysis aims to evaluate the strengths, weaknesses, opportunities and threats associated with the computational methodology proposed in this thesis. By critically examining the potential of NLP-based methodologies in the ESG realm, this assessment elucidates areas of improvement and further research scope.

The computational methods used for ESG report analysis and NLP are open-source and highly customisable. This means they can be easily accessed



S

- Open-source
- Scalable
- Automation

W

- Data Limitations
- Algorithmic Biases
- Interpretational Challenges

O

- Collaboration
- Advancements
- Further expansion
- Trends & Insights

T

- Data Privacy & Security
- External Factors
- Environmental Impact

Figure 29 : SWOT analysis

and adapted by other researchers for further scrutiny. Open-source based analytical methods are interoperable and also foster collaboration within research teams. The NLP pipelines can be scaled to handle large volumes of data including external sources without degradation in performance. This functionality is crucial as the amount of ESG data generated continues to grow exponentially, an adaptable NLP pipeline provides an opportunity to incorporate new training data and the latest developments in the realm of AI. Automation within NLP pipelines refers to automating diverse tasks encompassed in methodology, including data collection, document preprocessing, feature extraction, and model development. It can streamline the entire analysis process, reducing the need for manual intervention and iterations.

ESG-related data sources such as news articles, databases, ESG ratings, and sustainability reports are gaining in popularity amongst researchers to explore hypotheses through numerous perspectives. Albeit being scalable and able to incorporate additional data sources, many researchers claim that the results from the NLP analysis may be constrained by the availability and quality of the dataset (Crowston et al., 2012; Ehrhardt & Nguyen, 2021; Lee et al., 2022). Algorithmic biases are another well-discussed drawback amongst ESG disclosure and NLP researchers (Assael, 2023; Armbrust,2022). The unavailability of high-quality annotated datasets, LLM hallucinations, insufficient training data, and complexity of language are some of the main drivers leading to skewed interpretations and inaccurate conclusions. Diversity in influencing factors and the complexity of ESG disclosures raise additional challenges in the interpretation of data. As ESG reporting aims to provide information on a wide range of sustainability measures holistically, assessments and analysis are hard to interpret particularly for stakeholders lacking knowledge in the field.

Despite the many challenges, the integration of AI-based tools provides several opportunities to exploit and explore factors influencing ESG reporting practices. From the validity of insights to exploratory trend hunting, the applicability of AI tools is endless. A multi-faceted collaboration among industry experts, policymakers and researchers can enrich the AI analysis and ensure its applicability in practice. Advancements in NLP techniques and deep learning models offer opportunities to enhance the accuracy and efficiency of the results. Recent technological advancements in the field of AI are astronomical. It is plausible that a sophisticated NLP pipeline required to assess greenwashing practices will be available soon. The list delineating opportunities to utilise NLP techniques and AI-powered tools in the field of ESG is non-exhaustive, offering numerous opportunities for further exploration and innovation.

In addition to the opportunities presented by the integration of AI-based tools to gain insights in ESG reporting practices, it is imperative to consider several critical concerns. External factors such as regulatory changes, geopolitical instability, and socio-economic dynamics can significantly influence the efficacy and adoption of AI-powered tools in ESG reporting. There is a growing concern regarding the privacy and security of data while leveraging AI and NLP techniques for data analysis practices. Ensuring compliance with the local legal framework and data protection regulations are paramount considerations. Subsequently, the environmental impact of development and deployment of AI technologies should be considered, particularly the energy consumption associated with training large-scale deep learning models. According to an estimate, environmental impact for training of latest model of Generative pre-trained transformer (GPT-4) amounts to approximately 6,912 metric tons of CO_2 (Lacoste et al., 2019).

Addressing these concerns and challenges in a proactive manner is essential to harness the full potential of AI-driven approaches while ensuring ethical, sustainable, and socially responsible practices in ESG reporting endeavours.

Conclusions

Photo by Fateme Alaie on Unsplash

6. Conclusions

In the 21st century, sustainability has emerged as a significant and delicate concern, as the severe implications of climate change have come to light. Numerous institutions are now publishing their annual sustainability reports, which serve as pivotal references for comprehending their approaches and actions towards sustainability. Nonetheless, the abundance of information infused with forward-looking statements and business terminology poses a threat to individuals seeking to grasp the environmental impact of operations conducted by private institutions. Harnessing the AI based analytical techniques, this thesis proposes novel NLP pipelines to parse the complex information published through ESG reports and extract insights into environmental disclosures.

