Doctoral Thesis

Justus-Liebig-University Giessen

A Text-Based Approach to Sustainability Indicators

Author:
Elena Anna Tönjes

1. Supervisor:
Prof. Dr. Peter Winker

2. Supervisor:
Prof. Dr. Nicolas Pröllochs

Submitted in fulfillment of the requirements
for the degree of Doctor rerum politicarum

in the

Department of Statistics and Econometrics
Faculty of Economics and Business studies

June 26, 2024

Affidavit

I hereby declare that I completed the papers submitted and listed hereafter independently and only with those forms of support mentioned in the relevant paper. When working with the authors listed, I contributed no less than a proportional share of the work. In the analysis that I have conducted and to which I refer in the papers, I have followed the principles of good academic practice, as stated in the Statute of Justus Liebig University Giessen for ensuring good scientific practice.

Elena Anna Tönjes,
Giessen, June 26, 2024

Submitted Articles

- Funk, C., Tönjes, E., Teuber, R., & Breuer, L. (2024). Reading between the lines: The intersection of research attention and sustainable development goals. *Sustainable Development*, 1–22. https://doi.org/10.1002/sd.2906
- Funk, C., Tönjes, E., Teuber, R., & Breuer, L. (2023). Finding common development paths in voluntary national reviews reporting on sustainable development goals using aspect-based sentiment analysis. In: *Working Paper*.
- Auzepy, A., Tönjes, E., Lenz, D., & Funk, C. (2023). Evaluating TCFD reporting—A new application of zero-shot analysis to climate-related financial disclosures. *PLOS ONE 18*(11): e0288052. https://doi.org/10.1371/journal.pone.0288052
- Funk, C., Tönjes, E., & Haas, C. (2024). Exploring the Predictive Capacity of ESG Sentiment on Official Ratings: A Few-Shot Learning Perspective. In: *Working Paper*.

Acknowledgements

Many people have supported me over the past years, and I could not have completed this thesis without the support of my friends and family. I would like to thank the University of Giessen for making my journey possible in the first place. I would especially like to thank my supervisors, Peter Winker and Nicolas Pröllochs, who have been very supportive during this time. I am very grateful to Peter Winker, who has always supported me, encouraged me in my research, and made the writing of this thesis possible

I would also like to thank my two co-authors from the Faculty of Agricultural Sciences, Nutritional Sciences, and Environmental Management, Ramona Teuber and Lutz Breuer, who helped me greatly with their expertise in the field of Sustainable Development. Special thanks to my colleagues at the chair, Viktoria Naboka-Krell, Jenny Bethäuser, Albina Latifi, and Maykol Rodriguez, who made the time of my PhD feel not only like work, but also like being with friends. I would especially like to thank David Lenz, who was not only my co-author and mentor, but also a friend.

I would also like to thank Alix Auzepy, who introduced me to the topic that influenced the second part of this thesis and a large part of my work. A very special thanks goes to Christoph Funk, who has been a great help to me during my PhD, not only by providing me with academic expertise, but also by helping me through PhD in general. Without him, not only would it have been a million times harder to start the first paper for my thesis, but also to find new ideas for further research. I would also like to thank my co-author Christian Haas, who helped me finish the last paper for this thesis.

Many thanks to Carmen Hersener, who made coming to the office so much easier by always having an open ear and helping me navigate the bureaucracy. Last but not least, I would like to thank my wonderful family and my very dear friends who made all of this possible and supported me through this great but sometimes stressful time. Without any of you, this thesis would not have been possible. Thank you!

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

This PhD thesis consists of four papers that I wrote between 2020 and 2024. All papers share two common elements: the technique employed, namely natural language processing, and the topic of sustainability. I have chosen to focus on this topic because it affects us all, particularly in the context of climate change, which is one of the most pressing concerns facing humanity and requires immediate attention. This is why I and my co-authors attempted to provide alternative indicators to quantify valuable information incorporated in texts of different formats. It is my hope that the measures we create will facilitate the extraction of information in this field, which is of great importance. I hope that our work will save time for those who are striving to make this world a better place. The methods used in this thesis are all based on large language models, which have gained a great deal of attention since the first widely known model, Bidirectional Encoding Representations from Transformers (BERT) (Devlin et al., 2018), was introduced in 2018. During my master's studies, I was already intrigued by this model and aspired to utilize it and the subsequent years of my PhD to conduct further research and apply it more extensively, particularly in a field that affects us all. Since the start of the thesis, a number of new models have been developed, and we have striven to keep informed about and make use of the most advanced models over the course of this period. Consequently, this thesis employs a diverse range of variations of the model that initially piqued my interest and underwent training and use during this time.

The thesis is divided into four principal parts, each of which is further divided into chapters. The introduction, which constitutes part one of the thesis, is followed by two main parts, two and three, and then concluded by part four, which constitutes the thesis. A comprehensive list of all references is provided after the conclusion. Each part is comprised of chapters, which are further subdivided into sections. The

opening of the chapters in parts two and three provide an overview of the author's contributions and the current state of their published works. Please be advised that all papers presented in this thesis deviate slightly from the published versions due to formatting adjustments only, with no changes to the principal content.

The second major part of this thesis is devoted to the subject of sustainable development, with a particular focus on the systematic analysis and creation of indicators for the quantification of discourse in Sustainable Development Goals. The part is comprised of two papers. The first paper assesses the quantity of scientific discourse on Sustainable Development Goals. In recent years, research on Sustainable Development Goals has seen a notable increase in both quantity and quality. A tool is provided which allows the user to quantify the research attention to each goal in a systematic manner. The resulting index, generated by the tool, is designated as the Research Attention Index. The second paper develops a sentiment index that gauges the tone of countries reporting in voluntary national reviews. Although not obligatory, the reviews provide a forum for countries to report on their progress and challenges in achieving the Sustainable Development Goals. The index enables the identification of countries that may face similar problems and successes, suggesting that they may be on common development paths. This clustering may facilitate the identification of potential synergistic effects.

The third major part of the thesis continues to examine the topic of sustainability in company reports. The objective was to quantify the stances of companies on environmental, social, and governance aspects. Furthermore, the part comprises two papers. The initial paper seeks to determine whether companies report on all topics that should be addressed and that are provided by the Task Force for Climate-related Financial Disclosures. As some of the guidelines are specifically designed for banks, this study focuses on company reports provided by banks. The objective of this study is to ascertain whether certain topics from the guidelines are underrepresented in company reports. This would indicate the possibility of selective reporting. The second paper in this section addresses the broader issue of corporate reporting on environmental, social, and governance matters. The paper proposes an alternative index for assessing the tone of companies, which differs from the official rating scores. The final section presents a brief conclusion, an overview of the principal findings, and suggestions for future research.

Part II

Sustainable Development

Chapter 2

Reading Between the Lines: The Intersection of Research Attention and Sustainable Development Goals

The following chapter is based on the paper:	
Title:	Reading Between the Lines: The Intersection of Research Attention and Sustainable Development Goals
Authors:	Elena Tönjes (contribution: 60%), Christoph Funk (contribution: 20%), Ramona Teuber (contribution: 10%), Lutz Breuer (contribution: 10%)
Status:	Published: *Sustainable Development*, 2024, pp. 1–22
Available from:	`https://doi.org/10.1002/sd.2906`

Earlier versions of this paper were presented at:

- 24th International Conference on Computational Statistics, Bologna, Italy (presented by Co-Author)
- CSDA & EcoSta Workshop on Statistical Data Science, Bologna, Italy (presented by Co-Author)
- SDGnexus Network Seminar Lecture Series, Giessen, Germany (presented by Co-Author)

Reading Between the Lines: The Intersection of Research Attention and Sustainable Development Goals

Christoph Funk[*,†] Elena Tönjes[‡] Ramona Teuber[*,§] Lutz Breuer[*,¶]

Abstract

In September 2015, the United Nations (UN) adopted 17 Sustainable Development Goals (SDGs) to transform our world by 2030. The scientific discourse around these SDGs has expanded rapidly since then, highlighting the need for efficient analysis of the large amount of textual data using Natural Language Processing. Our research addresses this need by employing a zero-shot text classification for SDG-related scientific articles, which allows for a thorough examination of scholarly discourse and the relationship between research attention and SDG achievement. We introduce the Research Attention Index (RAI), a novel metric that quantifies the research attention each SDG receives within a specific country. Our study contributes to the existing literature by providing a holistic view of global research attention to the SDGs. It also demonstrates the effectiveness of zero-shot text classification for large-scale textual labeling, and underlines the relevance of abstract analysis in understanding SDG-related discourse. Moreover, we examine the (non)-linear relationship between the RAI and SDG achievement across countries. Our results indicate considerable variations in the scientific discourse across countries worldwide and reveal a complex, non-linear relationship between research attention and progress towards achieving the SDGs. This underscores the importance of understanding the dynamics between research attention and sustainable development outcomes.

Keywords: Natural Language Processing, Research Attention Index, Sustainable Development Goals, SDG

[*] Center for International Development and Environmental Research (ZEU), Justus Liebig University Giessen, Senckenbergstrasse 3, 35390 Giessen, Germany

[†] Corresponding author: Christoph.Funk@wi.jlug.de

[‡] Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

[§] Institute for Agricultural Policy and Market Research, Justus Liebig University Giessen, Senckenbergstrasse 3, 35390 Giessen, Germany

[¶] Institute for Landscape Ecology and Resources Management (ILR), Research Center for Bio Systems, Land Use and Nutrition (iFZ), Justus Liebig University Giessen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

2.1 Introduction

In September 2015, the United Nations (UN) launched an ambitious agenda aimed at transforming the world by 2030. To tackle urgent development issues in areas like gender equality, infrastructure, environment, and education, this plan emphasizes global collaboration among all nations. At the heart of this transformative 2030 Agenda are 17 Sustainable Development Goals (SDGs), consisting of 169 interrelated targets and monitored by 231 indicators, which every country worldwide is expected to achieve (Pedercini et al., 2019; United Nations, 2016b).

The scientific discourse around these SDGs has been expanding rapidly. While there is existing literature on the 2030 Agenda and the 17 SDGs established by the UN, this topic will be explored extensively in the 'Literature Review' section of this paper. It is important to note, however, that much of the current literature tends to focus on small text samples or relies on predefined numerical indicators set by the UN (Fuso Nerini et al., 2019; Pedercini et al., 2019).

Our study aims to contribute to this emerging field by applying state-of-the-art Natural Language Processing (NLP) techniques, specifically using pre-trained language models like Bidirectional Encoding Representations from Transformers (BERT) (Devlin et al., 2018), to a large corpus of SDG-related scientific articles. In doing so, we seek to uncover the underlying structure of the scientific discourse on the SDGs and to explore the relationship between research attention and SDG achievement.

Using a language model, we apply zero-shot text classification, a powerful NLP technique that can automatically classify texts into predefined categories without requiring any training data for those categories. This technique is particularly advantageous in the context of SDGs, where the classification categories (i.e., the 17 SDGs) are well-defined but labeled training data may be scarce or unavailable. Thus, we use zero-shot text classification to assign each scientific article in our dataset to one or more SDGs based on its content. The use of zero-shot text classification highlights the potential of NLP techniques to augment traditional methods in SDG research. By leveraging the capabilities of advanced language models, we can extract rich semantic features from text and conduct large-scale analysis that would be challenging, if not impossible, with manual methods. This method significantly differs from traditional text classification techniques in several key aspects. Primarily, it eliminates the need for a training phase, as it relies solely on the intrinsic semantics of the text and predefined labels. This approach inherently reduces biases commonly associated with training data and eliminates the costs and efforts involved in labeling such data. Moreover, its flexibility allows policymakers to easily tailor the model to suit specific requirements, ensuring that the classification process aligns closely with their evolving needs.

To better understand this relationship, we introduce a new Research Attention Index (RAI), to provide a quantitative measure of research attention dedicated to each SDG within a country, and thus providing a metric that can complement traditional indicators in capturing

the dynamics of SDG-focused scientific discourse. By quantifying research attention, the RAI offers a novel perspective to understand how global research attention is distributed across different SDGs and countries. It enables us to explore whether there is a correlation between the level of research attention and the progress made in SDG achievement, providing insights that could guide research funding and policy decisions. Furthermore, it allows us to determine whether the distribution of research attention aligns with global development priorities, thereby identifying potential gaps or imbalances.

Our research contributions are four-fold. First, we provide a comprehensive overview of global research attention on SDGs. Second, we demonstrate how zero-shot text classification can efficiently label extensive textual information, providing policymakers with information beyond the typical UN indicators in an efficient manner. Third, we introduce the novel RAI to quantify each SDG's research attention per country, and we analyze its (non)-linear relationship with SDG achievement. Fourth, we showcase that abstracts are sufficient for capturing the most relevant SDG-related information in scientific articles, thereby saving computational resources.

The remainder of this paper is organized as follows: The section 'Literature Review' provides an overview of the relevant literature related to our work. In the 'Data' section, we describe our dataset, providing a descriptive analysis that emphasizes key characteristics of our texts, such as the word counts related to each SDG, and the geographical distribution of our research articles. This section is followed by the 'Methodology' section, which explains the techniques we employed, including zero-shot classification, and presents an overview of the labels used for our analysis. The 'Abstracts versus Full Texts' section provides a comprehensive comparison between the use of abstracts versus full texts in the context of our RAI. In the 'Results' section, we present our primary findings followed by a discussion in the 'Discussion' section. Concluding remarks can be found in the 'Conclusion' section.

2.2 Literature Review

While the 2030 Agenda and its 17 SDGs, as launched by the UN, represent a global commitment to addressing pressing issues like gender equality, infrastructure, environment, and education, the scientific discourse on these topics is rapidly evolving. Existing literature often relies on small text samples or confines its assessment to predefined numerical indicators set by the UN (Fuso Nerini et al., 2019; Pedercini et al., 2019). For example, Bennich et al. (2020) manually assessed 70 peer-reviewed articles to provide an overview of SDG interactions in the scientific literature. Le Blanc (2015) conducted a network analysis to uncover relationships among SDGs based solely on the wording of the targets, while Bali Swain & Ranganathan (2021) performed a network analysis to reveal synergies and trade-offs exclusively based on the SDG targets. These small-scale or theory-based studies either capture only a subset of the scientific literature addressing the SDGs, potentially leading to a biased sample, or focus on theoretically assumed relationships among the SDGs.

Another segment of the existing literature consists of manual literature reviews that

analyze and discuss a limited number of papers on the SDGs in general or on specific SDGs (Hackl, 2018; Decouttere et al., 2021). While valuable, these studies do not fully exploit the potential of large-scale text analysis. NLP offers a compelling alternative, capable of handling vast textual data and extracting meaningful patterns (Smith et al., 2021). A small but growing number of studies have employed NLP techniques to analyze SDG-related texts, uncovering valuable insights and potential collaboration opportunities. Sebestyén et al. (2020) used NLP to perform a network analysis of the SDGs in Voluntary National Reviews, uncovering informative word pairs to draw SDG connections between countries. The Voluntary National Reviews share experiences, successes, challenges, and lessons learned to accelerate the implementation of the 2030 Agenda for Sustainable Development (United Nations Department of Economic and Social Affairs, 2019). Findings from studies like this can help reveal the SDG areas where countries face similar challenges and where they might collaborate (Sebestyén et al., 2020). Another study by Smith et al. (2021) used 85 UN progress reports on each SDG to measure the dependencies and interactions between SDGs and identify potential clusters. The authors also assessed whether their findings on SDG inter-dependencies based on the UN progress reports were reflected in scientific articles from the last two decades. Additionally, Chang et al. (2021) used cluster analysis and topic modeling on environmental education research journals from the Web of Science during 2011-2020 to analyze research topics. The techniques employed in Chang et al. (2021), such as encoding the text with term frequency-inverse document frequency and using a combination of Latent Dirichlet Allocation for topic modeling and K-means algorithm, are well-established concepts in NLP (Sebestyén et al., 2020; Smith et al., 2021; Chang et al., 2021).

Recent studies, such as Bellantuono et al. (2022), have begun integrating large language models like BERT for SDG analysis. They trained a BERT model for a multi-label classification task with the 17 SDGs as labels. Their training and analysis dataset consisted of Japanese texts from various sources. The primary aim was to harness the model's capabilities to predict connections among different SDGs using a dataset containing representative indicators. This approach aimed to uncover opportunities for uniting stakeholders interested in collaborating on SDG initiatives. Angin et al. (2022) also use a large language model, specifically utilizing an advanced RoBERTa model. This model excels in dissecting sustainability reports, identifying key sections related to SDGs. For development and validation, they employed the OSDG Community Dataset, which is extensively annotated with SDG-focused text by community members. Notably, this study leans towards a more technical and model-centric approach. It includes thorough comparative testing against both traditional machine learning and contemporary deep learning models, demonstrating the model's robustness and precision.

Despite the growing body of research employing NLP, there remains a need for more sophisticated, large-scale textual analysis that can capture the SDG discourse in the scientific community in depth. Our study takes a substantial step forward by employing the pre-trained language model BERT and the innovative technique of zero-shot text classification. By applying these techniques to an extensive corpus of SDG-related scientific articles, we aim to unveil the intricacies of SDG-focused scientific discourse, providing a detailed and

comprehensive understanding. To address these gaps, our research proposes the following questions:

1. How can advanced NLP techniques, particularly pre-trained language models and zero-shot text classification, be effectively utilized to analyze and interpret large-scale textual data related to the SDGs?

2. What novel insights and patterns can these advanced NLP techniques reveal in the scientific discourse on SDGs that are not evident in smaller-scale or traditional studies?

3. How does the application of these techniques inform the understanding of the relationship between research attention (as quantified by our RAI) and the progress in achieving SDGs?

By answering these questions, our study aims to contribute significantly to the field. We intend to provide an in-depth analysis of global research attention on SDGs and explore its implications for global development strategies. This approach not only enhances the understanding of SDG-focused discourse but also informs future research and policy-making in the realm of sustainable development.

2.3 Data

In our study, we utilized the Web of Science (www.webofscience.com) to conduct a comprehensive search for articles related to 'Sustainable Development Goals'. As of May 07, 2023, we identified 17,098 articles that matched our criteria. The methodology for determining our dataset is outlined in Figure 2.1. Since our model is specifically designed to analyze English text data, we limited our dataset to articles written in English, which reduced the dataset by 513 articles.

To ensure relevance and alignment with the Agenda 2030, we focused on articles published between the years 2015 and 2022. This decision was made based on the assumption that articles written before 2015 might not consistently align with the SDGs. As a result, we excluded 1,406 furthrer articles written prior to 2015. Furthermore, we chose not to incorporate articles from 2023 into our study, given that the corresponding SDG indicators for 2023 have not yet been released. The dataset utilized in our study comprises titles and abstracts provided by the Web of Science. These texts are machine-readable and available in English, even if the full articles are not, which made them easily accessible for our analysis.

However, we encountered a limitation in our dataset as 554 articles lacked abstracts in the Web of Science metadata. To maintain a concentrated focus on the SDGs, we included only articles that explicitly referenced the term in any form. We employed a regex function, detailed in the following section, to filter the articles accordingly. As a result, we excluded 1,487 articles from our dataset. This left us with a final count of 13,138 articles, which formed the basis for our analysis.

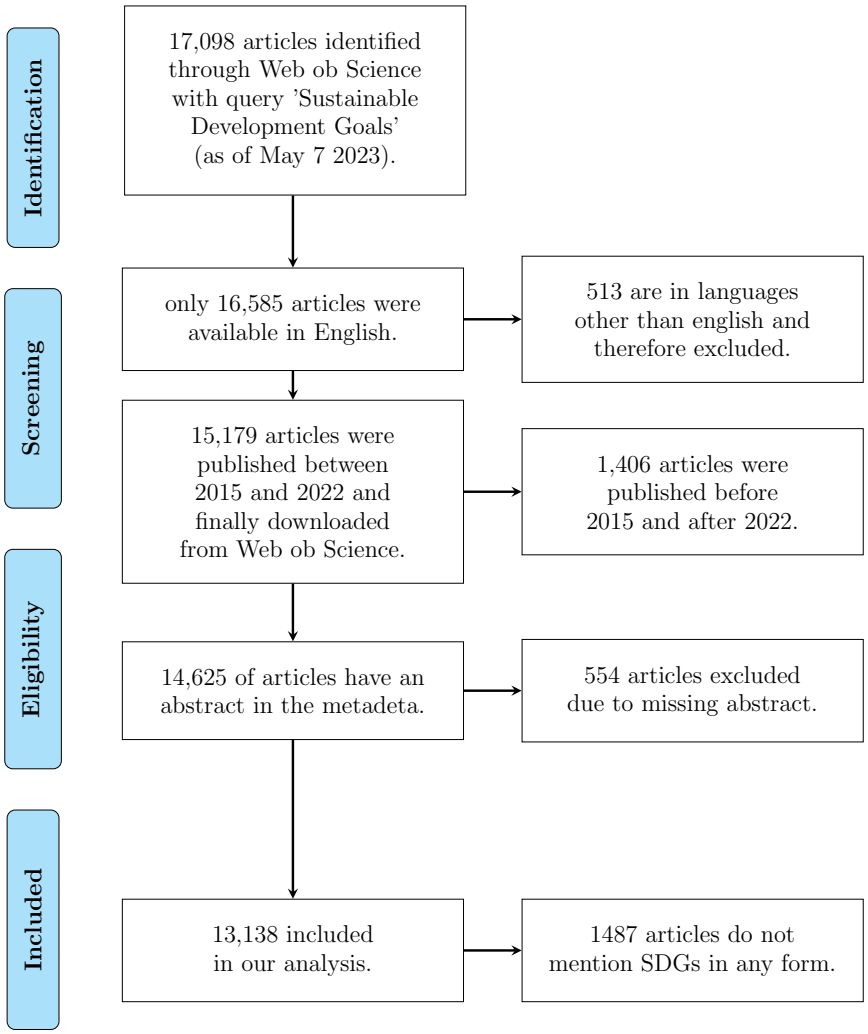


Figure 2.1: PRISMA Flow Diagram of SDG article data collection.

2.3.1 Descriptive Analysis

Descriptive analysis is crucial for understanding the distribution and prevalence of the 17 SDGs in our dataset. This enables us to provide an overview of the frequency distribution of each SDG, which can help researchers and policymakers identify the goals that are receiving more attention in the literature and potentially prioritize further research or interventions based on the analysis. To determine which of the 17 SDGs is most commonly mentioned in our dataset, we used regular expressions to count the occurrences of SDGs per article. We counted the acronym SDG and the tri-gram 'Sustainable Development Goal' as SDG, while the acronym SDGs and 'Sustainable Development Goals' were counted as SDGs. Furthermore, we included various combinations of each of the 17 SDGs, such as 'SDG1', 'SDG 1', 'SDG one', and combinations like 'SDG 1.1', 'SDG 1a', and so on, as SDG 1.

Table 2.1 presents a comprehensive overview of the total number of SDGs counted, the number of abstracts in which each specific SDG appears at least once, and the average number of mentions per abstract. As anticipated, the generic terms 'SDG' and 'SDGs' were most frequently used, appearing 36,366 times across 13,136 abstracts. On average, these

terms were mentioned 2.77 times per abstract and title.

We noticed a significant variation in the frequency of mentions for each SDG throughout the dataset. SDG 3 - *good health and well-being* was the most commonly cited goal, appearing in 395 abstracts with a total count of 564 mentions and an average of 1.43 mentions per abstract. SDG 6 - *clean water an sanitation* followed as the second most frequently mentioned goal, featured in 288 abstracts with a total of 453 mentions and an average count of 1.57 mentions per abstract.

Conversely, SDG 16 - *peace, justice, and strong institutions* was the least frequently cited goal in our dataset. It appeared in only 101 abstracts, with a total count of 150 mentions and an average of 1.49 mentions per abstract. This wide variation in the frequency of SDG mentions highlights the diverse range of research topics and focus areas within the context of sustainable development.

Table 2.1: Counted SDGs for all Abstracts.

SDG	Short Description	Total Count	Number of Abstracts	Average Count per Abstract
1	No Poverty	249	184	1.35
2	Zero Hunger	330	255	1.29
3	Good health and well-being	564	395	1.43
4	Quality education	301	212	1.42
5	Gender equality	206	160	1.29
6	Clean water and sanitation	453	288	1.57
7	Affordable and clean energy	340	257	1.32
8	Decent work and economic growth	233	193	1.21
9	Industry, Innovation and infrastructure	137	115	1.19
10	Reduced inequalities	147	125	1.18
11	Sustainable cities and communities	339	238	1.42
12	Responsible consumption and production	258	211	1.22
13	Climate action	329	277	1.19
14	Life below water	213	135	1.58
15	Life on land	269	187	1.44
16	Peace, Justice and strong institutions	150	101	1.49
17	Partnership for the goals	154	122	1.26
	SDG / SDGs	36,366	13,136	2.77

This table presents a comprehensive overview of the total number of SDGs counted, the number of abstracts in which each specific SDG appears at least once, and the average number of mentions per abstract.

2.3.2 SDG Co-Occurence

By examining the co-occurrence of SDGs within articles, we can identify possible relationships or interdependencies among the goals. This can provide insights into how researchers and policymakers might address multiple SDGs simultaneously, leading to more efficient and effective interventions.

Figure 2.2 illustrates the co-occurrence of SDGs in our dataset. The lines connecting the SDGs indicate the number of times two specific SDGs have been mentioned together in the same article. For clarity and readability, we have chosen to highlight only those connections that are present in 40 or more articles. This visualization aids in analyzing the interconnections among SDGs and identifying the most frequently co-occurring ones in the research discourse. It is apparent that certain SDGs are cited together more often, suggesting robust interrelationships or intersecting research interests. On the contrary, SDGs with fewer co-occurrences may reflect a lesser degree of interconnectivity or a narrower focus of research concerning those specific objectives.

For example, SDG 7 - *affordable and clean energy* and SDG 13 - *climate action* have been mentioned together in 107 abstracts. This frequent co-occurrence is not unexpected, as these goals are intrinsically linked in addressing global environmental challenges. Both goals aim to tackle the negative impacts of human activities on the environment and promote sustainable development. The frequent co-occurrence of SDGs 7 and 13 in research articles reflects the interconnected nature of these goals and the common understanding that addressing energy and climate issues simultaneously is crucial for sustainable development.

Additionally, SDG 1 - *no poverty* and SDG 2 - *zero hunger* have been mentioned together the second most times with 84 articles, followed by SDGs 12 and 13 with 79 times. While SDG 1 aims to end poverty in all its forms, SDG 2 focuses on ending hunger and malnutrition. Hunger and malnutrition are directly linked to poverty, as people living in poverty often lack the resources to access or produce enough nutritious food for themselves and their families. These interconnected issues require integrated solutions to promote sustainable development and ensure food security for vulnerable populations.

Moreover, the co-occurrence of SDG 12 - *responsible consumption and production* and SDG 13 - *climate action* highlights their intrinsic connection through the understanding that responsible consumption and production are essential components of climate action. Sustainable production and consumption patterns help reduce greenhouse gas emissions, minimize waste, and conserve natural resources, ultimately contributing to climate change mitigation and adaptation.

In general, our analysis of the co-occurence reveals that SDG 13 - *climate action* appears to be a central theme, as it exhibits strong interlinkages with SDGs 2 - *zero hunger*, 3 - *good health and well-being*, 6 - *clean water and sanitation*, 8 - *decent work and economic growth*, 11 - *sustainable cities and communities*, 12 - *responsible consumption and production*, and 15 - *life on land*. This central role of SDG 13 suggests that climate action and climate change as a global threat not only affects the environment but also has wide-ranging implications

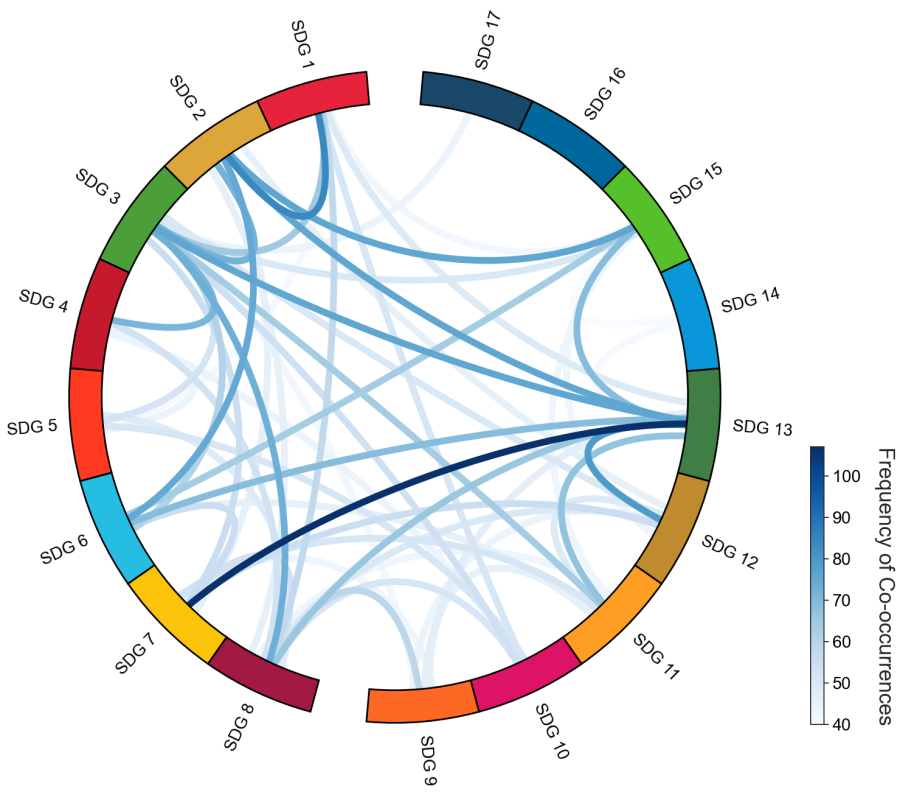


Figure 2.2: SDG Co-occurrence plot.
The lines connecting the SDGs represent the frequency at which two specific SDGs are mentioned together within the same abstract. The colorbar provides a visual representation of the strength of these relationships.

for human well-being, economic development, and social equity.

It is important to note that the number of mentions alone or the number of co-occurrences cannot conclusively determine the importance of individual SDGs or connections between them. For instance, an article may discuss actions required to combat climate change and its impacts without explicitly mentioning SDG 13 - *climate action*. Therefore, this descriptive analysis should be seen as a foundation for more advanced text analysis techniques, such as zero-shot text classification discussed in our 'Result' section. These methods can offer a more nuanced understanding of the significance of each SDG in the dataset, as well as the broader context in which they are discussed, thereby providing a more comprehensive insight into the research landscape surrounding the SDGs.

