INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Master's Thesis

University of Applied Science - Online

Study-branch: Master of Science Data Science

# Natural Language Processing for studying sustainability reports according to the Corporate Sustainability Reporting Directive

Mohammad-Marko Jibrini

Matriculation number: 32103720

79 Dryden Road, Gateshead, United Kingdom, NE9 5TR

related code: github link

Advisor: Prof. Frank Passing

Delivery date: 07.08.2023

# Contents

# I List of Figures

# II List of Tables

# 1 Introduction

In today's rapidly evolving world, sustainable development has become a pivotal concern for governments, businesses, and individuals. With the threats of climate change, resource depletion, environmental degradation, poverty, and social and economic inequalities, it has become evident that the unchecked status quo is unsustainable. In order to tackle any issue, we first need to understand it.

Sustainability reporting offers insight into the effects of large companies beyond their corporate activity. In the first part of this report, we will look into the motivation and historical development behind sustainability reporting. Concluding with the upcoming Corporate Sustainability Reporting Directive (CSRD).

As the number of companies publishing sustainability reports rises, analyzing them and extracting useful information becomes more expensive and time-consuming. Very few organizations have the required resources for the task, leading to issues with transparency and accessibility. This leads to the research question of this project:

"How large language models can be leveraged to assist with analyzing sustainability reports?"

Furthermore, the relevant framework for this thesis is the European Sustainability Reporting Standards (ESRS). Those standards will apply for 2024, with the first reports due in January 2025. This means that no sustainability reports explicitly written for this framework are available. Furthermore, there are no labeled datasets for classification tasks in this framework. Therefore, another question arises:

"What methodology can be used to build a text classifier with an unsupervised dataset?"

# 2 Sustainability reporting

## 2.1 Brief history of sustainability

We start by providing motivation and early developments that lead to sustainability reporting. At first, we will loosely follow the structure of ”The historical development of sustainability reporting: a periodic approach” [Goktenz et al., 2020], but diverge more and more as we approach the current state.

### 2.1.1 Motivations

The beginnings of sustainability reporting can be traced as far back as the 1960s. In 1962 Rachel Carson published ”Silent Spring”[Carson, 1962](ref silent spring).In this book, the author studies the effects of the widespread use of pesticides on human and animal health. She concludes that pesticides do not target pests exclusively, that Dichlorodiphenyltrichloroethane (DDT) can harm humans without direct exposure, and that biological (sustainable?) alternatives are needed. This book marks the effective beginning of ”toxic discourse” [Christie and Tansey, 2002]. It shows that decisions made solely to maximize growth/profit can have harmful environmental effects.

In 1966 Kenneth E. Boulding published the article ”Economics of the Coming Spaceship Earth”[Boulding, 1966]. In this article, Boulding argues that for most of history, humankind saw Earth as a virtually limitless plain or an open system. If a resource is lacking here, there would be more over there. If living conditions in a particular place got difficult, we could simply move elsewhere.

Gradually humanity transitioned to the idea of a spherical earth or a closed system. Especially after World War II, the global nature of the planet was unavoidable. Boulding emphasizes material, energy, and information interconnectivity in a closed system. He uses the spaceship metaphor to describe this closed-earth economy, where space and resources are limited, and one cannot escape from the unintended consequences of his actions. This leads to economics, where one must examine environmental and social effects to build a sustainable future. Bowling later suggests tax penalties for social damages as one way to address those issues.

In 1968 a team of economists, mathematicians, and other researchers was formed at the Massachusetts Institute of Technology (MIT) to study the relationship between economic development and the environment. The results were published in the 1972 report ”The Limits to Growth”[Meadows et al., 1972]. They examined five key factors- population increase, agricultural production, nonrenewable resource depletion, industrial output, and pollution generation and their interactions. Using computer models, they concluded that the global natural system cannot support the present economic growth rates in the long term. In other words, to create a sustainable society, humankind must self-impose economic growth limits.

The report was not ambiguously accepted and received its share of criticisms. In 1973 ”Thinking about the Future: A Critique of the Limits to Growth,” [Cole et al., 1973] points out that ”The Limits to Growth” ignores technological progress and the human’s ability to creatively problem solve. Other criticisms include overstating the scarcity of renewable energy sources and the potential of recycling.

So hopefully, we can reformulate the conclusion of ”The Limits to Growth” to a more positive one- To sustain healthy economic growth, we need to invest in new technologies, renewable energy, and a circular economy.

The growing concerns about the degradation of the environment led to the formation of The United Nations Environment Program (UNEP) at the United Nations Conference on Human Environment in Stockholm in 1972. This organization sought to bring global participation to the environmental challenges.

### 2.1.2 Strategy and standardization

In the late 1970s, there was a growing awareness of environmental dangers associated with economic activity. Research and popular books have been published on the topic, and the UNEP was formed, but there still needed to be concrete goals and methodologies for this task. One early attempt to set goals comes from the UN’s 1980 World Conservation Strategy [Kassas et al., 1980]. It sets the following goals:

1. To maintain essential ecological processes and life-support systems

2. To preserve genetic diversity

3. To ensure the sustainable utilization of species and ecosystems

To achieve its goals, The world Conservation Strategy outlines the following strategic principles:

1. **Integrate.** Use a cross-sectoral, interdisciplinary approach

2. **Retain options.** Be aware that our knowledge of ecosystems is lacking. Leave open options that will fit the newest research.

3. **Mix cure and prevention.** Adress both current and impending problems

4. **Focus on causes as well as symptoms.**

Furthermore, the World Conservation Strategy lists the main obstacles as the absence of conservation at the policy-making level; lack of environmental planning and rational use allocation; poor legislation and organization; lack of training and basic information; lack of support for conservation; and lack of conservation-based rural development. This report introduced the idea of sustainable development.

In 1983 the United Nations formed the: ”World Commission on Environment and Development” (WCED), also known as the ”Brundtland Commission,” after its leading member and former Norwegian Prime Minister Gro Harlem Brundtland. This commission is tasked with developing long-term strategies to ensure sustainable development in 2000 and beyond. The commission mandate was as follows:

1. to re-examine the critical issues of environment and development and to formulate innovative, concrete, and realistic action proposals to deal with them

2. to strengthen international cooperation on environment and development and to assess and propose new forms of cooperation that can break out of existing patterns and influence policies and events in the direction of needed change; and

3. to raise the level of understanding and commitment to action on the part of individuals, voluntary organizations, businesses, institutes, and governments.

The Brundtland Commission's research results were published in 1987 in a report titled- ”Our Common Future: Report of the World Commission on Environment and Development” [WCED, 1987]. In this report, the authors redefine the limits on growth imposed by sustainable development. Here the limits are not absolute but temporary, based on technology, social organization, and the ability of the biosphere to absorb human activity. Furthermore, the report adds a social dimension to the issue of sustainable development. For example, ”Our Common Future” addresses the effects that limiting economic growth could have on poorer nations. It emphasizes international cooperation to solve this and other problems.

## 2.2 Early reporting

In 1989, an oil supertanker Exxon Valdez spilled 37,000 tonnes of crude oil off the coast of Alaska. The spill left roughly 1500 km of coastline oiled in varying degrees. This accident caused extensive ecological harm, some of which is still relevant today. The spill was highly publicized. The public was saddened by the environmental damage and felt that important public trust had been broken [Shaw, 2009]. The public pushed for every possible clean-up effort regardless of cost or effectiveness.

As a result of the spill, the Exxon Shipping Company paid over 3.8 billion dollars for cleaning, compensation, and restoration. Investors did not consider this cost. The Exxon Valdez accident marks a turning point where a company's environmental impact can have significant financial consequences.

In the aftermath of the Exxon Valdez Accident, ”The Coalition for Environmentally Responsible Economies” (CERES) was established in 1989 by a group of environmentalists and investors. They sought to reassess the role of businesses as economic, environmental, and social entities. CERES published Valdez Principles [CERES, 1989], which aims to establish ethical, environmental behavior in business activities. The Valdez Principles were modeled after The Sullivan principles, which aimed to discourage South African investment, to protest the apartheid. CERES states its ten principles as follows: ”(1) protection of the biosphere, (2) sustainable use of natural resources, (3) reduction and disposal of wastes, (4) energy conservation, (5) risk reduction, (6) safe products and services, (7) environmental restoration, (8) informing the public, (9) management commitment, and (10) audits and reports.

### 2.2.1 Earth summit

In 1992, the Sustainable Development Commission was formed during the United Nations Conference on Environment and Development, also known as Earth Summit. This commission focuses on efforts to measure sustainability. During this summit, Agenda 21 was born. Agenda 21[UN, 1992] is an action plan developed by the United Nations and national governments that outlines how governments can combine poverty and pollution and develop while conserving national resources.

Agenda 21 is divided into four sections:

1. SECTION I. Social and economic dimensions
2. SECTION II. Conservation and management of resources for development
3. SECTION III. Strengthening the role of major groups
4. SECTION IV. Means of implementation

Following the Earth Summit 1993, John Elkington published ”Coming Clean: The Rise and Rise of the corporate environment report.” Here Elkington introduces the idea of environmental reporting. It is a voluntary report in which a business discloses its environmental impact to shareholders and investors. Later on, in 1998, John Elkington published ”Accounting for the Triple Bottom Line,” where he developed the three-dimensional measure for sustainability: people, planet, and profit. The TBL approach is also called the 3P measurement system.

CERES established the Global Reporting Initiative in 1997. This body aims to develop an environmental reporting framework. The following year, a steering committee of different stakeholders was formed under the GRI to define the framework's scope. It expanded the reporting framework to include economic and social impact and environmental concerns. In the following sections, we will first discuss the goals defined by the United Nations and then the frameworks for sustainability reporting.

## 2.3 Millennium Development Goals

Following the Earth Summit in 1992, development goals were also being formed. In 2000 The ”Millennium Summit” was held at The United Nations headquarters in New York City. This was among the largest gatherings of world leaders to date. The Millennium Development Goals (MDGs) were agreed upon during this meeting. The eight development goals are:

1. Eradicate extreme poverty and hunger
2. Achieve universal primary education
3. Promote gender equality and empower women
4. Reduce child mortality
5. Improve maternal health
6. Combat HIV/AIDS, malaria, and other diseases
7. Ensure environmental sustainability
8. Develop a global partnership for development

In 2015 the United Nations published ”The Millennium Development Goals Report” [UN, 2015a] to assess the progress toward the MDGs. According to the report, there has been unprecedented progress since 2000. Millions of lives were saved, and living conditions were improved for many more. On the other hand, there were uneven achievements and shortfalls in many areas.

Here is a quick summary of the main achievements in the report:

1. The number of people living in extreme poverty has declined by more than half. Furthermore, the proportion of undernourished people in developing regions has fallen by almost half.
2. The primary school net enrolment rate in the developing region has increased to 91% from 83% in 2000. The literacy rate of youth (15 to 24) has increased from 83% in 1990 to 91% in 2015
3. There are more girls in school, and the developing regions as a whole have achieved the targets to eliminate gender disparity in education. Women make up more of the paid workers outside of the agricultural sector. More women are in parliamentary representation in nearly 90% of the countries that report data. The average population of women has nearly doubled.

4. The global under-five mortality has reduced from 90 to 43 deaths per 1000 live births between 1990 and 2015. Since the early 1990s, the rate of reduction of under-five mortality has nearly tripled. About 84 % of children received at least one dose of the measles-containing vaccine in 2013, up from 73% in 2000. The number of globally reported measles cases declined by 67% in that period.

5. Maternal mortality ratio has declined by 45% since 1990. More than 71% of births were assisted by skilled health personnel, up from 59% in 1990.

6. New HIV infections fell by approximately 40% between 2000 and 2013. By June 2014, 13/6 million people living with HIV received antiretroviral therapy, an increase from 800,000 in 2003. Over 6.2 million malaria deaths have been averted between 2000 and 2015, and the global malaria incident rate has fallen by an estimated 37 % and mortality by 58%. The tuberculosis mortality rate fell by 45 % and the prevalence rate by 41% between 1990 and 2013, saving an estimated 37 million lives.

7. Ozone-depleting substances have been virtually eliminated, and the ozone layer is expected to recover by the middle of this century. In 2015, 91% of the population used improved drinking water, up from 76% in 1990. Worldwide, 2.1 billion people have gained access to improved sanitation. The proportion of the urban population living in slums in the developing regions fell from approximately 39.4 % to 29.7%.

8. Official development assistance from developed countries increased by 66%. In 2014, 79% of imports from developing to developed countries were admitted duty-free, up from 65% in 2000. As of 2014, 95% of the world population is covered by mobile-cellular signals. Internet penetration has grown from 6% to 43%, meaning 3.2 billion people are linked through this network.

On the other hand, progress has been uneven across regions. Millions are being left behind, especially the most disadvantaged. Gender inequality persists. Significant gaps exist between the poorest and wealthiest households and rural and urban areas. Climate change and environmental degradation undermine progress. Conflict remains the biggest threat to human development. Moreover, millions of poor people still live in poverty and hunger without access to essential services.

Furthermore, there are criticisms of the MDGs framework itself rather than extrinsic issues. Authors of ”Limitations of the Millennium Development Goals: a literature review” [Fehling et al., 2013] conducted a literal review of journals, news articles, editorials, and book chapters related to MDGs. The authors identified four potential challenges relating to the MDGs framework:

1. Limitations in the MDGs development process. In this section, the authors of Limitations of the Millennium Development Goals critique the MDGs’ formulation. The focus is on who identified the goals and targets and how the political agenda can influence the chosen goals. In ”The Millennium Development Goals: A Critique from the South” [Amin, 2006], the author claims that the initiative for the MDGs did not come from the South. Instead, it was pushed by the United States, Europe, and Japan and backed by the World Bank, the International Monetary Fund, and the Organization for Economic Cooperation and Development. This leads the author to question whether the MDGs are little more than a cover for a neoliberal agenda.

   From the opposite direction comes the criticism that the gender targets were restricted due to objection from Japanese representative [Eyben, 2006]. Furthermore, the [Hulme, 2010] claim that the Vatican and conservative Islamic states removed reproductive health goals.

Furthermore, [Langford, 2010] claims that there was a goal for affordable water, which was dropped to allow for the privatization of the sector.

2. Limitations in the MDG structure Another common criticism of the MDGs is that the goals are "overambitious" or "unrealistic," particularly when considering local capacities. This is especially true for low- and middle-income countries, while the opposite can be said for other countries. The list of goals approach was also criticized as it needs to include essential issues. According to [Zahar and Boerma, 2010], using goals and targets that are country specific gives little consideration to national baselines, contexts, and implementation capacities. Furthermore, [Norren and E., 2012] criticizes the lack of interconnectedness of the goals.

   Another point of weakness is the need for more accountability for each MDGs (except Goal 8). Furthermore, the insufficient involvement and participation of developing countries in drafting the goals led to a lack of national ownership of the MDGs

3. Limitations in the MDG content Here the authors of Limitations of the Millennium Goals list the omissions of the MDGs. Some of the omissions are: reducing inequality within and between countries [Fukuda-Parr, 2010], missing focus on the poorest by using metrics such as national averages [Bricki and Holder, 2006], objectives for gender-based violence, lack of political and human rights targets [Ziai, 2011], climate change objectives. Furthermore, some authors criticize present MDGs as insufficient or unfair.