The inquiry commenced by delving into the multifaceted realm of ESG, elucidating its significance in addressing contemporary challenges towards climate change, social equity, and corporate governance. The pivoted role of ESG criteria in making informed investment decisions, managing risk, and promoting sustainable development across diverse sectors is highlighted. As evidenced by initiatives such as the EU taxonomy and the CSRD, the regulatory landscape of ESG disclosures within the EU is evolving. Which in turn stresses the importance of standardisation and transparency.

Acknowledging the magnitude of the task, the research jotted in this thesis was methodically structured to address fundamental questions concerning the feasibility and effectiveness of computational analysis in extracting environmental variables from ESG disclosures. Leveraging the advanced AI and NLP techniques, customised approaches were derived to enable somewhat accurate parsing of the extensive ESG reports. The methodological journey entailed meticulous considerations regarding ethical and legal framework surrounding the applicability of AI in ESG. This was followed with curated ESG report directory and computer vision powered document preprocessing.

Two separate computational pipelines were discussed to address the intricacies and uniqueness in analysing textual and tabular based data parsed from ESG reports. Leveraging cutting-edge NLP architectures such as transformers and fine-tuning techniques, environmental disclosures in textual data was identified and classified. Somewhat simpler computational algorithms were employed to access the unstructured tabular data and glean insights.

The findings of this thesis highlights the transformative potential of AI and NLP in enhancing the efficiency and accuracy of ESG analysis. The solutions entailed in the methodology are scalable and could be subjected to diverse set of ESG report collections to identify major drivers and lacking environmental disclosures. Interpretations from the cross sectoral and cross regional comparison suggests that in comparison to water consumption disclosures related to GHG emissions and energy efficiency are more prominent in the

EU region. The findings also highlight the growing trend in inclusion of environment in ESG. With adaptation of new regulatory measures this trend is predicated to grow and more quantifiable environmental measures are foreseen to be included in next iteration of ESG report publications. Nonetheless, implementation of NLP pipelines also brought to light inherent challenges and limitations, necessitating ongoing refinement and innovation in computational methodologies to address the intricacies inherent in ESG reporting.

This thesis represents a pioneering endeavour combining 'E' in ESG with computational analysis. It paves a way to understand and manage ESG data effectively to gain foundational insights. Although the current open-source computational resources lacks ability to fully digest and understand ESG data to provide conclusive connections and identify cherry-picking, advancements in the field of AI hold tremendous potential to mitigate this gap.

References

References

Alexander S, (2019). Sustainable News – A Sentiment Analysis of the Effect of ESG Information on Stock Prices. Political Economy - Development: Environment EJournal. https://doi.org/10.2139/ssrn.3809657

Antoncic, M. (2020). Uncovering hidden signals for sustainable investing using Big Data: Artificial intelligence, machine learning and natural language processing. Journal of Risk Management in Financial Institutions. https://doi.org/10.69554/CIKJ7477

Armbrust, F. (2022). Deep Sustainable Finance: An End-to-End Text Analysis of the Financial and Environmental Narratives in Corporate Disclosures [Doctorate]. Universität Stuttgart, Stuttgart. Last accessed on 14.02.2024.

Assael, J. (2023). Machine learning for ESG data in the financial industry [Master].

https://theses.hal.science/tel-04138530/. Last accessed on 15.03.2024

Badenhoop, N., Hackmann, A., Mücke, C., & Pelizzon, L. (2023). Quo vadis sustainable funds? Sustainability and taxonomy-aligned disclosure in Germany under the SFDR (SAFE White Paper No. 94). Frankfurt a. M.: Leibniz Institute for Financial Research SAFE. https://www.econstor.eu/handle/10419/273716

Barrymore, N. (2021). Green or Greenwashing? How Manager and Investor Preferences Shape Firm Strategy. http://dx.doi.org/10.2139/ssrn.4555581

Bauckloh, T., Klein, C., Pioch, T., & Schiemann, F. (2022). Under Pressure? The Link Between Mandatory Climate Reporting and Firms' Carbon Performance. Organization & Environment, 108602662210833. https://doi.org/10.1177/10860266221083340

Bingler, J. A., Kraus, M., & Leippold, M. (2021). Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures. https://doi.org/10.2139/ssrn.3796152

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: [analyzing text with the natural language toolkit] (1. ed.). O'Reilly.