2.3.3 Geographical Patterns

A further analysis provides valuable insights into the global scientific discourse surrounding the SDGs. By examining our dataset in terms of the countries mentioned, we can discern geographical patterns and relationships between articles and their associated countries. To achieve this, we use the 'find_countries' function from the 'country_named_entity_recognition' library in Python, which identifies countries in the combined text of each article's title and abstract. The rationale behind this is to determine the main country or countries an article

focuses on. If a country is mentioned in either the title or abstract, we assume that the article is centered on that country. In cases where an article discusses multiple countries simultaneously, it will be counted towards all mentioned countries.

We use Python along with the 'pycountry' library to extract country objects from the tuples and append the country codes (ISO 2 and ISO 3) to the 'Articles' DataFrame. The 'country_named_entity_recognition' library is case-insensitive, ensuring accurate country detection regardless of text case variations. We manually verified the code's effectiveness by randomly sampling 30 abstracts, and it performed well in all instances.

In the Appendix, Table 2.5 provides a summary of research article counts by ISO2 country code. The top five countries with the most articles are China (CN) with 825, India (IN) with 623, the United States (US) with 309, South Africa (ZA) with 265, and Brazil (BR) with 238. The table includes counts for numerous other countries with at least five research articles, illustrating the global scope of research attention.

2.4 Methodology

2.4.1 Zero-shot Text Classification

Text classification is a crucial NLP task for leveraging knowledge in texts. It involves categorizing sentences, paragraphs, or entire documents into predefined labels. Deep neural models are commonly used for this task, but they require large amounts of labeled data for training. However, creating labeled data is expensive and time-consuming, and even human annotators can struggle to assign the most appropriate label to a text, despite fully understanding its semantics (Beltagy et al., 2019).

Zero-shot classification is a technique that facilitates text classification without necessitating training data. Traditional classification models typically undergo training, during which they learn patterns from the text, enabling them to categorize new, previously unseen texts. However, zero-shot classification distinguishes itself by its ability to classify texts with minimal training. This is achieved by leveraging the inherent semantics of both the text and the labels, instead of depending solely on learned patterns. One approach to achieve this is by embedding sentences and labels into the same latent space and computing the distance between them. This approach depends exclusively on the semantics of the sequence and label for classification (Socher et al., 2013). The model we employ, developed by Davison (2020), utilizes a unique embedding process outlined as follows:

- Select the top K words, referred to as V, from the most frequently used words in a word2vec model's vocabulary.
- Generate embeddings for these words through word2vec, represented as $\Phi_{\text{word}}(V)$.
- Similarly, produce embeddings for the same set of words using Sentence-BERT (S-BERT), indicated as $\Phi_{\text{sent}}(V)$.

- Develop a linear projection matrix Z, optimized by least-squares with $L2$ regularization, mapping $\Phi_{\text{sent}}(V)$ onto $\Phi_{\text{word}}(V)$.

Here, Z functions as an additional transformation layer, enhancing the S-BERT embeddings for both the sequence and the labels. However, our approach extends this concept by also interpreting sentences as labels, thereby allowing the mapping of entire sentences into the latent space. We adopt a variant of this technique that includes multi-natural language inference (MNLI) for advanced sentence classification. In this process, a given sequence is considered a premise, and the labels are formulated as hypotheses. These hypotheses are then evaluated against the premise to ascertain whether they entail, contradict, or are neutral to the premise. This method aligns with the architecture proposed by Yin et al. (2019). For instance, Davison (2020) provided an example where the sentence 'A soccer game with multiple males playing' is the premise, 'Some men are playing a sport' acts as the hypothesis, and the resulting label is 'entailment'. This example aptly demonstrates how MNLI can be applied to classify sentences based on their inferred relationships.

In previous work, Yin et al. (2019) transformed naturally classified pairs into binary categories, specifically 'entailment' versus 'non-entailment', to produce a binary outcome. They employed a pre-trained MNLI BERT model for zero-shot testing on this modified dataset. In our study, we have adapted the zero-shot classification model as developed by Davison (2020), which is available in the Hugging Face's Transformers library (Wolf et al., 2020). This model leverages the Bidirectional and Auto-Regressive Transformers (BART) framework, which is used for tokenizing both the input sequence and the label. The BART model, as described in Lewis et al. (2019), uniquely integrates a sequence-to-sequence translation framework, combining bidirectional encoding capabilities (similar to BERT) with unidirectional, left-to-right encoding (reminiscent of the GPT model). In context of sequence classification tasks like the MNLI problem addressed in our study, the BART model operates by feeding the same input into both its encoder and decoder components. The crucial aspect of this model is how it processes the input: the final hidden state token from the decoder is passed through a multi-class linear layer, producing a logit vector. Davison (2020) enhanced this model by applying a softmax function to the logit classification scores. This allows for the calculation of probabilities associated with 'entailment' and 'contradiction' while deliberately omitting the 'neutral' classification score. These probabilities are then interpreted as the likelihood that the labels accurately correspond to the sequences inputted into our system.

In summary, our zero-shot classification model relies on the pre-trained MNLI BART model, as developed by Davison (2020) and integrated in the transformer library (Wolf et al., 2020). This model's effectiveness lies in its ability to interpret and classify sequences by analyzing both the semantics of the input sequences and the labels fed into the model.

2.4.2 Model Validation

Zero-shot classification offers the advantage of not requiring a large labeled dataset for training the model. However, it also lacks a common validation procedure, such as splitting

the labeled dataset into train and test sets to evaluate the model's performance. To address this, we validated our model and labels using the UN's annual progress reports (United Nations Economic and Social Council, 2016, 2017, 2018, 2019, 2020, 2021) for each SDG. These reports should contain similar semantics about SDG performance as scientific articles that write about SDGs. We used these reports as a proxy to validate our labels and model by labeling each sentence in the progress reports with the corresponding SDG. This resulted in a labeled dataset of 2,301 sentences without any labeling effort. We then applied the zero-shot classification model with our defined labels to all sentences and verified whether the probabilities matched the labels.

Figure 2.3 illustrates our results, and Table 2.6 in the Appendix provides a detailed overview of the labels we used. To test the performance of different label lengths, we used two labels for each SDG, a shorter and a longer version. Figure 2.10 in the Appendix presents the results for both versions separately, and Figure 2.3 shows the mean across the long and short labels for each SDG. Initially, we included labels for all SDGs consisting of either the SDG number or the SDG number spelled out as a word (e.g., SDG 1 or SDG one), but we decided not to include these labels in our final analysis as they did not capture the semantic meaning of 'SDG' and performed poorly in our validation approach.

As we allow for multiple labels to be true, we have a probability distribution for each label over all pre-labeled sentences. The x-axis represents the sentences from the progress reports, grouped by the corresponding SDG and thus by the UN progress report label. The probability for a label should be highest for the sentence from the corresponding progress report, as the report should mostly contain sentences about the matching SDG. Therefore, for the version where we included the long and short label, both squares on the diagonal should have the darkest color, as a darker color indicates a higher probability. The same should hold for the average over both labels, where the squares on the diagonal should be darkest. In Figure 2.10 in the Appendix with the long and short versions, we can observe that the probabilities are typically highest on the diagonal, with one label usually performing slightly better than the other. Nonetheless, we opted to use the mean of both labels since we are only applying the model on a finite sample, and with a slightly different dataset, some results could be the opposite. Figure 2.3 presents a clearer picture of our results. In most cases, the diagonal is distinguished by the highest probability in the row. However, some labels have multiple high probabilities within their respective row. Notably, SDG 10 stands out as the probability of discussion is high for most SDG labels. This could be attributed to the fact that SDG 10 has numerous linkages to all the other UN SDGs (United Nations, 2016a). Moreover, "reduce inequality" can be applied to several other SDGs and can be framed as a general goal that needs to be achieved.

The validation heatmaps shown in Figure 2.3 reveal some interesting findings. Firstly, it is notable that sentences from all progress reports are related in some way to SDG 10, which aligns with the goals included in SDG 10 itself. Secondly, sentences related to SDG 7 are more likely to be classified as sentences related to SDG 12 and SDG 13, which can be explained by the linkages illustrated by the UN. The link between SDG 7 and SDG 13

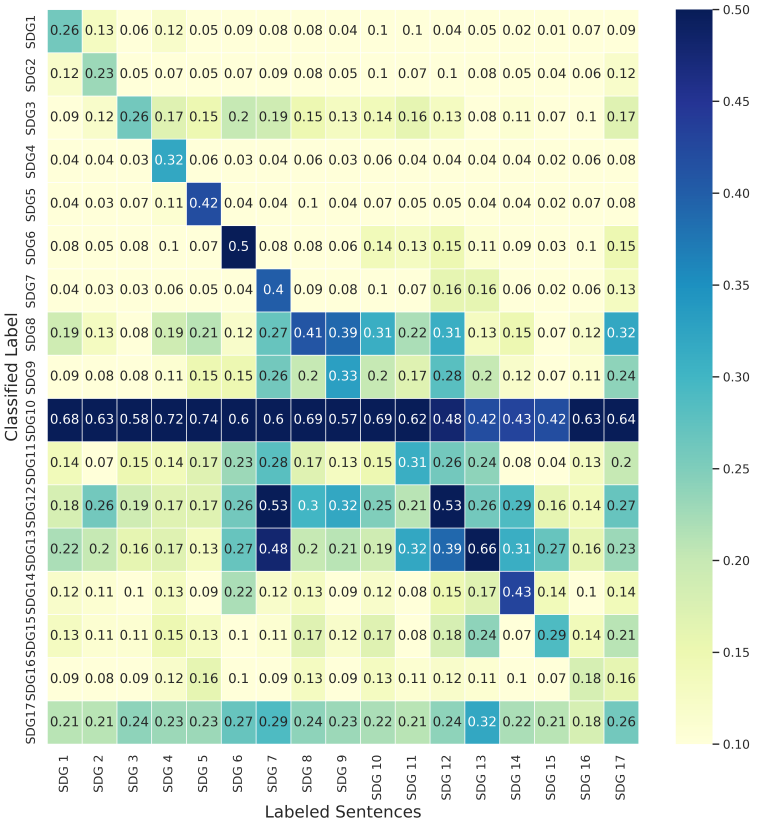


Figure 2.3: Model validation with mean values from short and long labels by comparing the sentences classified by the model on the y-Axis with the labels classified by humans on the x-Axis.

and the link between SDG 13 and SDG 12 can also be seen in the linkages found in our co-occurrences in Figure 2.2. However, sentences on affordable energy seem to fit *climate action* and *responsible consumption and production*, but not vice versa. Furthermore, in section 'SDG Co-Occurence' we find linkages between SDG 1 and SDG 2. Those linkages can be seen in Figure 2.3 as well. Sentences about SDG 1 are classified as SDG 2 with the second highest probability. The same is true vice versa for sentences about SDG 2.

Although the labels for SDGs 4, 5, 8 and 10 have the second highest probability after SDG 10, the label for SDG 16 does not work as well as the other labels, as the probability of sentences actually being about SDG 16 is not as high. Additionally, the correlations between SDGs 16, 11 and 14 are not captured by the label. The label for SDG 17 also does not show the highest probability for the sentences that specifically talk about it; instead, it has similar probabilities for all SDGs.

The validation results shown in Figure 2.3 highlight the importance of using a multi-class model instead of a single-class model. Our findings, which are consistent with the linkages provided by the UN (United Nations, 2016a), demonstrate that sentences often relate to multiple SDGs instead of only one. For example, SDG 17 is about partnership for the goals

and is therefore indirectly included in all other goals, which is reflected in our validation results. These results underscore that simply counting the word 'SDG' along with the numbers 1 through 17 would not adequately capture the semantics of sentences that discuss multiple SDGs without explicitly mentioning the term 'SDG'. Our personal experience in reviewing the gold standard labels also supports these findings, as humans would often assign two SDG labels to a single sentence.

2.5 Abstracts versus Full Texts

To strengthen and validate our study's conclusions, we expanded our analysis from abstracts to full texts. This expansion was undertaken to solidify the robustness of our findings, as relying solely on abstracts might not capture the full depth of the research related to SDGs. Our comparison revealed that the analysis of abstracts alone yielded results comparable to those obtained from full texts. For this comparison, we curated a distinct dataset focused exclusively on open-access publications from 2015 to 2022. Based on the 13,138 abstracts included in our initial analysis, we identified 8,503 open-access articles using Web of Science metadata. Of these, we successfully downloaded 7,257 full-text articles. After obtaining the full-text articles, the next step involves preparing these documents for subsequent processing, a crucial stage before employing the zero-shot text classification, as described in the next section.

The 'From PDFs to raw text data' section in the Appendix outlines our process of extracting text from research articles using the Visual Layout (VILA) tool. We successfully retrieved text from 7,229 PDFs, converting this data into CSV format for more efficient processing. Through subsequent cleaning steps, which included filtering for documents with a tagged main body and verifying the language as English using the Python Langdetect library, we prepared a final dataset of 7,090 articles for comparative analysis. Similar to the methodology used for the abstracts, we applied zero-shot text classification to the full texts of these articles to derive the RAI. To provide a clear comparison between the two datasets, Figure 2.4 presents the RAIs for the abstracts, from the 7,090 articles, and their corresponding extracted full texts. This comparative analysis allows us to evaluate any discrepancies or patterns emerging from the abstracts versus the full text, thereby enriching our understanding of the distribution of research attention across the SDGs.

As illustrated in Figure 2.4, we note that SDG 10 consistently exhibits the highest probability of attention across the abstracts and complete documents. This observation aligns with our validation findings, suggesting that SDG 10 is often referenced, even in contexts primarily discussing other SDGs.

We observed a significant similarity between results derived from abstracts and full texts, indicating that abstracts alone have the capacity to effectively encapsulate the core message of the articles. The average difference between the full texts and abstracts is minor, at just 0.0056 points. The most noticeable deviations were found for SDGs 10 and 17, with differences of 0.0334 and 0.035 points respectively. These findings, as illustrated in Figure 2.4,

have notable implications not only for the precision of analysis, but also for computational efficiency and the potential for significant time savings.

These findings underscore the value of our initial decision to focus on abstracts in our analysis, as our goal is to pinpoint sections where the scientific community explicitly discusses and gives attention to the SDGs. In fact, our results reaffirm the function of abstracts as concise summaries of the most salient points of an article. We observe that the RAI values are generally similar for abstracts compared to full articles, thereby validating our approach.

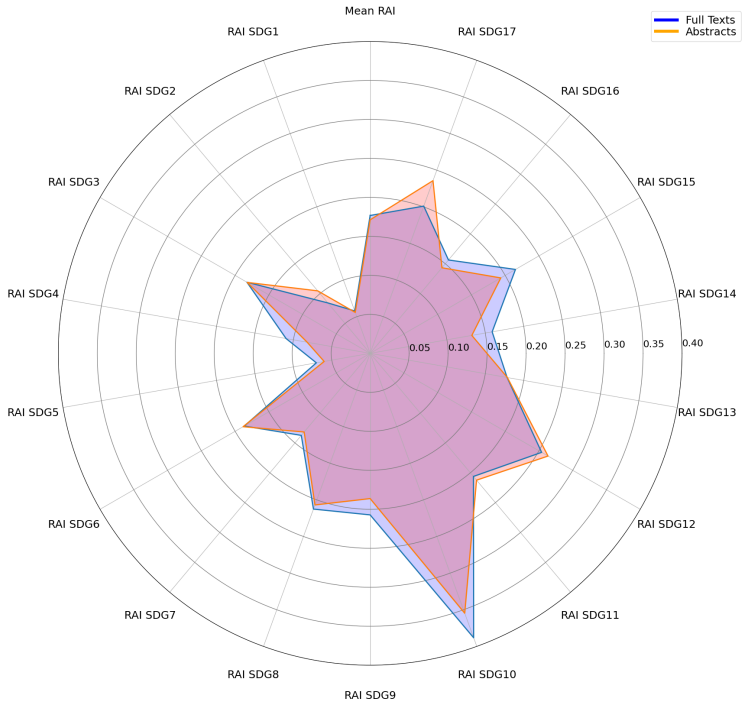


Figure 2.4: Comparison of the Research Attention Index (RAI) for all SDGs, utilizing both abstracts and full articles.

2.6 Results

2.6.1 Research Attention Index

We propose the novel RAI that enables us to compare the relationship between SDGs and research attention at a country level. To achieve this, we utilize the geographical distribution across countries explored in section 'Geographical Patterns' and combine it with the results of our zero-shot classification. We compute the mean probabilities of all sentences in an abstract and title, regarding the SDGs being discussed for a country, and use this as a proxy for the research attention or the SDG discourse in a given country. The zero-shot classification results yield a probability that can be interpreted as a natural index, with a spectrum from zero (indicating no research attention to an SDG in a country) to one (signifying maximum attention to an SDG).

Furthermore, it is important to consider that the RAI we propose here is a relative

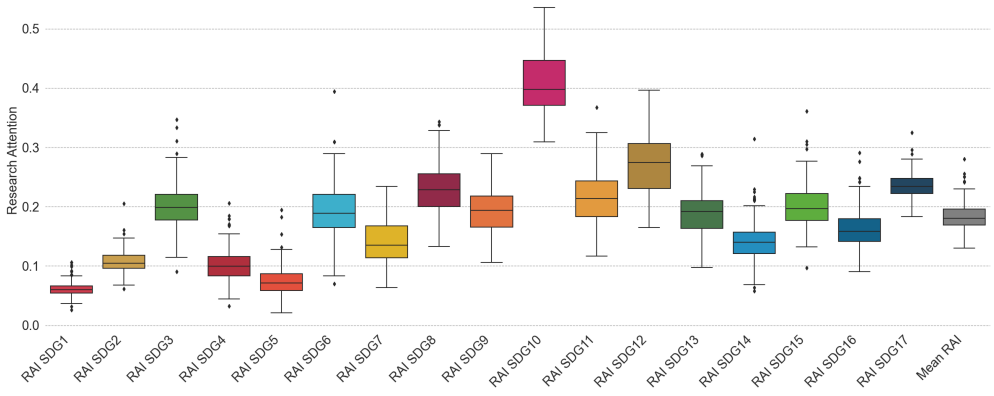


Figure 2.5: Boxplots representing research attention, based on the mean for countries with at least five research articles.

measure and may not accurately reflect the absolute levels of research attention an SDG receives within a country. Factors such as publication accessibility, language barriers, and regional research priorities can influence the distribution of research articles across countries. Consequently, the index should be perceived as a comparative tool to gauge research activities connected to SDGs across different nations, rather than an absolute measure of research attention an SDG receives.

Figure 2.5 presents a boxplot to summarize the distribution of research attention across all countries in our dataset with five or more research articles. It is essential to note that we reduced our dataset to 5,646 abstracts, but the results are qualitatively similar for the mean values of research attention provided for the full dataset, as illustrated in Figure 2.8 in the Appendix. This finding suggests that there is no apparent bias in the datasets with respect to the distribution of research attention. When comparing the distribution of mean values over countries obtained from the zero-shot classification (Figure 2.5), we find that SDGs 1, 2, 4, and 5, which pertain to *no poverty*, *zero hunger*, *quality education*, and *gender equality*, respectively, had the lowest research attention index, with less than 0.11 on average. These goals appear to be less prominent in the scientific discourse of articles indexed in Web of Science.

On the other hand, we find that discourses on SDGs 3, 6, 8, 9, 11, 13, and 15, which include the goals *good health and well-being*, *clean water and sanitation*, *decent work and economic growth*, *industry, innovation, and infrastructure*, *sustainable cities and communities*, *climate action*, and *life below water*, have a similar research attention score on average ranging between 0.18 and 0.22. In contrast, the top two SDGs in terms of research attention are SDG 10 - *reduced inequalities* and SDG 12 - *responsible consumption and production*. Overall, these findings underscore the reality that research attention towards the SDGs is not uniformly distributed across countries, and there is substantial variation in the level of research focus accorded to different SDGs.

A further comparison of the RAI with the number of articles for each SDG (Table 2.4 in the Appendix) highlights both similarities and differences in representation of SDGs across

research articles. The top five SDGs in the RAI are SDG 10, SDG 12, SDG 17, SDG 11, and SDG 8, while the top five SDGs with the highest number of articles from simple counting are SDG 3, SDG 6, SDG 2, SDG 11 and SDG 7.

Comparing these two methods highlights some similarities and differences in the representation of each SDG in the research landscape. For instance, both methods indicate high representation for SDG 11 - *sustainable cities and communities* and medium-to-high representation for SDG 3 - *good health and well-being*, while SDG 16 - *peace, justice, and strong institutions* consistently ranks low.

Contrastingly, SDGs 10 - *reduced inequalities* and 12 - *responsible consumption and production* rank highly in zero-shot results but lower in article mentions. SDG 6 - *clean water and sanitation* ranks second in article mentions but eighth in zero-shot results. Considering that the zero-shot classification is a more sophisticated method providing deeper insights, it offers a more nuanced understanding of research attention related to each SDG by capturing the broader context and relationships between SDGs, resulting in a more comprehensive understanding of the research landscape. This comparison emphasizes the importance of examining research coverage from multiple perspectives to better understand attention distribution across SDGs, guiding future efforts to address gaps and enhance overall understanding.

2.6.2 (Non)-Linear Relationship between Research Attention and SDGs

In the next step, we investigate the relationship between scientific discourse and actual development towards the SDGs. There are two plausible scenarios to consider: first, countries facing challenges in achieving the SDGs may be more likely receive attention towards them in the scientific community. Alternatively, it is possible that countries with higher levels of scientific discourse on a particular SDG may be more likely to achieve progress towards that SDG.

In our initial analysis, we delve into the official SDG Index scores and present our findings through Figure 2.9 in the Appendix, which features a boxplot for each SDG based on the official SDG Index score (United Nations, 2021) for countries within our sample. Several noteworthy observations emerge from these boxplots. Notably, many countries perform well in regards to SDGs 1, 4, 12, and 13, with the majority of countries achieving a median SDG Index score of 80% or higher. However, the remaining SDGs exhibit greater variability in SDG Index scores, indicating that achieving the development goals by 2030 poses a significant challenge for most countries in our sample. For instance, SDGs 5, 7, and 9 through 11 demonstrate the greatest variance in the average SDG Index score. However, it is important to note that interpretations of the SDG Index score must be approached with caution as they have been subject to criticism (Diaz-Sarachaga et al., 2018; Kroll et al., 2019).

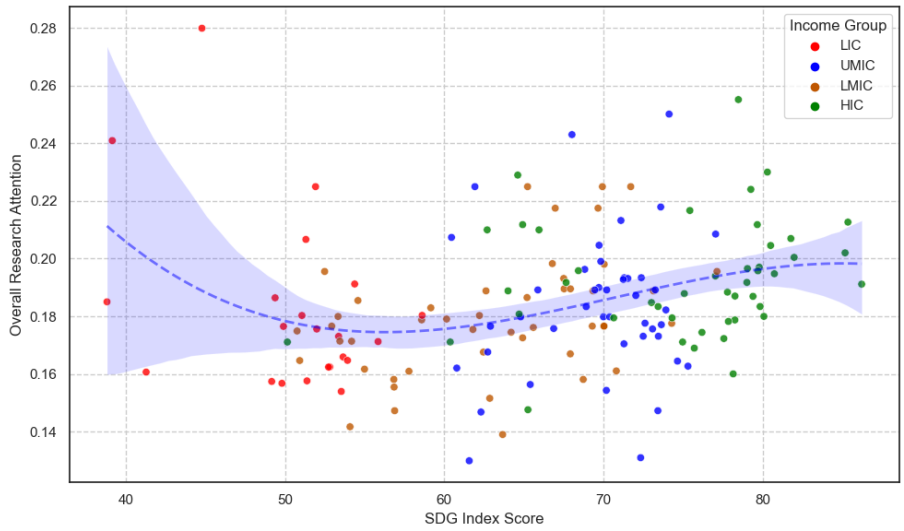


Figure 2.6: Scatter plot illustrating the relationship between mean RAI and SDG Index score.

Countries are categorized based on their income status into four classes: Low-Income Countries (LIC), Lower-Middle-Income Countries (LMIC), Upper-Middle-Income Countries (UMIC), and High-Income Countries (HIC). The blue dashed line represents the optimal order of the fitted regression polynomial using a leave-one-out cross-validation (order 3), while the light shaded area represents the 95% confidence band.

To investigate the degree of attention or focus that the issues related to each SDG receive in a country, we analyze the association between the RAI and their respective average SDG Index scores. Figure 2.6 illustrates a scatterplot of the overall research attention, categorizing countries based on their income status into four classes: Low-Income Countries (LIC), Lower-Middle-Income Countries (LMIC), Upper-Middle-Income Countries (UMIC), and High-Income Countries (HIC). The scatterplot is plotted against the SDG Index Score, which is determined as the average of the official SDG Index scores over the period from 2015 to 2022. Moreover, using the averages between 2015 to 2022 has the advantage that it smoothes out possible year-to-year fluctuations in both the RAI and the SDG Index scores. This mitigates the impact of potential data irregularities or anomalies in specific years. As a result, we can provide a more robust and stable estimate of the relationship between research attention and SDG progression, rather than focusing on a specific point in time.

By examining the scatter plot, we aim to identify any patterns or trends in the amount of attention, in terms of research output, that each SDG is receiving and the SDG achievement in a particular country. This analysis can help shed light on whether more attention is associated with higher SDG Index scores or vice versa. Additionally, it can also reveal non-linear relationships or varying degrees of correlation between research output and different SDGs Index scores, which can be valuable information for guiding future research and policy-making efforts in the field of sustainable development.

To conduct this analysis, we determine the optimal polynomial order for regression using

leave-one-out cross-validation. This method helps us identifying the best polynomial order for each pair of SDG research attention and average SDG Index score, ensuring the best fit for the data while minimizing overfitting. In the case of Figure 2.6, the blue dashed regression line has been fitted with a polynomial of third order, while the light shaded area represents the 95% confidence band.

By integrating these approaches, we can effectively analyze the relationship between research attention and SDG achievement, offering insights that can guide both research and policy directions in sustainable development. The results become particularly intriguing when considering the attention related to SDGs, their respective average scores, and the income group of countries. At an aggregated level, we observe a clear indication of a non-linear relationship, which is generally upward sloping. Research attention appears to exhibit a downward slope for LICs, while it demonstrates a slightly upward trend for LMICs, that increases for UMICs and HICs.

This observation underscores the necessity for further investigation into the factors contributing to these varying trends across income groups and their influence on progress towards achieving the SDGs. To obtain a more in-depth understanding, we will examine the disaggregated level, focusing on individual SDGs, from SDG 1 to SDG 17, as depicted in Figure 2.7. The optimal order of the fitted regression polynomial is observed to fluctuate across the SDGs. While a linear relationship (order 1) is evident for SDGs 1, 2, 4, 6, 7, 9, 14, and 17, multiple instances of non-linear relationships are also present. SDGs 5, 8, 11, 13, 15, and 16 appear to follow a quadratic relationship, whereas SDGs 3, 10, and 12 are described by a higher-order polynomial.

To examine the relationship between the RAI and the SDGs in greater detail, we employed Kendall's rank correlation coefficient as our analysis method, as illustrated in Table 2.2. This table showcases the Kendall rank correlation coefficients between research attention and the official average SDG Index score per country (United Nations, 2021). The table is organized into five columns, representing the overall correlation and correlations across four distinct income levels: LIC, LMIC, UMIC, and HIC. We chose Kendall's Tau over Pearson's correlation coefficient, as the latter presumes a normal distribution of the variables and a linear relationship—assumptions that do not align with our data. Moreover, we favored Kendall's Tau over Spearman's correlation coefficient, as Kendall's Tau is generally seen as more robust, especially when handling smaller sample sizes or datasets with a significant number of tied ranks.

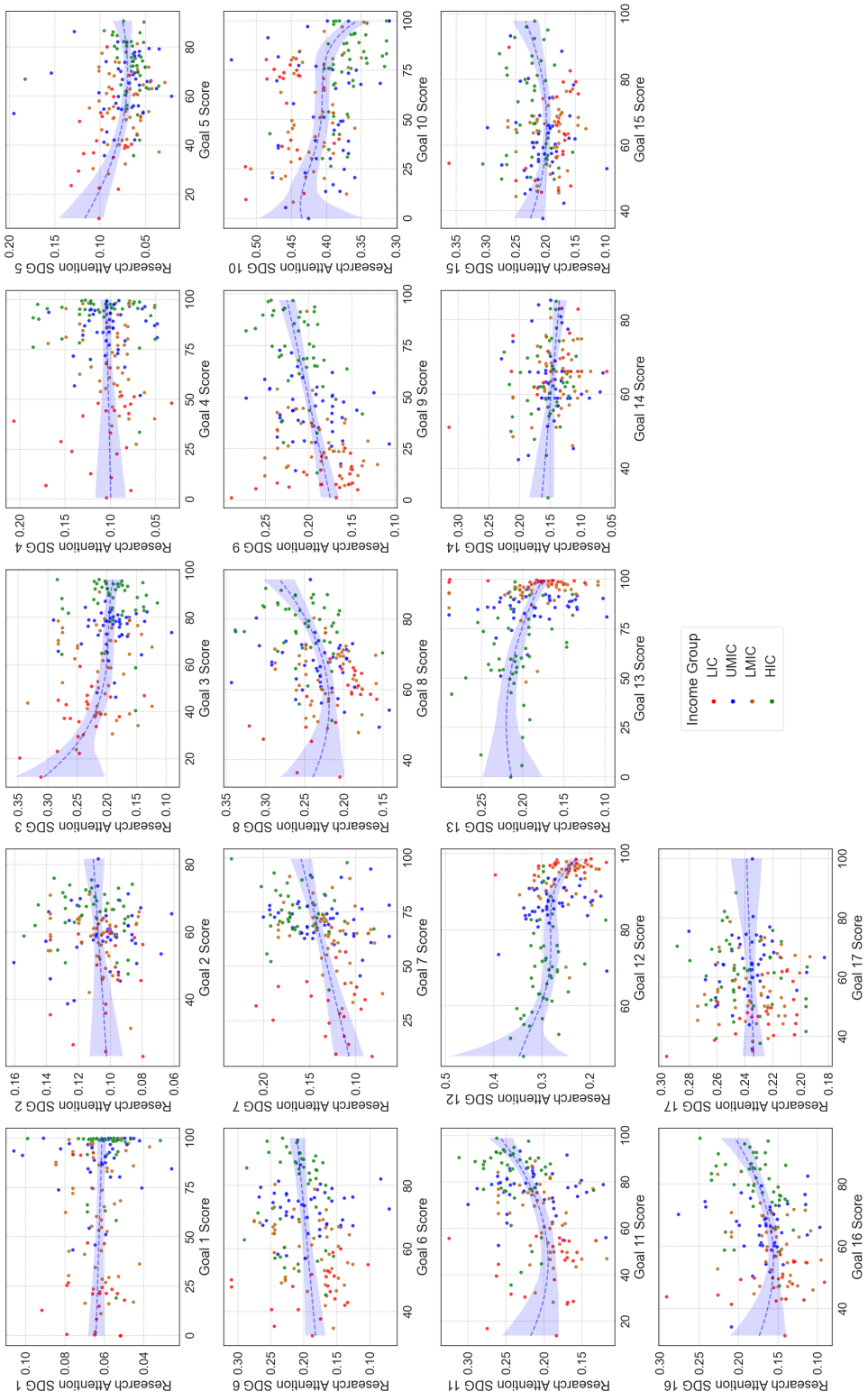


Figure 2.7: Scatter plots illustrating the relationship between research attention and individual SDG Index scores.