4. Limitations in the MDG implementation and enforcement Last but not least, we discuss the limitations in implementing and enforcing MDGs. This is the most relevant criticism of this paper as it deals with the lack of availability and reliability of the data regarding implementing the MDGs. The 1990 health-related baselines are often based on unreliable household surveys. The education enrollment data is often based on the beginning of the academic year and does not account for dropouts. Quantitative targets rely on tools that many countries lack. Furthermore, there needs to be more explicit guidance on policy changes that will achieve the MDGs.

   The authors also criticize the MDG framework for promoting short-term fixes to reach targets instead of long-term and structural change.

Overall it will always be a challenge to agree on a framework and a list of goals everyone can agree on. Sometimes, we have valid and opposing critiques on the same issue, where addressing one criticism could make the other worse. Furthermore, while there was undeniable progress in many of the goals, in some cases, whether there is a causation between the setting of the MDGs and the progress achieved is unclear. For example, the widespread of the internet would probably happen even without Goal 8.

## 2.4 Sustainable Development Goals

The outcome of the 2015 general assembly of the United Nations, where successes and shortcomings of the MDGs were discussed, was Agenda 2030 [UN, 2015b]. This marked the end of the MDGs and the beginning of the Sustainable Development Goals (SDGs). According to Agenda 2030, the 17 Sustainable and the 169 targets seek to build on the Millennium Development Goals and complete what they still need to achieve. All 17 SDGs can be seen in figure 2.1. In the MDGs, sustainability has been reduced to one of the eight goals (goal 7: "ensure environmental sustainability"), but sustainability takes center stage in the SDG. SDGs integrate social, ecological, and economic aspects of sustainability. Authors of

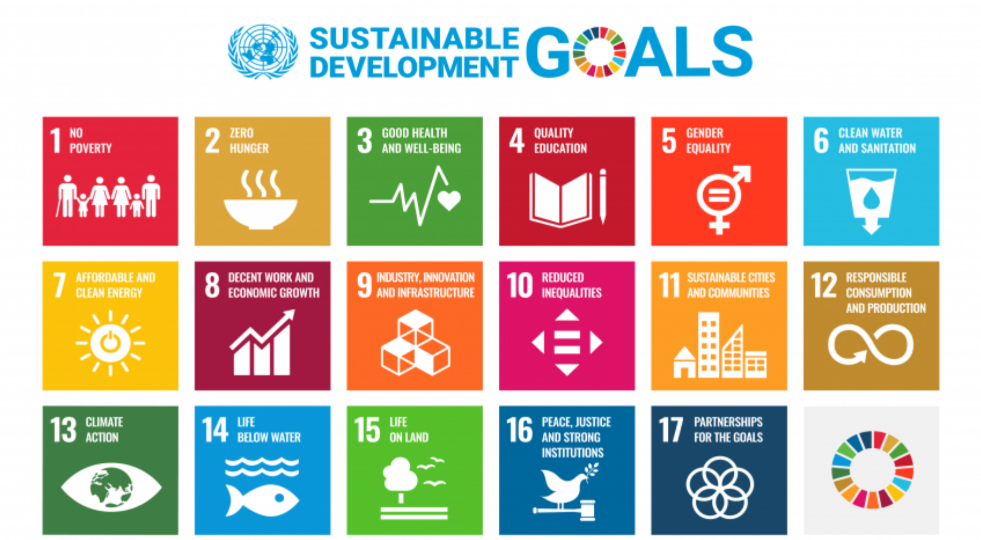
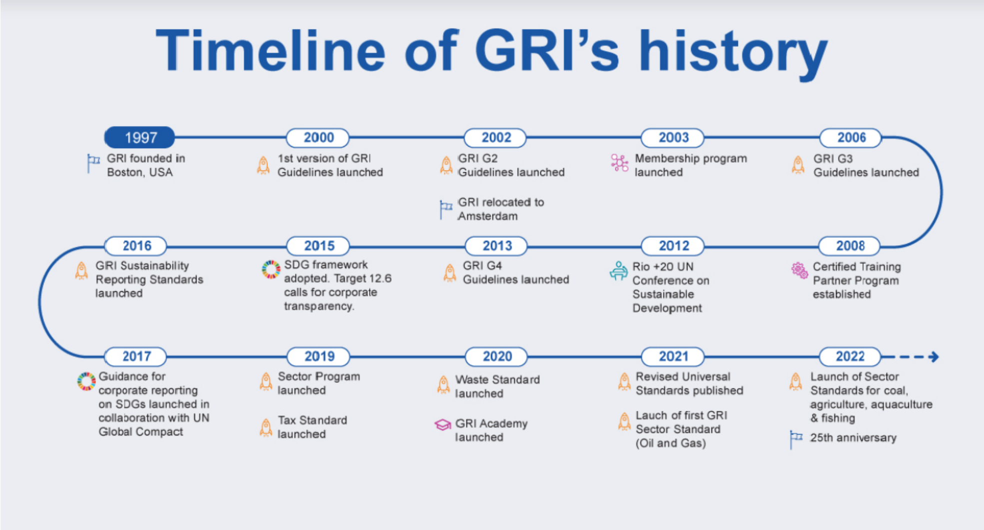

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Figure 2.1: Sustainable Development Goals

”From Millennium Development Goals to Sustainable Development Goals” [Kannengißer, 2023] claim that this change from in focus from development to sustainability requires a paradigm shift in the research of this field.

## 2.5 GRI Standards

At the turn of the millennium, the GRI became an independent organization and developed G1 guidelines. It then published G2, G3, and G4 guidelines in 2002, 2006, and 2013, respectively. Finally, in 2016 the GRI published the first set of global standards for sustainability reporting [GRI, 2016]. Figure 2.2 shows the history of GRI from its founding in 1997 to 2022.



Timeline of GRI’s history

1997 GRI founded in Boston, USA

2000 1st version of GRI Guidelines launched

2002 GRI G2 Guidelines launched; GRI relocated to Amsterdam

2003 Membership program launched

2006 GRI G3 Guidelines launched

2008 Certified Training Partner Program established

2012 Rio +20 UN Conference on Sustainable Development

2013 GRI G4 Guidelines launched

2015 SDG framework adopted. Target 12.6 calls for corporate transparency.

2016 GRI Sustainability Reporting Standards launched

2017 Guidance for corporate reporting on SDGs launched in collaboration with UN Global Compact

2019 Sector Program launched; Tax Standard launched

2020 Waste Standard launched; GRI Academy launched

2021 Revised Universal Standards published; Lauch of first GRI Sector Standard (Oil and Gas)

2022 Launch of Sector Standards for coal, agriculture, aquaculture & fishing; 25th anniversary

Figure 2.2: History of GRI

The 2016 GRI Standards are a modular system comprising three series of Standards: the GRI Universal Standards, the GRI Sector Standards, and the GRI Topic Standards. The three GRI universal standards apply to every company. GRI 1: *Foundation* outlines the purpose, clarifies concepts, and lists requirements needed for a report to meet GRI standards. GRI 2: *General Disclosures* gives insight into an organization’s scale and impact and relates to disclosure structure, reporting, and governance.GRI

3: *Material Topics* explains how an organization can determine its Material Topics. Here Material Topics are defined as "topics that represent the organization's most significant impacts on the economy, environment, and people, including impacts on their human rights." Then the organization uses the Sector Standards that apply to its sector to determine relevant material topics and what information to report. Finally, the Topic Standards inform the organization on what information it needs to report concerning a particular topic. Figure 2.3

As of January 2022, over 10,000 companies produce GRI reporters, including 70% of the Global Fortune 500 companies[GRI, 2022].

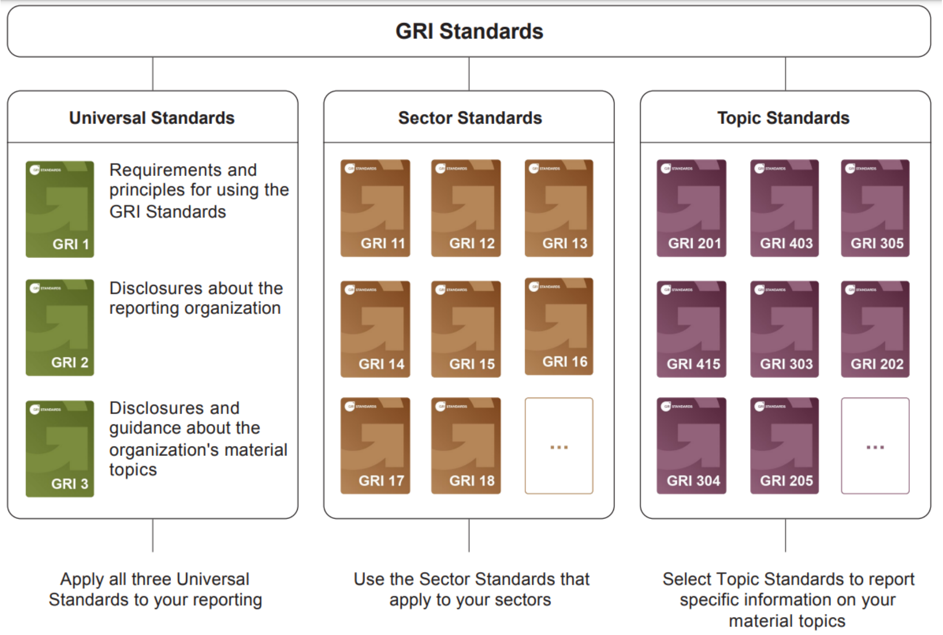

GRI Standards
Universal Standards
Sector Standards
Topic Standards
GRI 1
Requirements and principles for using the GRI Standards
GRI 2
Disclosures about the reporting organization
GRI 3
Disclosures and guidance about the organization's material topics
GRI 11
GRI 12
GRI 13
GRI 14
GRI 15
GRI 16
GRI 17
GRI 18
...
GRI 201
GRI 403
GRI 305
GRI 415
GRI 303
GRI 202
GRI 304
GRI 205
...
Apply all three Universal Standards to your reporting
Use the Sector Standards that apply to your sectors
Select Topic Standards to report specific information on your material topics

Figure 2.3: GRI Standards

One of the issues with the GRI and other contemporary reporting frameworks is that they are "Voluntary standards." Meaning that there is no mandate for any company to disclose relevant information. It was not until 2023 that the European Union's "Corporate Sustainability Reporting Directive" (CSRD) entered into force. This is the first mandatory set of rules regarding sustainability reporting.

## 2.6 CRSD

The European Sustainability Reporting Standards (ESRS) were developed by the European Financial Reporting Advisory Group (EFRAG), an independent body that brings together various stakeholders. The ESRS starts with two Cross-Cutting standards. ESRS 1 - General requirements, a set of requirements that must be followed. This includes: "Categories of Standards and disclosures under ESRS," "Qualitative characteristics of information," "Double materiality as the basis for sustainability disclosures," "Sustainability due diligence," "Value chain," "Time horizon," "Preparation and presentation of information," "Structure of sustainability statements," "Linkage with other parts of sustainability statements," and "Transitional provisions" [EFRAG, 2023a]. And ESRS 2 - General Disclosures defines the four-pillar approach of "Governance," "Strategy," "Impact, risk and opportunity management," and "Metrics and Targets." Furthermore, it defines outcomes and double materiality assessments, minimum disclosures, and the list of mandatory data points [EFRAG, 2023b]. They are followed by ten topical standards shown below. One might notice the similarity in the structure of the ESRS with the GRI. This is not an accidence as GRI served as a reference point in the development of the ESRS, and a high alignment between the two was intended.

- **Enviromental**
  - **E1** Climate change: subtopics - E1-1 Transition plan for climate change mitigation , E1-2 Policies related to climate change mitigation and adaptation, E1-3 Actions and resources in relation to climate change policies, E1-4 Targets related to climate change mitigation and adaptation, E1-5 Energy consumption and mix, E1-6 Gross Scopes 1, 2, 3 and Total GHG emissions, E1-7 GHG removals and GHG mitigation projects financed through carbon credits, E1-8 Internal carbon pricing, E1-9 Potential financial effects from material physical and transition risks and potential climate-related opportunities
  - **E2**: Pollution : subtopics - E2-1 Policies related to pollution, E2-2 Actions and resources related to pollution, E2-3 Targets related to pollution, E2-4 Pollution of air, water and soil, E2-5 Substances of concern and substances of very high concern, E2-6 Potential financial effects from pollution-related impacts, risks and opportunities
  - **E3**: Water and marine resources: subtopics - E3-1 Policies related to water and marine resources, E3-2 Actions and resources related to water and marine resources, E3-3 Targets related to water and marine resources, E3-4 Water consumption, E3-5 Potential financial effects from water and marine resources-related impacts, risks and opportunities
  - **E4**: Biodiversity and ecosystems : subtopics - E4-1 Transition plan on biodiversity and ecosystems, E4-2 Policies related to biodiversity and ecosystems, E4-3 Actions and resources related to biodiversity and ecosystems, E4-4 Targets related to biodiversity and ecosystems, E4-5 Impact metrics related to biodiversity and ecosystems change E4-6 Potential financial effects from biodiversity and ecosystem-related impacts, risks, and opportunities
  - **E5**: Resource use and circular economy: subtopics - E5-1 Policies related to resource use and circular economy, E5-2 Actions and resources related to resource use and circular economy, E5-3 Targets related to resource use and circular economy, E5-4 Resource inflows, E5-5 Resource outflows, E5-6 Potential financial effects from resource use and circular economy-related impacts, risks and opportunities
- **Social**
  - **S1** Own workforce: subtopics - S1-1 Policies related to own workforce, S1-2 Processes for engaging with own workers and workers' representatives about impacts, S1-3 Processes to remediate negative impacts and channels for own workers to raise concerns, S1-4 Taking action on material impacts on own workforce, and approaches to mitigating material risks and pursuing material opportunities related to own workforce, and effectiveness of those actions, S1-5 Targets related to managing material negative impacts, advancing positive impacts, and managing material risks and opportunities, S1-6 Characteristics of the undertaking's employees, S1-7 Characteristics of non-employee workers in the undertaking's own workforce, S1-8 Collective bargaining coverage and social dialogue, S1-9 Diversity indicators, S1-10 Adequate wages, S1-11 Social protection, S1-12 Persons with disabilities, S1-13 Training and skills development indicators, S1-14 Health and safety indicators, S1-15 Work-life

balance indicators, S1-16 Compensation indicators (pay gap and total compensation), S1-17 Incidents, complaints, and severe human rights impacts and incidents

  - **S2**: Workers in the value chain: S2-1 Policies related to value chain workers, S2-2 Processes for engaging with value chain workers about impacts, S2-3 Processes to remediate negative impacts and channels for value chain workers to raise concerns, S2-4 Taking action on material impacts on value chain workers, and approaches to mitigating material risks and pursuing material opportunities related to value chain workers, and effectiveness of those actions, S2-5 Targets related to managing material negative impacts, advancing positive impacts, and managing material risks and opportunities
  - **S3**: Affected communities : subtopics - S3-1 Policies related to affected communities, S3-2 Processes for engaging with affected communities about impacts, S3-3 Processes to remediate negative impacts and channels for affected communities to raise concerns, S3-4 Taking action on material impacts on affected communities, and approaches to mitigating material risks and pursuing material opportunities related to affected communities, and effectiveness of those actions, S3-5 Targets related to managing material negative impacts, advancing positive impacts, and managing material risks and opportunities
  - **S4**: Consumers and end-users: subtopics - S4-1 Policies related to consumers and end-users, S4-2Processes for engaging with consumers and end-users about impacts, S4-3Processes to remediate negative impacts and channels for consumers and end-users to raise concerns, S4-4 Taking action on material impacts on consumers and end-users, and approaches to mitigating material risks and pursuing material opportunities related to consumers and end-users, and effectiveness of those actions, ES-5 Targets related to managing material negative impacts, advancing positive impacts, and managing material risks and opportunities

- **Governance**
  - **G1** Business Conduct: subtopics - G1-1 Corporate culture and business conduct policies , G1-2 Management of relationships with suppliers, G1-3 Prevention and detection of corruption or bribery, G1-4 Confirmed incidents of corruption or bribery, G1-5 Political influence and lobbying activities, G1-6 Payment practices

The Corporate Sustainability Reporting Directive (CSRD) is the new EU legislation that requires companies to publish regular reports in accordance with the ESRS framework. The first wave of companies must submit their reports on the first of January 2025 for the year 2024. The CRSD aims to help increase money flow towards sustainable activities across the European Union. It will ensure that investors and other stakeholders have access to the information they need to assess the impact of companies on people and the environment and for investors to assess financial risks and opportunities arising from climate change and other sustainability issues [Commission, 2023]. Companies meeting two of the following three conditions will have to comply with the CSRD:

1. €40 million in net turnover
2. €20 million in assets
3. €250 or more employees

In addition, non-EU companies with a turnover of above €150 million in the EU must also comply. The EU is estimating that the CSRD will apply to nearly 50,000 companies.