Blanchy, G., Albrecht, L., Koestel, J., & Garré, S. (2023). Potential of natural language processing for metadata extraction from environmental scientific publications. SOIL, 9(1), 155–168. https://doi.org/10.5194/soil-9-155-2023

Bowen, F. (2014). After greenwashing: Symbolic corporate environmentalism and society. Organizations and the Natural Environment. Cambridge University Press.

Bowen, F., & Aragon-Correa, J. A. (2014). Greenwashing in Corporate Environmentalism Research and Practice. Organization & Environment, 27(2), 107–112. https://doi.org/10.1177/1086026614537078

Brettschneider, P. (2021). Neue rechtliche Regeln zum Text- und Data-Mining.

https://doi.org/10.5281/ZENODO.4592979

Bril, H., Kell, G., & Rasche, A. (2021). Sustainable investing: A path to a new horizon. Routledge. https://doi.org/10.4324/9780429351044

Camilleri, M. A. (2015). Environmental, social and governance disclosures in Europe.

Sustainability Accounting, Management and Policy Journal, 6(2), 224–242.

https://doi.org/10.1108/SAMPJ-10-2014-0065

Chagas, E. J. M. d., Albuquerque, J. d. L., Maia Filho, L. F. A., & Ceolin, A. C. (2022). Sustainable development, disclosure to stakeholders and the Sustainable Development Goals: Evidence from Brazilian banks' non-financial reports. Sustainable Development, 30(6), 1975–1986. https://doi.org/10.1002/sd.2363

Chen, M., Behren, R. von, & Mussalli, G. (2021). The Unreasonable Attractiveness of More ESG Data. The Journal of Portfolio Management, 48(1), 147–162.

https://doi.org/10.3905/jpm.2021.1.281

Chen, W., Alharthi, M., Zhang, J., & Khan, I. (2024). The need for energy efficiency and economic prosperity in a sustainable environment. Gondwana Research, 127, 22–35. https://doi.org/10.1016/j.gr.2023.03.025

Claudron, E. (2022). Measuring ESG Performance: A Text Mining Approach [Master]. Louvain School of Management.

Climate Bonds Initiative. (2023). Q1 2023 Market Update: Sustainable debt shows recovery. https://www.climatebonds.net/resources/reports/q1-2023-market-update-sustainable-debt-shows-recovery. Last accessed on 14.02.2024.

Consolandi, C., Phadke, H., Hawley, J., & Eccles, R. G. (2020). Material ESG Outcomes and SDG Externalities: Evaluating the Health Care Sector's Contribution to the SDGs. Organization & Environment, 33(4), 511–533.

https://doi.org/10.1177/1086026619899795

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. International Journal of Social Research Methodology, 15(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

Dai, J [Jiyuan], Ormazabal, G., Penalva, F., & Raney, R. A. (2023). Can Mandatory Disclosure Curb Greenwashing? First Evidence from the EU SFDR. https://doi.org/10.2139/ssrn.4564890

Delmas, M. A., & Burbano, V. C. (2011). The Drivers of Greenwashing. California Management Review, 54(1), 64–87. https://doi.org/10.1525/cmr.2011.54.1.64

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Dumitrescu, A., Gil-Bazo, J., & Zhou, F. (2022). Defining Greenwashing. SSRN Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.4098411

Dyck, A., Lins, K. V., Roth, L., & Wagner, H. F. (2019). Do institutional investors drive corporate social responsibility? International evidence. Journal of Financial Economics, 131(3), 693–714. https://doi.org/10.1016/j.jfineco.2018.08.013

Ehrhardt, A., & Nguyen, M. T. (2021). Automated ESG Report Analysis by Joint Entity and Relation Extraction. In (pp. 325–340). Springer, Cham. https://doi.org/10.1007/978-3-030-93733-1_23

Progress Report on Greenwashing, May 31, 2023. https://www.esma.europa.eu/sites/default/files/2023-06/ESMA30-1668416927-2498_Progress_Report_ESMA_response_to_COM_RfI_on_greenwashing_risks.pdf. Last accessed on 10.12.2023