Countries are categorized based on their income status into four classes: Low-Income Countries (LIC), Lower-Middle-Income Countries (LMIC), Upper-Middle-Income Countries (UMIC), and High-Income Countries (HIC). The blue dashed line represents the optimal order of the fitted regression polynomial using a leave-one-out cross-validation, while the light shaded area represents the 95% confidence band.

Table 2.2 shows varying degrees of correlation between RAI and the different SDGs overall and across different income classifications. The overall correlation between research attention and the SDG Index Score is positive and statistically significant, indicating that, in general, higher research attention is associated with a higher SDG Index score, and vice versa. A fact, that was also visible as a general indication on Figure 2.6. However, when looking at the correlations for each individual SDG, we observe mixed results. On the overall level, we find a significant positive correlation for SDGs 2, 6, 7, 8, 9, 11 and 16, which is in line with the overall trend. However, we find a significant negative correlation for SDGs 1, 3, 5, 10, 12, 13, 14. For SDGs 4, 15, and 17, the RAI does not exhibit a significant correlation with the SDG Index scores.

Table 2.2: Kendall rank correlation between RAI and SDG Index score per country

Research Attention vs SDG Index score	**Overall**	**LIC**	**LMIC**	**UMIC**	**HIC**
SDG Index Score	0.23^{***}	-0.11	0.28^{***}	0.03	0.22^{**}
SDG 1	-0.13^{**}	0.13	0.06	-0.19^{*}	0.05
SDG 2	0.09^{*}	0.02	0.11	-0.06	0.11
SDG 3	-0.19^{***}	-0.34^{**}	-0.05	-0.02	0.1
SDG 4	0.05	-0.24	0.04	-0.05	-0.06
SDG 5	-0.14^{***}	-0.35^{**}	-0.05	-0.15	0.25^{**}
SDG 6	0.13^{**}	-0.23	0.21^{**}	0.03	0.14
SDG 7	0.25^{***}	0.08	0.29^{***}	0.12	0.12
SDG 8	0.23^{***}	-0.26	0.04	0.22^{**}	0.19^{*}
SDG 9	0.29^{***}	-0.26^{*}	0.23^{**}	0.18^{*}	0.32^{***}
SDG 10	-0.25^{***}	0.11	-0.26^{***}	-0.06	-0.32^{***}
SDG 11	0.28^{***}	-0.05	0.23^{**}	0.09	0.22^{**}
SDG 12	-0.35^{***}	-0.17	-0.27^{***}	-0.22^{**}	-0.24^{**}
SDG 13	-0.27^{***}	0.13	-0.3^{***}	-0.16	-0.2^{*}
SDG 14	-0.08	-0.2	-0.06	0.05	-0.12
SDG 15	0.01	-0.28^{*}	-0.11	-0.09	0.12
SDG 16	0.24^{***}	-0.12	0.23^{**}	-0.02	0.27^{***}
SDG 17	0.04	-0.23	-0.0	-0.12	0.1

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

The overall correlation between RAI and SDG Index Score is significantly positive (0.23**). However, for SDG 1 - *no poverty*, we observe a contrary effect (-0.13**). This could imply that even in areas where research attention is high, poverty might remain a significant issue. On the other hand, it could mean that the awareness for the problems is high and more research is being conducted about these countries. This could stem from the specific focus of the research being conducted, or perhaps the benefits of heightened research attention

have not yet sufficiently mitigated poverty levels. Furthermore, in order to alleviate the dire situations caused by high poverty levels, it is imperative to dedicate more research attention to these countries. This focus on research could help improve the conditions and ultimately reveal a negative correlation. Additionally, the negative correlation with SDG 3 - *good health and well-being* (-0.19***), indicates that an increase in research attention could be associated with factors such as elevated stress levels, inadequate work-life balance, or other health-related concerns.

As suggested by Figure 2.2, SDGs 7 - *affordable and clean energy*, and SDG 13 - *climate action*, frequently co-occur in academic literature, underlining their importance for an effective transition towards achieving the 1.5-degree goal. However, intriguingly, we observe contrasting correlations for these two SDGs. SDG 7 exhibits a strong positive correlation (0.25***) between research attention and SDG achievement, suggesting that increased research focus may directly contribute to advancements in clean energy technology or strategies, thereby driving progress in this SDG. The progression of clean energy relies heavily on strategic investments in this sector, fostering increased attention in international discourse and consequently attracting greater research focus. While it is evident that research activities can stimulate advancements in SDG 7, it is equally plausible that progress made in SDG 7 can attract further research attention, demonstrating the bidirectional and dynamic nature of this relationship. In contrast, SDG 13 displays a strong negative correlation with research attention (-0.27***). However, it is worth noting that wealthier countries often score lower overall on SDG 13, despite possessing the resources and capabilities necessary to combat the climate crisis. This highlights the critical need to understand the relationships between income levels, research attention, and progress towards not just climate action but all other SDGs as well.

When examining the disaggregated effects, we can observe some interesting patterns. Particularly, in the LIC group, we generally observe either negative correlations or statistically insignificant correlations between research attention and SDG Index scores. For instance, the relationship between research attention and SDG 3 - *good health and well-being* and SDG 5 - *good health and well-being* is notably negative (-0.34** and -0.35**, respectively). These correlations suggest that in these countries, an increase in research attention does not necessarily lead to improvements in health or gender equality and could potentially even widen existing disparities. On the other hand, increasing disparities in health or gender equality could also lead to heightened research attention, which could be an attempt to understand the root cause of these disparities.

For LMICs, the correlations vary considerably across SDGs. Notably, there are strong positive correlations for SDG 7 - *affordable and clean energy* (0.29***) and SDG 11 - *sustainable cities and communities* (0.23**), suggesting that research attention in these countries might be particularly beneficial for these areas. Additionally, advancements in these areas, as indicated by the SDG Index Scores, could elevate their prominence and perceived significance within the research community. This heightened awareness could, in turn, stimulate increased research attention towards these SDGs. Conversely, significant

negative correlations exist with SDG 10 - *reduced inequalities* (-0,26***) and SDG 13 - *climate action* (-0,30***). This implies that growing inequalities and the escalating urgency of climate change might be attracting more research focus, as scholars strive to understand and address these pressing issues.

For UMICs, the results are generally weaker, with fewer statistically significant findings. However, SDG 8 - *recent work and economic growth* and SDG 9 - *industry, innovation, and infrastructure* (0,22** and 0.18*, respectively) show a positive correlation, suggesting that increased research attention might have a positive impact on economic growth and innovation in these countries, while economic growth could also provide an environment that provides more funding opportunities.

Finally, HICs predominantly exhibit positive correlations between research attention and SDGs, with the strongest being SDG 9 - *industry, innovation, and infrastructure* (0,32***). However, SDG 10 - *reduced inequalities* and SDG 12 - *responsible consumption and production* (-0.32*** and -0.24**, respectively) have significant negative correlations, suggesting that higher research attention might coincide with increased inequalities and less responsible consumption and production patterns.

Interestingly, we observe a trend towards a non-linear relationship between research attention and SDG Index scores, particularly for SDGs 5 and 9, and to some extent for SDG 8. For SDG 5 - *gender equality*, we find a negative and statistically significant correlation for LICs at -0.35**. This might imply that gender inequality issues could be attracting increased research attention as scholars seek to understand and propose solutions for these pressing issues. However, heightened research attention does not necessarily lead to improved gender equality. One explanation could be entrenched gender disparities or restricted access to education and opportunities for women. Conversely, the correlation for HICs is positive and significant (0.25***), which implies that increased research attention is associated with better gender equality outcomes. The disparity between income groups may stem from variations in socio-economic and cultural contexts, which could influence the impact of research attention on gender equality.

For SDG 9 - *industry, innovation, and infrastructure*, we observe a reversed pattern. While the overall correlation between research attention and SDG 9 is positive and significant (0.29***), it becomes negative and significant for LICs (-0.26*). This suggests that in low-income countries, increased research attention may not directly translate into advancements in industry, innovation, or infrastructure. This could be attributed to resource constraints, underinvestment in these sectors, or hurdles in implementing research outcomes. Moreover, the association between increased research attention and a lower score appears to be more pronounced in these particular countries. However, in LMICs, UMICs, and HICs, the correlations are positive and significant (0.23**, 0.18*, and 0.32***, respectively). This implies that in these countries, heightened research attention tends to be associated with improved performance in achieving SDG 9. Particularly in HICs, it seems reasonable to assume that they have the necessary resources and infrastructure to leverage research findings effectively, thereby driving advancements in industry, innovation, and infrastructure and

their successes become subjects of interest for researchers aiming to understand and replicate these advancements.

Lastly, we note a shift in sign for SDG 8 - *decent work and economic growth*, moving from a negative correlation for LICs (-0.26), albeit non-significant, to a significant positive correlation for UMICs and HICs (0.18* and 0.32***, respectively). This negative correlation in LICs could indicate that heightened research attention does not necessarily result in job creation or economic growth. This could be due to a misalignment between the focus of research and the needs of the local economy, or difficulties in implementing research findings. Conversely, for UMICs and HICs, the positive correlation might suggest that increased research attention tends to foster innovation, create jobs, and stimulate economic growth. These countries, with their more developed economies, robust institutions, and superior resources and infrastructure, are often better positioned to actualize the benefits of research.

In summary, our findings indicate substantial variation in research attention across different income groups and SDGs. This variation could reflect differing capacities to utilize research findings, differing research priorities, and unequal distribution of research resources. It could also reflect broader socio-economic trends and disparities. For instance, wealthier countries with more resources might be better equipped to translate research attention into progress towards SDGs. Their success in these areas might, in turn, draw further research attention. In contrast, in lower-income countries, other factors such as poverty, lack of infrastructure, and political instability might hinder the positive impacts of research. Paradoxically, these challenges could also attract increased research attention, as there is an urgent need to address these pressing issues.

2.7 Discussion

Overall, two contrasting theories could potentially explain the relationship between research attention towards SDGs and the actual SDG Index scores. The first theory suggests an negative relationship, suggesting that SDGs with lower scores receive more attention. The second theory suggests a positive correlation, implying that countries with higher SDG Index scores may draw more research and knowledge sharing attention. We observe evidence supporting both theories in our data, depending on the specific SDG. In the following, we delve into broader explanations for our results, building on the more detailed discussions in section '(Non)-Linear Relationship between Research Attention and SDGs'.

Some explanations for the first theory can be derived from theories about media coverage and human reactions to it. Soroka et al. (2019) suggest that humans tend to react more strongly to negative news coverage, a phenomenon labeled as 'negative biases' (Rozin & Royzman, 2001). Similarly, low SDG Index scores might receive more research coverage, while those with high SDG Index scores may not be as prominently featured, resulting in a negative relationship between research attention towards SDGs and the SDG Index scores.

Furthermore, general psychological concepts suggest that people tend to be more sensitive to negative information (Baumeister et al., 2001). Thus, low-scoring SDGs might trigger

stronger emotional responses and consequently attract more attention. Researchers might be more moved by lower SDG Index scores, thereby concentrating their research efforts on countries and SDGs with such scores.

Countries with lower SDG Index scores may indeed attract more funding and resources due to the perceived potential for improvement. This increase in financial support can lead to more attention being directed towards these countries and the specific challenges indicated by their lower scores. Recognition of the need for improvement can stimulate interest and investment from a variety of stakeholders, including governments, development agencies, NGOs, and philanthropies. As a result, research, initiatives, and collaborations focused on addressing the challenges and promoting sustainable development in these countries can receive heightened attention and support (OECD, 2002).

While we do not find such a negative relationship for the overall correlation between SDG Index Score and research attention, the arguments made above are plausible for some SDGs. In our analysis we find support for those theories with negative correlations between the SDG Index Scores and research attention for SDGs 1, 3, 5, 10, 12, 13, and 14. Thus, for these SDGs, countries with lower scores might attract more funding and attention, or the psychological explanation or any of the reasons stated above might hold true.

On the other hand, there is an alternative scenario suggesting a positive correlation between research attention and the SDG Index score. This theory implies that countries with higher SDG Index scores may receive greater attention, particularly in terms of research and knowledge sharing. Several explanations support this viewpoint:

Firstly, countries with higher SDG Index scores are generally more developed and possess greater resources, including technological, financial, and human capital (United Nations IATF, 2022). As a result, these countries are better equipped to invest in sustainable development, dedicating significant resources to research and development. Their commitment to sustainable development and their ability to collaborate with international partners contribute to an increased focus on the topic (King, 2004).

By leveraging their resources and engaging in research and development activities, countries with higher SDG Index scores demonstrate their commitment to advancing SDGs. This commitment and investment draw attention from researchers, policymakers, and practitioners who seek to understand and learn from their successes. As a result, these countries become focal points for research and knowledge sharing, providing valuable insights and best practices that can be replicated in other contexts (OECD, 2021).

Secondly, countries with higher SDG Index scores serve as compelling examples of success in implementing sustainable development policies and practices. Researchers, policymakers, and practitioners are often drawn to studying these countries to gain insights into the factors that contributed to their achievements, with the aim of replicating their success in other contexts (J. Sachs et al., 2022; J. Sachs, Kroll, et al., 2021; J. Sachs, Schmidt-Traub, et al., 2021; J. Sachs et al., 2019).

Another driving factor for the enhanced attention towards high-scoring countries is the recognition and support they receive from the international community, including development

agencies, non-governmental organizations, and academic institutions. These countries often serve as role models in various domains, such as digital transformation, thereby setting benchmarks for other nations. This recognition and admiration lead to enhanced research collaborations, increased funding opportunities, and greater knowledge sharing, all focused on promoting sustainable development practices (OECD, 2021).

Furthermore, countries with higher SDG Index scores often wield more influence in steering global sustainable development agendas and priorities. Their successful experiences and perspectives command greater attention in international forums and policy discussions, as they offer valuable insights and guidance for other nations (OECD, 2021).

Our results indicate a positive overall correlation between research attention and the SDG Index scores, which lends more weight to the arguments posited above. At the individual SDG level, a positive correlation can be found for SDGs 2, 6, 7, 8, 9, 11 and 16 as well. Consequently, for these SDGs, reasons stated above, such as being a role model as a high-scoring country or attracting more attention in international forums or policy discussions, might hold true.

Initially, we assumed that the reasons for a negative relationship would be more credible. However, our results reveal evidence supporting both sides of the debate, preventing us from unequivocally favoring one over the other.

2.8 Conclusion

In conclusion, this study has provided a new avenue for the exploration of the global research landscape surrounding SDGs. In addressing our first research question, we focused on the effective utilization of advanced NLP techniques, particularly pre-trained language models and zero-shot text classification, for analyzing large-scale scientific textual data related to the SDGs. We demonstrated that these techniques allow for a nuanced and efficient analysis, uncovering insights that are often hidden in traditional analysis methods. This approach has enabled us to better understand the complex dynamics within SDG-related research, providing a valuable tool for future studies in this field.

In response to our second research question regarding the novel insights these advanced NLP techniques could reveal, our study identifies a significant gap in the current literature: the underutilization of NLP methods in large-scale analysis of SDG-related scientific discourse. Despite a growing body of research using NLP, there is a need for more advanced textual analysis that is capable of capturing the SDG discourse within the scientific community. Using the power of zero-shot text classification as a modern NLP tool and the introduction of the RAI, we have provided a fresh perspective on the dynamics of SDG-focused scientific discourse, its relationship with SDG achievement and the correlation across different income levels in countries. This offers valuable insights for researchers, policymakers, and stakeholders involved in sustainable development efforts.

In relation to the third research question, our use of the RAI and NLP analysis aims to uncover the connections between SDG index scores and research attention for each specific

SDG. This valuable information can be utilized by policymakers to prioritize their attention towards particular SDGs and adjust their strategies accordingly. By identifying interesting relationships within scientific literature, policymakers can effectively concentrate their efforts on relevant studies, especially when it comes to unexpected and non-linear relationships like those observed for SDGs 5, 8, and 9. This highlights the potential of scientific literature to offer insights that may not be readily apparent, ultimately helping in improving policy implications. In summary, our study's primary contribution resides in equipping policymakers with knowledge and presenting them with a tool to efficiently navigate the vast amounts of scientific literature.

Nevertheless, it is crucial to acknowledge the limitations of our analysis and to use the RAI as a tool for relative comparison rather than an absolute measure of research attention. Future research should concentrate on addressing the limitations identified in this study. One such limitation is the complexity in comparing the composite measurement of the SDG Index scores and the RAI. Our analysis offers a composite measure of research attention spanning the publication years of 2015 to 2022, paralleling the UNs' SDG Index scores, which are computed by combining indices from the same time period. Consequently, our comparison reflects an aggregation of data from several years, as opposed to a time-series analysis. Although this method provides a general trend, a more granular, year-by-year comparison could yield a more in-depth analysis and provide a better understanding of the dynamic involved. At present, we have been able to identify a correlation between research attention and SDG achievement. However, the directionality of this relationship remains unclear. Therefore, a more detailed analysis could shed light on this complex interaction and enhance our understanding of its intricate mechanisms. However, obtaining the required data is challenging due to the complex structure of the SDG indices and the current non-stationary behavior of much of the data makes this level of analysis unattainable at present.

In addition, our data source was largely based on the abstracts of metadata provided by the Web of Science. While Web of Science is a comprehensive and well-regarded database for scientific research, it does not capture the entire global research output related to SDGs. There are potential other repositories such as ArXiv, various policy briefs, pre-print articles, and other machine-readable text formats could significantly contribute to the SDG research landscape. The used method of zero-shot text classification is robust and can be applied to any machine-readable text format. However, our analysis did not take these resources into account. Thus, future research could build on our approach, extending it to broader range of sources. This could lead to the creation of a more generalized and comprehensive version of the RAI, contributing a more holistic perspective on the global research attention devoted to the SDGs.

Moreover, while we opted for UN Indicator scores to bring transparency and simplicity to our analysis, alternative measures could also provide insightful and relevant perspectives on SDG achievement. Future research could consider incorporating other indicators, which provide a different perspective on SDG achievement.

Lastly, our analysis did not account for several potential confounding factors that could

influence the relationship between research attention and SDG performance. Factors such as government efficiency, research funding, and socio-economic context may all significantly impact this relationship. Thus, future research should seek to better understand the dynamics between research attention and SDG achievement by investigating the underlying mechanisms and factors driving the observed variations in correlations across different income groups. By deepening our understanding of these complex dynamics, we can more effectively develop and implement policies that facilitate progress towards achieving the SDGs across all income levels. We hope that our findings will stimulate a wider discourse on the correlation between research attention and SDG performance, as well as the factors driving this relationship.

A Appendix to Chapter 2

From PDFs to raw text data

In the world of scientific research, most articles are only available in the PDF format, and extracting text from these files is a crucial first step in any downstream NLP task. However, unlike common formats like txt or csv, accessing information from PDF files is not as straightforward. Many popular Python modules, such as 'PDFminer,' are not layout sensitive and can result in subpar text extraction quality, missing important information from the layout of PDF files. Our experience has shown that applying such non-layout sensitive models to our scientific PDFs can result in entire sentences being parsed incorrectly or being completely missed. To address this issue, several studies have attempted to improve text and layout extraction methods (e.g., Xu et al. (2020)). Utilizing layout information can help distinguish between actual text information in an article and text extracted from tables or figures, which we do not consider relevant for our analysis (Shen et al., 2021).

To extract text and layout information from scientific articles, we opted for a VILA-based model, specifically tailored to scientific texts, as proposed by Shen et al. (2021). This model fine-tunes a pre-trained language model, such as BERT, and does not require the costly pre-training process. The VILA model is based on the premise that scientific text can be grouped into blocks or lines, which can be identified using layout detection models or rule-based PDF parsing (Shen et al., 2021).

The authors Shen et al. (2021) present two different methods for incorporating the group structure, I-VILA (Injecting Visual Layout Indicators) and H-VILA (Visual Layout-guided Hierarchical Model). However, for our text and layout extraction, we opted for the H-VILA block trained on the grotoap2 training set using the layoutLM model (Xu et al., 2020). We chose this approach based on personal assessment of randomly selected articles, which showed slightly better results compared to other combinations.

The model provides a token classification alongside the parsed text, with classifications such as abstract, author, dates, body content, figure, keywords, title, and more. However, we only utilize the body content text blocks for our analysis. While parsing the PDF text yielded satisfactory results, manual cleaning was necessary to improve sentence semantics and text cleanliness for our topic model. Table 2.3 outlines the cleaning steps we performed.

Table 2.3: Cleaning steps of parsed raw texts.

Problem	**Fix**
Extra whitespaces	Replace extra whitespaces with a single whitespace.
Words separated by hyphen	Remove hyphen.
Words separated by whitespace	Remove whitespace.
Model parsed some ff, fi, and if characters as one special character	Replace special double characters by normal characters.
Whitespaces between word in sentence and punctuation	Remove whitespace.
URLs in text which do not have any semantic meaning	Remove all URLs from text.
Parsed some sentences character by character with whitespace between them, i.e., S E N T E N C E instead of Sentence. Problem mostly occurred with figure subtexts	As most figure subtexts do not have important semantic meaning, we removed such single characters.
Very short sentences or single tokens	Filter out sentences with less than 6 words or 50 characters.
Coding errors	Filters out sentences with a high percentage of unrecognized words using a spell checker.

Table 2.4: Counted SDGs for abstracts of countries with at least five research articles

SDG	Short Description	Total Count	Number of Articles	Average Count per Article
1	No Poverty	127	94	1.35
2	Zero Hunger	155	126	1.23
3	Good health and well-being	268	187	1.43
4	Quality education	138	98	1.41
5	Gender equality	81	70	1.16
6	Clean water and sanitation	231	145	1.59
7	Affordable and clean energy	155	118	1.31
8	Decent work and economic growth	93	78	1.19
9	Industry, Innovation and infrastructure	61	50	1.22
10	Reduced inequalities	69	55	1.25
11	Sustainable cities and communities	178	119	1.50
12	Responsible consumption and production	91	82	1.11
13	Climate action	142	117	1.21
14	Life below water	79	44	1.80
15	Life on land	115	81	1.42
16	Peace, Justice and strong institutions	52	33	1.58
17	Partnership for the goals	63	51	1.24
	SDG / SDGs	13,574	5,646	2.58

This table presents a comprehensive overview of the total number of SDGs counted, the number of abstracts in which each specific SDG appears at least once, and the average number of mentions per abstract.

Table 2.5: Counts by country code.

ISO2	Count	ISO2	Count	ISO2	Count	ISO2	Count	ISO2	Count
CN	825	IN	623	US	309	ZA	265	BR	238
NG	235	GH	203	BD	199	ES	175	ID	169
ET	166	KE	161	GB	160	PK	157	AU	145
JP	122	MY	113	DE	111	IT	103	TZ	103
IR	100	NP	99	RU	94	MX	91	UG	90
CO	86	CA	86	VN	85	UA	81	TH	78
MW	77	PL	71	SE	62	KR	59	RW	59
PH	59	TR	59	TW	58	MZ	56	EG	53
ZW	53	ZM	48	PT	47	PE	47	FR	46
FI	44	CL	44	CM	42	NL	42	SA	40
LK	40	MM	39	KH	39	IE	39	RO	37
NZ	37	EC	36	CD	35	CH	35	AE	34
AF	32	MG	31	GR	31	DK	31	SN	31
NO	30	AT	30	JO	28	SL	28	KZ	26
AR	26	LT	25	CZ	24	HU	23	NA	22
BF	22	SG	22	MN	22	EE	21	HK	20
LA	20	NE	19	MA	19	ML	18	QA	18
IS	18	RS	17	SI	17	LV	17	SK	17
BO	16	DZ	16	FJ	15	GT	15	AO	14
CR	14	SD	14	LB	14	CG	13	BG	13
BT	12	BE	12	BJ	12	BH	12	LR	12
HT	12	HR	11	PS	11	HN	11	KW	11
PG	11	TD	11	UY	11	CF	11	OM	10
CU	10	YE	10	BW	10	GN	9	TN	9
VU	9	JM	9	KG	9	TL	9	PY	8
IL	8	LS	8	LU	8	MU	8	SS	8
CY	7	SO	7	GM	7	LY	6	MO	6
TG	6	TJ	6	NI	6	SZ	6	IQ	6
BI	6	SB	5	MV	5	CI	5	MT	5
AZ	5	DO	5	GE	5	BZ	5	AL	5
TT	5	GY	5	ME	5	AM	5	BY	5

This table shows the ISO2 country code and the corresponding number of research articles included for the Research Attention Index (RAI) in each country.

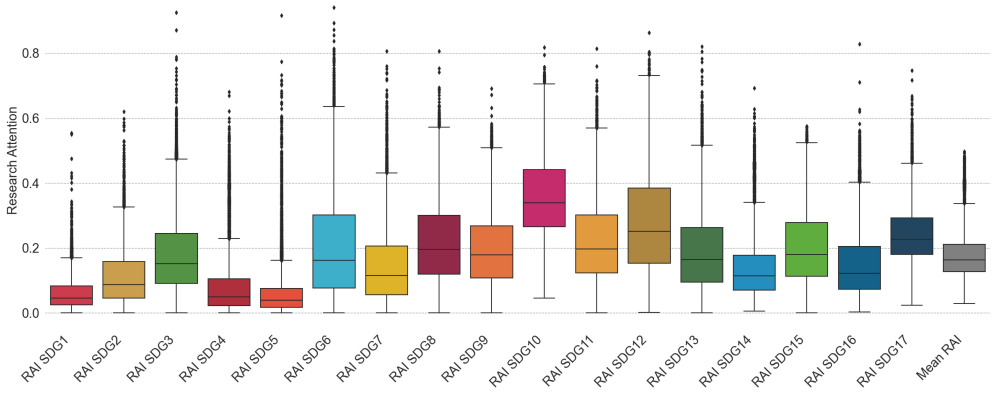


Figure 2.8: Boxplots representing Research Attention Index (RAI) based on the mean of all available abstracts.

The figure displays the results of the zero-shot classification at the sentence level for all 13,136 articles. Each sentence in the dataset has been assigned a probability between zero and one for each of the 17 SDGs.

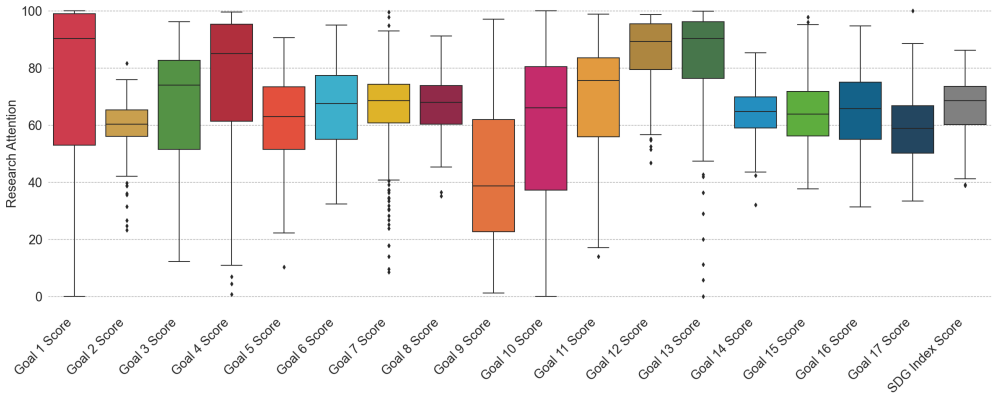


Figure 2.9: Boxplot of official SDG Index scores calculated as the mean from 2015 to 2022.

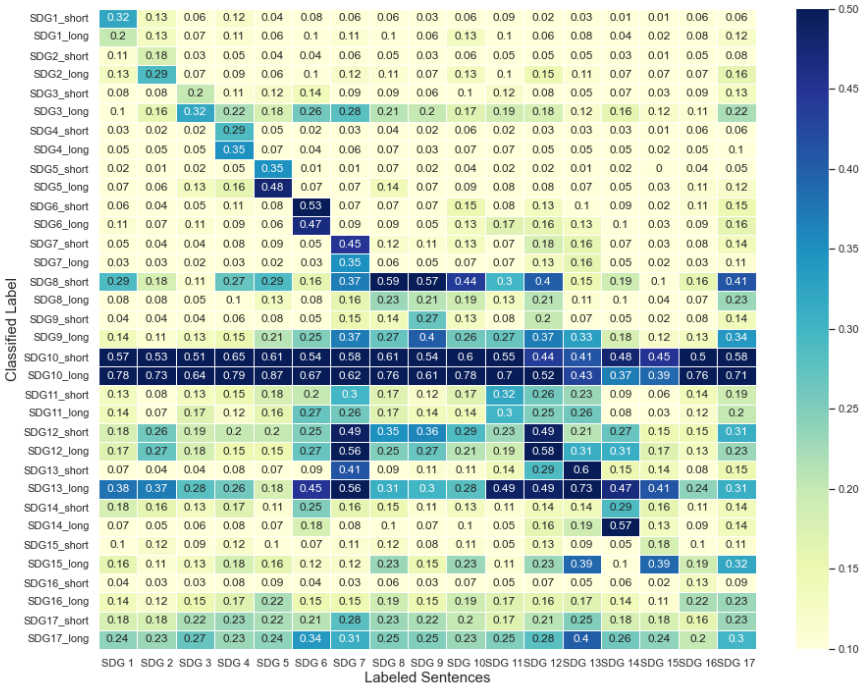


Figure 2.10: Model Validation with long and short labels.