# 3 Methodology

The sheer number of sustainability reports presents a challenge when studying and analyzing them. It can be costly for humans to process all this information. Therefore, the idea of using large language models to aid in such analysis emerges. This section will examine how we can use natural language processing to help in this task. To do so, we integrate the CRoss Industry Standard Process for Data Mining (CRISP-DM) [Shearer, 2000] with Desing Science Research (DSR) [Henver et al., 2004]. We keep the relevance and rigor cycles of DSR while leveraging the iterative phases of CRISP-DM in the Design Cycle, see figure 3.1

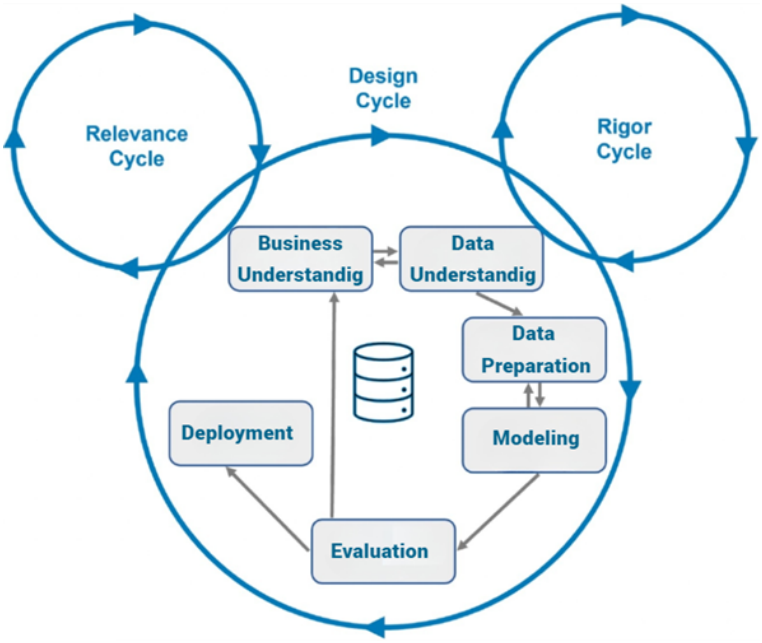


Figure 3.1: Methodology

There have been successes in labeling textual data according to the relevant SDG by fine-tuning a pre-trained BERT model for text classification [Pukelis et al., 2022]. This is an equivalent of labeling text according to its 2-level ESRS (E1, E2, E3...). For more specificity, getting level 3 labels (sub-labels) or extracting specific targets and metrics, an approach using question answering and prompt engineering might be more appropriate [Ni and et al., 2023].

There are additional challenges when analyzing reports on a framework before any reports are written specifically for that framework. Firstly, as the writers did not tailor their reports to ESRS, the reports might not be organized in a manner that corresponds to ESRS, and critical information might be missing altogether. More importantly, there is a lack of labeled data. To address this, we adopt a method similar to ”data programming” outlined in ”Data Programming: Creating Large Training Sets, Quickly” [Ratneri and et al., 2016]

## 3.1 Data Understanding

For this thesis, we use sustainability reports of German DAX companies. The DAX (German) stock index consists of the 40 major German blue chip companies trading on the Frankfurt Stock Exchange, all of which fall under the CRSD.

At this stage, we face the first challenge of studying sustainability reports. There is no comprehensive and easily accessible database for such reports. Therefore, the reports used in this study have been manually downloaded from companies' sites. While great care was taken to have as complete a dataset as possible, some reports might be missing. Out of the 40 DAX companies, no reports were found, only for Badische Anilin- und Sodafabrik (BASF).

In total, 316 reports were gathered. Some of the reports were excluded from this study for the following reasons:

1. We could not find pdf reports for Fresenius; instead, pdfs were created by printing their website sustainability information - 11 reports
2. Pdfs from the same company published in the same year. For example, a company had a corporate report that included sustainability information published the same year they had a sustainability report published - 7 reports
3. Porcshe SE published short non-financial reports that do not match the requirements of this project - 4 reports
4. Reports not compatible with text extractor ( space between lines too large, making each line appear as a separate paragraph) – 7 reports, including all reports from Commerzbank

After the exclusions, we are left with 287 reports from 36 companies (40 DAX companies minus BASF, Fresenius, Porche SE, and Commerzbank). Those reports' median number of pages is 104; the mean is 125.4. Refer to figure 3.2 for full distribution.

Interestingly, some companies have a history of publishing sustainability reports as far back as 2000. But, it is not until 2016 that we have reports from most companies. Figure 3.3 shows the number of companies publishing sustainability reports for each year starting in 2000. Henkel published a sustainability report in 2000, Deutsche Post DHL released an environmental report in 2003, and E.ON Corporate Responsibility report in 2004. For more information on the number of companies that are publishing reports for a given year, refer to 3.3

Finally, meta information on the pdfs was stored in a ”reports_links” CSV file. In the format: report path, Company, Year, status. The first three columns are self-explanatory, and status refers to whether the report will be used for the rest of the project (status=1) or why it is excluded.

## 3.2 Text extraction

Now that we have PDF reports, the next step is to extract the text from them in a usable format. For this task, the following modules were considered. PyPDF2, pytesseract, and PyMuPDF from the fltz module. A simple script was written for each consisting of the following steps to test the modules' performance. Read data from a PDF, break it down into paragraphs, discard short paragraphs, and write the results into a CSV file. The 2016 Adidas sustainability progress report was used.

PyMuPDF from the fltz approach was the fastest, with a running time of 0.4 seconds, followed by the PdfReader from PyPDF2, which took 4.9 seconds. Finally, the pytesseract took 206.5 seconds

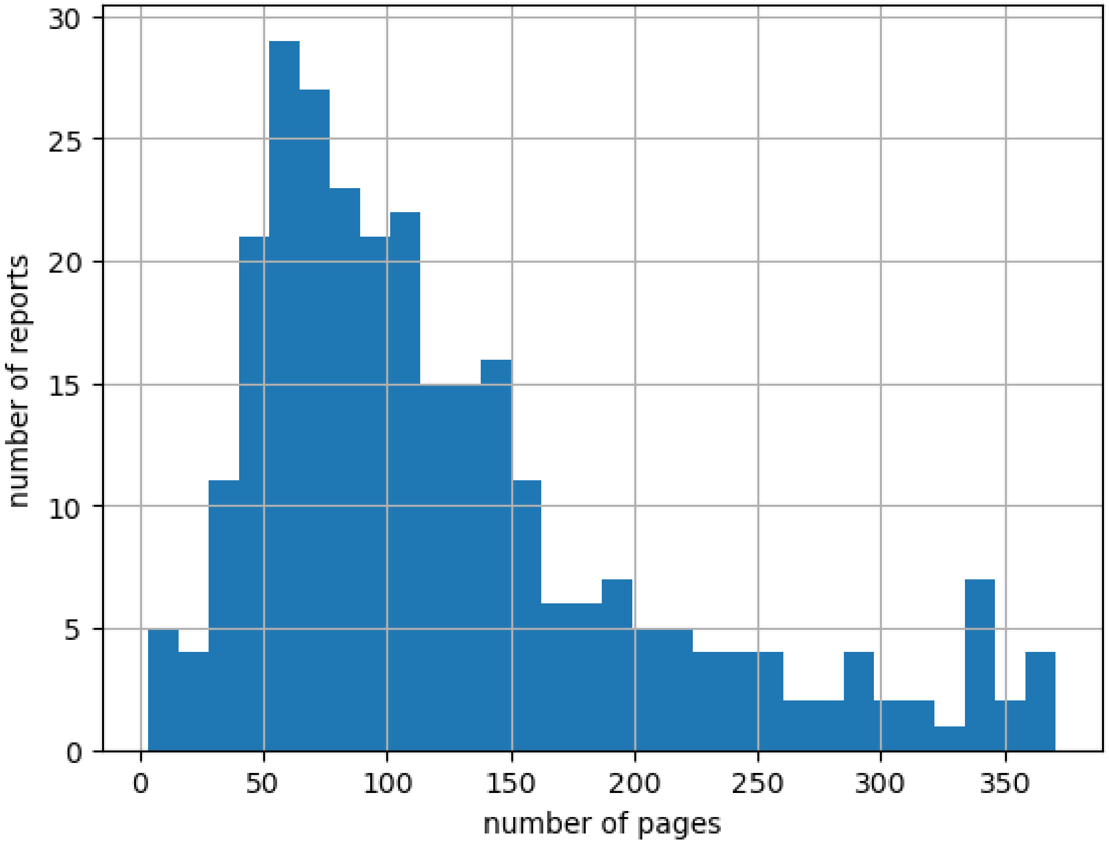


Figure 3.2: Number of pages in reports

to complete the task. Pytesseract is an optical character recognition (OCR) tool. It first converts the PDF file to images and then reads them optically, which explains its slowness. This can be very powerful when using different types of input files, for example- scanned pdfs, but it was not needed for this use-case.

PdfReader extracts the text of each page in a pdf as a string of characters. It is a simple approach and has extra utility when it comes to using password-protected files and creating pdfs.

PyMuPDF was not only the fastest performer but also had capabilities that made it particularly useful for this project. It discriminates between textual and non-textual data; text can be accessed block by block (paragraph by paragraph) or even line by line, where we get the text, font, and font size. These options were helpful for further data cleaning and, combined with their speed, made PyMuPDF the obvious choice for this project.

The extraction pipeline is outlined below:

1. Extract textual data from a pdf into a list using the Fitz package. The PDF is read page by page, block by block, and line by line. An extracted row of data represents one line and contains the following: text, font size, font, block number, page number, and block coordinates.

2. Convert the list from step 1 into Pandas DataFrame

3. Discard font sizes that appear infrequently (less than 5% of total text).

   a) Create the rounded font size column

   b) Count the frequency of each font size appearing in the pdf

   c) Keep lines with font sizes that appear more than 5% of the time

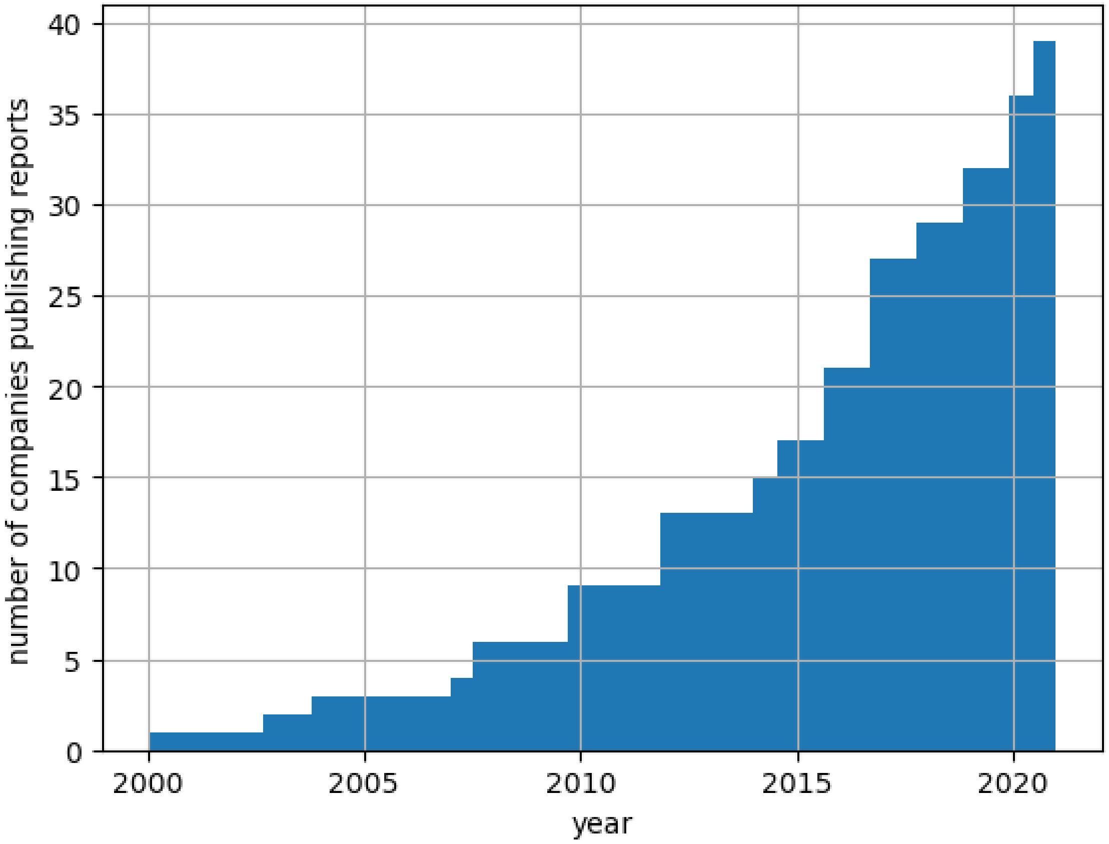


Figure 3.3: Number of companies publishing reports

4. Remove non-ASCII characters from the text
5. Combine lines of text back into paragraphs by grouping them according to block number
   a) Identify lines with the same block number
   b) Add them together while inserting a new line in between using .groupby(['block_number'])['text'].transform(lambda x: '\n '.join(x))
   c) Remove duplicates. A paragraph made of four lines would be created four times
6. Identify paragraphs that are broken by page end and combine:
   a) Identify paragraphs that do not end with a punctuation mark
   b) Scan the rest of the page and the next page for paragraphs that do not start with a capital letter
   c) combine them into one paragraph
7. Remove tables. Traditional tables will be excluded in Step 1. Here we remove table-like paragraphs by measuring the frequency of new lines
8. Add Company name and year as new columns to the data

We first read the "reports_links" CSV file to extract the textual data. Then we go through it row by row. First, we check the status; if it is 1, we run it through the pipeline outlined above and write the results into a CSV, whose name is based on the company's name and the paper's year. Figure 3.4 shows the number of paragraphs extracted each year. The total number of paragraphs extracted is

149083. The lowest number of paragraphs extracted are from Rheinmetall, with only 141, while the highest belongs to Bayer, with 13706.