Esposito, F., Malerba, D., & Semeraro, G. (20--). A knowledge-based approach to the layout analysis. In Third International Conference on Document Analysis and Recognition (Volume 2), Montréal, Canada, 14.08-15.08.1995 (pp. 466–471). IEEE Comput. Soc. Press. https://doi.org/10.1109/ICDAR.1995.599037

The European Green Deal (2019). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1588580774040&uri=CELEX%3A52019DC0640. Last accessed on 22.12.2023

Renewed sustainable finance strategy (2021). https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12635-Renewed-sustainable-finance-strategy_en

European Commission. (2023, March 1). Sustainable Finance: Commission welcomes political agreement on European green bond standard [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/mex_23_1301. Last accessed on 22.12.2023

Sustainable Finance Disclosure Regulation (2019). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019R2088. Last accessed on 22.12.2023

EU Taxonomy, June 18, 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32020R0852. Last accessed on 22.12.2023

Corporate Sustainability Reporting Directive (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464. Last accessed on 22.12.2023

FRECAUTAN, I., & Andreea NITA (2022). Who Is Going To Win: The Eu Esg Regulation Or The Rest Of The World? A Critical Review. Annals of Faculty of Economics, 2(2), 109–120. https://ideas.repec.org/a/ora/journl/v1y2022i2p109-120.html. Last accessed on 23.03.2024.

Garvey, T. G., Iyer, M., & Nash, J. (2018). Carbon footprint and productivity: does the “E” in ESG capture efficiency as well as environment. Journal of Investment Management, 16(1), 59–69. https://joim.com/wp-content/uploads/emember/downloads/p0569.pdf. Last accessed on 03.10.2023

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021, July 18). YOLOX: Exceeding YOLO Series in 2021. https://doi.org/10.48550/arXiv.2107.08430

Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG), February 24, 2022. https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html#p0018. Last accessed on 23.04.2024.

GHG Protocol (2001). A Corporate Accounting and Reporting Standard. GHG Protocol. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf. Last accessed on 23.04.2024.

Global Sustainable Investment Alliance. (2020). GLOBAL SUSTAINABLE INVESTMENT REVIEW 2020. Last accessed on 31.10.2023

GRI (2022). Consolidated Set of the GRI Standards. GRI. https://www.globalreporting.org/

how-to-use-the-gri-standards/gri-standards-english-language/. Last accessed on 09.10.2023

Henisz, W., Koller, T., & Nuttall, R. (2019). Five ways that ESG creates value. http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/2319/1/five-ways-that-esg-creates-value.pdf. Last accessed on 31.10.2023

Herbohn, K., Walker, J., & Loo, H. Y. M. (2014). Corporate Social Responsibility: The Link Between Sustainability Disclosure and Sustainability Performance. Abacus, 50(4), 422–459. https://doi.org/10.1111/abac.12036

Höck, A., Klein, C., Landau, A., & Zwergel, B. (2020). The effect of environmental sustainability on credit risk. Journal of Asset Management, 21(2), 85–93.

https://doi.org/10.1057/s41260-020-00155-4

Hugging Face. (2024). Hugging Face – The AI community building the future. https://huggingface.co/. Last accessed on 12.12.2023

Internationale Agentur für Erneuerbare Energien. (2020). Global renewables outlook: Energy transformation 2050. International Renewable Energy Agency. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2020/Apr/IRENA_Global_Renewables_Outlook_2020.pdf. Last accessed on 14.11.2023

Jiang, L., Gu, Y., & Dai, J [Jun] (2023). Environmental, Social, and Governance Taxonomy Simplification: A Hybrid Text Mining Approach. Journal of Emerging Technologies in Accounting, 1–21. https://doi.org/10.2308/JETA-2022-041

Jurczenko, E. (Ed.). (2018). Factor Investing: From Traditional to Alternative Risk Premia. Elsevier.