Table 2.6: Explanation of labels (Part 1)

Classified Label	**Label Explanation**
SDG1_long	Eradicate extreme poverty
SDG1_short	Extreme poverty
SDG2_long	End hunger, achieve food security and improved nutrition and promote sustainable agriculture
SDG2_short	Eradicate extreme hunger
SDG3_long	Ensure healthy lives and promote well-being for all at all ages
SDG3_short	Good health and well-being
SDG4_long	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
SDG4_short	Quality education
SDG5_long	Promote gender equality and empower women
SDG5_short	Gender equality
SDG6_long	Ensure availability and sustainable management of water and sanitation for all
SDG6_short	Clean water and sanitation
SDG7_long	Ensure access to affordable, reliable, sustainable and modern energy for all
SDG7_short	Affordable and clean energy
SDG8_long	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
SDG8_short	Decent work and economic growth

Table 2.7: Explanation of labels (Part 2)

Classified Label	**Label Explanation**
SDG9_long	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
SDG9_short	Industry, innovation and infrastructure
SDG10_long	Reduce inequality within and among countries
SDG10_short	Reduced inequalities
SDG11_long	Make cities and human settlements inclusive, safe, resilient and sustainable
SDG11_short	Sustainable cities and communities
SDG12_long	Ensure sustainable consumption and production patterns
SDG12_short	Responsible consumption and production
SDG13_long	Take urgent action to combat climate change and its impacts
SDG13_short	Climate Action
SDG14_long	Conserve and sustainably use the oceans, seas and marine resources for sustainable development
SDG14_short	Life below Water
SDG15_long	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
SDG15_short	Life on Land
SDG16_long	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels
SDG16_short	Peace, justice and institutions
SDG17_long	Strengthen the means of implementation and revitalize the global partnership for sustainable development
SDG17_short	Partnerships for the goals

Chapter 3

Finding common development paths in voluntary national reviews reporting on sustainable development goals using aspect-based sentiment analysis

The following chapter is based on the paper:	
Title:	Finding common development paths in voluntary national reviews reporting on sustainable development goals using aspect-based sentiment analysis
Authors:	Elena Tönjes (contribution: 60%), Christoph Funk (contribution: 20%), Ramona Teuber (contribution: 10%), Lutz Breuer (contribution: 10%)
Status:	*Working Paper;* submitted to *Plos One*; 2nd revise and resubmit round

Finding common development paths in voluntary national reviews reporting on sustainable development goals using aspect-based sentiment analysis

Christoph Funk*,† Elena Tönjes‡ Ramona Teuber*,§ Lutz Breuer*,¶

Abstract

Voluntary National Reviews (VNRs) provide a platform for participating countries to share their experiences, failures, and successes in achieving the United Nations (UN) Sustainable Development Goals (SDGs). The objective of this study is to gain a deeper understanding of the narrative elements, particularly the sentiment, in VNRs in order to more effectively assess and support global SDG progress. A total of 232 VNRs from 166 countries are analyzed using Aspect-Based Sentiment Analysis (ABSA) to extract each country's sentiment toward the 17 SDGs. The sentiment scores are then compared to the corresponding official UN SDG scores, and countries are grouped by their sentiment toward all 17 SDGs to identify potential common development pathways. The analysis uncovers a notable positive correlation between the reported sentiment and official SDG scores for SDG 2 (zero hunger) and SDG 11 (sustainable cities and communities), and a negative correlation for SDG 5 (gender equality). Conversely, this relationship is not significant for the majority of SDGs, suggesting that VNR narratives may not directly reflect actual progress. A t-distributed stochastic neighbor embedding (t-SNE) approach indicates a consistent sentiment score among developed countries. In contrast, there are greater differences in reporting sentiment among Emerging Markets, Frontier Markets, and Least Developed Countries (LDCs), where there is greater dispersion (especially among LDCs) and sentiment in reporting on SDG progress that appears to have changed from one reporting year to another. These findings highlight the need to interpret VNRs in the context of each country's unique situation and challenges specific to each country.

* Center for International Development and Environmental Research (ZEU), Justus Liebig University, Giessen, Germany

† Corresponding author: Christoph.Funk@wi.jlug.de

‡ Department of Economics, Chair of Statistics and Econometrics, Justus Liebig University, Giessen, Germany

§ Institute for Agricultural Policy and Market Research, Justus Liebig University Giessen, Germany

¶ Institute for Landscape Ecology and Resources Management (ILR), Research Centre for Bio Systems, Land Use and Nutrition (iFZ), Justus Liebig University Giessen, Germany

3.1 Introduction

The United Nations' (UN) 2030 Agenda, which outlines 17 Sustainable Development Goals (SDGs) and 169 targets, provides a structured framework for its member states to collectively achieve a better future. Voluntary National Reviews (VNRs) serve as a unique platform, allowing member states to share experiences, both successes and setbacks, as they navigate their paths towards achieving the 17 SDGs (United Nations, 2016b). Consequently, VNRs provide textual insights into each country's progress toward the SDGs. However, despite the structured framework and shared experiences, there is a growing concern that many countries may not fulfill the targets set for 2030, especially in light of the challenges caused by the COVID-19 pandemic (Pradhan, 2023).

This raises the question: How can VNRs be systematically evaluated to understand the progress of each nation towards the SDGs without manually reviewing the extensive VNR texts? Addressing this question is imperative, given the increasing evidence of delays and the need for transformative changes to redirect the world towards sustainable development (J. D. Sachs et al., 2019; Murphy et al., 2023).

To address this gap, we employ Natural Language Processing (NLP) techniques to extract and analyze the textual data from VNRs. These methods are powerful for uncovering insights into each country's SDG performance, identifying patterns, and understanding the interactions between different goals. The SDGs can have limited efficacy due to selective implementation that often overlooks their complex interactions. Such interactions, depending on specific contexts and locations, can manifest both as synergies, where one SDG positively influences another, or as trade-offs, where progress in one area may come at the expense of another (Pradhan, 2023). This may extend beyond the interactions between SDGs to encompass those between countries. The existence of shared challenges or achievements may provide an opportunity for countries to collaborate and accelerate the achievement of SDGs. Consequently, the concept of synergies and trade-offs is not limited to the relationship between SDGs, but also extends to that between countries. In the event that countries encounter analogous challenges or attainments, they are said to share a common developmental paths. In this context, we proceed with our analysis, which involves examining potential synergies and trade-offs between countries, rather than between SDGs, when addressing them in our analysis. We aim to uncover trends in the way countries report on individual SDGs, shedding light on shared developmental paths on a global level.

While the existing literature on the topic of identifying common development paths is limited (Moinuddin, 2017; Sebestyén et al., 2020), a number of studies have already revealed the dynamics between the SDGs and identified trade-offs and synergies (Le Blanc, 2015; Hegre et al., 2020; Pradhan et al., 2017; Pham-Truffert et al., 2020; Nilsson et al., 2018; Xiao et al., 2023). Network analysis techniques were employed by Le Blanc (2015) to identify that some SDGs are interlinked with numerous other SDGs through multiple targets, whereas other SDGs exhibit a comparatively weaker connection to the wider system.

Hegre et al. (2020) employ a clustering algorithm, namely principal component analysis (PCA). This methodology enables the authors to identify trends, synergies, and trade-offs between SDGs. Their findings indicate the existence of synergies within and between SDGs, with the exception of SDG 10, in terms of both levels and temporal change. Pradhan et al. (2017) also examine the interrelationships between the SDGs. In contrast with the previously discussed methodologies, the present approach involves the application of correlation analysis between SDG scores for over 200 countries. In addition, Kroll et al. (2019), also rely primarily on the UN's official SDG metrics to identify trends and patterns. They analyze how trade-offs and synergies have evolved globally in the recent past. Furthermore, Pham-Truffert et al. (2020) seek to identify potential synergies and trade-offs between SDGs, employing a comprehensive literature review as the foundation for their analysis. Nilsson et al. (2018) also address SDG interactions and trade-offs, although they adopt a somewhat more theoretical approach. The authors develop a new conceptual framework for understanding the interactions of SDGs, specifically by using case studies from the energy, health, and ocean sectors. This framework allows for the understanding of how interactions might differ between different factors, such as the time horizon or resource endowments. The authors posit that their framework has the potential to enhance scientific research and policy-making decisions by establishing a SDG Interactions Knowledge Platform, which would facilitate the exchange of knowledge about SDG interactions. In contrast to the papers previously mentioned, which examine direct synergistic or trade-off relationships between the SDGs, Xiao et al. (2023) utilise methods such as a plus-minus decision-making trial and evaluation laboratory model in order to also examine the indirect effects of SDGs on each other. The results demonstrate that when indirect effects are considered, the synergy effect is predominant.

Research on SDGs encompasses not only the identification of synergies and trade-offs, but also the development of scores based on data that can be collected at the regional level. For example, D'Adamo et al. (2021) utilize SDG scores provided by the Italian Institute of Statistics (ISTAT). The authors examine Italy in particular and find that the northern regions outperform the southern regions with regard to achieving the SDGs. Furthermore, Anselmi et al. (2023) utilise numerical data in the form of the official SDG scores provided by the UN, with the intention of facilitating geographical comparisons. However, the comparison is conducted at the country level and is limited to European countries. The study also indicates that countries in the northern hemisphere, such as Sweden, Denmark, and the Netherlands, appear to excel in specific areas. Murphy et al. (2023) also focus on the European Union in particular, developing a composite index where European countries are measured against the worst and best performing countries in terms of achieving the SDGs. In addition, the majority of extant literature employs the official SDG scores without raising any concerns regarding their quality (Diaz-Sarachaga et al., 2018; Mugellini et al., 2021). As the quality of the scores is open to question, we also examine whether the UN-provided SDG scores reflect the countries' assessment on reaching the goals.

As the literature indicates, research on interactions and regional differences of SDGs is relatively extensive. However, there are clear gaps yet to be addressed. As the majority

of existing literature focuses on the analysis of official SDG scores, there is a deficiency in research focusing on the systematic analysis of textual data on SDGs at the country level. Similar to our study, Sebestyén et al. (2020) apply NLP techniques to VNRs. The authors primarily focus on network analyses, which is based on the co-occurrence of important keywords from the VNRs. Like this study, they attempt to cluster countries in order to gain insights into which countries may face similar challenges in achieving the SDGs. The authors employ a multiplex network analysis, where important keywords form central nodes in a network, and describe the significant fields of sustainable development. They then utilize the similarities between the networks to cluster countries. While numerical data can provide valuable insights, it is not always sufficient to fully capture the complete picture. Consequently, this may result in the overlooking of the nuances present in textual reports shared by the countries themselves. The purpose of our study is to address this research gap. A sentiment analysis of VNRs is being conducted with the objective of gaining a deeper understanding of how countries perceive their journey towards achieving the SDGs. By integrating the network analysis techniques of Sebestyén et al. (2020) with our sentiment-based methodology, we present a novel text-based sentiment score for each SDG, offering a more comprehensive perspective than studies that rely solely on numerical data.

While the SDGs are crucial for global sustainability, evaluating how nations perceive their progress towards the SDGs remains a challenge. To gain insight into how countries perceive each SDGs progress in their VNRs, we employ a sophisticated sentiment analysis method, namely Aspect-Based Sentiment Analysis (ABSA). This allows us to generate a text-based SDG-related sentiment score for each country. To the best of our knowledge, this study is the first to apply fine-grained sentiment analysis to VNRs. In addition, given the significant financial demands of achieving the SDGs, especially for developing countries, and the critical role of capital markets, we further categorize countries based on their market strength and economic stage. We adopted the MSCI Inc. market classification framework (MSCI Inc, 2021) to categorize countries into Developed Markets (DM), Emerging Markets (EM), or Frontier Markets (FM), with Standalone Markets being subsumed under Frontier Markets. Additionally, we utilized the UN's classification for Least Developed Countries (LDCs) (United Nations, 2019). This specific classification, while differing from traditional systems, was rooted in the updated M49 categorization (United Nations Statistics Division, 2006). All countries that do not fall into these classifications are referred to as 'Others'. Our categorization approach also highlights the growing role of capital markets in achieving the SDGs (United Nations Global Compact, 2019; The World Bank, 2020).

We aim to address the following research questions: 1. What countries are facing similar challenges and successes in their efforts to achieve the SDGs? 2. To what extent does the sentiment expressed in VNRs correlate with actual progress towards the SDGs? Specifically, does the sentiment align with or diverge from the countries' perspectives on their progress? The remainder of this paper is organized as follows. Section 3.2 provides a comprehensive description of the data used. Section 3.3 presents the methodology to extract the text data and the ABSA model. Section 3.4 outlines the model validation, our main findings,

common development paths and a comparison between sentiment score and actual country performance. A discussion of our results is provided in Section 3.5 followed by a conclusion in Section 3.6.

3.2 Data

The VNRs used in our study can be downloaded from the sustainable development platform (`https://sustainabledevelopment.un.org/vnrs/`) where 168 countries currently publish VNRs. We used all available reports between 2016 and 2021. The UN does not provide guidelines for the structure of VNRs; therefore, reports vary widely in length and layout (Sebestyén et al., 2020). The reports are published in PDF format only and they were extracted using the **VI**sual **LA**yout (VILA) layout parser described in Section 3.3.

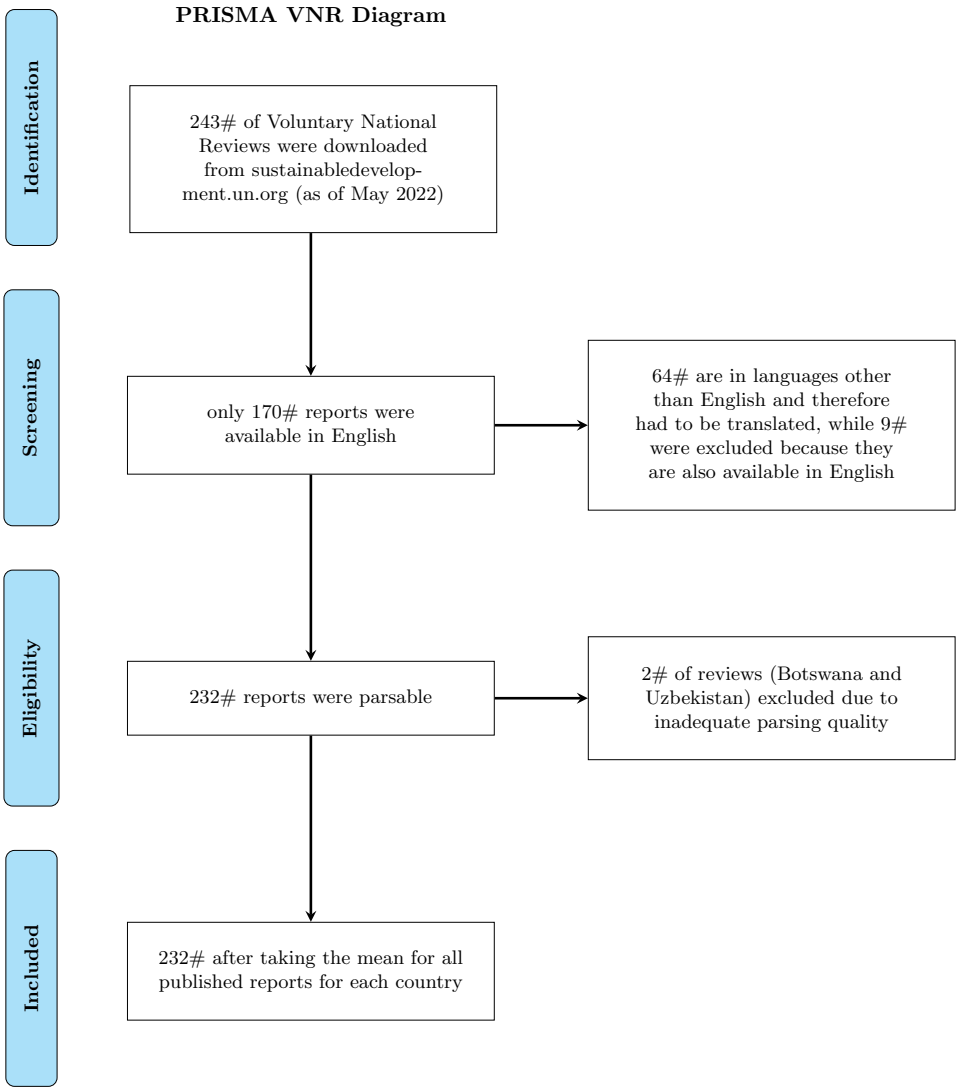


Figure 3.1: PRISMA Flow Diagram for Dataset Generation.
This diagram outlines the systematic process employed for the creation of our final dataset, depicting the stages of selection, screening, eligibility, and inclusion. It provides a transparent and methodical overview of data curation specific to our research needs.

Figure 3.1 provides an overview of the steps involved in creating our dataset. After screening the available VNRs, we downloaded 243 VNRs published between 2016 and 2021. Out of these, 73 reports were not available in English, including a report from Bolivia which, despite being listed as English on the UN's homepage, was written in Spanish. To integrate these into our dataset, we implemented the MarianMTModel and MarianTokenizer from the Helsinki-NLP group, which are part of the transformers library, to systematically translate sentences from French and Spanish into English. These non-English reports had to be translated for our analysis since our ABSA model operates exclusively on English-language text. Out of these 73 non-English reports, we did not use 9, because they were available in both English and another language (Spanish or French), and for consistency, we included only their English versions in our analysis. Of the remaining 170 PDFs that were parsed, reports from Botswana and Uzbekistan were excluded due to inadequate quality. In these cases, 'insufficient quality' refers to one report being a slideshow with minimal text, and the other being text in image form rather than in a machine-readable PDF format. Hence, we used 232 reports in our analysis. Table 3.1 provides an overview of the number of VNRs based on the MSCI Inc. market classification.

Table 3.1: Number of Voluntary National Reviews (VNRs) in our dataset based on country classification.

Country Classification	**Number of VNRs**
Developed Markets (DM)	29
Emerging Markets (EM)	40
Frontier-Markets (FM)	54
Least Developed Countries (LDC)	37
Others	72
Sum	232

An overview of all countries, their ISO3 code, length, and year of publication of our entire dataset can be found in Table 3.5. The length of VNRs varies widely, ranging from an average of 28 pages over the years in Switzerland and the Maldives to 430 pages in Indonesia. The maximum number of pages for a single report in our sample is from Indonesia with 786 pages. Although most countries (110) have published one report so far, 46 have already published two reports, and ten countries have published three reports in the last 6 years. For each year we have the following numbers of reports available: 2016: 20; 2017: 41; 2018: 46; 2019: 44; 2020: 42; and 2021: 39 VNRs. For our analysis, we took the mean per country over all its published reports. As a result, we have 166 observations (i.e. countries) available for our analysis.

3.3 Methodology

3.3.1 PDF Parsing

The VNRs are only available in PDF format, which requires text extraction prior to detailed analysis. Despite the popularity of Python modules like 'PDFminer' or 'Py2PDF', we found them inferior due to their inconsistent text quality extraction and omission of important layout details (Xu et al., 2020; Shen et al., 2021; Auzepy et al., 2023). In search of a more robust extraction tool, we employed a model originally intended for scientific articles, only to discover that it was also well-suited for VNRs. Developed by Shen et al. (2021), the model uses VILA groups, it offers two model variants: the Visual Layout-guided Hierarchical Model (H-VILA) and the Injecting Visual Layout Indicators (I-VILA), both rooted in pre-trained language models like BERT (Devlin et al., 2018). The VILA model operates on the premise that text can be systematically segmented into lines or blocks, facilitating easier extraction using an OCR-backed, rule-based PDF parser.

For our study, we chose the H-VILA model pre-trained on the grotoap2 dataset (Tkaczyk et al., 2012), integrated with the layout-aware BERT model (layoutLM (Xu et al., 2020)), a combination that demonstrated superior extraction capabilities. Among the 232 VNRs, 14 – presented in landscape mode – were split for effective extraction. The extraction yielded text classified by types: abstract, body content, and so on. Since our ABSA approach heavily relies on full sentences, we concentrated only on the main body content, deliberately excluding text from figures or tables. This method, though superior to its peers, was not foolproof and required additional text cleaning, elaborated in Table 3.5. Furthermore, we modified certain source codes in the VILA site packages. To ensure text parsing even in the event of decoding errors, we implemented error-ignoring functionalities within the decoding functions of 'PDFminer' and 'PDFplumber' (Auzepy et al., 2023).

3.3.2 Aspect-Based Sentiment Analysis

To discern sentiment from the extracted VNR texts, we employed ABSA as proposed by Smith et al. (2021). This methodology, augmented with insights from various studies (Zhang et al., 2018; Saeidi et al., 2016; Jo & Oh, 2011; Pontiki et al., 2015, 2016) and methodologies (Sun et al., 2019; Brulhardt, 2021), allowed us to pinpoint specific sentiments towards individual SDGs. Unlike vanilla sentiment analysis, ABSA is designed to capture a text's sentiment toward a specific entity, such as a company, an individual, or a location (B. Liu, 2020).

An ABSA example introduced by Li et al. (2019) is an end-to-end BERT layer model. Although powerful, this approach has its limitations; it relies on the explicit mention of the entity in the sentence. Given that VNRs often imply sentiments toward SDGs without naming them directly, we chose a sequence classification model (Sun et al., 2019; Brulhardt, 2021). Rooted in the 'SentiHood' dataset introduced by Saeidi et al. (2016), this model

emphasizes targeted aspect-based analysis. In our application, we focused on the target sentiment analysis (Jiang et al., 2011; Vo & Zhang, 2015), considering the SDG as the target and classifying the sentiment as positive, negative, or neutral. The code for the model we used and slightly modified can be found at `https://github.com/mwbrulhardt/yelp-absa`.

For our model's training, we utilized all UN progress reports from 2016 to 2021. We partitioned 80% of the data for training and 20% for testing. To mitigate overfitting, we allocated 10% of the training data for validation (United Nations Economic and Social Council, 2016, 2017, 2018, 2019, 2020, 2021). Our ABSA required each sentence to be tagged with the appropriate SDG (from 1 to 17) and a sentiment. Given the structured nature of UN progress reports, divided by SDG progress, we effectively had a pre-tagged dataset. Manual reviews ensured accurate sentiment and SDG tags. For example, one section in the 2019 report covers SDG 1 and its progress. We reviewed each sentence by hand and tagged most of the sentences in this section with the appropriate entity 'SDG1'. However, in a minority of cases, the sentences referred to a different goal and they were tagged accordingly. We (i.e. three people) then tagged the sentiment for each sentence and selected the majority sentiment for each of the resulting 2,079 sentences.

Addressing our dataset's imbalance — specifically in the sentiments for SDGs 1, 2, 4-7, 11, and 16 — was crucial. We tackled this by doubling the positive sentences, using the ContextualWordEmbsAug function from nlpaug and the 'distilbert-base-uncased' model (Ma, 2019). This augmentation process increased our training dataset from 2,079 to 2,265 sentences.

In terms of model selection, we faced a range of choices. We settled on BERT-base uncased for a few reasons. While more extensive models might offer minor improvements, they come with longer training times. On the other hand, choosing a much smaller model could compromise the results. Leveraging the capabilities of an NVIDIA RTX A5000 GPU, we found that the BERT-base uncased struck the right balance between performance and efficiency.

We initially trained our model using the unbalanced dataset to evaluate the effectiveness of our data augmentation strategy. This comparison between the up-sampled model and the original was instrumental in validating our approach. To optimize the model's performance, we conducted a systematic grid search on hyperparameters, focusing on 'learning rate', 'batch size', and 'epochs'. The results indicated that the up-sampled model, when paired with the optimal hyperparameters, yielded superior performance over the model trained on the unbalanced dataset. Specifically, using a learning rate of 3e-5, a batch size of 16, and 15 epochs, the up-sampled dataset achieved an MCC of approximately 0.74489, whereas the non-augmented set secured an MCC of 0.70998.

Figure 3.2 shows the confusion matrix for the model with up-sampling. In contrast, without up-sampling, we only obtained lower values (true positive: 0.570; true negative: 0.602; true neutral: 0.525). The model with up-sampling predicts the minority class of positive sentiment better than the model without up-sampling does. The 'none' label is included in the matrix due to the specific output format the model generates. The model's

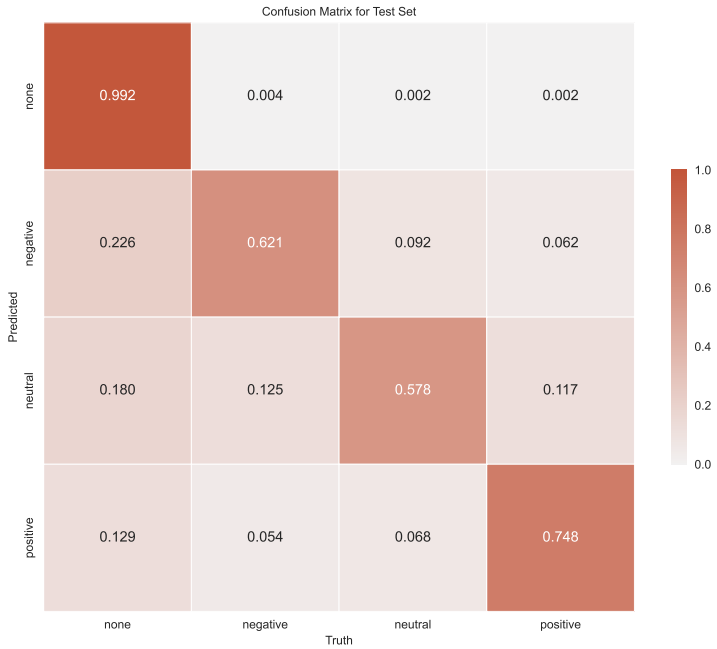


Figure 3.2: Confusion Matrix for the training results of the ABSA Model. This matrix contrasts the predicted sentiment labels (y-axis) against the true human-annotated labels (x-axis) across sentences. The sentiments are categorized into negative, positive, and neutral, with an additional 'none' category representing sentences where other sentiments not under discussion are present. A correct model prediction aligns a sentence with its respective SDG sentiment while categorizing all non-relevant SDGs as 'none'.

output for a sentence consists of one label for each SDG. The SDG, which the model predicts to be the entity discussed in the sentence, is assigned with the labels 'positive,' 'negative,' or 'neutral.' The other SDGs that the model does not predict to be discussed in the particular sentence are labeled 'none'. Therefore, each sentence receives one polarity label for the SDG spoken about and 16 'none' labels. The 'none' category has the highest value in the confusion matrix because there are far more 'none' labels in the test set. This is because each sentence has 16 'none' labels attached and only one polarity label. Therefore, the model correctly predicted 99.2% of the 'none' labels, which can be interpreted as correctly predicting the specific entity addressed by the sentence.

The results produced by the model comprise entities, in this case one of the SDGs and the associated sentiment, which may be negative, positive, or neutral. Consequently, each sentence for each country and report is allocated to all SDGs, with either a "none" designation if the sentence does not address the SDG or one of the sentiments. In order to derive a sentiment score for our analysis, we assign numerical values to different sentiment labels. These values are as follows: a value of 1 for positive sentiment, a value of -1 for negative sentiment, and a value of 0 for neutral sentiment. Sentences labeled as 'none' are considered

to be missing data. The sentiment score, which ranges between -1 and 1, is calculated by computing the mean sentiment score for each SDG based on all sentences within a single report per country. The mean is calculated by averaging the assigned numerical values, with missing data (NaNs) being excluded from the calculation. The resulting values represent the sentiment score that will be utilized for subsequent analysis.

Figure 3.3 summarizes the sentiment score for each SDG in each country in our dataset. For the 56 countries that published more than one report, we took the average of the countries' sentiment score, leaving us with 166 countries and their respective scores. We also calculated the median of these averages for comparison. We find that the median of the average sentiment scores is positive across all 17 SDGs at 0.209. However, we find that the sentiment score varies considerably across countries around the world. Medians of the mean sentiment score range from 0.0734 for SDG 8 to 0.416 for SDG 10. In particular, we note that SDGs 7, 10, 13, and 17 appear to be rated more positively than the median, while SDG 16 has a more negative sentiment score, as indicated by its scores falling outside the interquartile range of SDGs.

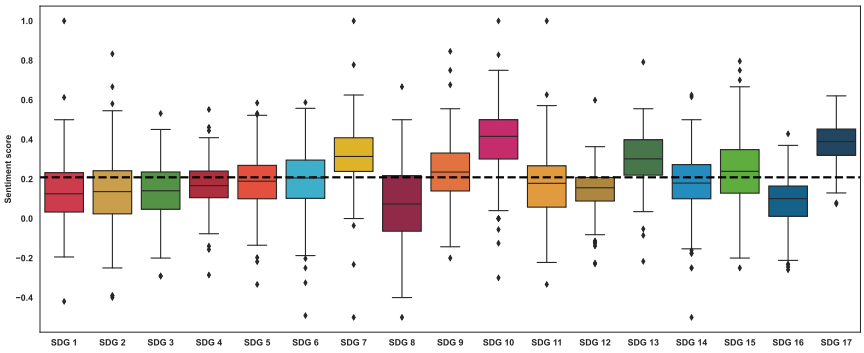


Figure 3.3: Boxplots of model-predicted sentiment scores for each SDG.
This series of boxplots illustrates the distribution of the model-predicted sentiment scores for each of the 17 Sustainable Development Goals (SDGs). A black dashed line across the boxplots marks the overall mean of the median scores across all SDGs, which serves as a benchmark for comparison.

In addition to estimating VNR tone at the entity level, we applied three vanilla SA sentiment models to assess sentiment across all 17 SDGs within the VNRs, and determined the median sentiment score for each. However, we did not train the model; instead, we used the 'cardiffnlp/twitter-roberta-base-sentiment-latest', 'Souvikcmsa/BERT_sentiment_analysis' and 'Souvikcmsa/SentimentAnalysisDistillBERT' models from the transformers pipeline. Figure 3.4 provides a comparison between the vanilla SA models and our ABSA results, averaged across all 17 SDGs. Overall, the sentiment predictions do not appear to differ qualitatively regarding median sentiment score (ABSA: 0.2088, RoBERTa: 0.1352, BERT-base: 0.1344, DistilBERT: 0.1133) and the standard deviation (ABSA: 0.1086, RoBERTa: 0.1024, BERT-base: 0.1170, DistilBERT: 0.1078). However, the ABSA model offers a more

fine-grained analysis.

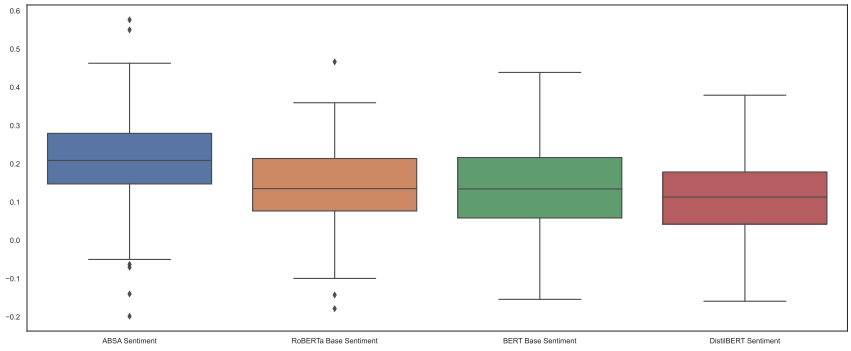


Figure 3.4: Comparative Analysis of Sentiment Scores: ABSA Model vs. Vanilla SA Models.