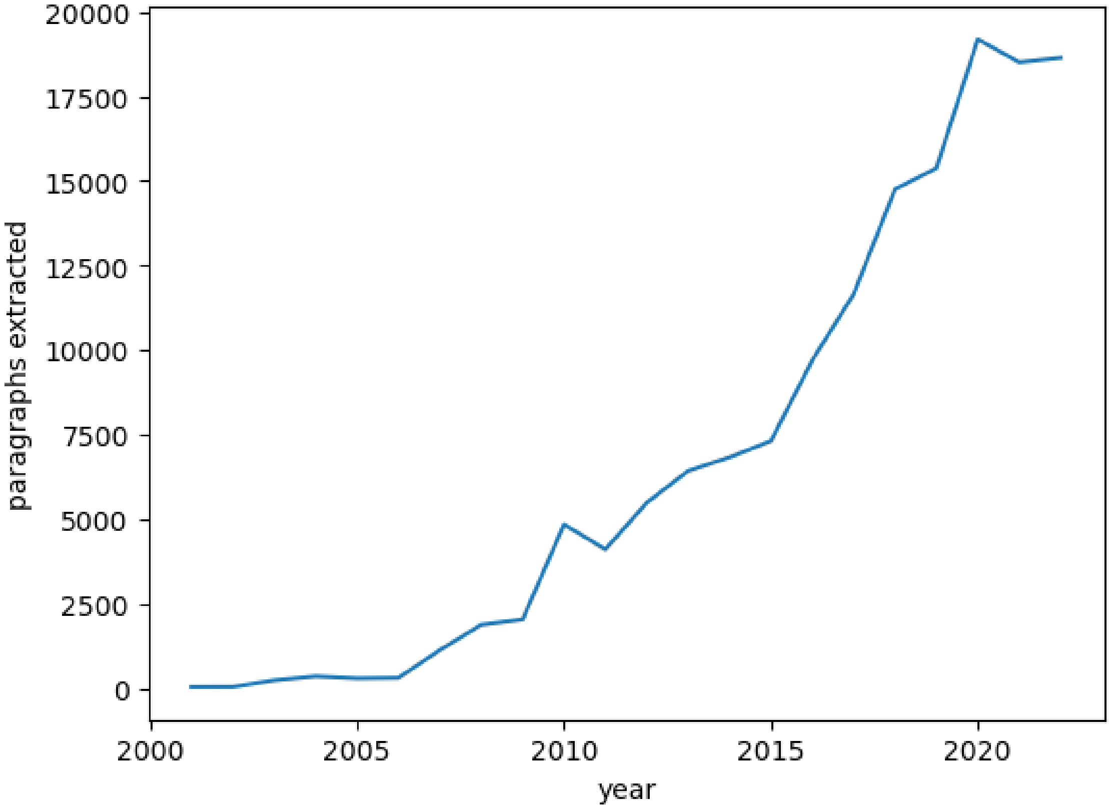


Figure 3.4: Paragraphs extracted by year

## 3.3 Data processing

Now that we have extracted our textual data, the next step is processing it and preparing it for training purposes. The major challenge here is that our data is not labeled. In order to create a labeled data set that can be used for training, we will rely on text embeddings from OpenAI to act as a zero-shot label. The results of such labeling are not very accurate, but the hope is that we can select a subset of data for which this labeling works particularly well.

### 3.3.1 ada_002 embeddings

Now that we have our data in the required format, first, we will build a simple zero-shot classifier. We achieve this by utilizing text embeddings. Embeddings aim to represent a meaning of a text by representing them as multidimensional vectors (ref). In this project, we use the state-of-the-art text-embedding-ada-00 from OpenAI. Ada-002 embeddings represent any text into a 1536-dimensional vector of floating point numbers. The distance between two vectors measures the relatedness of the text they represent. We start by writing a brief description for each of our output classes. The complete list of that description is below:

1. **E1**:"Climate Change. Transition plan, policies, actions, resources, and targets relating to climate change. Energy consumption, Scope 1, 2, 3, and total GHG emissions and Greenhouse gas removal. Carbon pricing and financial effects related to climate change."

2. **E2**:"Pollution. Policies, Actions, and Targets related to pollution. Pollution of air, water, and soil. Substances of concern and potential financial effects from pollution."

3. **E3**:"Water and marine resources. Policies, Actions, and Targets related to water and marine resources. Water consumption. Financial effects relating to water and marine resources."

4. **E4**:"Biodiversity and ecosystems. The transition plan, policies, actions, resources, and targets related to biodiversity and ecosystems. Impact on biodiversity and ecosystems. Financial effects related to biodiversity and ecosystems."

5. **E5**:"Resource use and circular economy.Policies, Actions, and Targets related to recycling. Resources inflows and Resource outflows. Financial effects related to Resource use and circular economy."

6. **S1**:"Own workforce. Policies, processes for engagement, and action to remediate negative impact on own workforce. Characteristics of employees and non-employees. Collective bargaining, diversity, adequate wage, social protection, training, and health and work-life balance indicators. Compensation and incident reporting."

7. **S2**:"Workers in the value chain. Policies, targets, processes for engagement and action to remediate negative impact on Workers in the value chain."

8. **S3**:"Affected communities. Policies and processes for engaging and remediating negative impact for Affected communities. Action and targets related to affected communities."

9. **S4**:"Consumers and end-users. Policies and processes for engaging and remediating negative impact for Consumers and end-users. Action and targets related to Consumers and end-users."

10. **G1**:"Business Conduct. Corporate culture and business conduct policies, relationship with suppliers. Prevention and detection of corruption or bribery. Political influence and lobbying. Payment practices "

Now we use the OpenAI API to give each output class its ada-002 embeddings. Next, we get the embeddings for Our test-set data and calculate the cosine similarity between our data embeddings and each output class embedding. For any given text, the output class with the highest cosine similarity score is the zero-shot classifiers result for that text. First, we check whether the zero-shot classifier represents all classes. Figures 3.5 and 3.6 show the distribution of classes in human and zero-shot labeling, respectively. We can see that some high-frequency classes are overrepresented in the zero-shot classifier, particularly G1 and E1. This leaves less frequent classes to be under-presented. Furthermore, the zero-shot classifier has no 0 or none class, as there will always be a class with the highest cosine similarity.

We do the zero-shot classification for two main reasons. First, it gives us a simple baseline model. The accuracy of this model is 0.681, and the cohen kappa score for the test set and the zero-shot classifier is 0.616. This is interpreted as a substantial agreement (although at the low end of it) between a human and the zero-shot classifier. The accuracy by class can be seen in the table 3.1 The second purpose of this exercise is more interesting. The hypothesis is that we can use a data embedding classifier to select data that is highly likely to be correctly classified and use it as training data for building a neural network. One approach is only to use data with exceptionally high cosine similarity. To test this, we used different cosine similarity thresholds and then calculated the Cohen kappa only for data with

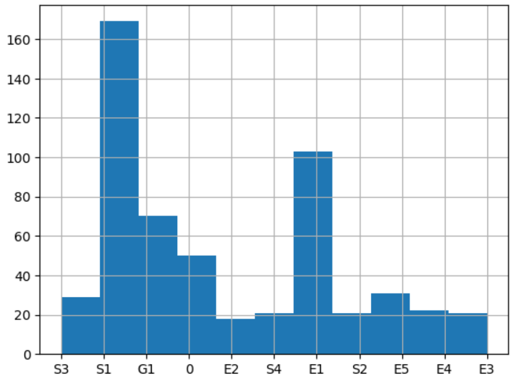


Figure 3.5: Human Labeling

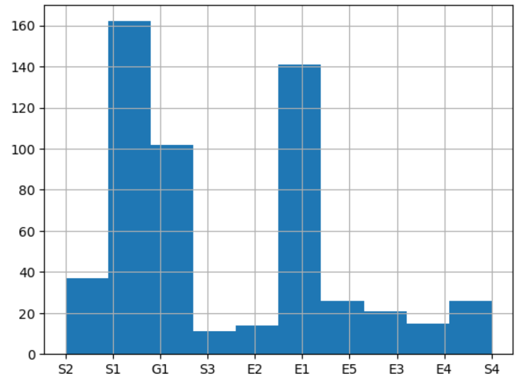


Figure 3.6: Zero Shot Labeling

| Class | Accuracy |
|---|---|
| 0 | 0.000000 |
| E1 | 0.970874 |
| E2 | 0.555556 |
| E3 | 0.809524 |
| E4 | 0.636364 |
| E5 | 0.677419 |
| G1 | 0.685714 |
| S1 | 0.822485 |
| S2 | 0.476190 |
| S3 | 0.344828 |
| S4 | 0.428571 |

Table 3.1: Accuracy by class

cosine similarity higher than this threshold. The results can be seen in 3.7. For example, if we choose only to include results with cosine similarity equal to or better than 0.82, we get the following results: Cohen kappa = 0.8, Accuracy 0.848, and data reduction = 0.189. Note that as the dataset is reduced, so is the statistical significance of the results. The table of accuracy per class for the reduced dataset is shown in 3.2

| Class | Accuracy |
|---|---|
| 0 | 0.000000 |
| E1 | 0.976190 |
| E2 | 0.500000 |
| E3 | 1.000000 |
| E4 | 0.833333 |
| E5 | 1.000000 |
| G1 | 0.818182 |
| S1 | 0.952381 |
| S2 | 0.571429 |
| S3 | 0.250000 |
| S4 | 0.000000 |

Table 3.2: Accuracy by class threshold = 0.82

Using the similarity score threshold method, we can produce more accurate data. Figures 3.8 and 3.9 show that this dataset's over-representation of high-frequency classes is reduced. Furthermore, the 0 or

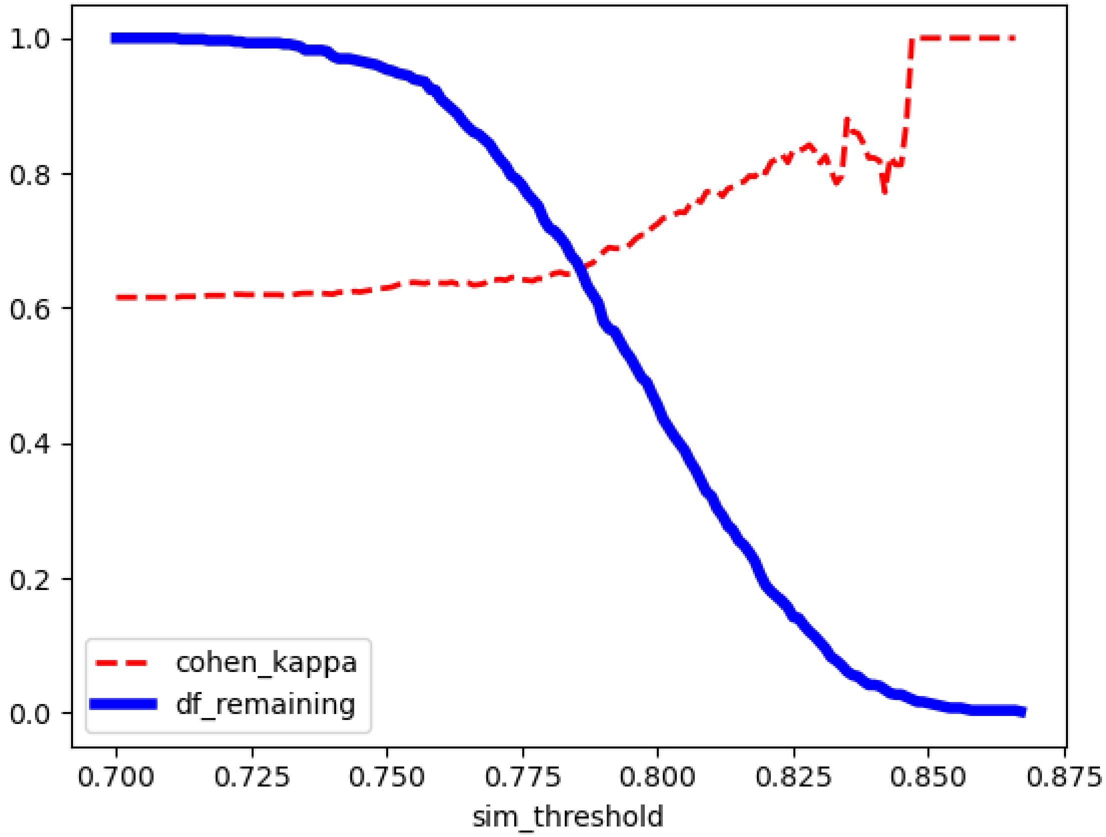


Figure 3.7: Cosine Similarity Threshold

none class is almost eliminated from the data. This makes sense as we would not expect many data points with the ”none” label to have high similarity scores with any particular class.