Kang, H., & Kim, J. (2022). Analyzing and Visualizing Text Information in Corporate Sustainability Reports Using Natural Language Processing Methods. Applied Sciences, 12(11), 5614. https://doi.org/10.3390/app12115614

Khaled, R., Ali, H., & Mohamed, E. K. (2021). The Sustainable Development Goals and corporate sustainability performance: Mapping, extent and determinants. Journal of Cleaner Production, 311, 127599. https://doi.org/10.1016/j.jclepro.2021.127599

Kölbel, lbel, J., Leippold, M., Rillaerts, J., & Wang, Q. (2020). Does the CDS Market Reflect Regulatory Climate Risk Disclosures? https://doi.org/10.2139/ssrn.3616324

Kölbel, J. F., Leippold, M., Rillaerts, J., & Wang, Q. (2022). Ask BERT: How Regulatory Disclosure of Transition and Physical Climate Risks Affects the CDS Term Structure. Journal of Financial Econometrics, Article nbac027. https://doi.org/10.1093/jjfinec/nbac027

KPMG. (2022). Survey of Sustainability Reporting. https://kpmg.com/xx/en/home/insights/2022/09/survey-of-sustainability-reporting-2022/climate-risk.html. Last accessed on 31.10.2023

Krueger, P., Sautner, Z., & Starks, L. T. (2020). The Importance of Climate Risks for Institutional Investors. The Review of Financial Studies, 33(3), 1067–1111. https://doi.org/10.1093/rfs/hhz137

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019, October 22). Quantifying the

Carbon Emissions of Machine Learning. http://arxiv.org/pdf/1910.09700

Le Luo, & Tang, Q. (2023). The real effects of ESG reporting and GRI standards on carbon mitigation: International evidence. Business Strategy and the Environment, 32(6), 2985–3000. https://doi.org/10.1002/bse.3281

Lee, O., Joo, H., Choi, H., & Cheon, M. (2022). Proposing an Integrated Approach to Analyzing ESG Data via Machine Learning and Deep Learning Algorithms. Sustainability, 14(14), 8745. https://doi.org/10.3390/su14148745

Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., & Wei, F. (2022, March 4). DiT: Self-supervised Pre-training for Document Image Transformer. https://arxiv.org/pdf/2203.02378

Li, T., Wang, K., Sueyoshi, T., & Wang, D. D. (2021). ESG: Research Progress and Future Prospects. Sustainability, 13(21), 11663. https://doi.org/10.3390/su132111663

Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate Environmental Disclosure under Threat of Audit. Journal of Economics & Management Strategy, 20(1), 3–41. https://doi.org/10.1111/j.1530-9134.2010.00282.x

Lyon, T. P., & Montgomery, A. W. (2015). The Means and End of Greenwash. Organization & Environment, 28(2), 223–249. https://doi.org/10.1177/1086026615575332

M. Lukinović, & L. Jovanović (2019). Greenwashing – fake green/environmental marketing. Fundamental and Applied Researches in Practice of Leading Scientific Schools, 33(3), 15–17. https://doi.org/10.33531/farplss.2019.3.04

Marquis, C., Toffel, M. W., & Zhou, Y. (2016). Scrutiny, Norms, and Selective Disclosure: A Global Study of Greenwashing. Organization Science, 27(2), 483–504. https://doi.org/10.1287/orsc.2015.1039

Melas, D., Nagy, Z., & Kulkarni, P. (2018). Factor Investing and ESG Integration. In E. Jurczenko (Ed.), Factor Investing: From Traditional to Alternative Risk Premia (pp. 389–413). Elsevier. https://doi.org/10.1016/B978-1-78548-201-4.50015-5

Moodaley, W., & Telukdarie, A. (2023). Greenwashing, Sustainability Reporting, and Artificial Intelligence: A Systematic Literature Review. Sustainability, 15(2), 1481. https://doi.org/10.3390/su15021481

Ni, J., Bingler, J., Colesanti Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stammbach, D., Vaghefi, S., Wang, Q., Webersinke, N., Wekhof, T., Yu, T., & Leippold, M. (2023). Paradigm Shift in Sustainability Disclosure Analysis: Empowering Stakeholders with Chatreport, a Language Model-Based Tool. https://doi.org/10.2139/ssrn.4476733

Nugent, T., Stelea, N., & Leidner, J. L. (2020, October 16). Detecting ESG topics using domain-specific language models and data augmentation approaches. https://arxiv.org/pdf/2010.08319

Parris, T. M. (2006). Corporate Sustainability Reporting. Environment: Science and Policy for Sustainable Development, 48(5), 3. https://doi.org/10.3200/ENVT.48.5.3-3

Perazzoli, S., Joshi, A., Ajayan, S., & Santana Neto, J. P. de (2022). Evaluating Environmental, Social, and Governance (ESG) from a Systemic Perspective: An Analysis Supported by Natural Language Processing. SSRN Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.4244534