This figure presents boxplots that compare the overall sentiment scores predicted by our Aspect-Based Sentiment Analysis (ABSA) model with those derived from three baseline vanilla sentiment analysis models—BERT, RoBERTa, and DistilBERT. The boxplots illustrate a slightly higher mean sentiment score for our ABSA model, indicating robust performance. Notably, our ABSA model offers the added benefit of entity-specific sentiment scores, providing a more granular analysis than the general sentiment outputs of the baseline models.

3.4 Results

3.4.1 Sentiment Analysis Results

Figure 3.5 summarizes the sentiment score for each SDG by market classification. We find that LDCs and FM countries tend to report more positively than DM countries. In most cases, either FM countries or LDCs lead in terms of positive reporting on SDGs 3, 4, 5, 6, 7, 10, 14, 15, 16, and 17. Since VNRs are also tools for communicating a country's progress on each SDG, these findings seem plausible. FM countries and LDCs are generally further away from fully achieving the SDGs, but they appear to be making good progress toward achieving them.

In addition, EM countries report the most positively on SDG 13 'climate action', while the largest differences are found on SDG 1 'no poverty' and SDG 8 - 'decent work and economic growth'. The latter, in particular, is not surprising, given that emerging economies, such as China and India are experiencing, are showing higher economic growth, increased productivity and technological innovation, and, most importantly, a growing middle class (Kravets & Sandikci, 2014; Claessens & Yurtoglu, 2013; Kearney, 2012).

Finally, we find that DM countries on average report less positively on most SDGs than EM countries, FM countries, and LDCs do. The exceptions are SDGs 2, 9, 11 and 12,

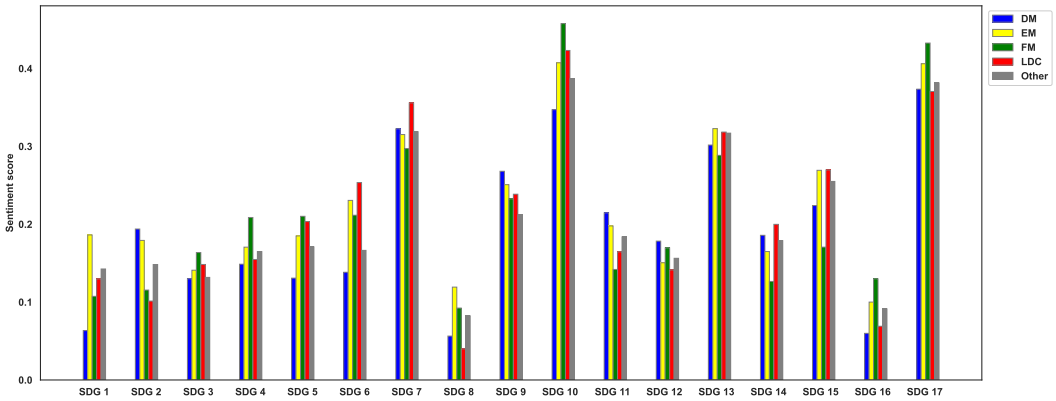


Figure 3.5: Sentiment score per SDG by MSCI country classification.
This figure represents the results of our sentiment analysis for each of the 17 Sustainable Development Goals (SDGs), by country classification. For the 56 countries that published more than one report, we took the average of the countries' sentiment score, leaving us with 166 countries and their respective scores.

although the differences from second place are negligible. One possible explanation for the results could be that the countries that are already close to achieving some goals are less to report improvements than countries that are at a much lower level of achievement. This possible explanation, according to Allen et al. (2019), applies to Australia, where closing the gap to reach 100% of targets seems to be the most difficult. Other DM countries may face the same problems.

3.4.2 Common Development Paths

Descriptive statistics alone provide only a summary measure and therefore do not describe the spatial (and temporal) variation within VNRs. To explore common development paths and the potential for cross-country collaboration, we used the t-distributed stochastic neighbor embedding (t-SNE) algorithm. This allows us to compress the multidimensional sentiment data related to the 17 SDGs within each VNR into a singular two-dimensional representation suitable for visualization in a scatterplot. This allows us to assess which countries report similarly – positive or negative – about achieving the SDGs and whether other potentially interesting patterns emerge. Clusters of countries on the plot suggest analogous sentiments towards the goals, indicating common challenges or successes related to the 2030 Agenda.

Figure 3.6 provides an overview of all 17 SDGs and 232 VNR reports, which are embedded in a two-dimensional space. Additionally, we report the results of a PCA, a linear dimension reduction technique, that maximizes the variance by preserving large pairwise distances as a robustness check with qualitatively similar results in Figure 3.7. One would expect that for countries issuing multiple reports, these documents would cluster closely on the t-SNE plot, reflecting a consistent sentiment score over time. In addition, a narrower time interval between report publications should correlate with proximity on the plot, as minimal changes in sentiment score are expected over shorter periods. This pattern appears to be supported

when examining Figure 3.6.

In general, DM countries appear to be more concentrated in the plot's center and exhibit a smaller spread compared to EM countries, FM countries, and LDCs. Table 3.2 presents the statistics for the nearest neighbor distances within each country classification. This analysis reveals that DM countries exhibit the smallest mean and median distances, indicating that the VNRs in this category are generally closer to their nearest neighbors compared to other categories. This suggests a tighter clustering. Furthermore, the standard deviation for DM countries is the lowest, indicating that the distances between points in this category are more consistent than in the other categories. In contrast, we observe greater differences between EM countries, FM countries, LDCs, and the 'Others,'where we find considerably greater dispersion (especially among LDCs). Consequently, LDCs exhibit the highest standard deviation. This indicates that they exhibit the most variability in distances to the nearest neighbor with some LDCs being very close and others much further apart. Moreover, the median of the LDCs is considerably lower than its mean, which suggests that the distribution of nearest neighbor distances is skewed, with a greater number of points having a distance less than the mean. The 'Others' category has the highest mean distance, indicating that points in this category are, on average, the most distant from their nearest neighbors. This could indicate a greater dispersion within the category or a more heterogeneous composition of points.

Figure 3.6 also illustrates potential synergies between specific countries. For instance, Canada, Israel, Australia, Great Britain, Ireland, and Singapore have expressed similar sentiments regarding the SDGs in reports spanning from 2018 to 2020. Germany, Austria, and Japan appear to share comparable challenges or achievements in their progress toward the SDGs. Additionally, a cluster is evident for Spain, Switzerland, Finland, Norway, Sweden, and Belgium. Denmark's reports exhibit greater variation over time. However, in 2017, its report is comparable to those of Finland and Latvia, with the latter two countries' reports dating to 2016 and 2018, respectively. Further groups can be formed based on the clustering approach, but naming them all would be impractical, particularly given the greater dispersion of FMs, EMs, and LDCs. Nevertheless, in order to identify potential policy implications for a specific country, it is necessary to examine the country in question and the cluster to which it belongs, in order to identify any potential synergies.

In summary, DMs form the most cohesive and uniform cluster, whereas that of LDCs and the 'Others' category is more dispersed and variable. The median distances, particularly for LDCs, suggest that the distribution of distances is not symmetrical, and that there may be outliers influencing the mean.

3.4.3 Sentiment Scores vs. SDG Scores

To compare VNR sentiment scores with actual country performance, we employed Kendall's rank correlation coefficient as the analytical framework, as illustrated in Table 3.3. The correlation between the outcomes of our ABSA model and the official UN SDG scores

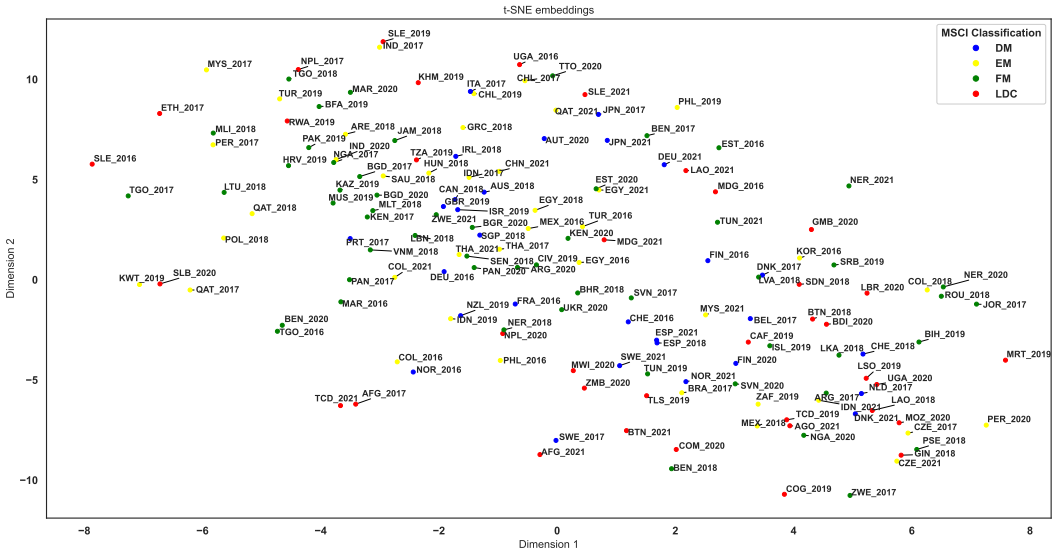


Figure 3.6: t-SNE embedding of country sentiments towards the 17 SDGs, excluding 'Others'.

This figure illustrates the results of our t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction by embedding the sentiment towards all 17 Sustainable Development Goals (SDGs) of the countries into a two-dimensional space. The country labels are abbreviated using their ISO3 codes, accompanied by the corresponding publication year. It should be noted that the category 'Others' has been intentionally excluded from this figure for illustrative purposes. For the complete representation including the category 'Others', please refer to Figure 3.8.

was determined. We opted for Kendall's Tau instead of Pearson's correlation coefficient, as the latter assumes a normal distribution of data and a linear relationship between variables, conditions that may not hold true due to the nature of our sentiment dataset. Due to the brevity of some VNRs or the limited scope of their discussion of specific SDGs, it is not uncommon for the sentiment score to be recorded as 1 or -1 at the country level. Furthermore, we selected Kendall's Tau over Spearman's correlation coefficient due to its robustness. In order to conduct our analysis, we utilized a cross-data sample that included all countries for which we had access to VNRs. As the scores are derived from a combination of indicators from different years, the mean sentiment score for each country was calculated across all years that the countries reported, ensuring that the period was matched. This procedure yielded a mean SDG sentiment score for each country and an SDG score per SDG. Our hypothesis is that a country underperforming in achieving a SDG will exhibit a low SDG score. This is likely to be reflected in their reports with a negative tone. Consequently, we hypothesize that there is a positive correlation between the sentiment scores and the SDG scores.

Table 3.3 reveals a relationship that differs from the initial hypothesis, particularly when analyzing the collective data from all countries (Overall). The findings indicate that there is no relationship for the majority of SDGs, as indicated by the lack of significant coefficients at

Table 3.2: Statistics for the nearest neighbor distances within each Country Classification.

This table provides a descriptive analysis of the nearest neighbor distance, which is a measure of proximity between points in a dataset. In the context of our country classifications, these distances quantify the closeness or similarity between countries within the same group, according to the selected criteria.

Country Classification	mean	median	std
DM	0.475	0.473	0.221
EM	0.533	0.531	0.303
FM	0.553	0.529	0.283
LDC	0.591	0.469	0.382
Others	0.632	0.602	0.320

conventional significance levels. Nevertheless, there are exceptions. Notably, SDGs 2 and 11 exhibit significant and positive correlations. This suggests that countries with higher scores on these SDGs tend to report more positive towards them, and vice versa. Conversely, SDG 5 exhibits a significant negative correlation, indicating that countries with lower SDG scores for SDG 5 tend to report more positively on this goal, and vice versa.

Upon examination of the disaggregated correlations by country classification, it becomes evident that the results exhibit a greater degree of distinct patterns. In the case of the DMs, a pronounced negative correlation is observed for SDG 3 and SDG 17, which is significant at the 1% level. This suggests that higher SDG scores on these SDGs may be associated with more negative sentiment score. Additionally, we observed negative correlations for SDGs 5, 7, 9, and 11 at the 5% significance level, and to some extent for SDG 10 at the 1% level, which provides additional support for the notion that greater SDG achievements might correspond with more negative sentiment score within these categories.

For EMs, the correlations are largely non-significant. However, there are exceptions: SDGs 4, 5, and 11 show significant negative correlations at the 10% level. For FMs, SDG 15 is noteworthy for its robust positive correlation, indicating that higher SDG scores are associated with more positive sentiment score. Similarly, LDCs demonstrate positive correlations for SDGs 2, 3, and 11, with SDG 2 exhibiting a particularly robust positive relationship. Furthermore, countries classified as 'Others' also exhibit significant positive correlations for SDGs 11 at the 5% and 17 at the 10% level.

These findings, which are disaggregated by country classification, reveal that the relationship between SDG scores and sentiment scores is complex and varies considerably across different country classifications. This highlights the importance of considering these contexts when analyzing sentiment towards SDGs. This is consistent with the results presented in Figure 3.5, which indicate that countries that are perceived to be less capable of achieving

the goals (LDCs and FM countries) tend to write more positively about them in their reports.

Table 3.3: Kendall rank correlation between sentiment score and SDG score.

This table displays the outcomes of a Kendall rank correlation analysis that compares the sentiment score derived from our ABSA model for each Sustainable Development Goal (SDG) with the official UN SDG Index scores. The 'Overall category' refers to the inclusion of all countries in the analysis, while DM (Developed Markets), EM (Emerging Markets), FM (Frontier Markets), LDC (Least Developed Countries), and 'Others' represent the specific country classifications consistently used throughout the paper.

SDG	Overall	DM	EM	FM	LDC	Others
SDG 1	0.004	0.068	0.035	−0.234	0.169	0.086
SDG 2	0.114**	0.176	0.000	0.149	0.236**	0.036
SDG 3	0.005	−0.361***	0.004	−0.160	0.197*	0.046
SDG 4	−0.009	0.110	−0.216*	−0.043	0.112	−0.019
SDG 5	−0.124***	−0.330**	−0.237*	−0.202	0.123	0.007
SDG 6	−0.081*	0.087	−0.149	0.155	−0.111	0.018
SDG 7	−0.010	−0.285**	−0.142	0.080	−0.102	0.071
SDG 8	0.045	−0.142	−0.104	−0.208	0.126	0.099
SDG 9	0.083*	−0.281**	0.132	0.032	0.103	0.105
SDG 10	−0.064	−0.249*	−0.200	0.066	0.105	−0.004
SDG 11	0.102**	−0.303**	−0.236*	0.097	0.192*	0.191**
SDG 12	−0.046	0.016	−0.015	0.096	0.021	−0.033
SDG 13	0.014	0.129	−0.144	0.213	0.154	−0.047
SDG 14	−0.059	−0.122	0.078	−0.146	−0.097	−0.038
SDG 15	−0.035	−0.086	−0.111	0.396**	−0.090	−0.011
SDG 16	0.051	−0.222	−0.033	−0.064	0.313***	0.093
SDG 17	0.002	−0.390***	−0.156	0.144	0.028	0.138*

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

3.5 Discussion

Our study makes a distinctive contribution by conducting a comprehensive sentiment analysis of VNRs. In order to explore each country's perspective on its progress toward the SDGs, we employed the ABSA method. While previous research has shed light on the dynamics of sustainable development, the trade-offs, synergies, and geographical patterns associated with SDGs (Le Blanc, 2015; Hegre et al., 2020; Pradhan et al., 2017; Pham-Truffert

et al., 2020; Nilsson et al., 2018; Xiao et al., 2023; D'Adamo et al., 2021; Anselmi et al., 2023; Murphy et al., 2023), our use of ABSA to analyze VNRs introduces a novel and pivotal angle to the field. In contrast to existing studies that primarily focus on textual similarities (Sebestyén et al., 2020) or rely on UN-provided indices (Le Blanc, 2015; Hegre et al., 2020; Pradhan et al., 2017; Xiao et al., 2023; D'Adamo et al., 2021; Anselmi et al., 2023; Murphy et al., 2023), our methodology places emphasis on the sentiment expressed by countries regarding their SDG progress. In this way, we prioritize the self-evaluations of nations, thereby capturing a more complementary view of their challenges and successes than is possible with traditional indices. The t-SNE algorithm has been employed in an analysis of the data, which has yielded insights into the patterns of reporting by countries on their progress towards the SDGs. By reducing the multi-dimensional sentiment data to a two-dimensional space, we have created a visual representation of the similarities and differences in countries' sentiments towards the SDGs. The observed clustering in the scatter plot suggests that countries with similar sentiment scores may face shared challenges or successes, which could be pivotal for cross-country collaboration within the 2030 Agenda framework. This appears to be particularly the case for DM countries, as they are more centralized in the visualization, indicating a more uniform reporting sentiment, whereas EM, FM and LDCs show a broader dispersion. This could reflect the varying degrees of challenges faced by these countries, with LDCs appear to exhibit the most variability, suggesting a diverse range of experiences and responses to the SDGs. The close grouping of countries, including Switzerland, Sweden, Norway, Denmark, and Finland in the two-dimensional embedding, all of which are among the top 10 most sustainable countries worldwide (Dias et al., 2017), aligns with findings from Sebestyén et al. (2020).

In particular, we observe analogous clusters in the context of reported sentiments, which is consistent with the keyword clustering approach proposed by Sebestyén et al. (2020). For instance, as demonstrated in our clustering approach, Australia is situated in close proximity to Canada and Ireland. Consequently, not only do these countries share similar sentiments, but they also face similar specific challenges in achieving the SDGs. This suggests that they could potentially benefit from exploring synergies and collaborating on strategies to overcome shared obstacles. Additionally, Switzerland, Norway, and Finland are identified as a group in both the results of Sebestyén et al. (2020) and our study. While we find some similar groupings to those in Sebestyén et al. (2020), there are also differences. For example, Italy, Spain, Montenegro, Portugal, and Greece do not form a close group in our results. This suggests that there are differences between some countries with regard to the topics covered in VNRs regarding the SDGs and the sentiment towards the SDGs' progress. This suggests a consistency in reporting sentiments that correlates with actual sustainability performance. It would be advantageous for policymakers to establish forums where these exemplary countries can disseminate best practices and insights, with the objective of accelerating progress towards the achievement of the 2030 Agenda.

However, the variability in sentiment scores for countries publishing multiple reports, especially in the context of the COVID-19 pandemic, which brought significant additional

challenges to achieving the SDGs by 2030 (Abidoye et al., 2021), indicates that external factors such as global crises can significantly influence the narrative of progress. This is particularly evident for approximately half of the countries, for which two or more reports are available before and after the year 2020. The availability of reports from the same country from different periods allows us to conduct a comparative analysis that differs from that of Sebestyén et al. (2020) who do not include the time horizon in their analysis. For instance, VNRs from Peru, Nepal, and Argentina in South America; Benin, Chad, Nigeria, Uganda, and Zimbabwe in Africa; as well as Malaysia, Laos, Indonesia, and Qatar in Asia, demonstrate that the sentiment expressed in reporting has undergone a significant transformation over time. This underscores the necessity for policies that are sufficiently flexible and robust to withstand external shocks. In light of this knowledge, policymakers must utilize it to create resilient strategies that ensure continuous progress towards the SDGs, even in the face of unforeseen challenges.

Similarly to Pradhan et al. (2017), we employ a correlation analysis. However, we do not examine correlations between SDG scores to identify potential synergies or trade-offs. Instead, our objective is to ascertain whether the SDG scores align with the reporting sentiment of the countries themselves. The Kendall rank correlation coefficient was employed to compare the sentiment expressed in VNRs with the actual SDG performance scores. In contrast to our initial hypothesis, which predicted a positive correlation between VNR sentiment scores and SDG scores, we found an absence of significant correlations for the majority of SDGs on a collective level. Notably, SDGs 2 and 11 show significant positive correlations, while SDG 5 presents a negative one, indicating a complex relationship between reported sentiments and actual performance. A further analysis by country classification revealed some interesting patterns. It can be observed that DM countries often report higher SDG achievements with a critical tone, as indicated by the negative correlations. This could be indicative of the fact that despite their progress, such countries may report with a critical perspective, which may be reflective of higher self-imposed standards or expectations. Conversely, FMs and Least LDCs expressed a more positive tone in general. This positivity may be attributed to the perception that making advancements from a lower starting point seems less daunting than the challenging endeavor of approaching the SDG finishing line for countries that are already on the brink of achieving these goals (Allen et al., 2019). In contrast, the data indicates that there are significant negative correlations between the DMs and certain SDGs, while the FMs and LDCs exhibit positive correlations for others, such as SDG 16 and SDGs 2, 3, and 11, respectively. The critical tone observed in the VNRs of DM countries, despite their high SDG scores, suggests that self-assessment can be a powerful tool for continuous improvement. One could argue that policymakers in these countries should maintain a critical perspective to identify areas for further development and avoid complacency.

In summary, the findings indicate that the sentiment expressed in VNRs may not consistently reflect a country's actual progress on the SDGs. Rather, they appear to be shaped by a variety of factors, including political, economic, and social contexts, as well as the distinct reporting styles and strategies of the countries in question. This complexity

suggests that VNRs require a more nuanced interpretation that considers the broader context in which these reports are produced and also indicates the need for differentiated policy approaches. We recommend that policymakers consider the unique contexts of each country classification when formulating SDG-related policies. In particular, for LDCs, it is crucial to develop policies that foster positive sentiment and active engagement with the SDGs, as our findings indicate that positive reporting may not always correlate with high performance. However, the quality of SDG scores is also open to question (Diaz-Sarachaga et al., 2018; Mugellini et al., 2021). Consequently, the results of our study may also be attributed to the lack of SDG score quality. Therefore, it is recommended that policy-makers consider using VNRs as an addition to traditional numerical data in the form of SDG scores.

The findings of our study offer new insights into the field, yet it is essential to acknowledge the limitations of the study and the avenues they open for future research. Firstly, the t-SNE algorithm, while effective for visualization, may not fully capture the multi-dimensional nature of sentiment data. Although we have supplemented this with a PCA dimension reduction, which yields similar results, the full complexity of the data might still be underrepresented. Moreover, the use of Kendall's Tau, despite its robustness, may not fully capture all the nuances of the relationship between sentiment and SDG scores. Additionally, the issue of outliers due to truncated SDG texts in some reports presented a further challenge. The correlation analysis between sentiment scores and the official UN SDG scores reveals significant variations across different country classifications, further highlighting the inherent complexities in conducting such analyses. The inconsistency in VNR lengths and the selective focus on certain SDGs further complicate a comprehensive analysis. We recommend that the UN to provide more comprehensive guidelines regarding the length and structure of reports. Such measures would not only facilitate a more standardized comparison but also aid in recognizing shared pathways for development across different nations.

One significant challenge was the language barrier, as approximately one-third of VNRs were not published in English and thus required translation using NLP tools. While this approach enabled the inclusion of these reports, it is possible that translations may not fully capture the nuances of sentiment and specific terminology, which could result in the loss of crucial information. As other authors, such as Sebestyén et al. (2020), who face the same issue and therefore only had 75 VNRs available in their study, this reinforces our recommendation that VNRs be published in English in addition to the national languages. This will ensure inclusivity and a wider scope for analysis. Such a measure would facilitate research and promote global understanding and collaboration with respect to the SDGs.

The absence of a consistent annual publication requirement for VNRs limited our ability to undertake a detailed time-series analysis from 2016 to 2021. Once more countries have published a second report, a time-series analysis could be a viable option to provide further details on how and whether the tone of VNRs has changed over the years. A comparable issue was encountered when attempting to utilize the official UN scores as indicators, which serves to reinforce the aforementioned commentary on the quality of SDG scores. The aggregated UN score for a single SDG is not published as a time series; rather, it is presented

as a combination of several sub-indicators from various years. Previous studies (Le Blanc, 2015; Hegre et al., 2020; Pradhan et al., 2017; Xiao et al., 2023; D'Adamo et al., 2021; Anselmi et al., 2023; Murphy et al., 2023) might have benefited from a more comprehensive dataset, which would have included a continuous time series. Consequently, future research could concentrate on utilising alternative indicators (e.g., GDP per capita) to facilitate a comparative analysis of sentiment across reports. Future research could, similar to D'Adamo et al. (2021), but using NLP techniques, perform an analysis on a regional basis if sufficient data is available, as the model we used is applicable to any text. Therefore, if texts on the progress of the SDGs are available on a regional level, our analysis could also be conducted on a disaggregated level. One disadvantage of employing the sentiment of reports in place of keywords, as exemplified by Sebestyén et al. (2020), is that it precludes the formulation of targeted policy recommendations for specific countries with respect to the SDGs. However, our advantage lies in the fact that the sentiment may be able to capture progress, rather than merely identifying similar SDG themes across countries.

The insights from our study also have implications for future research and policy-making. The patterns observed in sentiment reporting could facilitate cross-country collaborations and facilitate the sharing of best practices and the devising of strategies to tackle common challenges. Moreover, the discrepancies between sentiments and actual SDG performance underscore the necessity for a more nuanced approach to the analysis and interpretation of VNRs. Future research should examine the factors that influence the sentiment of reports, including the impact of global events such as the COVID-19 pandemic. An investigation into the reasons behind the tendency of DMs to present a critical outlook despite achieving high scores on the SDGs provide insight into the intricacies of self-assessment within the SDG framework. This investigation could reveal insights into the dynamics of reporting and the actualization of sustainable development goals, thereby providing a more comprehensive understanding for for policy-makers and stakeholders.

3.6 Conclusion

In our analysis, we used an advanced sentiment analysis technique, ABSA, to examine the sentiment expressed in VNRs concerning the SDGs. This is the first comprehensive sentiment analysis of VNRs, offering an in-depth analysis of a country's distinctive perspective as they report on their progress towards the 2030 Agenda. We trained an ABSA model on UN progress reports and then applied the model to the reviews. Consequently, we generated sentiment score for each country on each of the 17 SDGs. First, we applied t-SNE to our results in order to identify similarities across countries that have published VNRs. Our findings indicate that while there is a consistent sentiment exhibited by DMs over time, there is a notable dispersion in the sentiments of FMs and LDCs. This reflects the varied challenges and developmental stages that these countries face. Secondly, we examined the overall sentiment expressed by the countries in question. Our findings indicated that FM countries and LDCs tend to report more positively than DM countries in their assessments. This phenomenon, which may be

described as a 'progress paradox,' suggests that initial advancements are often reported with optimism, while the final strides towards full SDG achievement are met with a more critical assessment. One potential explanation for this finding is that achieving an SDG of 100% is typically more challenging than making progress when starting from a lower point (Allen et al., 2019). Thirdly, the absence of a significant correlation between the sentiment expressed in VNRs and the actual SDG performance scores for the majority of the SDGs indicates that VNRs may serve more as a reflection of a country's narrative and outlook rather than as a direct measure of their progress. Consequently, the interpretation of VNRs must be contextualised within the broader framework of each country's distinctive circumstances and challenges.

Upon returning to our research question, which countries face similar successes and challenges, it can be postulated that a number of countries, including Finland, Switzerland, Spain, Norway, Sweden, and Belgium, among others, appear to be making comparable progress in achieving the SDGs. This suggests that they may benefit from collaboration. Furthermore, the clustering approach can be utilized to identify common development pathways between countries that are closely clustered. The second question, whether SDG scores correlate with reporting sentiments in VNRs, can be answered in the negative. There is no uniform correlation between the sentiment expressed by countries in VNRs and official measures for all SDGs and countries classified according to market.

While our study provides insights into the countries reporting on the progress of the SDGs, there are some drawbacks. Language barriers and translation issues might compromise the accuracy of the data. Furthermore, the variance in sentiment scores across multiple reports, especially when influenced by external shocks such as the COVID-19 pandemic, highlights the sensitivity of narrative sentiment to global crises. In addition, the lack of consistent annual publication of the VNR limits the ability to conduct detailed time-series analyses, which may affect the robustness of conclusions about changes in sentiment over time.

As the 2030 deadline approaches, our research emphasizes the necessity for a nuanced interpretation of VNRs, one that considers the complex interplay of optimism, critical self-assessment, and the tangible realities of SDG implementation. Future research should employ advanced NLP techniques to further explore the sentiment and narratives embedded within VNRs. Furthermore, there is a necessity for the collection of more detailed textual data at the local level in order to facilitate the formulation of nuanced policy recommendations for collaboration. Another potential future research aim could be to understand why there are discrepancies between sentiment reported in VNRs and actual performance on the SDGs, especially in developed markets, where critical tones might indicate higher self-imposed standards or a deeper self-assessment process. Such investigations will not only enhance our comprehension of global SDG advancement but also facilitate more efficacious policy-making and international collaboration towards a sustainable future for all.

B Appendix to Chapter 3

Table 3.4: Cleaning steps of parsed raw texts.

Problem	**Fix**
Extra whitespaces	Replace extra whitespaces with a single whitespace.
Words separated by hyphen	Remove hyphen.
Words separated by whitespace	Remove whitespace.
Model parsed some ff, fi, and if characters as one special character	Replace special double characters by normal characters.
Whitespaces between word in sentence and punctuation	Remove whitespace.
URLs in text which do not have any semantic meaning	Remove all URLs from text.
Parsed some sentences character by character with whitespace between them, i.e., S E N T E N C E instead of Sentence. Problem mostly occurred with figure subtexts	As most figure subtexts do not have important semantic meaning, we removed such single characters.
Very short sentences or single tokens	Filter out sentences with less than 6 words or 50 characters.
Coding errors	Filters out sentences with a high percentage of unrecognized words using a spell checker.

This tables presents the various cleaning steps completed to extract text from parsed PDF documents.