Data was selected based on tabel 3.3. It was then used to fine-tune ”roberta-base-finetuned-sdg,” a RoBERTa-based model by Jonas Nothnagel trained to do SDG classification. We got Cohen kappa = 0.66 and accuracy= 0.717. That shows that we can use embeddings to select training data for a model that will outperform the embeddings themselves.

| Class | number selected | selection method |
|---|---|---|
| E1 | 2000 | nlargest sim_score |
| E2 | 100 | nlargest sim_score |
| E3 | 100 | nlargest sim_score |
| E4 | 200 | nlargest sim_score |
| E5 | 200 | nlargest sim_score |
| G1 | 1500 | nlargest sim_score |
| S1 | 1000 | nlargest sim_score |
| S2 | 500 | nlargest sim_score |
| S3 | 100 | nlargest sim_score |
| S4 | 100 | nlargest sim_score |

Table 3.3: nlargest

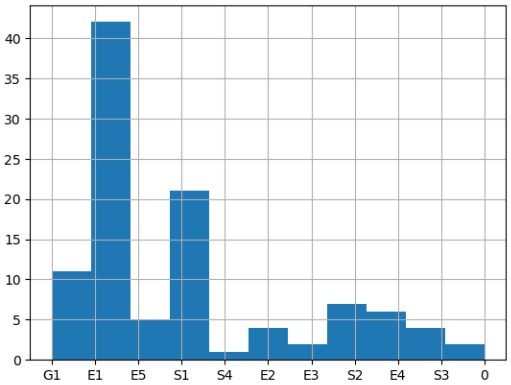


Figure 3.8: Human Labeling threshold = 0.82

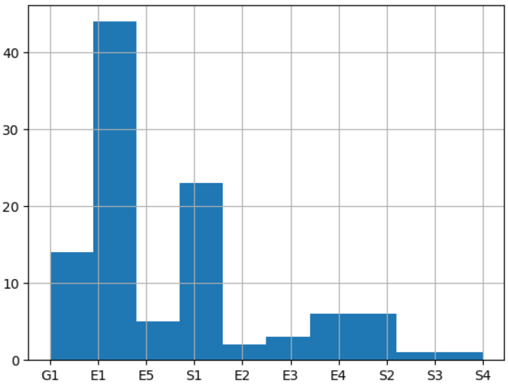


Figure 3.9: Zero Shot Labeling threshold = 0.82

### 3.3.2 Centroid Labeling

In the previous step, we compared text embeddings to our label embeddings. However, the label embeddings were not much more than embeddings of the descriptions of each class, i.e., they did not represent the actual text we would expect to find in any of the reports. To introduce a slightly different approach, we introduce centroids. To create a centroid for any given class we:

1. For each class, select data points corresponding to that class according to ada_002 classification
2. Take 1000 datapoints with the highest cosine similarity from step I
3. Calculate the average embedding of those data points.

We repeat this process for each class. The results are sudo-embeddings that do not necessarily correspond to any text but are an average of 1000 text embeddings. Those sudo-embeddings are derived from textual data itself and could give more of an indication of how companies write about any particular subject.

Then we calculate cosine similarities for each text again, but this time we compare them to the centroids instead of label embeddings. This gives us another zero-shot label. We can reproduce the process used to create figure 3.7 to get 3.10.

### 3.3.3 Well Formness of the data

While great care was taken in extracting the data, some data points do not represent paragraphs as we would expect - a sequence of one or more sentences dealing with a singular point. To measure how well any data point represents an actual paragraph, we introduce the ”Well Formedness” score. To evaluate the ”Well Formedness ” score, we use the query_wellformedness_score model [Salesken.ai, 2021] from huggingface. This is another model trained on RoBERTa. The authors define ”Well Formedness” as a non-fragment, grammatically correct sentence score. Histogram 3.11 shows the frequency of ”Well Formedness” in our data.

Now the question is, how well-formed is well-formed enough? The only way to find out is to take a look. For each 0.1 increment, we select a random sample of 10 data points. At the lower end of ”Well Formedness” between 0 and 0.1, we have text such as: ”ANTI-CORRUPTION / ANTI-FRAUD Preventing bribery, corruption and fraud through global policies and monitoring.” and ”Risk Management System 214 20.3.2 Opportunities and Risks 217 21 . Takeover-Relevant Information

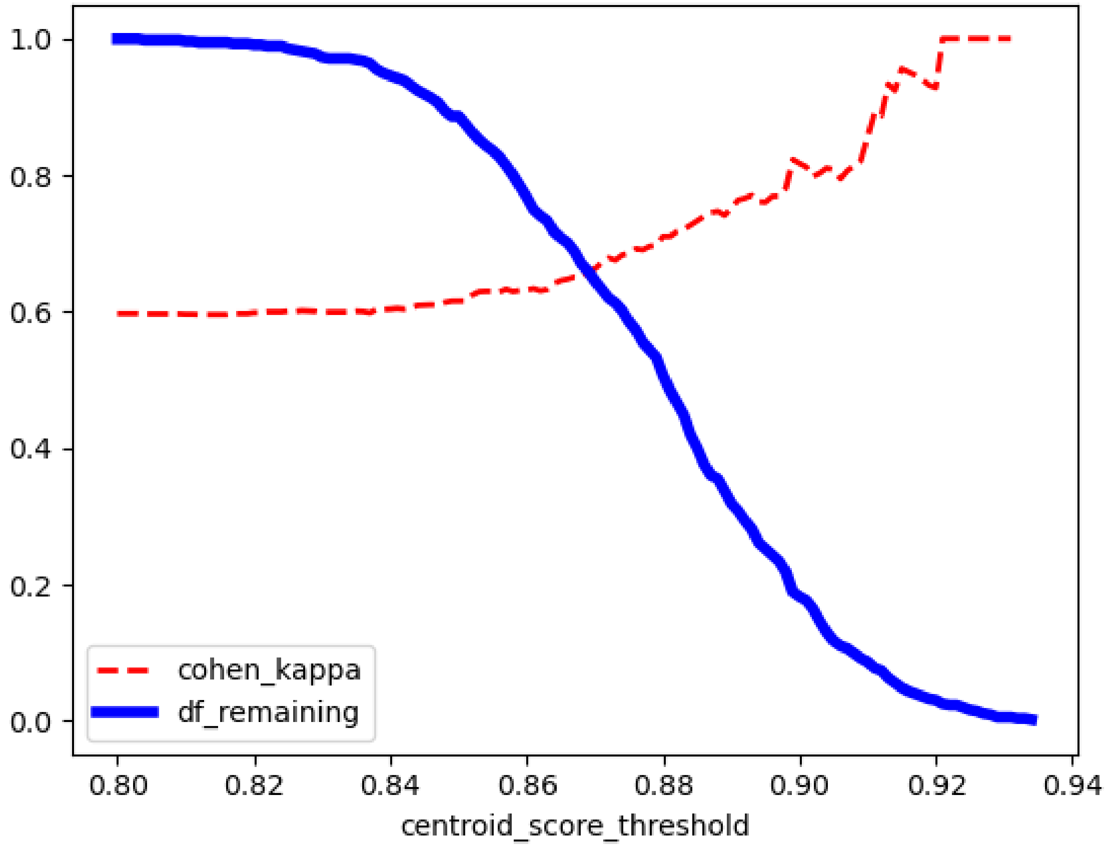


Figure 3.10: Centroid Threshold

224" which is most likely a label of a figure and should be excluded. In the 0.1 to 0.2 range, we have "Reproduction in full or in part only with the publisher prior written approval; photos and copy to be credited to Deutsche Bank AG." which should be excluded but also text like: "Hamburg In terms of energy efficiency, the Hamburg sites focus was on the comprehensive introduction of LED tech- nology. More than 12,000 lights were replaced over an area of 29,000 square metres, allowing approximately 1,200,000 kWh in electricity savings per year." which contains useful information, but perhaps a header snuck in, lowering the "Well Formedness" score. Text with scores between 0.2 and 0.4 continues this trend. Some have helpful information, while others less so. We need to consider this data if we want to extract comprehensive information on Company practices. However, here, we want to select clean training data and, therefore, can be picky. For the next section, we only use data with a Well Formedness score of 0.4 or higher. When we apply this filter, we are left with 94532 out of the original 149083.

We use the wellformness score to select training data in our model. However, it is not used in the final pipeline when analyzing reports as we don't want to miss out on any information.

## 3.4 Data Selection

Now our data has two zero-shot labels attached to each data point. One comes from a cosine similarity to an artificial label text embedding, while the other comes from a cosine similarity to a centroid pseudo-embedding. Each label has a similarity score that can be interpreted as a confidence measure. The idea here is that we can use the combination of the two to select a subset of the data which has a confident label. This approach would work best if the two approaches were completely independent of

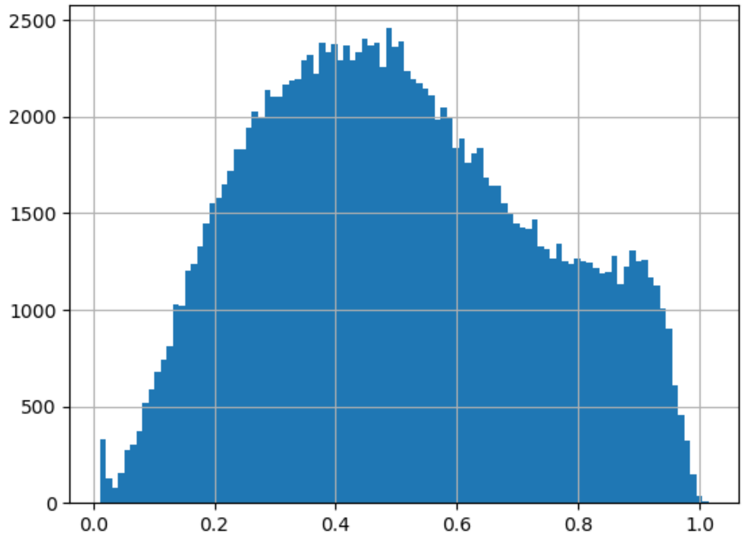


Figure 3.11: Well Formedness

each other, which is not the case here. However, they are different enough that additional knowledge can be obtained from combining the two. When comparing the two approaches, they agree on only 38.4% of the data and have a Cohen kappa of 0.317.

### 3.4.1 The threshold approach

One way to select a clean subset of data would be to apply thresholds to both confidence scores. For one of the training models, we used the following conditions to select the data:

1. original similarity score> 0.8
2. centroid similarity score> 0.9
3. original label == centroid label

This yielded 8936 data points ( this was done before implementing the ”Well Formness” filter, so repeating the process would yield slightly fewer data points). To check the accuracy of this dataset, we manually labeled a random sample with n=369. The results are accuracy = 0.8 and Cohen kappa = 0.75.

| Human Label | number of errors | accuracy per class |
|---|---|---|
| 0 | 33 | 0% |
| E1 | 4 | 96.6% |
| E2 | 0 | 100% |
| E3 | 1 | 95 % |
| E4 | 1 | 91% |
| E5 | 1 | 96% |
| G1 | 7 | 90% |
| S1 | 17 | 76% |
| S2 | 5 | 0% |
| S3 | 2 | 0% |
| S4 | 1 | 75% |

Table 3.4: accuracy per class

Table 3.4 shows each class's number of errors and accuracy. The dominant source of error is the 0 label. Neither of our zero-shot approaches accounts for the possibility of a datapoint being in none of the labels; they simply measure which class the text is close to. Furthermore, the CRSDs are still in development, and there are no rigid guidelines for what is required for each label at this level. On a second inspection of the 0 labeled data, the conclusion was that it could have really gone either way. For example, the following paragraph:

"Law and Compliance Ethical conduct is a matter of essential impo rtance for society. Many stakeholders evaluate companies accordi ng to whether they conduct themselves not just legally but also legitimately. The Covestro Group is committed to sustainable dev elopment in all areas of its commercial activity. Any violations of this corporate commitment can result in adverse media report ing and thus lead to a negative public perception of the Covestr o Group. We counter this risk through responsible corporate mana gement that is geared toward generating not only economic but al so ecological and societal benefit."

It was marked as 0 by a human as it was deemed too vague, but both zero-shot models gave it a G1 label. And the same dilemma applies to the other 0 labels in this subset. This applies to other points that were labeled 0. Out of which G1 was most dominant with 20 entries, five belonged to S1, 3 to S4 and E2 each, and 2 to E1. Given the unambiguity of the CRSDs labels and my own unsureness, we can think of this error as insignificant to verify the accuracy of this dataset. Remember that the conditions for selecting this data included high similarity scores, both original and centroid based. But, to select a training dataset the 0 label issue is persistent and needs to be addressed.

The next most common error is associated with the S1 label. All 17 errors received the G1 label from the zero-shot. There is a significant overlap between the two categories. Here are a couple of examples of the text that was labeled as S1 by a human and G1 by the models.

"Managerial employees have a vital part to play in implementing t he Corporate Compliance Policy. As role models, they must help t o ensure that this important code of conduct is adhered to in pr actice. Man- agers may lose their entitlement to variable compen sation components and be subject to disciplinary measures if sys tematic violations of applicable law entailing loss or damage to Bayer have occurred in their sphere of responsibility and could have been prevented if they had taken appropriate action. Com- pliant and lawful conduct forms part of the performance evaluati ons of all managerial employees."

"Training for all employees on Business Conduct Guidelines which reflect our commitment to respect and uphold international human rights. Every new employee is automatically signed up for a web -based training session or required to physically attend trainin g. Every employee must revisit the training sessions on a regula r basis."

On a second inspection of these paragraphs, they represent the intersection of the two classes. Classification could go either way but with a slight advantage towards the G1 label. In these particular paragraphs, the zero-shot models outperformed a human if the human in question is me.

Another issue with this dataset is that S4 class is underrepresented, while E2, S2, and S3 are not represented at all.

When using this dataset to train a model, we get slightly worse results than if we just used the data according to 3.3 table, with an accuracy score of 0.71 and a Cohen kappa of 0.65. The model trained with this approach is quite accurate on the training data but does not generalize well. This could be due to the selection process selecting only data that is easy to label.

To summarize, key takeaways when selecting data purely based on high similarity scores are the following:

1. The model does not learn how to handle more difficult data and does not generalize well
2. The model does not learn anything about the 0 label
3. Some labels will be under-represented in the training data

### 3.4.2 Categorical Approach

To better understand the two zero-shot models, we can look at figure 3.12. It plots the centroid score against the original similarity score, with red points where two modes agree and blue where they disagree.

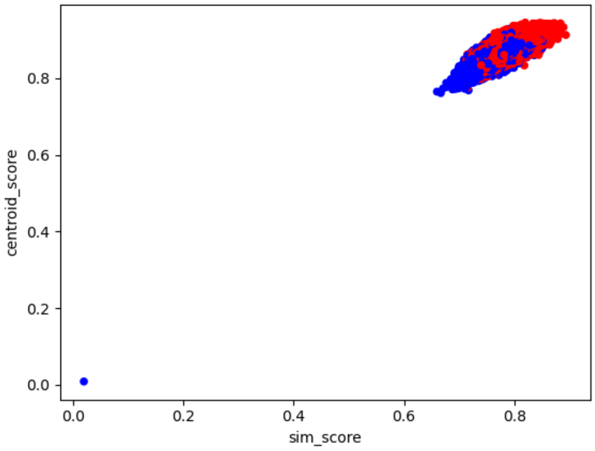


Figure 3.12: Outlier

We notice a single outlier in the bottom left of the figure 3.12. The text of it is:

"'Within the framework of our sustainability management system, we steer our sustainability program in a manner that enables us to verify the implementation of its objectives and thus ensure continuous improvement. Our management and organizational structures support this process by establishing clear lines of responsibility in all business divisions. Our sustainability objectives and their management are central components of our corporate governance system and are also incorporated into the target agreements between employees and managers.'

Interestingly, it got such low similarity scores (0.02 and 0.009) given that it has many of the key words associated with multiple labels, such as "corporate governance" and "employees." Never the less it, the models are correct about not being confident about any label for it.

Figure 3.13 gives us a better look at the rest of the data. We can see three rough categories of data. On the top left, we have the data with high similarity scores in both approaches. This data is related to the dataset discussed in the previous section. We can select data with high precision from here, but that could cause a model that dose not generalize well. The second section is the body of our dataset. It is messy data with some agreements between the two models and some disagreements. We must select some correctly labeled data from here for our model to generalize well. Lastly, there is the blue area at the bottom left. This area is a candidate for some 0-label data.