Pham, V., Pham, C., & Dang, T. (2020, October 28). Road Damage Detection and Classification with Detectron2 and Faster R-CNN. http://arxiv.org/pdf/2010.15021.pdf

Plastun, O. L., Makarenko, I. O., Khomutenko, L. I., Osetrova, O., & Shcherbakov, P. (2020). Sdgs and ESG disclosure regulation: Is there an impact? Evidence from Top-50 world economies. https://essuir.sumdu.edu.ua/handle/123456789/82717

Pörtner, H. O., Roberts, D., Tignor, M., & Poloczanska, E. (2022). Climate Change 2022 : Impacts, Adaptation and Vulnerability Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change: Impacts, Adaptation and Vulnerability. Cambridge University Press. https://doi.org/10.1017/9781009325844

Ramus, C. A., & Montiel, I. (2005). When Are Corporate Environmental Policies a Form of Greenwashing? Business & Society, 44(4), 377–414. https://doi.org/10.1177/0007650305278120

Ruberg, N. (2021). BERT goes sustainable: an NLP approach to ESG financing. UNIVERSITÀ DI BOLOGNA.

sb-ai-lab. (2023). ESGify [Cloud]. Huggingface. https://huggingface.co/ai-lab/ESGify/

Scheitza, L., & Busch, T. (2023). SFDR Article 9: Is It All About Impact? https://doi.org/10.2139/ssrn.4505637

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021, March 29). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. http://arxiv.org/pdf/2103.15348.pdf

Sherwood, M. W., & Pollard, J. (2023). Responsible investing: An introduction to environmental, social, and governance investments (Second edition). Routledge, Taylor & Francis Group. https://doi.org/10.4324/9781003213666

Silva Lokuwaduge, C. S. de, & Silva, K. M. de (2022). ESG Risk Disclosure and the Risk of Green Washing. Australasian Accounting, Business and Finance Journal, 16(1), 146–159. https://doi.org/10.14453/aabfj.v16i1.10

Simionescu, L. N., Gherghina, Ș. C., Sheikha, Z., & Tawil, H. (2020). Does Water, Waste, and Energy Consumption Influence Firm Performance? Panel Data Evidence from S&P 500 Information Technology Sector. International Journal of Environmental Research and Public Health, 17(14). https://doi.org/10.3390/ijerph17145206

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020, April 20). MPNet: Masked and Permuted Pre-training for Language Understanding. https://arxiv.org/pdf/2004.09297

Statista. (2023). Top 100 Unternehmen: Europäische Union | Statista. https://de.statista.com/statistik/studie/id/50705/dokument/top-100-unternehmen-eu/

TCFD. (2017). Recommendations of the Task Force on Climate-related Financial Disclosures. TCFD. https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf

Tonidandel, S., Lebreton, J. M., & Johnson, J. W. (2009). Determining the statistical significance of relative weights. Psychological Methods, 14(4), 387–399. https://doi.org/10.1037/a0017735

Torelli, R., Balluchi, F., & Lazzini, A. (2019). Greenwashing and Environmental Communication: Effects on Stakeholders' Perceptions. SSRN Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.3470659

Tsang, A., Frost, T., & Cao, H. (2023). Environmental, Social, and Governance (ESG) disclosure: A literature review. The British Accounting Review, 55(1), 101149. https://doi.org/10.1016/j.bar.2022.101149

United Nations. (2004). Who Cares Wins : Connecting Financial Markets to a Changing World [Press release]. https://www.unepfi.org/fileadmin/events/2004/stocks/who_cares_wins_global_compact_2004.pdf

United Nations. (2015, December 13). Paris Agreement [Press release]. https://unfccc.int/documents/9096

Unstructured Technologies. (2023). Unstructured documentation. https://unstructured-io.github.io/unstructured/

van der Jerome, E. (2021). Extracting ESG data from business documents. Ecole polytechnique de Louvain.