Table 3.5: Analyzed Voluntary National Reviews - Part 1

Country	**Abb.**	**Pages**	**Year**	**Country**	**Abb.**	**Pages**	**Year**
Afghanistan	AFG	81	2017	Afghanistan	AFG	54	2021
Albania	ALB	108	2018	Algeria	DZA	176	2018
Andorra	AND	116	2018	Angola	AGO	237	2021
Antigua and Barbuda	ATG	144	2021	Argentina	ARG	130	2017
Argentina	ARG	252	2020	Armenia	ARM	82	2018
Armenia	ARM	60	2020	Australia	AUS	132	2018
Austria	AUT	116	2020	Azerbaijan	AZE	138	2021
Azerbaijan	AZE	70	2017	Azerbaijan	AZE	137	2019
Bahamas	BHS	159	2018	Bahrain	BHR	106	2018
Bangladesh	BGD	82	2017	Bangladesh	BGD	201	2020
Belgium	BEL	95	2017	Belize	BLZ	51	2017
Benin	BEN	80	2017	Benin	BEN	92	2018
Benin	BEN	94	2020	Bhutan	BTN	116	2021
Bhutan	BTN	86	2018	Bolivia	BOL	115	2021
Bosnia and Herzegovina	BIH	96	2019	Brazil	BRA	41	2017
Brunei Darussalam	BRN	113	2020	Bulgaria	BGR	132	2020
Burkina Faso	BFA	117	2019	Burundi	BDI	138	2020
Cabo Verde	CPV	86	2018	Cabo Verde	CPV	168	2021
Cambodia	KHM	116	2019	Cameroon	CMR	200	2019
Canada	CAN	148	2018	Central African Republic	CAF	120	2019
Chad	TCD	116	2019	Chad	TCD	100	2021
Chile	CHL	258	2019	Chile	CHL	133	2017
China	CHN	84	2021	Colombia	COL	74	2016
Colombia	COL	136	2018	Colombia	COL	166	2021
Comoros	COM	112	2020	Congo	COG	119	2019
Democratic Republic of the Congo	COD	124	2020	Costa Rica	CRI	119	2017

This table offers a detailed overview of all the Voluntary National Reviews included in this analysis, detailing their ISO3 country codes, the number of pages, and the year of publication.

Table 3.6: Analyzed Voluntary National Reviews - Part 2

Country	Abb.	Pages	Year	Country	Abb.	Pages	Year
Costa Rica	CRI	156	2020	Croatia	HRV	109	2019
Cuba	CUB	124	2021	Cyprus	CYP	148	2021
Cyprus	CYP	81	2017	Czech Republic	CZE	82	2021
Czech Republic	CZE	40	2017	Cote divoire	CIV	153	2019
Denmark	DNK	140	2017	Denmark	DNK	294	2021
Dominican Republic	DOM	274	2018	Dominican Republic	DOM	124	2021
Ecuador	ECU	202	2018	Ecuador	ECU	238	2020
Egypt	EGY	59	2016	Egypt	EGY	72	2018
Egypt	EGY	92	2021	El Salvador	SLV	50	2017
Estonia	EST	60	2016	Estonia	EST	104	2020
Eswatini	SWZ	79	2019	Ethiopia	ETH	52	2017
Fiji	FJI	108	2019	Finland	FIN	64	2016
Finland	FIN	172	2020	France	FRA	53	2016
Gambia	GMB	191	2020	Georgia	GEO	16	2016
Georgia	GEO	69	2020	Germany	DEU	144	2021
Germany	DEU	59	2016	Ghana	GHA	118	2019
Greece	GRC	160	2018	Guatemala	GTM	277	2017
Guatemala	GTM	459	2019	Guinea	GIN	122	2018
Guyana	GUY	178	2019	Honduras	HND	52	2017
Honduras	HND	108	2020	Hungary	HUN	85	2018
Iceland	ISL	151	2019	India	IND	41	2017
India	IND	188	2020	Indonesia	IDN	786	2021
Indonesia	IDN	298	2019	Indonesia	IDN	138	2017
Iraq	IRQ	95	2019	Iraq	IRQ	123	2021
Ireland	IRL	300	2018	Israel	ISR	430	2019

Table 3.7: Analyzed Voluntary National Reviews - Part 3

Country	Abb.	Pages	Year	Country	Abb.	Pages	Year
Italy	ITA	50	2017	Jamaica	JAM	163	2018
Japan	JPN	258	2021	Japan	JPN	52	2017
Jordan	JOR	70	2017	Kazakhstan	KAZ	159	2019
Kenya	KEN	124	2020	Kenya	KEN	76	2017
Democratic Peoples Republic of Korea	PRK	66	2021	Republic of Korea	KOR	34	2016
Kuwait	KWT	115	2019	Kyrgyzstan Republic	KGZ	182	2020
Lao PDR	LAO	130	2018	Lao PDR	LAO	156	2021
Latvia	LVA	63	2018	Lebanon	LBN	94	2018
Lesotho	LSO	145	2019	Liberia	LBR	147	2020
Liechtenstein	LIE	77	2019	Lithuania	LTU	70	2018
Luxembourg	LUX	46	2017	Madagascar	MDG	42	2016
Madagascar	MDG	86	2021	Malawi	MWI	104	2020
Malaysia	MYS	82	2017	Malaysia	MYS	144	2021
Maldives	MDV	28	2017	Mali	MLI	70	2018
Malta	MLT	122	2018	Marshall Islands	MHL	147	2021
Mauritania	MRT	96	2019	Mauritius	MUS	140	2019
Mexico	MEX	112	2016	Mexico	MEX	162	2018
Micronesia	FSM	140	2020	Moldova	MDA	171	2020
Monaco	MCO	70	2017	Mongolia	MNG	97	2019
Montenegro	MNE	163	2016	Morocco	MAR	11	2016
Morocco	MAR	214	2020	Mozambique	MOZ	68	2020
Namibia	NAM	82	2021	Namibia	NAM	44	2018
Nepal	NPL	52	2017	Nepal	NPL	104	2020
Netherlands	NLD	44	2017	New Zealand	NZL	63	2019
Nicaragua	NIC	105	2021	Niger	NER	75	2018
Niger	NER	131	2020	Niger	NER	85	2021
Nigeria	NGA	100	2017	Nigeria	NGA	120	2020
North Macedonia	MKD	148	2020	Norway	NOR	29	2016
Norway	NOR	124	2021	Pakistan	PAK	81	2019
Palau	PLW	107	2019	State of Palestine	PSE	128	2018
Panama	PAN	106	2017	Panama	PAN	386	2020
Papua New Guinea	PNG	63	2020	Paraguay	PRY	100	2018
Paraguay	PRY	432	2021	Peru	PER	67	2017

Table 3.8: Analyzed Voluntary National Reviews - Part 4

Country	Abb.	Pages	Year	Country	Abb.	Pages	Year
Peru	PER	146	2020	Philippines	PHL	27	2016
Philippines	PHL	50	2019	Poland	POL	106	2018
Portugal	PRT	89	2017	Qatar	QAT	184	2021
Qatar	QAT	44	2018	Qatar	QAT	52	2017
Romania	ROU	94	2018	Russian Federation	RUS	228	2020
Rwanda	RWA	124	2019	Saint Lucia	LCA	51	2019
Samoa	WSM	87	2020	San Marino	SMR	109	2021
Saudi Arabia	SAU	96	2018	Senegal	SEN	153	2018
Serbia	SRB	103	2019	Seychelles	SYC	132	2020
Sierra Leone	SLE	81	2021	Sierra Leone	SLE	56	2016
Sierra Leone	SLE	54	2019	Singapore	SGP	84	2018
Slovakia	SVK	33	2018	Slovenia	SVN	78	2017
Slovenia	SVN	100	2020	Solomon Islands	SLB	106	2020
South Africa	ZAF	130	2019	Spain	ESP	181	2018
Spain	ESP	352	2021	Sri Lanka	LKA	115	2018
Sudan	SDN	62	2018	Sweden	SWE	148	2021
Sweden	SWE	88	2017	Switzerland	CHE	28	2018
Switzerland	CHE	28	2016	Tajikistan	TJK	46	2017
Tanzania	TZA	186	2019	Thailand	THA	94	2017
Thailand	THA	83	2021	Timor Leste	TLS	198	2019
Togo	TGO	32	2016	Togo	TGO	44	2017
Togo	TGO	36	2018	Tonga	TON	84	2019
Trinidad and Tobago	TTO	100	2020	Tunisia	TUN	148	2019
Tunisia	TUN	254	2021	Turkey	TUR	149	2019
Turkey	TUR	50	2016	Turkmenistan	TKM	79	2019
Uganda	UGA	118	2016	Uganda	UGA	104	2020
Ukraine	UKR	117	2020	UAE	ARE	73	2018
UK	GBR	235	2019	Uruguay	URY	386	2017
Uruguay	URY	246	2018	Uruguay	URY	110	2021
Vanuatu	VUT	97	2019	Venezuela	VEN	285	2016
Viet Nam	VNM	94	2018	Zambia	ZMB	112	2020
Zimbabwe	ZWE	58	2017	Zimbabwe	ZWE	144	2021

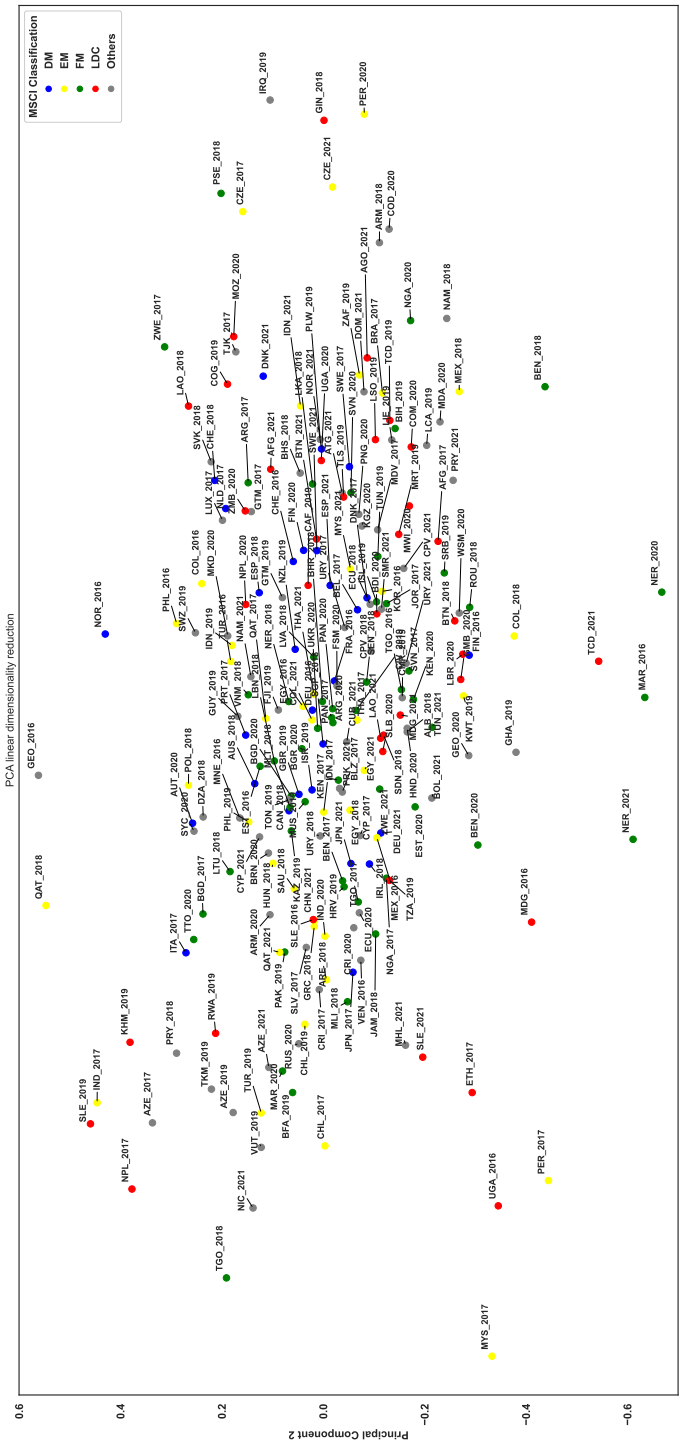


Figure 3.7: PCA linear dimensionality reduction.

This figure illustrates the results of a Principal Component Analysis (PCA), a linear dimensionality reduction technique that maximizes variance and preserves large pairwise distances, by projecting the sentiment towards all 17 Sustainable Development Goals (SDGs) of the countries into a two-dimensional space. The labels for countries are abbreviated using their ISO3 codes, accompanied by the corresponding publication year.

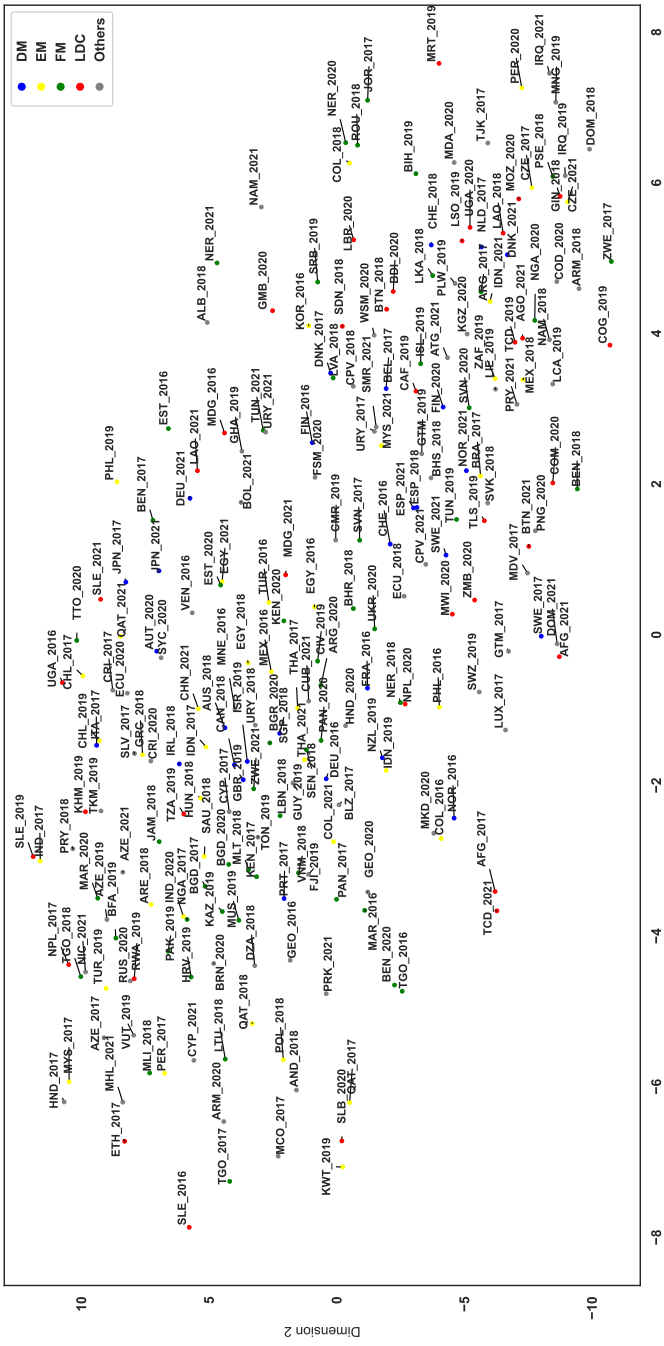


Figure 3.8: t-SNE embeddings of country sentiments towards the 17 SDGs, including the category 'Others'.

This figure illustrates the results of our t-distributed stochastic neighbor embedding (t-sne) dimensionality reduction by embedding the sentiment towards all 17 Sustainable Development Goals (SDGs) of the countries into a two-dimensional space. The labels for countries are abbreviated using their ISO3 codes, accompanied by the corresponding publication year. This figure provides a comprehensive view, incorporating the category 'Others' to present the full spectrum of data.

Part III

Sustainability in Company Reports

Chapter 4

Evaluating TCFD reporting - a new application of zero-shot analysis to climate-related financial disclosures

The following chapter is based on the paper:	
Title:	Evaluating TCFD reporting - a new application of zero-shot analysis to climate-related financial disclosures
Authors:	Elena Tönjes (contribution: 40%), Christoph Funk (contribution: 10%), Alix Auzepy (contribution: 40%), David Lenz (contribution: 10%)
Status:	Published: *Plos One*, 2023, vol. 18, nr. 11, e0288052
Available from:	`https://doi.org/10.1371/journal.pone.0288052`

Earlier versions of this paper were presented at:

- 16th International Conference on Computational and Financial Econometrics
- AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges, Arlington, Virginia, USA (online, presented by Co-Author)

Evaluating TCFD reporting - a new application of zero-shot analysis to climate-related financial disclosures

Alix Auzepy[*,†] Elena Tönjes[‡] David Lenz[‡,§] Christoph Funk[‡,¶]

Abstract

We examine climate-related disclosures in a large sample of reports published by banks that officially endorsed the recommendations of the Task Force for Climate-related Financial Disclosures (TCFD). In doing so, we introduce a new application of the zero-shot text classification. By developing a set of fine-grained TCFD labels, we show that zero-shot analysis is a useful tool for classifying climate-related disclosures without further model training. Overall, our findings indicate that corporate climate-related disclosures increased after the launch of the TCFD recommendations and following individual endorsements. However, there are marked differences in the extent of reporting by recommended disclosure topic, suggesting that some recommendations have not yet been fully met. Our findings yield important conclusions for the design of climate-related disclosure frameworks.

[*] Chair of Banking & Finance, Justus Liebig University, Giessen, Germany
[†] Corresponding author: alix.auzepy@wi.jlug.de
[‡] Department of Economics, Chair of Statistics and Econometrics, Justus Liebig University, Giessen, Germany
[§] istari.ai, Mannheim, Germany
[¶] Center for International Development and Environmental Research (ZEU), Justus-Liebig-University, Giessen, Germany

4.1 Introduction

Published in 2017, the recommendations of the Financial Stability Board's Task Force on Climate-related Financial Disclosures (TCFD) have been described by the Government of the United Kingdom (UK) as "one of the most effective frameworks for companies to analyse, understand and ultimately disclose climate-related financial information" (Department for Business, 2021).

The TCFD recommendations, which have been formally endorsed by more than 4,000 companies worldwide to date, are a set of voluntary disclosure guidelines aimed at providing consistent climate-related information to investors and other key company stakeholders (Friederich et al., 2021). Compared to other reporting frameworks (e.g. Carbon Disclosure Project, Global Reporting Initiative), a particular focus of the TCFD recommendations is on disclosing information about the integration of climate-related considerations into risk management, control structures and strategic aspects of business operations (Beyene et al., 2022). Overall, the recommendations are organized into four primary disclosure categories (Governance, Strategy, Risk Management, Metrics and Targets) and eleven corresponding recommended disclosures associated with each category. For an overview of the TCFD recommendations, see Table 4.1.

Since reliable information on climate-related exposure is critical for making informed investment decisions and appropriately pricing risks, an increasing number of investors have been exerting pressure on companies to issue reports that include comprehensive climate-related disclosures (Krueger et al., 2020). In addition, several countries, including the UK and Switzerland, have taken steps to make TCFD reporting mandatory for large companies in their jurisdictions. From a company perspective, the disclosure of climate-related information often signals awareness and preparedness for climate-related issues, while the absence of disclosure may, on the contrary, indicate that such issues are not being sufficiently addressed by the company (Sullivan & Gouldson, 2012; Bingler et al., 2022). Illustrating these arguments, Jung et al. (2018) and Subramaniam et al. (2015) show that firms are more likely to integrate risks associated with climate change into their overall risk management when such firms also disclose climate-related information.

Against this background, it is surprising to find that research on climate-related disclosures remains sparse. More importantly, prior studies on the TCFD recommendations mostly focus on the quantity of information disclosed at the aggregate TCFD category level, and have left an analysis of the reported content within each category

largely untouched. A more in-depth analysis within each of the 4 TCFD pillars is essential, as the quality and financial materiality of the disclosed information may vary (Friederich et al., 2021). For example, Bingler et al. (2022) investigate climate-related disclosures based on the four core TCFD disclosure categories from 2015 to 2020. Their findings reveal that firms selectively disclose climate-related information, mainly concentrating on Governance and Risk Management, which the authors regard as the least material categories. Similarly, Ding et al. (2022) analyze how carbon emissions affect voluntary climate-related disclosures at the TCFD category level. Their results show that firms with higher levels of carbon emissions disclose more climate-related information. Specifically, they report a positive relationship between carbon emissions and disclosures at the category level for strategy, risk management and metrics and targets. Finally, Friederich et al. (2021) analyze the types of climate risks (physical or transition) reported in corporate annual reports. Their study provides evidence that disclosures related to transition risks have experienced a more pronounced increase compared to disclosures concerning physical risks, which are still lagging behind. While the authors take a more granular approach with regard to the materiality of climate-related disclosures, their findings do not encompass all the recommended TCFD disclosures, as climate risks make up only a small portion of the multifaceted issues to be addressed in TCFD reporting.

A shared characteristic of these studies is to rely on computerized textual analysis techniques, such as natural language processing (NLP), for the evaluation of climate-related disclosures. An important contribution in this regard is the work by Bingler et al. (2022), who introduced "ClimateBERT", a BERT model specifically fine-tuned to identify climate-related information within company disclosures based on the broad TCFD categories. Similarly, Friederich et al. (2021) employ multiple pretrained BERT-related models, including DistilBERT and RoBERTa, to examine references to climate risks in company reports. Overall, these studies present mixed results regarding the effectiveness of climate-related disclosures in delivering high-quality and material information, primarily due to challenges such as greenwashing, a lack of transparency, and insufficient availability of quantitative data (Mehra et al., 2022).

Motivated by the importance of a detailed and comprehensive analysis of climate-related disclosures not only at the broad TCFD category level, but also within each individual category, we contribute to the literature in two ways. First, we provide novel insights into the state of disclosures related to the TCFD recommendations by examining a sample of 3,335 reports published by TCFD-supporting banks between 2010 and 2021. Our focus on the banking industry is due to two key factors. On the

one hand, it aligns with the fact that the TCFD recommendations primarily target financial institutions. On the other hand, it acknowledges the varying materiality of sustainability-related information across different industries. Within the financial sector, the identification and reporting of carbon-related asset concentrations hold particular importance due to the impact of climate change on credit risk and the potential risks associated with stranded assets (Beyene et al., 2021; FSB, 2015). Thus, we build a sample of TCFD-supporting banks by retrieving all banks that have publicly declared their support to TCFD and are listed as official supporters on TCFD's website. To leverage information on the TCFD recommendations beyond the 4 core disclosure categories, we develop a set of 14 fine-grained labels that are designed to capture the most central aspects of the TCFD recommendations for banks using similar semantics.

Second, we contribute to the literature on the use of NLP in the context of climate-related financial disclosures by introducing the zero-shot text classification as a new method for systematic and automated extraction of textual information from large amounts of reports including climate-related data. The zero-shot approach offers a critical advantage over other language models as it allows for sentence classification based on labels for which it has received no specific prior training. Our model relies on a multi-label approach and assigns probabilities (ranging from 0 to 100%) to each extracted text sequence in our sample of 3,335 reports. These probabilities represent the likelihood that a given text sequence aligns with a specific label. A higher probability suggests that the semantics of a text sequence match the semantics of the corresponding label, indicating that the sequence addresses the topic specified by the label. Thus, when a text sequence explicitly and precisely discusses a topic related to a label, it receives a higher label probability through the zero-shot text classification. In our analyses, we interpret higher label probabilities as proxys of disclosure quality. Furthermore, a higher label probability can also serve as an indication of the extent of disclosure on a particular topic, as a more detailed coverage of the topic is likely to result in a higher probability. As our method does not require any labeled training data, it also does not impose any restrictions on the number of labels (or classes), which allows us to perform a more detailed analysis of the underlying TCFD recommended disclosures. Additionally, the TCFD recommendations are well-suited for the zero-shot analysis, as they provide us with an already-existing framework and semantics (Ding et al., 2022).

In contrast, a weakness of ClimateBERT (Webersinke et al., 2021) and, more generally, of algorithms trained to identify and classify climate-related content is

that such models require an extensive training set of human-labeled sentences. Manual labeling of sentences is not only time-consuming, but can also be error-prone. Therefore, for quality and consistency reasons, highly-trained and specialized "labelers" are required, which can also make the labeling process costly. Furthermore, the more classes (or categories of labels) to be included into the classification scheme of the model, the more labeled data is needed to ensure that each class comes with a reasonable amount of examples attached to it, which can be a limiting factor in some scenarios.

Our paper yields the following sets of findings. First, we investigate the level of disclosure within our sample of TCFD-supporting banks at the broad category level. Specifically, we develop two different types of labels: "general labels" that cover topics that are not specifically related to climate, as well as "climate-related labels" that correspond to the broad TCFD categories (i.e., climate-related governance, climate-related strategy, climate-related risk management and climate-related metrics and targets). We find that the mean probabilities relating to the general labels remain stable over the period from 2010 to 2021, while we observe an increase in all of the probabilities for the climate-related labels at category level over the same time period. In particular, we report that the disclosures related to "climate-related strategy" and "climate-related metrics and targets" grew particularly dynamically, reaching mean probabilities of up to 22% and 20% respectively in 2021, compared to 12% in 2010. Overall, this suggests an increased attention and priority given to the development of climate-related business strategies with corresponding targets among TCFD-supporting banks.

As a next step, we analyze the mean probabilities associated with our fine-grained labels, which cover the underlying recommended disclosures. This approach enables us to provide a more nuanced assessment of reporting quality beyond each broad TCFD category. Our results indicate substantial variation and notable gaps in reporting. In the strategy area, which is the most comprehensive category and contains several specific recommended disclosures for banks, we find that label probabilities are particularly low for disclosures related to financing and investment activities for carbon-intensive industries such as the fossil fuel industry. Similarly, in comparison to other strategy-related topics, TCFD-supporting banks exhibit a reduced likelihood of explicitly addressing the use of climate-related scenario models in their reporting. Under metrics and targets, we find that the incorporation of climate-related performance metrics into remuneration policies is associated with a lower label probability compared to labels related to carbon footprints and emissions

reduction targets. In the governance area, TCFD-supporting banks demonstrate comparable levels of disclosure regarding the board's responsibility in overseeing climate-related issues and the management's role in assessing and managing such matters.

Third, as joining TCFD necessitates internal capacity and preparation (e.g., due to data collection), not all of the banks joined directly in 2017. We follow the approach in Bingler et al. (2022) and investigate whether climate-related disclosures increased after the official launch of the TCFD recommendations in 2017 and after banks individually endorsed the TCFD recommendations. Overall, we find that the individual support of the TCFD recommendations goes along with an increase in climate-related reporting, which is statistically significant but economically modest. In terms of magnitude, we find a total average increase of 2.72% across all labels, which is in line with Bingler et al. (2022) who report an increase of approximately 2.2 percentage points. Examining the disclosures of the banks that became supporters after the official TCFD launch, the most notable differences are observed in the Metrics and Targets category, and pertain to carbon footprints as well as emissions reduction targets. Consequently, our results indicate that TCFD-supporting banks enhance their level of disclosures relating to carbon emissions following their official TCFD endorsement, which is consistent with the findings of Ding et al. (2022).

Altogether our findings are robust to various labels evaluation tests. We also examine whether our results are consistent with existing literature on the relationship between company size and CSR activities (e.g., Gillan et al. (2021)). In line with our expectations, we report that larger banks exhibit higher disclosure probabilities compared to medium or small banks.

The remainder of this paper is organized as follows. First, we present our data, followed by our methods and model performance evaluation. Our results section is twofold. In the first part, we present the results of the zero-shot classification at the category level. In the second part, we analyze the results for the fine-grained labels covering the TCFD recommended disclosures. The results are summarized and discussed in the last section.

Table 4.1: The TCFD disclosure categories and underlying recommended disclosures. Source: TCFD (2017b)

Broad disclosure categories:			
Governance	**Strategy**	**Risk Management**	**Metrics and Targets**
Disclose the organization's governance around climate-related risks and opportunities.	Disclose the actual and potential impacts of climate-related risks and opportunities on the organization's businesses, strategy, and financial planning where such information is material.	Disclose how the organization identifies, assesses, and manages climate-related risks.	Disclose the metrics and targets used to assess and manage relevant climate-related risks and opportunities where such information is material.
Underlying Recommended Disclosures:			
a) Describe the board's oversight of climate-related risks and opportunities.	a) Describe the climate-related risks and opportunities the organization has identified over the short, medium, and long term.	a) Describe the organization's processes for identifying and assessing climate-related risks.	a) Disclose the metrics used by the organization to assess climate-related risks and opportunities in line with its strategy and risk management process.
b) Describe management's role in assessing and managing climate-related risks and opportunities.	b) Describe the impact of climate-related risks and opportunities on the organization's businesses, strategy, and financial planning.	b) Describe the organization's processes for managing climate-related risks.	b) Disclose Scope 1, Scope 2, and, if appropriate, Scope 3 greenhouse gas (GHG) emissions, and the related risks.
	c) Describe the resilience of the organization's strategy, taking into consideration different climaterelated scenarios, including a 2°C or lower scenario.	c) Describe how processes for identifying, assessing, and managing climaterelated risks are integrated into the organization's overall risk management.	c) Describe the targets used by the organization to manage climate-related risks and opportunities and performance against targets.

4.2 Data

We apply the zero-shot classification to a sample of 3,335 hand-collected reports between 2010 and 2021. Due to the large differences between the homepages of the TCFD-supporting banks in our sample, a fully automated scraping approach is not possible. As a preliminary step, we extract the names of the TCFD-supporting banks from TCFD's website, based on the industry categories "banks", "central

banks" and "capital markets". After eliminating banks that could not be identified or lacked online annual reports, we are left with 188 TCFD-supporting banks. Table 4.2 presents the bank level data.

As a subsequent step, we proceed to categorize the banks in our sample based on two criteria: the region where their headquarters are located and their total asset size. In our analysis, we designate banks with total assets exceeding USD 500 billion as "large", those with total assets ranging between USD 50 billion and USD 500 billion as "medium", and banks with less than USD 50 billion as "small". Interestingly, nearly half of our sample comprises banks from the Asia-Pacific region. European banks constitute around one-third of the sample, while North American banks represent approximately 10%. In terms of asset size, the majority of banks fall into the mid-sized category, with total assets ranging between USD 50 billion and USD 500 billion.

Table 4.2: Size and region of TCFD-supporting banks

Region	Large	Medium	Small	$\sum$
Asia Pacific	15	51	24	90
Europe	23	26	17	66
Latin America	0	4	3	7
Middle East & Africa	0	3	2	5
North America	9	8	3	20
$\sum$	47	92	49	188

Next, we follow the approach in Bingler et al. (2022) and collect available bank reports for the period 2010 to 2021 to capture textual data both before and after the publication of the TCFD recommendations in June 2017. The reports are classified according to the following categories: annual reports, CDP reports, corporate governance reports, integrated reports, remuneration reports, sustainability reports, and TCFD reports. Our analysis extends beyond relying solely on TCFD reports since most TCFD supporters do not publish standalone reports specifically dedicated to climate-related disclosures but rather integrate key information into their annual and sustainability reports. This aligns with the TCFD recommendations, which indicate that climate-related disclosures should be included in "mainstream (i.e., public) annual financial filings" (TCFD, 2017b). The majority of reports in our

sample consists of annual and sustainability reports.