Figure 3.14 shows this division with the equations of the two lines being $y = -1.637x + 2.21$ and $y = -1637x + 5.05$ First, let us consider the top right part. We can select the data from this part with the command:

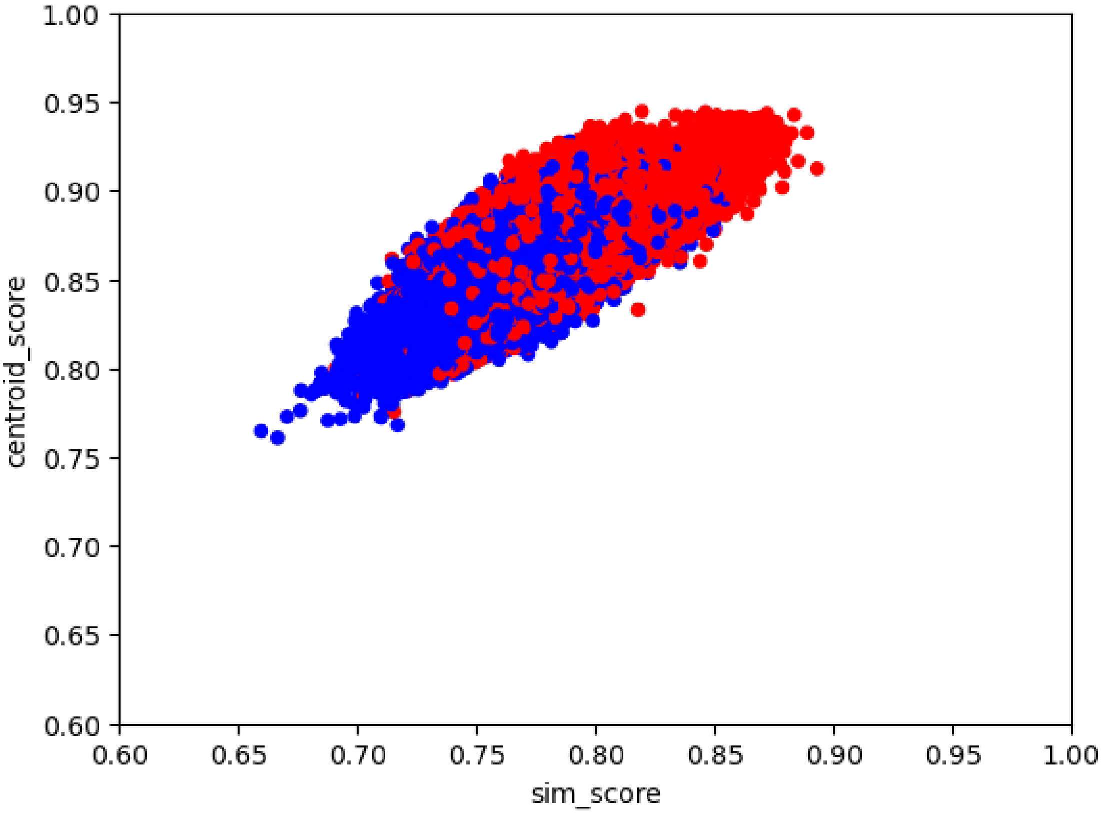


Figure 3.13: Original vs Centroid

```
low_df=df[df["sim_score"]*1.637+df["centroid_score"] <= 2.04]
```

This group consists of 16712 data points or 17.7% of our data. From this section, we can extract data with high accuracy. However, we need to avoid some of the downfalls from the earlier section. To increase the accuracy of the data, we will impose the condition of the two labels being equal. However, before that, we take a look at the results individually in tabel 3.5. The first thing that stands out is the complete absence of the S2 label in the centroid approach when there are 1090 instances in the original method. We get two dominant categories when we look at what centroid label was assigned to data labeled S2 with the original approach. The most dominant is the S4 label, with 555 instances, followed by the S1 label, with 222. To understand what is happening, we extract a random sample of 100 data points and take a closer look emphasizing S4 and S1 labels. For the data labeled S4, there is no straightforward correct label. Some of the data leans towards S4, some S4, some are too vague, and some would be the 0 label. Therefore, these data points cannot be used. On the other hand, data labeled as S1 should have been labeled S2. Both categories focus on workers, and my guess is that the S1 centroid simply dominates the S2 one. So this data can be used in training with the S2 label.

Next, we look at the S3 label, which only has 115 instances in the original method. Again we look at all the individual original labels where the centroid label is S3. Out of all the categories, E1 stands out. We have 55 paragraphs discussing mitigating climate change's impact on affected communities. This can be converted to the S3 label.

The "final agreement" column in table 3.5 reflects the state of the agreements after these changes were made.

Now we move on to the mid section in figure 3.14.Data here is defined by:

```
mid_df=df[(df["sim_score"]*1.637+df["centroid_score"]> 2.05) & (df["sim_score"]*1.637+df["cen
```

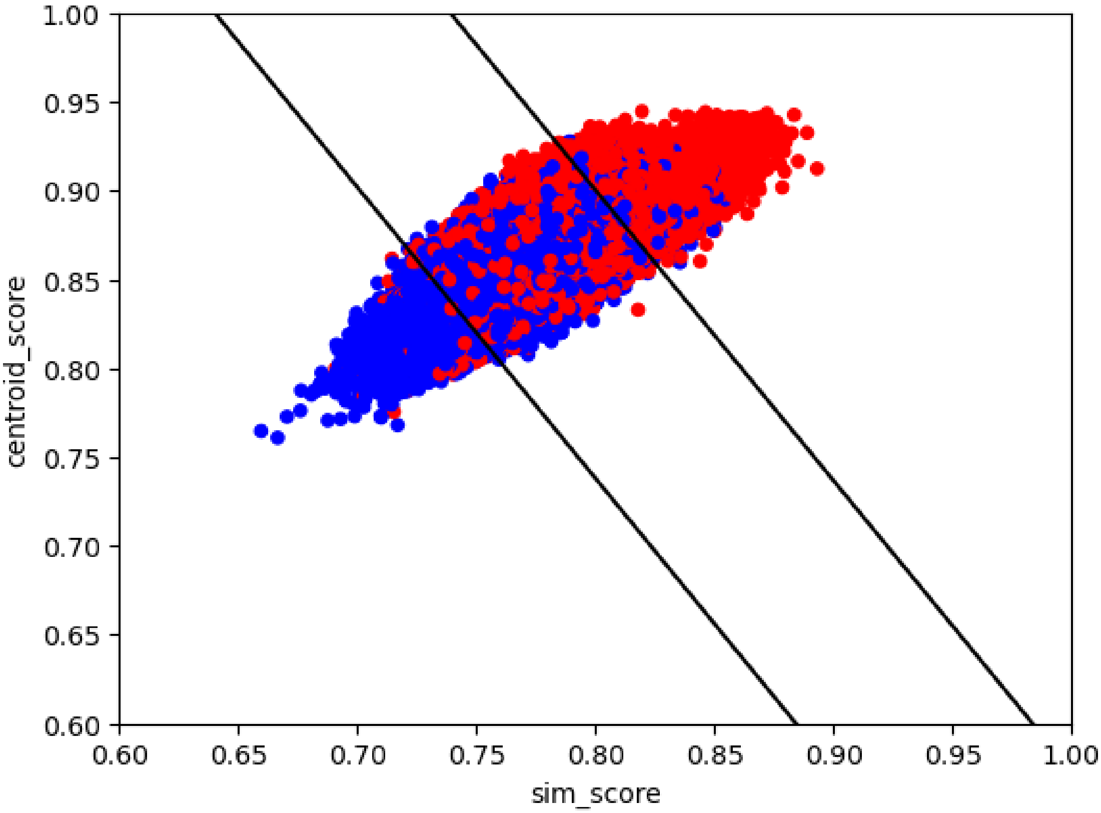


Figure 3.14: Clusters

| Label | original count | centroid count | agreement | final agreement |
|---|---|---|---|---|
| E1 | 5518 | 4599 | 4580 | 4580 |
| E2 | 228 | 499 | 190 | 190 |
| E3 | 324 | 390 | 321 | 312 |
| E4 | 250 | 323 | 232 | 232 |
| E5 | 434 | 767 | 418 | 418 |
| S1 | 3736 | 3609 | 3203 | 3203 |
| S2 | 1090 | 0 | 0 | 222 |
| S3 | 115 | 358 | 88 | 140 |
| S4 | 402 | 2611 | 333 | 333 |
| G1 | 4624 | 3565 | 32464 | 3246 |

Table 3.5: top data

Table 3.6 shows us that the differences between the two approaches become more drastic here. First of all, they only agree on 33.8% of the data. Secondly, the distribution of the categories is vastly different. These differences work in our favor; as the differences between model increase, the accuracy of shared predictions also increase. However, the agreement between the two predictions is not accurate enough for us to use this as training data. One way to select more accurate data would be only to use data points with high similarity scores. However, that would defeat the point of what we are trying to do here. We want to include correct labels with lower scores to increase our final model's generalization ability.

Another issue is that, again, as was the case with top data, the centroid approach is missing the S2 label completely. On closer inspection of entries given the S2 labels in the original approach, we notice that they all talk about the value chain but not necessarily about workers in the value chain. To increase the accuracy of our results, we introduce a simple keyword search. Here is the list of keywords for each label:

- **E1**: "gross scope","climate change","global warming","ghg","carbon"
- **E2**: "pollution","toxic","chemicals","substances"
- **E3**: "water consumption","water","marine"
- **E4**: "biodiversity","ecosystems","species"
- **E5**: "circular","recycling","recycled"
- **S1**: "employee"," worker representative","workforce","diversity","wages","social","health","training","collecti bargaining"
- **S2**: "worker","employee"
- **S3**: "community","communities"
- **S4**: "consumer","user","customer"
- **G1**:"corporate culture","tax","bribery","corruption"

So now the restriction for each class is that the two approaches must agree and that the text needs to contain at least one of the keywords in the list above, case insensitive. The only exception to this is the S2 class, where we ignored the centroid labeling.

| Label | original count | centroid count | agreement | final agreement |
|---|---|---|---|---|
| E1 | 17695 | 4057 | 4003 | 1073 |
| E2 | 1349 | 4780 | 1017 | 298 |
| E3 | 1037 | 1439 | 607 | 565 |
| E4 | 621 | 4288 | 530 | 171 |
| E5 | 1502 | 3961 | 1153 | 676 |
| S1 | 14509 | 9426 | 7903 | 5852 |
| S2 | 7574 | 0 | 0 | 218 |
| S3 | 1039 | 20367 | 872 | 249 |
| S4 | 3690 | 14249 | 2187 | 1084 |
| G1 | 19024 | 5473 | 4757 | 446 |

Table 3.6: mid data

Finally, we move on to the bottom right area in the 3.14 figure. This data is defined by:

```
low_df=df[df["sim_score"]*1.637+df["centroid_score"] <= 2.04]
```

From this subset, we want to select data corresponding to zero labels. The low similarity scores indicate this is the case for most of the data points here, combined with the very low agreement rate of only 6.54 %. To maximize the chances of selecting data points that do not correspond to any label, we run the same but opposite process of the one for the data in the middle section. First, we only select data where the two models disagree. Then we select data that does not contain the keywords corresponding to either label. We achieve this by running the not search for each label individually and then taking the intersection of the two sets. The result is 6388 data points; we can assign the zero label to.

## 3.5 Model

Before the introduction of Transformers, the dominant models for Natural Language Processing (NLP) were based on recurrent or convolutional neural networks (RNN and CNN). In the 2014 paper ”Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation” [Cho and et al., 2013], the authors proposed a new network model called the RNN Encoder-Decoder consisting of two RNNs used for machine translation. Here, one RNN encodes the text into fixed-length vector representations while the other decodes this vector representation into text. The networks are jointly trained to maximize the conditional probability of the decoded text given the encoded text. This approach improved results on text translation compared to state-of-the-art models.

Building on that work was the paper titled ”Neural machine translation by jointly learning to align and translate” [Bahdanau et al., 2014]. In this paper, the authors claim that the fixed length vector that the encoder produces in an Encoder-Decoder network is the bottleneck for improving the performance. Instead, The authors propose an architecture where the model automatically searches for relevant parts of the source for predicting the target. In other words, the decoder decides which parts of the code to pay attention to.

In the 2017 groundbreaking research paper ”Attention Is All You Need” [Vaswani and et al., 2017], the author proposes the transformers model based solely on attention mechanisms. It dispenses recurrence and convolutions entirely. Transformers models perform better in translation tasks, require significantly less training time and generalize well to other tasks. The transformer's architecture can be seen in figure 3.15 In 2019 in an article titled” ”BERT: Pre-training of Deep Bidirectional Transformers

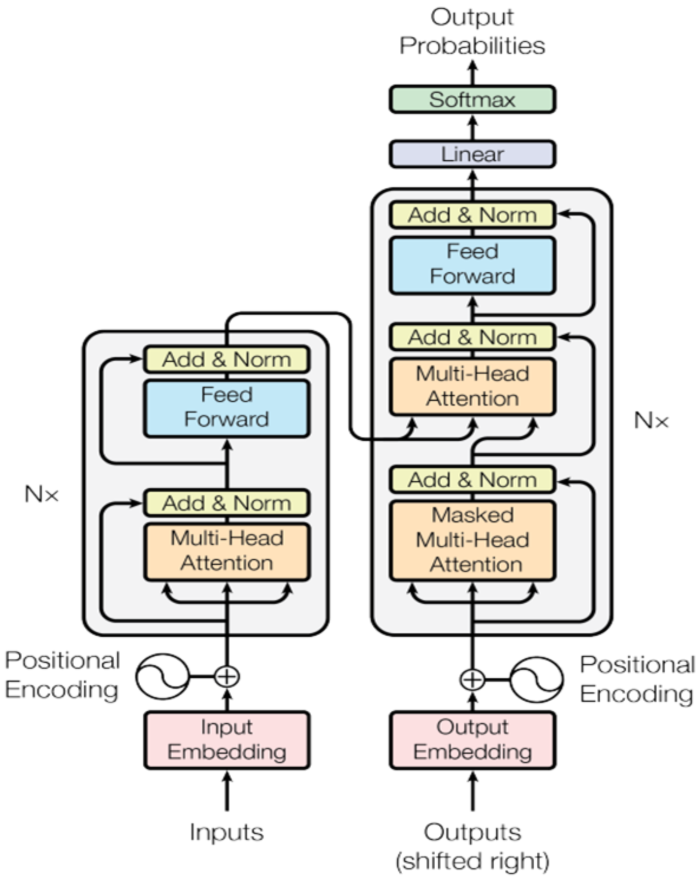


Figure 3.15: Transformers Arhitecture

for Language Understanding" [Devlin et al., 2019], the authors introduce a new language model BERT or Bidirectional Encoder Representations from Transformers. BERT is a deep bidirectional model that is pre-trained using two unsupervised tasks. First is Masked LM, where some percentage of input tokens are masked at random, then the task is to predict those masked tokens. The second task is Next Sentence Prediction (NSP), where the model is trained to predict a sentence based on the previous sentence. The result is a model that can be easily fine-tuned to a wide range of tasks with just one additional output layer.