Varini, F. S., Boyd-Graber, J., Ciaramita, M., & Leippold, M. (2020, December 1). ClimaText: A Dataset for Climate Change Topic Detection. https://arxiv.org/pdf/2012.00483

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention Is All You Need. http://arxiv.org/pdf/1706.03762

Wackernagel, M., & Galli, A. (2007). An overview on ecological footprint and sustainable development: a chat with Mathis Wackernagel. International Journal of Ecodynamics, 2(1), 1–9. https://doi.org/10.2495/ECO-V2-N1-1-9

Widianingsih, L. P., Kohardinata, C., & Vlaviorine, E. (2024). Renewable Energy Consumption, ESG Reporting, and Fixed Asset Turnover: Does it Work in Asia? International Journal of Energy Economics and Policy, 14(1), 552–558. https://doi.org/10.32479/ijeep.15325

Yang, Y., UY, M. C. S., & Huang, A. (2020, June 15). FinBERT: A Pretrained Language Model for Financial Communications. https://arxiv.org/pdf/2006.08097

Yu, E. P., van Luu, B., & Chen, C. H. (2020). Greenwashing in environmental, social and governance disclosures. Research in International Business and Finance, 52, 101192. https://doi.org/10.1016/j.ribaf.2020.101192

Zhi Yang, T. Nguyen, Hoang-Nam Nguyen, Thi Thien Nga Nguyen, & Thi Thanh Huong Cao (2020). Greenwashing behaviours: Causes, taxonomy and consequences based on a systematic literature review. https://doi.org/10.3846/jbem.2020.13225

Ziohos, E. (2023). ESG Factors (Environmental, Social, Governance): Institutional framework and impact on credit institutions. https://dspace.lib.uom.gr/bitstream/2159/29385/4/ZiogosEvangelosMsc2023.pdf

Appendices

Appendix 1 : Python Modules

Python is a popular programming language known for its simplicity, readability, and versatility. It's widely used across various domains such as web development, data analysis, artificial intelligence, scientific computing, and more. One of the key reasons for Python's popularity is its extensive ecosystem of modules and libraries. Modules in Python are files containing Python code that define functions, classes, and variables. These modules can be imported into other Python programs, which allows programmers to developers to reuse code and organise their projects more efficiently. Complete guides to Python program development are available freely on:

https://docs.python.org/3.10/tutorial/index.html

https://www.w3schools.com/python/

In addition to the standard library, Python also supports a vast collection of open-source libraries and packages developed by the community. These libraries extend Python's functionality to tackle specific tasks and domains. To deploy the various pipelines outlined in the thesis, multiple Python modules were employed. These modules with its version and general introductions are listed in the array formatted as

module(version) : Functionality

Modules

- IOpath (0.1.8) : iopath is a lightweight I/O abstraction library that provides a common interface across storage backends.
- torch (2.1.2+cu118) : Torch package contains data structures for multi-dimensional tensors and defines mathematical operations over variety of tensors.
- portalocker (2.8,2) : Portalocker is a library to provide an easy API to file locking
- transformers (4.36.2) : Transformers provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio.
- nltk (3.8.1) : NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers.
- flair (0.13.1) :Flair is NLP model repository that provides functionality to perform named entity recognition (NER), sentiment analysis, part-of-

speech tagging (PoS).

- numpy (1.26.2) : NumPy is the fundamental package for scientific computing in Python.
- opencv-python (4.9.0.80) : Wrapper package for OpenCV python bindings.
- detectron2 (0.5) : Detectron2 is Facebook AI Research developed library that provides state-of-the-art detection and segmentation algorithms.
- pycocotools (2.0.7) : pycocotools is a large image dataset designed for object detection, segmentation, person keypoints detection, stuff segmentation, and caption generation.
- pillow (10.2.0) : pillow is a Python Imaging Library that adds image processing capabilities to the Python interpreter.
- omegaconf (2.3.0) : OmegaConf is a YAML based hierarchical configuration system, with support for merging configurations from multiple sources
- pyyaml (6.0.1) : YAML is a data serialization format designed for human readability and interaction with scripting languages. PyYAML is a YAML parser and emitter for Python.
- fvcore (0.1.5.post20221221) : fvcore is a light-weight core library that provides the most common and essential functionality shared in various computer vision frameworks developed in FAIR, such as Detectron2, PySlowFast, and ClassyVision.
- tabulate (0.9.0) : Tabulate enables printing small tables in python.
- torchvision (0.16.2) : The torchvision package consists of popular datasets, model architectures, and common image transformations for computer vision.
- psutil (5.9.7) : psutil (process and system utilities) is a cross-platform library for retrieving information on running processes and system utilisation (CPU, memory, disks, network, sensors) in Python.
- matplotlib (3.8.2) : Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- datasets (2.16.1) : Datasets is a lightweight library providing two main features : one-line dataloaders for many public datasets and efficient data pre-processing
- scipy (1.11.4) : SciPy is a collection of mathematical algorithms and convenience functions built on NumPy .
- cloudpickle (3.0.0) : cloudpickle makes it possible to serialize Python

constructs not supported by the default pickle module from the Python standard library.