After parsing the reports to ensure they are in a suitable raw text format for the zero-shot classification, we are left with a total sample of 3,335 bank reports, as illustrated in Table 4.3. In comparison to prior TCFD-related studies, the zero-shot allows us to examine a comparatively large sample of reports. As a comparison, Ding et al. (2022) apply computerized textual analysis to a sample of 140 reports from TCFD signatories, while Demaria & Rigot (2020) examine the reference documents of a sample of 40 French firms between 2015 and 2018.

Table 4.3: Sample composition

Report Category	Number of reports	Average pages	Average number of sentences
Annual Report	1,869	207.98	2,711.21
CDP Report	75	63.43	699.79
Corporate Governance Report	148	69.44	1,014.25
Integrated Report	183	163.98	2,354.95
Remuneration Report	83	36.88	494.42
Sustainability Report	896	81.01	1,158.54
TCFD Report	81	36.68	544.37
	$\sum = 3,335$	$\bar{x} = 94.20$	$\bar{x}$= 1,282.50

4.3 Methodology

4.3.1 Parsing PDFs

The reports utilized in our study are in PDF format. Extracting and converting textual information from PDF documents for subsequent analysis using NLP techniques is not as straightforward as working with text stored in CSV or TXT files. To address this challenge, we employ a layout-parsing model designed to detect and extract the actual text from PDF documents. In particular, we include the actual text from the reports and deliberately omit text from tables and graphs, which not only increases the quality of our data, but also saves computation time.

Our parsing model is based on Visual-Layout (VILA) groups introduced by Shen et al. (2021). VILA converts textual data into groups of tokens (text lines or blocks) and assigns a layout tag to these tokens. There are several variants of VILA, such as

Table 4.4: Cleaning steps of parsed raw texts.

Problem	**Fix**
Extra whitespaces	Replace extra whitespaces by single whitespace
Words separated by hyphen	Remove hyphen
Words separated by whitespace	Remove whitespace
Model parsed some ff, fi and if characters as one special character	Replace special double characters by normal characters
Whitespaces between word in sentence and punctuation	Remove whitespace
URLs in text which do not have any semantic meaning	Remove all urls from text
Parsed some sentences character by character with whitespace between them, i.e., S E N T E N C E instead of Sentence. Problem mostly occurred with figure subtexts	As most figure subtexts do not have important semantic meaning, we removed such single characters

H-VILA (Visual Layout-guided Hierarchical Model) and I-VILA (Injecting Visual Layout Indicators). After conducting several trials, we select the H-VILA block variant trained on grotoap2 using the layoutLM model (Xu et al., 2020) since it delivers the best extraction and tokenization results. The output consists of the extracted text as groups of tokens together with the corresponding layout tags. Depending on the training set, the layout tags can be figures, body content, abstract and title. For our analysis, we keep the parts tagged as body content and abstract. In order to further improve our extraction results, we take further cleaning steps, which are summarized in Table 4.4.

4.3.2 Zero-shot text classification

A widely used and important NLP task is text classification (Belinkov & Glass, 2019). Text classification is used to organize and analyze very large amounts of textual data by assigning classes, or so-called "labels", based on the topic of individual text sequences, which can consist of sentences, paragraphs, or entire pages. In general, text classification is carried out using neural network models, which can be as simple as basic neural networks or more sophisticated language models equipped with classification end-layers. These models (e.g., BERT or BART) undergo pre-training

on extensive text data to acquire semantic understanding (Mehra et al., 2022). These pretrained models can subsequently be employed for various NLP tasks and fine-tuned for a specific task. The fine-tuning process involves combining different combinations of pre-trained language models (i.e., the base model) with task-specific end-layers.

Fine-tuning a base model can be accomplished by training it with labeled training data. As a result, the accuracy of the model often relies on the size and quality of the training data. However, creating a training set for sentence classification presents several drawbacks: First, the process of manually labeling large amounts of text is extremely time-consuming and requires significant human resources. Second, the labeling process must be repeated when classes need to be changed or new labels need to be added. Third, assigning the correct label to a sentence can be challenging even for humans, as certain sentences can be interpreted in different ways. Another drawback is the potential bias introduced by human labelers, making it difficult to obtain a representative training set (Beltagy et al., 2019). Lastly, finding suitable training data poses a challenge since the training data cannot be the same as the data used for actual analysis. For instance, in the analysis of TCFD reports, the labeled reports used for model training cannot be reused. Moreover, due to the limited number of TCFD reports published by banks and the imperfect nature of TCFD reporting by companies, there is a limitation in acquiring high-quality training data.

In this paper, we address the aforementioned drawbacks by employing a zero-shot text classification model introduced by Davison (2020). The model is able to classify text sequences based on the semantics of the input sequences and the labels without further requiring additional training. A simplified structure of our model architecture is shown in Figure 4.1. To perform the zero-shot classification, we employ BART as a base model. The model has been pre-trained on approximately 160GB of text from the English Wikipedia and BookCorpus dataset to develop an understanding of textual semantics (Mehra et al., 2022). Given that the TCFD recommendations do not specifically focus on financial language but rather general semantics, we consider BART to be well-suited for our analysis. Additionally, since the reports in our sample are in English, we can leverage the fact that the model was pre-trained on a large English language corpus. Compared to other models like BERT, the BART model (Lewis et al., 2019) utilizes a sequence-to-sequence translation architecture with bidirectional encoders (BERT) and a left-to-right autoregressive decoder (GPT model), combining the strengths of both. When used as a base model in conjunction with zero-shot text classification, BART demonstrates good performance results (Davison, 2020).

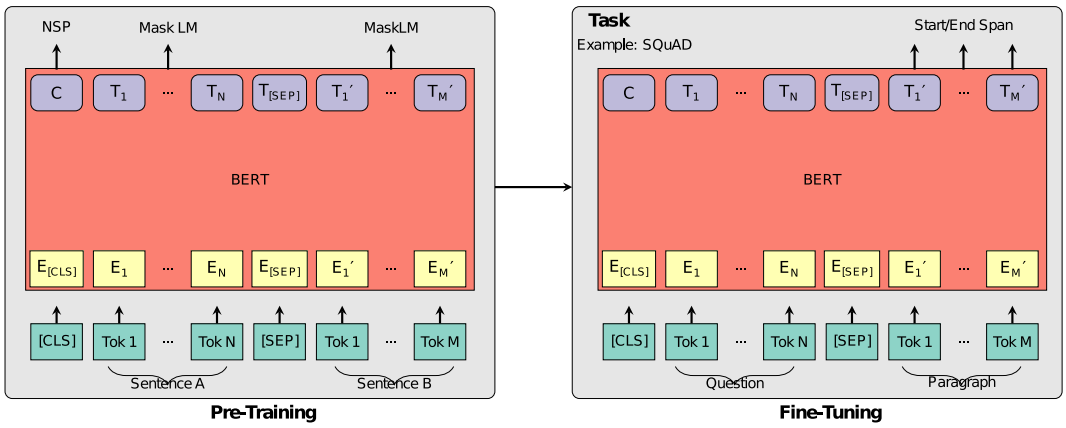


Figure 4.1: Model architecture overview.

On the left hand side, the BART model is pre-trained on all English Wikipedia articles and the BooksCorpus dataset. By masking parts of sentences (`[MASK]`), the model is trained to learn the semantics and to predict the missing parts. The process is repeated for all sentences in the pre-training dataset. On the right hand side, the model is fine-tuned on the MNLI task and returns probabilities for entailment, contradiction and neutral, as shown on the left hand side. Source: Own representation.

For our zero-shot classification, we employ a specific NLP task known as multi-natural language inference (MNLI) (Davison, 2020). This approach embeds both the sentences of a text (sequence of words) and the labels themselves into a shared latent space. In this latent space, the proximity between the sentence and the label can be computed, indicating the probability of a match. The closer the label is to the sentence, the higher the probability that the label matches the sentence. In the context of zero-shot text classification, labels are therefore used to assess the probability that a text sequence addresses one (or more) labeled topics. While the zero-shot text classification yields probabilities from 0 to 100% for each label, the ClimateBERT model used in Bingler et al. (2022) does not yield probabilities as output. Instead, it produces,but a binary output where one label is considered true and all others are false. For example, the classification output of one governance-related paragraph could be 1 for the label governance, and 0 for all other labels, i.e. 0 for strategy, 0 for risk Management etc. Compared to our approach, the authors measure the proportion of TCFD-related content in corporate reports by summing up the results for all labels and putting them in relation to the number of paragraphs (Bingler et al., 2022).

In contrast, our text classification model treats text sequences as premises and labels as hypotheses. It tokenizes both the sentence and the label, leveraging the underlying language to embed them. These embedded representations are then processed through the pre-trained MNLI layer. The MNLI end-layer consists of

a simple fully-connected neural network that produces a vector of logit scores for three possible outcomes: "neutral", "contradiction" and "entailment" (Davison, 2020; Lewis et al., 2019). Consequently, the hypothesis is evaluated against the premise, resulting in a classification of entailment, contradiction, or neutrality. The score for "neutral" is discarded and a softmax function is applied to the "contradiction" and "entailment" scores in order to be able to interpret them as a probability on a scale from 0 to 100%. In our analyses, the scores shown are for the "entailment" only. They can be interpreted as the probability that a text sequence matches a given label, or in other words, the probability that the given label is true. This fine-tuned end-layer can subsequently be used for zero-shot classification without any additional training.

Finally, due to the interconnected nature of the TCFD recommendations and the possibility for a sentence to align with multiple recommended disclosures simultaneously, we adopt a multi-label approach. Consequently, we do not constrain the zero-shot text classification to return probabilities that sum up to one, as in the single-label approach. Instead, we employ an approach where the model can assign probabilities ranging from 0 to 1 for each label (multi-label approach), accounting for the potential overlap between a sentence and multiple labels. As a result, the probabilities assigned to all labels for each sentence do not necessarily add up to one.

4.3.3 Fine-grained TCFD labels

The TCFD has structured its recommendations along four core disclosure categories: Governance, Strategy, Risk Management, Metrics and Targets (see Fig 1. in the Appendix). Each category comprises two to three recommended disclosures, accompanied by detailed descriptions specifying the information to be included. Recognizing the varying materiality of information across industries, the TCFD has also developed additional guidance tailored to the financial sector, including banks, insurance companies, asset managers, and asset owners. In its guidance for banks, the TCFD emphasizes that climate-related disclosures from credit institutions should facilitate the identification of large concentrations of carbon-related assets and provide a better understanding of the financial system's exposure to climate-related risks (TCFD, 2017b).

We proceed by creating a set of general labels covering each main category: **GO.1** *Climate-related Governance*, **ST.1** *Climate-related Strategy*, **RM.1** *Climate-related Risk Management*, **MT.1** *Climate-relate Metrics and Targets.* The inclusion of "climate-related" ensures that the zero-shot classification primarily captures sentences

addressing climate-related topics. For a comprehensive overview of our fine-grained labels, please refer to Table 4.5. Next, we develop a set of targeted labels based on the recommended disclosures, their description, and the additional guidance provided for the financial sector. Under the governance pillar, we summarize the recommended disclosures, "describe the board's oversight of climate-related risks and opportunities" and "describe management's role in assessing and managing climate-related risks and opportunities" into two labels: **GO.1.1** *Board's responsibility for overseeing climate-related issues* and **GO.1.2** *Executive management's role related to the assessment and management of climate-related issues*. We include the terms "executive" and "strategic role" to highlight that these recommendations refer to the assignment of strategic responsibilities at executive management level. In addition, the TCFD often uses the expression "climate-related issues" to refer to climate-related risks and opportunities. Hence, we incorporate this expression in our labels.

In the strategy area, we employ the labels **ST.1.1** *Climate-related transition risks such as policy, legal, technology, market and reputation risks emerging from climate change* and **ST.1.2** *Climate-related physical risks such as acute weather events and chronic shifts in weather patterns* to capture the recommended disclosure that states to "describe the climate-related risks and opportunities the organization has identified over the short, medium, and long-term". Since the TCFD specifically recommends discussing examples of transition and physical risks, we distinguish between the two types of risks in our labels. Next, we turn to the second recommended disclosure in the strategy area, which is "describe the impact of climate-related risks and opportunities on the organization's businesses, strategy, and financial planning". Given that issues related to "businesses, strategy, and financial planning" are particularly difficult to summarize into one label, we rather focus on the first part related to the financial impact. We aim to capture whether financial institutions perceive climate change as having a significant financial impact on their operations. Based on this, we develop the label **ST.1.3** *Material financial impact of climate-related issues*. In addition, the TCFD encourages the description of climate-related scenarios if such scenarios are used (TCFD, 2017b). Therefore, we create the label **ST.1.6** *Use of climate-related scenario models to analyze the impact of climate-related risks* to assess whether the banks in our sample conduct such analyses and disclose related information.

Additionally, the TCFD recommends banks to describe significant concentrations of credit exposure to carbon-related assets. This recommendation overlaps with a similar recommended disclosure in the metrics and targets category, which requires banks to "provide the amount and percentage of carbon-related assets relative to

total assets as well as the amount of lending and other financing connected with climate-related opportunities". Considering these recommendations, we create the labels **ST.1.4** *Credit exposure to carbon-related assets* and **ST.1.5** *Financing and investment for carbon-intensive industries such as fossil fuel industry.* Finally, we turn to the last recommended disclosure under the strategy category, "describe the resilience of the organization's strategy, taking into consideration different climate-related scenarios, including a 2°C or lower scenario" and add the label **ST.1.7** *Resilience of the bank's strategy under different climate-related scenarios.*

Within the risk management pillar, the TCFD advises to "describe the organization's processes for identifying and assessing climate-related risks", to "describe processes for managing climate-related risks", and to "describe how processes for identifying, assessing, and managing climate-related risks are integrated into the organization's overall risk management". Using the label **RM.1.1** *Processes to identify, assess and manage climate-related risks and integrate them into overall risk management*, we combine the above recommended disclosures into one label. With respect to the additional guidance for banks, the TCFD advises to consider characterizing climate-related risks in the context of traditional banking industry risk categories such as credit risk, market risk, liquidity risks and operational risks. We therefore add the label **RM.1.2** *Relationship between climate-related risks and financial risks such as credit risk, market risk, liquidity risk and operational risk.*

Under the metrics and targets pillar, the TCFD recommends to "disclose the metrics used by the organization to assess climate-related risks and opportunities in line with its strategy and risk management process". In this context, organizations are encouraged to provide key metrics used to measure and manage climate-related risks, and banks are particularly advised to disclose metrics employed to assess the impact of transitional and physical climate-related risks on their lending and other financial activities (TCFD, 2017b). Additionally, banks are required to disclose "the amount and percentage of carbon-related assets relative to total assets as well as the amount of lending and other financing connected with climate-related opportunities". These recommendations are closely related to the recommended disclosures under the strategy and risk management pillars, as they also involve the disclosure of metrics related to climate-related transition and physical risks and their impacts (e.g., labels **ST.1.1**, **ST.1.2**, and **ST.1.3**). Thus, we focus on a specific sub-element within the recommended disclosure, i.e. "where climate-related issues are material, organizations should consider describing whether and how related performance metrics are incorporated into remuneration policies". We label this **MT.1.2** *Incorporation*

of climate-related performance metrics into remuneration policies. This label aims to assess whether banks report on the inclusion of climate-related metrics in their compensation practices, reflecting the growing importance of sustainability-related performance measures in executive remuneration (TCFD, 2017b). Next, in response to the recommended disclosure "disclose Scope 1, Scope 2, and, if appropriate, Scope 3 greenhouse gas (GHG) emissions, and the related risks", we create the label **MT.1.1** *Carbon footprint, direct and indirect GHG.* Lastly, the TCFD recommends to "describe the targets used to manage climate-related risks and opportunities and performance against targets". Since the TCFD encourages the disclosure of GHG emissions targets, as it also places great emphasis on gaining a better understanding of the concentrations of carbon-related assets in the financial sector, we focus on this particular type of target and create the label **MT.1.3** *Emissions reduction and carbon neutrality targets.*

Table 4.5: Overview of TCFD labels

Governance	GO.1.	Climate-related Governance
	GO.1.1	Board's responsibility for overseeing climate-related issues
	GO.1.2	Executive management's strategic role related to the assessment and management of climate-related issues
Strategy	ST.1.	Climate-related Strategy
	ST.1.1	Climate-related transition risks such as policy, legal, technology, market and reputation risks emerging from climate change
	ST.1.2	Climate-related physical risks such as acute weather events and chronic shifts in weather patterns
	ST.1.3	Material financial impact of climate-related issues
	ST.1.4	Credit exposure to carbon-related assets
	ST.1.5	Financing and investment for carbon-intensive industries such as fossil fuel industry
	ST.1.6	Use of climate-related scenario models to analyse the impact of climate-related risks
	ST.1.7	Resilience of the bank's strategy under different climate-related scenarios
Risk Management	RM.1.	Climate-related Risk Management
	RM.1.1	Processes to identify, assess and manage climate-related risks and integrate them into overall risk management
	RM.1.2	Relationship between climate-related risks and financial risks such as credit risk, market risk, liquidity risk and operational risk
Metrics & Targets	MT.1.	Climate-related metrics and targets
	MT.1.1	Carbon footprint, direct and indirect greenhouse gas emissions
	MT.1.2	Incorporation of climate-related performance metrics into remuneration policies
	MT.1.3	Emissions reduction and carbon neutrality targets

4.4 Label evaluation

The zero-shot text classification model used in our study does not require specific training with pre-labeled data. As a result, the model cannot be validated in the conventional manner of splitting a dataset into training and test sets and evaluating the model's performance on the test set. Nonetheless, we can still conduct a series of evaluation tests to assess the text sequence recognition and classification performance of our zero-shot text classification model.

To create a dataset for evaluation, we manually collect sentences that align with the TCFD recommendations and assign them specific labels. These sentences are extracted from the TCFD good practice handbook (CDSB, 2019, 2021), which features examples of best practice disclosures selected by the TCFD. Additionally, we manually extract sentences from TCFD reports of companies outside the banking sector. We also incorporate sentences from the training repository provided by Webersinke et al. (2021). All of these sentences have been assigned a label by the authors, which is either "governance", "strategy", "risk management", "metrics and targets" or "none". We reassign one of our fine-grained labels to these text sequences.

In the first step, we apply the zero-shot classification to the dataset using a multi-label approach. This approach allows the zero-shot model to provide results for each label independently, acknowledging that a text sentence may correspond to multiple labels. Consequently, the results returned by the zero-shot classification for each label do not sum up to one. We utilize the fine-grained labels listed in Table 4.5. We also include the label "none" in the classification task. The purpose of this label is to capture non-climate-related text sequences, i.e., text sequences that do not fit any of our fine-grained labels. Including the "none" label ensures that the labels assigned by the zero-shot are based on the semantics of the text sequences rather than by pure chance. To assess the model's performance with respect to the "none" label, we use sentences labeled as "none" in the aforementioned training repository, as well as additional sentences labeled as "none" by us. The results of the zero-shot text classification applied to the dataset are shown in Figure 4.2. The x-axis represents the manually annotated sentences, while the y-axis shows the results from our zero-shot analysis based on the climate-related labels. The results represent the mean probability returned by the zero-shot model that the sentences in a given column deal with the topic represented by the label in the corresponding row. Darker entries represent higher likelihoods classified by the model. In an optimal zero-shot text classification, the entries along the diagonal would be the darkest.

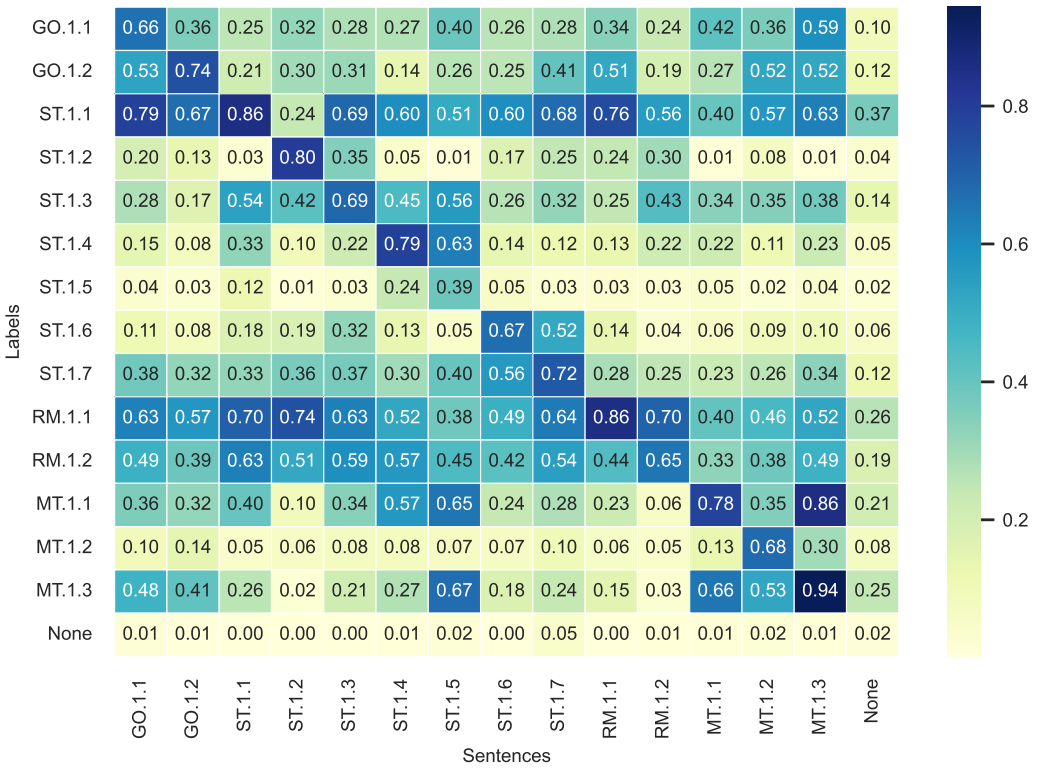


Figure 4.2: Label evaluation matrix based on test dataset.
This matrix presents the results of the zero-shot text classification applied to our test dataset. The fine-grained TCFD labels are based on the recommended disclosures and the supplemental guidance for the financial sector.

These results demonstrate that our model accurately assigns high probabilities to the appropriate labels. At the same time, we also observe that labels ST.1.1 (Climate-related transition risks), ST.1.3 (Material financial impact of climate-related issues), ST.1.7 (Resilience of the bank's strategy), RM.1.1 (Processes to identify, assess and manage climate-related risks and integrate them into overall risk management), and RM 1.2 (Relationship between climate-related risks and financial risks) perform worse than other labels in terms of text sequence recognition. For example, both the groups of sentences that we labeled as risk management sequences as well as other sentence groups were assigned the labels RM.1.1 and RM1.2 with a probability of 60% or higher.

There could be several reasons for these results. First, the abstract nature of some topics covered may render them less suitable for zero-shot text classification. Terms such as "resilience" encompass a wide range of interpretations, posing challenges in precise reporting and potentially causing overlaps with multiple text sequences. This observation also brings attention to weaknesses in the design of the TCFD recommendations. Second, the TCFD recommendations often encompass closely

interconnected themes, leading to situations where text sequences can align with multiple labels simultaneously. For example, the only label which does not have the highest probability for its own group of sentences is label MT.1.1 (Carbon footprint, direct and indirect greenhouse gas emissions), where a higher probability was assigned by the zero-shot to MT.1.3 (emissions reduction and carbon neutrality targets). Given the close relationship between these topics, it is unsurprising to observe high probabilities assigned to both sets of sentences. However, in order to maintain consistency with the semantics of the TCFD recommendations, we made a deliberate choice not to modify the labels.

The probabilities for the "none" label are consistently low across all sentences, indicating that the model does not simply label by chance, but rather incorporates the semantics of the labels in its classification. Sentences labeled as "none" received low probabilities for nearly all labels. The exception is label "climate-related transition risks" (ST.1.1), which has a slightly higher probability of 37% for the "none" sentences. This may be linked to the fact that this label encompasses many different types of risks, such as political and legal risks, technological risks, market risks, and reputation risks, all of which belong to the "transition risks" category. For some sentences describing these risks, the zero-shot classification may not directly identify the link to climate change. Furthermore, we also notice that more abstract labels such as RM.1.1 and RM.1.2 have higher values for "none" sentences as well.

In addition to our graphical analysis, we evaluate the model performance by examining the overall F1 scores and the individual F1 scores of our labels, as illustrated in Table 4.6. We evaluate the model based on test data previously used in Figure 4.2, focusing on our fine-grained labels. In contrast to the previous matrix, we calculate the F1-scores using a single-label approach, as we also performed a single-label approach by manually attributing a specific label to the text sequences in our dataset. Overall, our model obtains a micro F1 score of 0.60 and a macro F1 score of 0.57, which is satisfactory considering the presence of 14 classes. In addition, we observe that material financial impact of financial issues (ST.1.3) is the most challenging to identify (F1-score of 0.34) and the incorporation of climate-related performance metrics into remuneration policies (MT.1.2) the easiest (F1-score of 0.78). This discrepancy may be explained by the fact that financial material impact is a broader concept, and the task of identifying corresponding sentences may be more challenging, even for humans. We also observe that our governance labels exhibit a comparatively high precision. In contrast, the labels pertaining to transition risks (ST.1.1), the relationship between climate-related risks and financial risks (RM.1.2)

and emissions-reduction targets (MT.1.3) display relatively lower precision, suggesting a higher number of false positives. Additionally, our labels exhibit relatively high recalls, with the exception of material financial impact of financial issues (ST.1.3.) and financing and investment for carbon-intensive industries (ST.1.5).

Altogether, our zero-shot text classification does not appear to assign probabilities purely by chance. The TCFD recommendations appear to be intertwined, which is an argument for the multi-label approach we use. We also find that zero-shot classification yields better results when labels are based on well-delineated and precisely defined concepts.

Table 4.6: Comparison of performance based on F1 scores

Label	GO.1.1	GO.1.2	ST.1.1	ST.1.2	ST.1.3	ST.1.4	ST.1.5	ST.1.6	ST.1.7	RM.1.1	RM.1.2	MT.1.1	MT.1.2	MT.1.3
Recall	0.53	0.74	0.66	0.61	0.28	0.52	0.31	0.57	0.51	0.73	0.59	0.58	0.68	0.93
Precision	0.97	0.72	0.29	0.79	0.46	0.54	1.00	0.77	0.50	0.61	0.29	0.90	0.91	0.24
F1-Score	0.69	0.73	0.40	0.69	0.34	0.53	0.48	0.65	0.51	0.66	0.39	0.70	0.78	0.39

The overall performance scores are: Micro F1-score: 0.6029, Macro F1-score: 0.5668, Weighted F1-score: 0.6281.

4.5 Results

4.5.1 Climate-related disclosures by broad TCFD categories

e main objective of the TCFD recommendations is to guide companies in disclosing consistent and decision-useful information for key stakeholders (Bingler et al., 2022). These climate-related disclosures aim to reduce information asymmetry between reporting firms and stakeholders and demonstrate companies' awareness of climate-related issues (Krueger et al., 2020; Jung et al., 2018; Ding et al., 2022). In a first step, we examine the probability that corporate reporting addresses climate-related issues and relates to one of the four main TCFD pillars. Table 4.7 presents the probabilities associated with the general labels "Governance", "Strategy", "Risk Management", and "Metrics and Targets" as well as the probabilities for our labels "climate-related Governance" (GO.1), "climate-related Strategy" (ST.1), "climate-related Risk Management" (RM.1) and "climate-related Metrics and Targets" (MT.1), respectively. We intentionally include both types of labels to facilitate a comparison between disclosures on general topics and disclosures specifically related to climate-related matters.

Table 4.7: Mean of label probabilities at category level per financial year

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Governance	0.31	0.31	0.30	0.30	0.31	0.31	0.31	0.31	0.31	0.31	0.32	0.32
GO.1	0.11	0.11	0.12	0.11	0.12	0.12	0.12	0.13	0.14	0.15	0.17	0.19
Strategy	0.40	0.40	0.40	0.40	0.39	0.40	0.39	0.40	0.39	0.40	0.40	0.40
ST.1	0.12	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.15	0.17	0.20	0.22
Risk management	0.23	0.23	0.24	0.23	0.24	0.23	0.23	0.23	0.23	0.23	0.24	0.24
RM.1	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.10	0.11	0.12	0.15	0.16
Metrics & Targets	0.31	0.31	0.31	0.31	0.31	0.32	0.31	0.32	0.32	0.32	0.33	0.34
MT.1	0.12	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.14	0.16	0.18	0.20

The table presents the mean of label probabilities (on a scale from 0 to 1) for the general and climate-related labels at category level based on the full sample of 3,355 reports.

Several observations can be made based on these results. First, the mean probabilities for the general labels, which do not specifically mention climate-related topics, are higher compared to the probabilities for the specific climate-related labels. For instance, the probability of reporting on "Governance" is consistently higher than

the probability of reporting on "climate-related Governance" (GO.1) throughout the entire sample period. The same result holds true for the other main categories. This result is reasonable considering that our text sequences are extracted from various reports, including corporate governance reports and annual reports that cover a wide range of governance-related topics, not solely focused on climate-related governance. Thus, the zero-shot text classification model appears to effectively distinguish between climate-related and non-climate-related textual data.

Second, we observe that the mean probabilities associated with the general labels (without explicit mention of climate) remain stable over the sample period from 2010 to 2021, while there is an increase in all probabilities for the climate-related labels after 2017. Among the general labels, "Strategy" exhibits the highest mean probability compared to the others, maintaining a consistent probability of around 40% over time. In contrast, the probability of "climate-related Strategy" shows a dynamic increase, growing from 12% in 2010 to 22% in 2021. This indicates that the probability of text sequences in our sample relating to "climate-related Strategy" was only 12% in reports from 2010 but reaches 22% for reports published in 2021. In addition, we find that the label "Metrics and Targets" has the second highest mean probability over the sample period compared to the other labels with a mean probability between 31% and 34%. When examining the mean probability of "climate-related Metrics and Targets", we find an increase from 12% in 2010 to 20% in 2021, surpassing the probabilities of "climate-related Governance" and "climate-related Risk Management" in 2021.

To further examine climate-related disclosures at the category level, we examine the trends in reporting before and after the publication of the TCFD recommendations in 2017. Fig 4.3 provides a visual representation of these trends for all four TCFD categories over the sample period. Specifically, the blue lines represent the label probabilities of the general labels, while the orange lines illustrate the probabilities of the climate-related category labels, denoted as GO.1, ST.1, RM.1 and MT.1 in Table 4.5. We observe an overall increase in all climate-related label probabilities, with a more pronounced change occurring around 2017.