Finally, in July 2019, the world met RoBERTa [Liu and et al., 2019]. Authors of RoBERTa: or A Robustly Optimized BERT Pretraning Approach replicated the study that led to BERT. They found that BERT was significantly undertrained and proposed modifications to BERT pre-training procedures. RoBERTa was trained on dynamic masking, where the masking pattern was generated every time sequence was fed into the model. Additionally to (NSP) ROBERTa was trained to predict whether a segment comes from the same or distinct documents. Furthermore, the ROBERTa model was trained on a much larger dataset than BERT.

## 3.6 Level 3 labels

Previously, we have trained an LLM classifier that can assign a two-level label to a paragraph of text. For example, E1 is a two-level label where the first level is E (environmental), and the second level is E1 (climate change).

The next step would be to extend this process to three levels. Example: E1-1 represents "Transition plan for climate change mitigation."

At first, we tried the same approach used for two-level labeling, but none of the zero-shot approaches outlined in previous sections worked. Next, we tried to use question answering, but the answers provided were often nonsensical.

The approach that showed the most promise is to provide the paragraphs of a particular two-level label as the context in a question-answering pipeline. Ask questions relating to each sublabel in that two-level label. Then we would disregard the answer itself but use the confidence score of the answer as an indicator that relevant information is included in that paragraph.

To demonstrate this process for the "E1" label, we run the following pipeline on E.ON Sustainability Report 2021:

1. Extract paragraphs from PDF
2. Label each paragraph with the fine-tuned model
3. Extract paragraphs with "E1" label
4. Ask a set of questions with each "E1" as context
5. for each question, use the paragraph with the highest score as an answer

Here is the complete set of questions for each sublabel and the corresponding answer and score:

1. **sublabe E1-1** Transition plan for climate change mitigation
   - **Question:** What is E.ON Transition plan for climate change mitigation?

- **Answer:** ”As mentioned above, in 2021 E.ON further embedded sustainability and climate action into its decision-making processes by introducing an ESG Reporting Manual and a new carbon ma agement plan. Both will enhance our ability to monitor progress toward our climate targets and to take sufficient action to meet them.”, score = 0.299167275

2. **sublabe E1-2** Policies related to climate change mitigation and adaptation
   - **Question:** What policies were adopted by E.ON?
   - **Answer:** ”As mentioned above, in 2021 E.ON further embedded sustainability and climate action into its decision-making processes by introducing an ESG Reporting Manual and a new carbon management plan. Both will enhance our ability to monitor progress toward our climate targets and to take sufficient action to meet them.”, score = 0.118998587

3. **sublabel E1-3** Actions and resources in relation to climate change policies
   - **Question 1:** What did EON do to mitigate climate change?
   - **Answer1:** ”In 2021 E.ON joined Science Based Target initiatives (SBTi) Business Ambition for 1.5C and committed to setting science-based emissions-reduction targets that are consistent with keeping global warming to 1.5C above pre-industrial levels. E.ON also joined the Race to Zero, a global campaign to accelerate progress toward a decarbonised economy. [ GRI 103-2/3 ]”, score= 0.241040558
   - **Question 2:** How much resources did E.ON invest to mitigate climate change?
   - **Answer 2:** ”Over the next five years, E.ON will invest around 27 billion in the energy transition: to expand and upgrade our networks and to offer new services to our customers. In addition, our customer solutions business supplies energy to around 51 million customers and helps them reduce their carbon footprint.”, score=0.407071322

4. **sublabel E1-4** Targets related to climate change mitigation and adaptation
   - **Question:** What are E.ON’s GHG emission reduction targets?
   - **Answer:** ”Strategy Although E.ONs business operations obviously cause carbon emissions, these operations also help millions of customers avoid emissions. Our two core businesses Energy Networks and Customer Solutions make the energy system more efficient, increase the proportion of renewables in the energy mix, and therefore help prevent GHG emissions. Moreover, our current climate strategy includes emission-reduction targets for 2030 through 2050. In 2020 we set new climate targets and intend to be climate-neutral by 2040 (see Goals and performance review).”, score= 0.863179147

5. **sublabel E1-5** Energy consumption and mix
   - **Question 1:** How much fuel consumption from coal?
   - **Question 2:** How much fuel consumption from natural gas?
   - **Question 3:** How much fuel consumption from nuclear?
   - **Answer for questions 1,2 and 3:** Quantitative IF-EU-000.D Owned generation by energy source in percentages Natural gas/oil: 4.8 Nuclear (Non-Core Business): 87.1 Coal: 0.1 Other (includes biomass, wind and solar): 8.0, scores =0.485947698, 0.36838311, 0.194994807
   - **Question 4:** How much fuel consumption from renewable sources?

- **Answer 4:**" As countries step up their decarbonisation efforts, the proportion of greener and thus more sustainable energy will steadily rise. Alongside this trend, the energy world will also become increasingly decentralised. Electricity Networks had and further have to be adapted to fulfill their role in being the central platform to make the transition towards a more sustainable and decarbonised society a success. Applying smart technologies and consequently fostering the digitalisation of grid infrastructure allows us to manage our existing grids at high efficiency and at the same time expand our networks in a resource-efficient way. E.ON considers its Electricity Networks, in line with the described eligibility criteria for green financing, applying asset values. Projects at our Customer Solutions segment include investments in integrated embedded energy solutions for cities and businesses, electricity generation from renewable sources, manufacturing and storage of hydrogen, and charging stations for electric cars. One example is our ectogrid project MIND (Milano Innovation District). MIND will transform the 900,000 square-metre site of Expo 2015, located about 15 kilometres north of Milan, Italy, into a multi-use urban space. In December 2021 international property developer Lendlease and E.ON signed a project cooperation agreement to design and manage a sustainable energy solu- tion for MIND based on ectogrid technology. Installation started in 2021 and will continue in 2022. Considering the nature of the different projects and assets, these are considered in the Renewable Energy, Energy Efficiency or Clean Transportation category.", score =0.26971665

6. **sublabel E1-6** Gross Scopes 1, 2, 3 and Total GHG emissions
   - **Question:** How much gross Scope 1 GHG emissions?
   - **Answer:** "Greenhouse gas emissions (total CO equivalents in million metric tonnes, location-based) E03-01 107.99 116.27 129.08", score= 0.065873474
   - **Question 2:** How much gross Scope 2 GHG emissions.?
   - **Answer 2:** "Energy from wastewater In March 2021 a state-of-the-art office building was unveiled in east-central Berlin. The build- ing, which is actually a repurposed department store, meets about half of its heating and cool- ing needs sustainably. E.ON made this possible by installing a 200-metre heat exchanger in a nearby underground wastewater canal. Our solution enables the building to displace around 400 metric tonnes of carbon dioxide each year." , score= 0.063862711
   - **Question 3:** How much gross Scope 3 GHG emissions?
   - **Answer 3:** "Our 2021 Scope 3 emissions of 100.38 million metric tonnes made up the lions share of our total carbon footprint. We recorded a slight reduction compared with 2020 and expect the car- bon intensity of purchased power to continue to decline further as the European countries in which we purchase power decarbonise their energy mixes.", score= 0.702646852

7. **sublabel E1-7**– GHG removals and GHG mitigation projects financed through carbon credits
   - **Question 1:** How much GHG removed?
   - **Answer 2:**"Truly green beer E.ON is enabling Knig brewery in Duisburg, Germany, to brew green, climate-neutral beer. The plan, which Knig approved in June 2021, is to pipe waste heat from a nearby power plant to provide thermal energy for brewing processes. This

means that Knig wont have to consume fuel and emit carbon to produce the heat itself. E.ONs role is to install and manage the pipeline, which we expect to be operational in the second quarter of 2022. The project, which will be funded by the Federal Ministry of Economic Affairs and Energy under its energy-effi- ciency programme, will displace about 7,000 metric tonnes of carbon dioxide annually.", score= 0.075596698

- **Question 2:** How much GHG mitigated due to carbon credits?
- **Answer 2:** "Brand campaign: Time for Action Both of the new E.ONs core businesses energy networks and customers solutions make a tangible difference in the fight against climate change. For our new brand campaign, Time for Action, we teamed up with renowned mountaineer and environmentalist Reinhold Messner, who was joined by E.ON CEO Leonhard Birnbaum and 25 of our employees, customers, and partners on a glacier in Austria to share stories and experiences. We filmed their discussions, and the resulting advertisement premiered in the United Kingdom, Italy and Hungary in the fourth quarter of 2021. In the spirit of the campaign, we made the ad itself climate-neutral by purchasing certified offsets for the emissions associated with its production in Austria and broadcast in the United Kingdom.", score= 0.118613645

8. **sublabel E1-8** Internal carbon pricing
   - **Question:** What is E.ON's internal carbon pricing?
   - **Answer:** "E.ONs updated strategy aims to empower over 51 million customers to switch from fossil to green energy sources and to progress toward net zero. Well help companies and cities convert to renewable and recycled energy, including green gas and hydrogen, which will be essential for industry and heavy transport to decarbonise. Well offer solutions that enable residential cus- tomers to generate their own renewable energy, use it efficiently, make their home smart, and embrace eMobility. These solutions will enable our residential customers to displace more than 3 million metric tonnes of CO each year by 2026.", score= 0.016249374

9. **sublabel E1-9** Potential financial effects from material physical and transition risks and potential climate-related opportunities
   - **Question:** What are financial risks associated with climate change?
   - **Answer:** "Both climate change and the energy transition aimed at slowing it could create risks as well as opportunities for our business. In 2021 we performed a qualitative scenario analysis to model how the key value drivers of E.ON and some of its business units might be affected under three different climate scenarios conservative, ambitious, and fully determined between now and 2050. The conservative scenario foresees unhurried decarbonization that lags behind Paris Agreement targets leading to global warming of well above 2 C by 2100. The ambitious scenario reflects current commitments under the Paris Agreement and results in global warming of around 2C by 2100. The fully determined scenario, which is in line with the Paris Agreement, limits global warming to 1.5C by 2100. The findings, which will be factored into E.ONs ongoing strategy development, will be available in early 2022. We intend to repeat the scenario analysis on an annual basis.", score= 0.270954967
   - **Question 2:** What are financial opportunities associated with climate change?

- **Answer 2:** ”Task Force on Climate-related Financial Disclosures E.ON is committed to acting sustainably in all respects. This includes making steady progress toward our climate targets, effectively managing our climate-related risks, seizing climate-re- lated opportunities that fit with our corporate strategy, and reporting transparently on all these matters. The recommendations of the Task Force on Climate-related Financial Disclosures (TCFD) provide important guidance for our reporting. Established in 2015, the TCFD aims to develop consistent, comparable, and accurate climate-related financial risk disclosures that companies can use to provide information to investors, lenders, insurers, and other stakehold- ers. E.ON became an official TCFD supporter in 2019, which marks the start of our TCFD reporting below. Going forward, well continue to expand our TCFD reporting. One consequence of TCFD reporting is that E.ON has developed a qualitative scenario analysis to assess how E.ONs businesses might be affected under different climate scenarios (for more information, see Strategy below.In addition, our TCFD reporting is supported by additional information in”, score= 0.311751753

The questions above were derived from the ”Draft ESRS E1 Climate Change November 2022” [EFRAG, 2022] document. While the answers are hit-and-miss, we can see that the score represents a good indicator of their validity. Here we only showed the top results for each question; a better approach is to include all the paragraphs with a sufficiently high score. By the process of trial and error, we deduced that score over 0.1 indicates that some knowledge is relevant to the question, while scores of over 0.25 indicate a good match. And a low score for the top result could indicate a lack of specific information However, a more rigorous study must be done to determent relevant scores for each topic.

Furthermore, document [EFRAG, 2022] contains more specificity in what information is needed and how it should be presented. We can catholicize the questions into two categories. First, concrete questions with numerical or categorical answers, such as GHG emissions and energy mix questions. And second, more descriptive topics, like the transition plan or the actions and resources, which need to include references to the questions from the first category.

Therefore, it might be better to reframe the question of tree-level labeling from classification to knowledge extraction. Here the fine-tuned Roberta model for two-level classification would be the first step in the pipeline. Then the complete set of paragraphs with a specific two-level label would move into their respective knowledge extraction phase with techniques like prompt engineering and named entity recognition.

# 4 Results

Data discussed above was used to fine tune a pre-trained RoBERTa model with the following training arguments:

- num_train_epochs=4
- per_device_train_batch_size=8
- per_device_eval_batch_size=8
- learning_rate=5e-5
- weight_decay=0.01
- warmup_steps=500

Only a little was done regarding parameter tuning (other than ensuring training does not crash the system). The focus was on comparing results given different ways of selecting training data.

## 4.1 Evaluation

The models trained in this project were evaluated against two separate test sets. One is derived from the top right, and the other from the middle section of figure 3.14, referred to as the top-test set and mid-test set, respectively. Both comprise 369 randomly selected data points from the corresponding section, manually labeled. Table 4.1 shows the distribution of classes in each test dataset. Note the difference in the distribution of the two tables. This suggests that environmental classes might be easier to classify than social ones. We will find that this is true for both human and machine labeling.

| Label | top test set | mid test set |
|---|---|---|
| 0 | 36 | 167 |
| E1 | 116 | 47 |
| E2 | 6 | 7 |
| E3 | 22 | 5 |
| E4 | 11 | 6 |
| E5 | 29 | 15 |
| S1 | 74 | 56 |
| S2 | 4 | 18 |
| S3 | 2 | 14 |
| S4 | 3 | 16 |
| G1 | 66 | 18 |

Table 4.1: evaluation data

## 4.2 Data with high scores only

Data selection for Model 1 was entirely based on the model score. Training data was selected according to table 3.3, with The results for this model are:

- Test set ”top”: Cohen kappa = 0.778, accuracy= 0.821. If we exclude ”0” labels: Cohen kappa= 0884, accuracy= 0.91
- Test set ”mid”: Cohen kappa = 0.349, accuracy= 0.407. If we exclude ”0” labels: Cohen kappa= 0.660, accuracy= 0.713
- Combined score : Cohen kappa = 0.557, accuracy= 0.614. If we exclude ”0” labels: Cohen kappa= 0.798, accuracy= 0.836

## 4.3 Combined data

Data selection for Model 2 was all the data that passed the criteria outlined in the data selector. This was the biggest training dataset, with over 30000 entries. To compensate, we reduced the number of epochs to 3.