- onnx (1.15.0) : ONNX is an open format built to represent machine learning models. ONNX defines a common set of operators - the building blocks of machine learning and deep learning models - and a common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers.
- pandas (2.1.4) : pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool.
- beautifulsoup4 (4.12.2) : Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.
- unstructured (0.11.6) : The unstructured library is designed to help preprocess and structure unstructured text documents for use in downstream machine learning tasks.

Appendix 2 : Experimentation

Textual NLP Pipeline

LLM fine-tuned models

The available options of pre-trained and fine-tuned large-language models in the open-source domain are vast and not limited to a specific set. In our research, we evaluated numerous models to determine their suitability for our specific use case. Due to the extensive range of available models, it is impractical to provide information on each model. Therefore only tested models are described and discussed.

ESG-BERT

ESG-BERT is a domain-specific BERT model tailored for text mining in sustainable investing. Developed by Mukut Mukherjee, Charan Pothireddi, and Parabole.ai, this language model offers a specialised approach to understanding and analysing textual data within the realm of sustainable investing. While its direct applications encompass text mining for sustainable investment purposes, the versatility of the model was deemed not compatible with the research objective and text classification task. The model exhibited higher biases in comparison and several hallucinations were noted in assigning classification labels.

FinBERT

FinBERT is also BERT based fine-tuned model tailored to identify sentiments in financial text. Although the model performs adequately in sentiment classification, it needed further fine-tuning to fit the user case undertaken in this thesis. Since the computational resources were limited, the model was not chosen for further scrutiny.

CPU_Transport_GHG_Classifier

This model is fine-tuned on MPNet-base-v2 large language model by utilising a dataset provided by Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). It serves as a multi-class text classifier specifically designed to analyse text sourced from national climate policy documents. The model assigns one of three classes, ‘GHG,’ ‘NOT_GHG,’ or ‘NEGATIVE,’ to passages from the documents based on their alignment with GHG-related targets. While the model performs adequately in classification of GHG emissions related text segment, the overall objective and expectations from the model was deemed not fit for the adopted NLP pipeline.

Tabular Analysis

Zero-shot learning

Zero-shot learning is a method in machine learning where a model is trained to understand and extract information from data that it hasn't been explicitly trained on. In simple terms, zero-shot learning for parsing tables works by training a model on labelled examples of table structures and associated information. The model learns to understand the general layout and patterns of tables, as well as the types of information typically found within them. During training, the model is provided with examples of tables and the information contained within them, along with annotations that describe the structure of the tables and the types of data they contain. The model learns to recognise patterns in the data and associate them with the corresponding labels.

To train the model, a small portion of the tabular data was annotated for a transformer based model provided by SpaCy. After a few iterations, the results observed from the Zero-shot learning models were sub-par at best.

Random Forest Classification

Similar to Zero-shot learning methodology, random forest classifier is used to train a model to understand and extract information from data that it hasn't been explicitly trained on. This machine learning methodology is applied in following phases : training phase, feature selection, ensemble of decision trees, voting mechanism and information extraction. Without dwelling into much details basic functionality and process flow is described below.

- Training Phase: In this phase, the random forest classifier is trained using a labelled dataset of tables according to research purpose.
- Feature Selection: Before training, the classifier selects a subset of features (rows or columns) from the tables that are most informative for predicting the labels.
- Ensemble of Decision Trees: The random forest classifier is made up of an ensemble of decision trees. Each decision tree is trained independently on a random subset of the training data.
- Voting Mechanism: During prediction, each decision tree in the random forest independently predicts the label for a given table. The final prediction is then determined by a majority vote among all the decision trees.
- Information Extraction: Once the decision tree classifier is trained, it can be used to parse new tables and extract information from them.

After the training with various degree of labelled data, the random forest classifier was unable to detect the multi-hierarchical tabular structure common in the ESG reports. Since there are no open-source NLP models or classifiers are available to parse complex tabular structure, a simpler pattern matching approach was employed which performed better than the models fine-tuned using zero-shot learning or random forest classifier.