- Test set ”top”: Cohen kappa = 0.787, accuracy= 0.829. If we exclude ”0” labels: Cohen kappa= 0.896, accuracy= 0.919
- Test set ”mid”: Cohen kappa = 0.486, accuracy= 0.604. If we exclude ”0” labels: Cohen kappa= 0.576, accuracy= 0.634
- Combined score : Cohen kappa = 0.661, accuracy= 0.717. If we exclude ”0” labels: Cohen kappa= 0.770, accuracy= 0.811

## 4.4 Combined and trimmed data

Training data for model 3 was a subset of model 2. We selected the highest scoring data points from the top-right section and random points from the midsection, and no points from the bottom left section (no zero label training data)

- Test set ”top”: Cohen kappa = 0.784, accuracy= 0.827. If we exclude ”0” labels: Cohen kappa= 0.892, accuracy= 0.916
- Test set ”mid”: Cohen kappa = 0.370, accuracy= 0.420. If we exclude ”0” labels: Cohen kappa= 0.725, accuracy= 0.767
- Combined score : Cohen kappa = 0.571, accuracy= 0.623. If we exclude ”0” labels: Cohen kappa= 0.828, accuracy= 0.860

## 4.5 Accuracy per class

The table 4.2 below shows the accuracy per class for each model using the combined evaluation dataset.

| Label | Model 1 | Model 2 | Model3 |
|---|---|---|---|
| 0 | 0 | 0.468 | 0 |
| E1 | 0.926 | 0.896 | 0.926 |
| E2 | 0.923 | 0.769 | 0.923 |
| E3 | 0.889 | 0.926 | 0.926 |
| E4 | 0.882 | 1 | 1 |
| E5 | 0.886 | 0.886 | 0.909 |
| S1 | 0.792 | 0.831 | 0.831 |
| S2 | 0.545 | 0.136 | 0.318 |
| S3 | 0.5 | 0.125 | 0.562 |
| S4 | 0.579 | 0.789 | 0.947 |
| G1 | 0.857 | 0.821 | 0.869 |

Table 4.2: Results

## 4.6 Errors

In this section, we look at non-zero data mislabeled by model 3.

For the data labeled ”E1” by a human, there were 12 total errors. Out of those, seven are mislabeled as one of the other environmental categories (”E2”, ”E3”, ”E4”, ”E5”). All the paragraphs in this category talk about multiple environmental issues and have elements of both the labels given by a human and the model. Four paragraphs were given the ”S4” label, two of which are clear errors, and the other two talk about energy solutions that are in some way available to consumers. Out of the latter, one is an error by the model, while a human mislabeled the other.

The rest of the environmental categories combine for additional seven errors. One ”E2” paragraph was ”mislabeled” and ”E4”. It is about the effects of pollution on biodiversity (both labels apply). One ”E3” is ”mislabeled” as ”E5” about circular economy and the ocean (again, both labels apply). Moreover, one ”E3” to ”S4”, here S4 is the correct label, the model outperformed the human. The ”E5” label was ”mislabeled” to one of the other E labels three times, ”E2”, ”E3,” and ”E4” once each. Those again represent intersections of the categories where both labels make sense. Another mislabeling of ”E5” to ”S4” was a clear error.

There were 22 errors associated with the ”S1” label. For the ”E1”, ”S3,” and ”S4” categories, there was one mislabel each due to a human error, while one ”E3” was a model error. The model gave two paragraphs to the ”S2” label. Depending on the context, those could go either way, but as it is not stated explicitly, we assume they were about their own workforce. The most significant error here was the ”G1” mislabel that occurred 16 times. Paragraphs here are about employees, but their role in establishing and promoting the Business or Ethics Conduct. This fits mainly with the ”G1-1” sublabel as it is not directly about employee well-being but establishing corporate culture. Remarkably, the model was better at picking this up than a human.

The ”S2” class was the hardest for the model to handle, with 15 total errors. Two paragraphs were given the environmental labels ”E2” and ”E5” (model error). Six were mislabeled as ”S1”, again clear errors by the model; this error could be addressed by implementing a keyword detector in the pipeline. But most errors came from a ”G1” mislabel, seven total. Those errors have intersectionality with the ”G1-2” label: ”Management of relationships with suppliers,” but conversely to the S1 to G1 error, here, the well-being of the workers was the central theme. Therefore, the error was in the model. Finally, one point for humanity!

There were seven mislabels in the ”S3” category. Four of them received an environmental label. Those paragraphs were about remedying the effects of environmental issues on an affected community. Two received the ”S4” labels by the model, referring to paragraphs representing an intersection of the two categories. Finally, one ”S1” mislabel was a clear error on the side of the model.

There were 11 errors in the ”G1” category. Seven times the mislabel was in the environmental category. Those paragraphs are relatively vague and could have been a zero label, but the ”G1” was closer to the truth than any of the E labels. Therefore, we will consider them clear errors on the part of the model. The four errors that were mislabeled with one of the social categories are also mistakes on the part of the model.

There were only 35 clear errors on the second inspection, with a whopping 20 coming from the ”S2” and ”S3” labels and 11 from the ”G1” label. If we only account for those as ”true” errors, the model's accuracy jumps to over 93 %. Note that we are ignoring the zero label in those calculations. More on the relevance of the zero error can be found in section 6.3.

# 5 Conclusion

In this project, we successfully extracted textual data from pdfs, a notoriously unstructured format. Then we used ada_002 embeddings and cosine similarity between embeddings to build two approaches for a zero-shot classifier. In one, we compared textual data embeddings to embeddings of a written description of the labels, and in the other, the centroid of the most similar data embeddings to each class. The resulting data were classified into three categories based on how easy a paragraph of text is to label. Then we used different techniques to extract valuable data points from each category. This gave us an accurate training set that sufficiently represented the overall data. This dataset was used as training data for building a classifier that outperforms each zero-shot approach. We managed to build a reasonably accurate model out of an unsupervised dataset. Note that the figure 4.2 underestimates the accuracy of the model, as discussed in the 4.6 section; most of the errors here come due to a human mislabeling and not an error in the model itself. At the very least, we can say that the resulting classifier outperforms human labeling.

In doing the above, we demonstrated two things:

1. Methodology that can be used to create training textual data out of an unsupervised set quickly.
2. As a feasibility study, we show that as the ESRS guidelines come into play and labeled datasets can be gathered, building a precise two-level classifier is more than possible.

Moreover, in the 4.6, we pointed out the reasoning behind the errors of the model when applied to the test datasets. The model is highly precise for all the environmental and the ”S1” and ”S4” categories. It performs reasonably well in the ”G1” category but struggles with the ”S2” and ”S3” labels.

Furthermore, we have touched on examining the data on the third-label level. Here we suggest that there are better approaches to studying data than classification at this level. We demonstrate one method applied to the ”E1” label. This method would need to be improved, and then a different approach needs to be created for each of the ten classes. Much more work needs to be done to develop an AI tool that would open up the space of sustainability reporting analysis.

# 6 Reflections, limitations, and future work

## 6.1 Reflections

This section will discuss what could have been done better without changing this project's scope.

First, more emphasis, time, and effort should have gone into creating a test set. As we used unlabeled data, we needed to label a representative subset manually. The ideal test set should have been labeled by multiple humans. Then only matching data should be used to avoid typing errors.

As the project evolved, so did my understanding of the topic, and I could have given different answers at different times. This would have been a significant issue if this project aimed to train a precise classifier. Instead, the focus was on the feasibility of building such a network with unlabeled data. However, this was still a pain point in the project.

Next, the data was stored in CSV format, and different parts of the project were communicated by reading and writing CSV files. This was not memory efficient at all. Instead, a simple relational-database scheme should have been used, such as MySQL.

Moreover, too much time was spent building different models for a marginal gain in accuracy 4.2. This time might have better served developing non-LLM solutions to S2 and S3 labels or studying the third-level data. Furthermore, as shown in section 4.6, some paragraphs contain information regarding multiple labels. Building a multi-label classifier would address this.

Finally, if I were to start this project again today, I would first re-train Roberta for an MLM task with the textual data extracted. Then use that model as a foundation for downstream tasks. This could improve the results as the model would be better at capturing language characteristics of the data.

## 6.2 Limitations

Here we discuss limitations in the methodology represented in this project and what should be added to make a better overall tool.

An obvious limitation comes from studying a framework yet to be implemented. The guidelines are incomplete and subject to change. Furthermore, no labeled data was present, and the reports were not tailored for this framework.

The second limitation is focusing solely on the text in paragraphs. This can lead to the mislabeling of data. For example, two identical paragraphs could be in the ”S1” or ”S2” category, depending on whether they refer to their own workforce or value chain workers. Integrating contextual and metadata from headings, titles, and the table of contents could clarify this. Furthermore, critical information and measures are often displayed in graphs or tables. For a comprehensive analyzing tool, this information must be incorporated.

## 6.3 Future work

For best results, future work on this topic should gather a comprehensive labeled dataset. Furthermore, such work should implement changes outlined in the Reflections and Limitations sections. Namely, incorporate data from graphs and tables, make use of contextual metadata, use better data storage solutions, and use re-trained Roberta as the base model.

However, the grunt of the work yet to be done is studying the data at the third level. Each of the ten labels represents a unique puzzle to be solved in order to extract data according to ESRS guidelines. Those puzzles are to be solved using expert knowledge and prompt engineering. Furthermore, having two different information extraction methods for each label could be helpful. A comprehensive one for companies with matching materiality and a ”minimal requirements” otherwise.

The complete pipeline of a comprehensive analytical tool could look like this:

1. Input: Sustainability Report in pdf format
2. Determine the materiality of the company
3. Extract textual data. Alongside the paragraph text, we would include the most recent title, heading, and table of contents entry.
4. Classify each paragraph corresponding to its two-level label.
5. Pass all of the data from a particular label to its corresponding information-extraction algorithm
6. Apply a set of questions, with the level of scrutiny depending on the materiality of the company
7. Output: a set of answers and text extract relevant to points in ESRS guidelines.

In this pipeline, a multi-label output would pass on the paragraphs to all the corresponding sections. Furthermore, this might be a better place to address the persistent ”zero” label issue. An irrelevant paragraph would contain no useful information to be extracted in a downstream task. The consequence of not addressing the ”zero” label would be a higher running time of the software but no decrease in the accuracy of the solution.

# Bibliography

[Amin, 2006] Amin, S. (2006). The millennium development goals: A critique from the south.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

[Boulding, 1966] Boulding, K. E. (1966). The economics of the coming spaceship earth.

[Bricki and Holder, 2006] Bricki, N. and Holder, A. (2006). Mdg4-hope or despair for africa? revista de economia mundial.

[Carson, 1962] Carson, R. (1962). *Silent Spring.* Houghton Mifflin.

[CERES, 1989] CERES (1989). Valdez principless. https://changeoracle.com/2015/03/02/valdez-principles-ceres-principles/,.

[Cho and et al., 2013] Cho, K. and et al. (2013). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

[Christie and Tansey, 2002] Christie, D. A. and Tansey, E. M. (2002). Environmental toxicology - the legacy of silent spring. (transcript of a Witness Seminar, London, 12 Marech 2002).

[Cole et al., 1973] Cole, H. S. D., Freeman, C., Jahoda, M., and Pavitt, K. L. R. (1973). *Thinking about the future - A Critique of The limits to growth.* Sussex University press.

[Commission, 2023] Commission, E. (2023). Corporate sustainability reporting. https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en,.

[Devlin et al., 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[EFRAG, 2022] EFRAG (2022). [draft] esrs e1 climate change.

[EFRAG, 2023a] EFRAG (2023a). Glimpse into draft esrs 1 general requirements. https://www.youtube.com/watch?v=a1pdAO62bHO,.

[EFRAG, 2023b] EFRAG (2023b). Glimpse into draft esrs 2 general disclosures. https://www.youtube.com/watch?v=G_uCqFXK7qU,.

[Eyben, 2006] Eyben, R. (2006). The road not taken: international aid's choice of copenhagen over beijing.

[Fehling et al., 2013] Fehling, M., Nelson, B. D., and Venkatapuram, S. (2013). Limitations of the millennium development goals: a literature review.

[Fukuda-Parr, 2010] Fukuda-Parr, S. (2010). Reducing inequality – the missing mdg: A content review of prsps and bilateral donor policy statements.

[Goktenz et al., 2020] Goktenz, S., Ozerhan, Y., and Gokten, P. O. (2020). The historical development of sustainability reporting: a periodic approach.

[GRI, 2016] GRI (2016). Gri 101: Foundation.

[GRI, 2022] GRI (2022). The gri perspective. (business case for environment & society).

[Henver et al., 2004] Henver, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research.

[Hulme, 2010] Hulme, D. (2010). Lessons from the making of the mdgs: Human development meets results-based management in an unfair world.

[Kannengißer, 2023] Kannengißer, S. (2023). From millennium development goals to sustainable development goals: Transforming development communication to sustainability communication.

[Kassas et al., 1980] Kassas, D., Tolba, M. K., and Loudon, J. J. (1980). *World conservation strategy - Living Resource Conservation for Sustainable Development.*

[Langford, 2010] Langford, M. (2010). A poverty of rights: Six ways to fix the mdgs.

[Liu and et al., 2019] Liu, Y. and et al. (2019). Roberta: A robustly optimized bert pretraining approach.

[Meadows et al., 1972] Meadows, D. H., Meadows, D. L., Randers, J., and Behrenes, W. W. (1972). *The limits to frowth.* Potomac Associates. (A Report for THE CLUB OF ROME'S Project on the Predicament of Mankind).

[Ni and et al., 2023] Ni, J. and et al. (2023). Paradigm shift in sustainability disclosure analysis: Empowering stakeholders with chatreport, a language model-based tool.

[Norren and E., 2012] Norren, V. and E., D. (2012). The wheel of development: The millennium development goals as a communication and development tool.

[Pukelis et al., 2022] Pukelis, L., G.Statulevičiūtė, Statulevičiūtė, V., Dikmener, G., and Akylbekova, D. (2022). Osdg 2.0: a multilingual tool for classifying text data by un sustainable development goals (sdgs).

[Ratneri and et al., 2016] Ratneri, A. and et al. (2016). Data programming: Creating large training sets, quickly.

[Salesken.ai, 2021] Salesken.ai (2021). query_wellformedness_score. `https://huggingface.co/salesken/query_wellformedness_score`,.

[Shaw, 2009] Shaw, D. (2009). The exxon valdez oil-spill: Ecological and social consequences.

[Shearer, 2000] Shearer, C. (2000). Crisp-dm the new blueprint for data mining.

[UN, 1992] UN (1992). Agenda 21. (United Nations Conference on Environment & Development Rio de Janerio, Brazil, 3 to 14 June 1992).

[UN, 2015a] UN (2015a). The millennium development goals report 2015.

[UN, 2015b] UN (2015b). Transforming our world: The 2030 agenda for sustainable development.

[Vaswani and et al., 2017] Vaswani, A. and et al. (2017). Attention is all you need.

[WCED, 1987] WCED (1987). *Our Common Future*. Oxford University Press. (Report of the World Commission on Environment and Development: Our Common Future).

[Zahar and Boerma, 2010] Zahar, C. A. and Boerma, T. (2010). Five years to go and counting: Progress towards the millennium development goals.

[Ziai, 2011] Ziai, A. (2011). The millennium development goals: Back to the future?

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

## Declaration of Authenticity

I hereby declare that I have completed this Master's thesis on my own and without any additional external assistance. I have made use of only those sources and aids specified and I have listed all the sources from which I have extracted text and content. This thesis or parts thereof have never been presented to another examination board. I agree to a plagiarism check of my thesis via a plagiarism detection service.

04|08|2023 ............ Hibrini ............

Date Signature