4.5.2 Climate-related disclosures by fine-grained TCFD labels

To more accurately assess the quality of climate-related disclosures, it is important to go beyond the quantity of reporting for each broad TCFD category and instead focus on examining corporate reporting on the specific recommended disclosures within each category. Therefore, as a next step, we consider the fine-grained labels

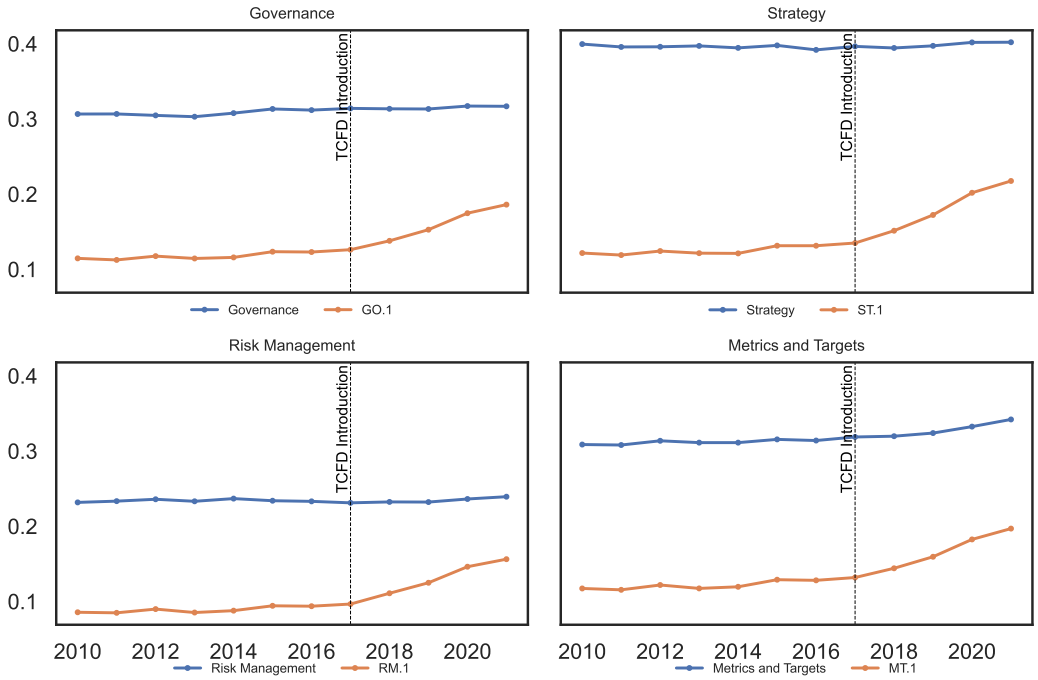


Figure 4.3: Climate-related disclosures by broad TCFD categories.

that address specific topics related to the TCFD recommendations, rather than the climate-related category labels (GO.1, ST.1, RM.1, and MT.1). As highlighted earlier, the TCFD provides additional guidance for the financial sector, including banks, insurances, asset managers and asset owners (TCFD, 2017b). The additional guidFance for banks particularly emphasizes disclosures related to strategy, risk management, and metrics and targets. We consider a higher label probability to serve as a proxy for disclosure quality. When a text sequence explicitly and accurately addresses a topic expressed in a label, it is associated with a higher label probability. A higher label probability also suggests a more comprehensive disclosure on a particular topic, as labeling is more likely to have a higher probability if a text sequence provides detailed information about the topic.

Figure 4.4 displays the results of the zero-shot text classification for the fine-grained labels applied to the entire sample of reports. As can be seen, there is considerable variation in the extent of disclosure within each TCFD category. The strategy category is the most comprehensive, encompassing several specific recommended disclosures for banks, which explains the presence of a greater number of labels compared to the other pillars. The reporting quality appears to be lower for disclosures related to financing and investment in carbon-intensive industries such as the fossil fuel industry (ST.1.5), as indicated by a probability of only 7%. This suggests that among all the text sequences extracted from our full sample of

reports and classified by the zero-shot model, there is only a 7% probability of some of them matching the semantics of the ST.1.5 label. Similarly, TCFD-supporting banks seem to provide relatively limited disclosure on climate-related physical risks (ST.1.2) and the use of climate-related scenario models (ST.1.6), with both labels attaining probabilities of only 17% and 14%, respectively.

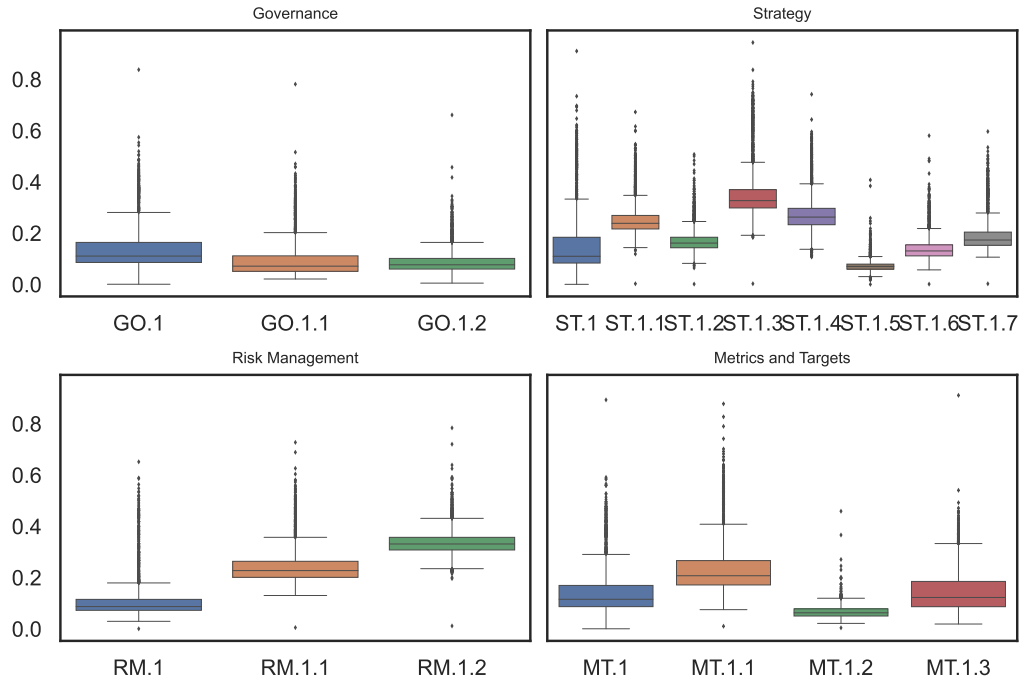


Figure 4.4: Climate-related disclosures by fine-grained TCFD labels.

There could be several interpretations for these results. First, it is possible that the reports in our sample only address to a limited extent issues linked to the use of scenario analyses or climate-related physical risks. This may be because some of the TCFD-supporting banks are still in the early stages of developing the necessary tools and expertise to conduct such analyses or identify such risks. This finding aligns with the study conducted by Bingler et al. (2022), which highlights that only 10 out of 16 existing climate scenario tools allow for the assessment of climate-related physical risks. Similarly, Friederich et al. (2021) find that disclosures on transition risks have seen more significant growth compared to disclosures on physical risks.

The low label probability for disclosures related to the fossil fuel industry (ST.1.5), indicates that banks tend to provide limited information on this topic. Nevertheless, recent research highlights that financing for fossil fuel firms by international banks has not decreased since the Paris Agreement, and these banks continue to provide funding regardless of the associated stranded asset risk (Beyene et al., 2021). In fact, Beyene et al. (2021) specifically identified several TCFD-member banks, including JP

Morgan (TCFD-member since December 2017), BNP Paribas SA (TCFD-member since June 2017) and Wells Fargo & Co (TCFD-member since November 2019), as among the top lenders to fossil fuel firms between 2007 and 2018. This suggests that TCFD-supporting banks may engage in selective disclosure, as suggested by Bingler et al. (2022), potentially omitting certain information due to both reputational concerns and the absence of specific guidelines for measuring such exposures (Beyene et al., 2022). On the other hand, it is worth noting that banks appear to disclose more information regarding their credit exposure to carbon-related sectors. This is not surprising since this category encompasses a broader definition that includes sectors like transportation and utilities.

In the risk management area, we observe that banks tend to address processes for identifying, assessing, and managing climate-related risks and integrating them into overall risk management (RM.1.1) less frequently, on average, compared to the relationship between climate-related risks and financial risks (RM.1.2). The median value for RM.1.1 is also lower (23%) than in the case of RM.1.2 (33%). However, the probabilities provided by the zero-shot text classification could be slightly inflated compared to the actual reporting since the zero-shot performed less well for RM.1.2 (F1-score: 0.39). As shown previously, several sentence groups achieved a high probability of fitting into RM.1.1 and RM.1.2. Notably, several recommended disclosures within other pillars, such as the role of management in assessing and managing climate-related issues (GO.1.2), are also related to risk management topics.

In the metrics and targets category, we find that the incorporation of climate-related performance metrics into remuneration policies (MT.1.2) is associated with a lower label probability compared to metrics related to carbon footprints (MT.1.1) and emissions reduction targets (MT.1.3). The mean label probability for MT.1.2 is only 7%, while it is 24% for MT.1.1 and 15% for MT.1.3. This result aligns with our expectations, as financial institutions are less likely to align their compensation policies with climate-related performance metrics compared to making symbolic commitments to carbon neutrality goals, even if they may fail to meet these commitments (see e.g., (Gibson et al., 2022; Bolton & Kacperczyk, 2021)).

In the governance area, the TCFD-supporting banks appear to report at comparable levels on the board's responsibility for overseeing climate-related issues (GO.1.1) and the management's role in assessing and managing climate-related issues (GO.1.2). Both labels exhibit a relatively low average probability of 9%, with median values of approximately 7% for GO.1.1 and 8% for GO.1.2. However, the maximum values

for GO.1.1 are higher, indicating more comprehensive disclosures on the role of the board in overseeing climate-related issues.

In order to better assess the evolution of reporting on the underlying recommended disclosures, we report in Table 4.8 the mean label probabilities of the fine-grained labels in each financial year. We observe that an increase in mean label probabilities can be observed for most labels between 2017 and 2018, and in particular between 2018 and 2019.

Table 4.8: Mean of label probabilities for fine grained labels per financial year

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
GO.1.1	0,07	0,07	0,08	0,07	0,07	0,08	0,08	0,08	0,09	0,11	0,13	0,14
GO.1.2	0,08	0,07	0,08	0,07	0,08	0,08	0,08	0,08	0,09	0,10	0,11	0,12
ST.1.1	0,24	0,24	0,24	0,23	0,24	0,24	0,24	0,24	0,25	0,26	0,28	0,28
ST.1.2	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,18	0,19	0,19
ST.1.3	0,34	0,34	0,34	0,33	0,33	0,34	0,33	0,33	0,34	0,36	0,40	0,40
ST.1.4	0,25	0,25	0,26	0,25	0,26	0,26	0,27	0,26	0,27	0,28	0,31	0,32
ST.1.5	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,08	0,08
ST.1.6	0,13	0,13	0,13	0,12	0,13	0,13	0,13	0,13	0,14	0,15	0,16	0,16
ST.1.7	0,18	0,17	0,18	0,17	0,17	0,18	0,18	0,18	0,19	0,20	0,22	0,22
RM.1.1	0,22	0,22	0,23	0,22	0,23	0,23	0,23	0,23	0,24	0,25	0,27	0,28
RM.1.2	0,34	0,34	0,34	0,33	0,33	0,33	0,33	0,32	0,33	0,34	0,35	0,35
MT.1.1	0,21	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,23	0,25	0,28	0,29
MT.1.2	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,07	0,07	0,07	0,07	0,07
MT.1.3	0,12	0,12	0,13	0,12	0,13	0,13	0,14	0,14	0,15	0,16	0,18	0,19

The table presents the descriptive statistics of label probabilities (on a scale from 0 to 1) for the fine-grained TCFD labels based on the full sample of 3,355 reports.

4.5.3 Climate-related disclosures after individual TCFD support

In the following step, we examine whether the observed increase in climate-related disclosures following the introduction of the TCFD recommendations is statistically significant for both the overall fine-grained labels and after individual banks declared their support. As the process of joining the official TCFD supporters was gradual, with banks joining at different times, we can investigate whether climate-related

disclosures truly increased after banks individually began supporting the TCFD recommendations. Our approach builds upon the study of Bingler et al. (2022), but extends the analysis by focusing on the specific recommended disclosure topics rather than the broad category level.

Table 4.9 presents the results of a paired t-test (in percentage points), in which we compare the mean differences of label probabilities before and after the official TCFD introduction in 2017 (column 1), as well as before and after the year of individual TCFD support (columns 2 to 6). As a robustness check, we also performed permutation p-value tests using a Monte Carlo simulation and bootstrap confidence intervals, which yielded qualitatively similar results. As an illustration, if banks became TCFD supporters in 2018, we compare the mean of each label probability for all of these banks (26 in total) by taking the mean per bank from 2010 to 2017 and comparing it to the mean from 2018 to 2021 after the banks became supporters. For the full sample, in column 1, we compare the mean difference up to the publication of the official TCFD recommendations (mean of years 2010 to 2016) and after they were published (mean of years 2017 to 2021).

In column 1, we observe a small but statistically significant increase in climate-related reporting probabilities for the full sample following the official introduction of the TCFD in 2017. The mean differences are statistically significant at the 1% level, except for RM 1.2, which is significant at the 10% level. On average, we find a total increase of 2.72% across all labels, which aligns with the findings of Bingler et al. (2022), who report an increase of approximately 2.2 percentage points. In columns 2 and 3, we report that the groups of banks that became supporters in 2017 and 2018 exhibit higher disclosure levels after their individual support, as illustrated by the relatively larger and statistically significant mean differences. For the group of banks who became official supporters in 2019, as shown in column 4, we do not find any significant change in mean label probabilities for two labels. This result is consistent with Bingler et al. (2022) who report the largest nominal effects for companies that supported the TCFD recommendations directly in 2017 and 2018 as compared to companies that supported the TCFD in 2019 and 2020.

In terms of magnitude, the greatest differences in reporting probabilities for banks that joined the TCFD in 2017 are found in the fine-grained labels MT.1.1 (carbon footprint, 6.43%), MT.1.3 (emissions reduction and carbon neutrality targets, 5.34%), and GO.1.1 (board oversight of climate-related issues, 5.02%). For the banks that joined in 2018, we find the largest mean differences for the labels MT.1.1 (carbon footprint, 5.44%), ST.1.3 (financial impact of climate-related issues, 5.36%) and

GO.1.1 (board oversight of climate-related issues 4.43%). Similarly, for banks that joined in 2019, 2020, and 2021, the largest differences are found in the labels MT.1.1 (carbon footprint), MT.1.3 (emissions reduction and carbon neutrality targets), and ST.1.3 (financial impact of climate-related issues).

Altogether, the largest differences tend to be oftentimes observed in the Metrics and Targets category and pertain to carbon footprints as well as emissions reduction targets. Thus, TCFD-supporting banks appear to increase their level of disclosures related to these topics in the course of their official TCFD endorsement. One possible reason for this observation could be that the importance that stakeholders, including investors, place on carbon risk (Bolton & Kacperczyk, 2021). This is also consistent with Ding et al. (2022) who show that there is a positive relationship between carbon emissions and climate-related disclosures.

Table 4.9: Mean differences in percentage points of climate-related disclosures

		TCFD support since				
	Full Sample	2017	2018	2019	2020	2021
	n = 188	n = 38	n = 26	n = 25	n = 30	n = 53
	(1)	(2)	(3)	(4)	(5)	(6)
GO.1	3.59***	5.69***	4.72***	4.25***	4.51***	3.24***
GO.1.1	3.29***	5.02***	4.43***	3.74***	4.34***	3.12***
GO.1.2	2.21***	3.34***	2.88***	2.61***	2.98***	2.25***
ST.1	4.80***	7.70***	6.51***	5.61***	6.16***	4.23***
ST.1.1	2.41***	4.15***	3.41***	2.59***	3.22***	2.54***
ST.1.2	1.47***	2.57***	1.89***	1.80***	2.10***	1.75***
ST.1.3	2.74***	4.84***	5.36***	4.05***	4.52***	3.28***
ST.1.4	2.45***	4.60***	3.98***	2.53***	3.25***	3.10***
ST.1.5	0.50***	1.19***	0.96***	0.28	0.65***	0.71**
ST.1.6	1.59***	3.24***	2.19***	1.77***	1.75***	1.93***
ST.1.7	2.44***	4.13***	3.54***	3.02***	3.24***	2.45***
RM.1	3.39***	5.76***	5.02***	3.42***	4.25***	3.09***
RM.1.1	2.82***	4.72***	3.84***	3.11***	3.56***	2.74***
RM.1.2	0.33*	1.12**	2.20***	0.20	1.40**	1.29**
MT.1	3.90***	6.22***	5.03***	4.57***	4.78***	3.62***
MT.1.1	4.06***	6.43***	5.44***	5.28***	5.46***	3.75***
MT.1.2	0.63***	1.06***	0.37*	1.03***	0.55**	0.86**
MT.1.3	3.74***	5.34***	4.19***	4.94***	4.56***	3.38***

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.
This table presents the mean differences (in percentage points) of climate-related disclosures and the significance of the corresponding paired t-Test based on the full sample of 3,355 reports.

4.6 Discussion and Conclusion

This paper examines the climate-related disclosures of TCFD-supporting banks using the zero-shot text classification as a novel computerized approach for textual analysis of climate-related disclosures. By combining the TCFD recommendations with additional guidance specific to the financial sector, we create fine-grained labels that enable a more detailed examination of climate-related reporting. Our findings reveal significant variation in climate-related disclosures, not only across the broad TCFD categories but also within each category. Specifically, we observe that banks have a lower probability of reporting on topics such as climate-related physical risks, financing and investments in fossil fuel activities, the use of climate-related scenario models, and the integration of climate-related performance metrics into remuneration policies. These results indicate that the TCFD-supporting banks in our sample have not yet implemented all the recommendations to the same extent. Our research contributes to the expanding body of literature on voluntary climate-related corporate reporting (e.g., (Friederich et al., 2021; Bingler et al., 2022; Ding et al., 2022)).

Our study also entails some limitations, which warrant careful consideration and indicate potential areas for future research. First, although we observe an overall increase in climate-related reporting following the release of the TCFD recommendations, it is important to note that this does not necessarily imply that banks are taking more substantial internal actions to address the identified issues. Simply disclosing more information does not guarantee a corresponding increase in efforts to address climate-related challenges. This highlights the need for further research regarding the factors influencing reporting decisions. Additionally, there is a possibility that some banks intentionally do not disclose certain information and engage in selective disclosure (i.e., greenwashing) to improve their public image. This emphasizes the need to consider potential motivations and biases behind the disclosed information in the context of voluntary disclosures. Addressing these limitations and exploring these areas in future research can contribute to a deeper understanding of the relationship between climate-related reporting, internal actions, and the effectiveness of voluntary disclosure frameworks such as the TCFD. Furthermore, our research also reveals weaknesses in the TCFD recommendations, such as a lack of precise concepts and overlaps in recommended disclosures. Future research could therefore further investigate the determinants of high-quality climate-related disclosure frameworks and assess their impact in delivering material and decision-useful information.

Despite the limitations mentioned, our study has important practical implications. It underscores the necessity for precise and specific recommendations within climate-related disclosure frameworks. Without such consistent methodologies and explicit definitions for recommended disclosure topics, there is a potential for significant variation in the scope and depth of reporting. In the banking sector, while the TCFD recommendations are a positive step forward, the lack of concrete guidance hinders accurate assessment of banks' exposure to the fossil fuel sector and potential stranded assets. Addressing these issues can enhance the effectiveness and reliability of climate-related reporting frameworks, facilitating informed decision-making by stakeholders.

C Appendix to Chapter 4

Data: Dataset underlying the results described in this manuscript. Available at `https://doi.org/10.1371/journal.pone.0288052.s001`

Chapter 5

Exploring the Predictive Capacity of ESG Sentiment on Official Ratings: A Few-Shot Learning Perspective

The following chapter is based on the paper:	
Title:	Exploring the Predictive Capacity of ESG Sentiment on Official Ratings: A Few-Shot Learning Perspective
Authors:	Elena Tönjes (contribution: 80%), Christoph Funk (contribution: 10%), Christian Haas (contribution: 10%)
Status:	Working Paper

Exploring the Predictive Capacity of ESG Sentiment on Official Ratings: A Few-Shot Learning Perspective

Christoph Funk[*,†] Elena Tönjes[‡] Christian Haas[‡,§]

Abstract

Environmental, social, and governance (ESG) criteria are increasingly central to corporate reporting. This study applies natural language processing (NLP) techniques, specifically a RoBERTa-based few-shot model, to conduct aspect-based sentiment analysis (ABSA). Our analysis targets ESG-related entities and their sentiments within EUROSTOXX 50 company reports to assess their impact on ESG ratings. The ratings data are sourced from established providers, including Refinitiv, S&P, and Bloomberg. Furthermore, to explore the potential reciprocal influences on these variables, we employ a vector auto-regressive (VAR) model, which facilitates the modeling of bidirectional interactions. This combination of advanced NLP methods and comprehensive data integration aims to provide detailed insights into the dynamics between company disclosures and rating providers' ESG scores. The results of our study indicate that in general there is no discernible relationship between the ESG sentiment as reflected in company reports on the EUROSTOXX50 and the ESG ratings provided by the rating agencies. Nevertheless, our tool can provide an alternative, fine-grained measure of companies' own views on ESG-related matters.

[*] Centre for International Development and Environmental Research (ZEU), Justus-Liebig-University, Giessen, Germany

[†] Corresponding author: Christoph.Funks@wi.jlug.de

[‡] Department of Economics, Chair of Statistics and Econometrics, Justus Liebig University, Giessen, Germany

[§] Frankfurt School of Finance & Management, Frankfurt am Main, Germany

5.1 Introduction

Environmental, social, and governance (ESG) factors have become increasingly prominent in corporate reporting since their introduction in 2004 (United Nations, 2004). This trend highlights the increasing importance of ESG considerations for both companies and stakeholders, as evidenced by the development of reporting standards such as those proposed by the Task Force on Climate-related Financial Disclosures (TCFD) (TCFD, 2017a). The manner in which companies report on ESG issues can significantly impact perceptions, observable in metrics like stock prices and, most notably, ESG ratings. The improvement of companies' ESG standards is often evaluated through official ratings from agencies like Standard & Poor's (S&P). However, the reliability and consistency of these ESG ratings have been questioned, paving the way for alternative, text-based indicators. Company reports contain extensive information about their ESG measurements and perspectives (Berg et al., 2022). Natural language processing (NLP) offers a powerful solution for efficiently analyzing large volumes from these reports without the need for manual review (Schimanski et al., 2024).

As NLP has gained traction as a tool in recent years, a significant body of literature has emerged on ESG-related topics in various sources, including corporate reports, academic papers, or news data. To date, the most common methods for text classification involve fine-tuned transformer models for classification, generative prompt-based models such as GPT 3.5 and unsupervised methods like Latent Dirichlet Allocation (LDA). These models are generally employed to estimate the extent of ESG reporting and to relate it to other measures, such as ESG ratings or stock returns. A common text format utilized in this context is news data. For instance, Fischbach et al. (2023) employ NLP techniques to identify ESG-related news headlines. Subsequently, a BERT model, designated as the ESG-miner, is trained to identify company headlines and categorize them as ESG-relevant or not. An ESG score is then calculated based on the sentiment of the related headlines.

Moreover, the advent of OpenAI's ChatGPT has led to a surge in interest in prompt-based generative models. For instance, Jain et al. (2023) employ GPT-3.5 as an ESG classifier. The authors demonstrate a 20% correlation between company stock returns and ESG news, suggesting that that their query-based ESG classifier can accurately identify ESG factors. This capability can assist investors making more informed decisions. Moreover, the authors of Föhr et al. (2023) examine whether ChatGPT can be used as an auditing tool for sustainability reports, with the objective

of assessing their compliance with the EU taxonomy.

Another frequently employed NLP technique in this context is topic modeling. Goloshchapova et al. (2019) applied LDA to the Corporate Social Responsibility (CSR) reports of several companies listed on major stock market indices in 15 industrialized countries. The findings indicate that certain topics, such as 'employee safety,' are more frequently addressed by companies in the UK and Europe. However, they also identify sectoral biases, with certain sectors focusing more on certain issues than others. In their study, Lee et al. (2023) employed BERTopic as a topic model to gain insight into ESG discourse. In contrast to the focus on corporate reporting, the authors employ news data from LexisNexis and academic papers from the Web of Science to identify differences in ESG discourse between these two sources.

Our study focuses on reports published by the companies themselves. These reports are generally broader in nature, providing an overview of the company's performance or events, such as annual reports. Alternatively, they may be more specific in their focus, addressing ESG-related issues in greater depth, such as dedicated ESG reports.

One format of ESG-related reporting is the TCFD report. In 2017, the TCFD published guidelines that include specific questions which can be addressed either in a dedicated TCFD report or within a company's broader report. These guidelines are recommendations and therefore not mandatory. Companies that support the TCFD guidelines are not required to address all of the issues identified by the TCFD. Consequently, numerous studies have employed NLP techniques to assess the extent to which companies adhere to the TCFD guidelines and broader ESG-related issues. For instance, Luccioni et al. (2020) sought to identify sections in corporate reports that address climate-related issues by training a RoBERTa Question-Answering Model. This model leverages the 14 TCFD questions and the corresponding text sections that answer these questions as training data. The model was then employed to ascertain whether there were any differences in the extent to which companies in different sectors addressed the 14 TCFD questions. Additionally, Bingler et al. (2022) utilized a BERT model to investigate whether companies might engage in selective reporting with regard to the 14 TCFD questions. The findings indicate that companies tend to omit information on strategy, metrics, and targets, suggesting a selective reporting strategy that prioritizes non-material risks while neglecting material risks. This behavior suggests that companies are engaging in cherry-picking when it comes to TCFD reporting. Additionally, Auzepy et al. (2023) employ a zero-shot model to analyze TCFD reports. The researchers developed fine-grained

labels that align with the TCFD recommendations. Their findings revealed an increase in climate-related disclosures, although they also identified instances of selective reporting, indicating that some recommended topics may not have been fully addressed.

In their study, Friederich et al. (2021) trained a RoBERTa model and applied it to the company reports of 337 firms over a 20-year period. The model identified an overall increase in risk disclosure, with a particularly dynamic growth observed in transitional risks compared to physical risks. This conclusion was based on the observation that the mentions of risks, especially transitional risks, exhibited a pronounced increase around the year 2015.

This study builds upon the research presented in Schimanski et al. (2024), which examines the relationship between ESG ratings and reporting. The authors employ fine-tuned RoBERTa and DistilRoBERTa models to determine the relative amount of ESG reporting in corporate documents. To achieve this, one model is trained for each aspect of ESG on 2,000 sentences.Moreover, the authors of Schimanski et al. (2024) employ a fixed-effects panel data model for their time series analysis.

The objective of this study is to enhance the detection of ESG reporting through the use of an Aspect-Based Sentiment Analysis (ABSA) Few-Shot model. This approach not only requires less training data but also outperforms existing models on the same datasets. It offers a more detailed analysis by examining ESG subcategories and their respective sentiments.

In contrast to Schimanski et al. (2024), we challenge the assumption of a unidirectional effect—from reporting to ratings—by utilizing a panel Vector AutoRegressive (pVAR) model for our time series analysis. This allows for the examination of reciprocal effects between variables, providing a more comprehensive understanding of the dynamic relationship between ESG reporting and ratings.

In addition, instead of merely examining the quantity of ESG reporting, we utilize a qualitative measure of sentiment toward ESG issues. This approach is more plausible given the limitations of a quantitative measure. For instance, if ESG reporting increases on sustainability, it is possible that only negative reporting increases. However, an increase in negative reporting should result in a decrease in the ESG score, rather than an increase. As mentioned by Schimanski et al. (2024), the relationship between reporting quantity and ESG scores is found to be positive, regardless of whether the reporting is negative or positive in tone. In contrast, an analysis of the tone of reporting indicates that a more negative sentiment should have a negative effect on ESG scores, while a more positive tone should have a positive

effect on ESG scores. This aligns with the approach taken in this study, where we analyze the relationship between reporting sentiment and ESG scores.

Moreover, our ESG entity-based approach is more granular and provides a tool that outputs specific ESG entities with the corresponding sentiment. Stakeholders can use this tool to gain insights from company reports without the need to manually read them and form impressions of which ESG entities might be particularly negatively or positively annotated for a specific company in a given year. Consequently, stakeholders are provided with a tool that offers insights into the specific ESG entities discussed in a report and their respective positive or negative sentiment. This contrasts with previous approaches, which only allowed for the analysis of the extent to which a report addressed the pillars of E, S, or G.

In essence, our study diverges from previous research in several key respects. We adopt a detailed approach by identifying ESG-related entities in corporate reports and categorizing them into subcategories, thereby moving beyond the analysis of the three main pillars of ESG reporting (environmental, social, and governance). Rather than focusing on the quantity of reporting, we examine the tone of ESG reporting, creating sentiment timelines for each ESG subcategory. This analysis covers reports from EUROSTOXX 50 companies from 1999 until 2023 across all report types. Furthermore, we investigate the bidirectional relationship between reporting and ratings from three major rating agencies (Refinitiv, Bloomberg, and S&P) through a panel vector auto-regressive (pVAR) model, recognizing that ratings can influence subsequent reporting. The results of our time series analysis indicate that there is no discernible relationship between our sentiment scores and the ESG scores provided by the rating agencies. This could be attributed to the questionable quality of ESG ratings (Berg et al., 2022; Schimanski et al., 2024), an insufficiently large data set, or a model misspecification. Nevertheless, our sentiment scores remain a valuable tool for gaining insights into the perspectives of companies on ESG issues.

Our contribution is four-fold: First, we show that our few-shot model works better on a much smaller training sample than larger models. Thus, we save human and computational resources due to less annotation and training time. Second, our model is more fine-grained than existing models and provides detailed insights into firms' views on ESG issues. Third, we demonstrate that ESG ratings issued by rating agencies fail to accurately reflect the sentiment of firms with regard to ESG issues. Furthermore, our findings suggest that other variables exert a more significant influence on the composition of ESG ratings, while the quality and consistency of ESG ratings are open to question. Fourth, we create a measure to extract the

information on ESG issues contained in ESG reports, which can be used to analyze the reports on an ESG entity basis.

The remainder of this paper is organized as follows: In Section 5.2, we describe the data used to generate the sentiment indices and the ESG scores employed. In Section 5.3, we provide a brief overview of the methodologies employed for parsing the text data, the key functionality of the SetFit few-shot model, aspect-based sentiment analysis, and the integration of the two. Section 5.4 describes the training process and the performance of the model. Section 5.5 presents the results of our ABSA model and pVAR, while Section 5.6 provides a conclusion.

5.2 Data and Methodology

5.2.1 ESG ratings

For our analysis, we used ESG scores from three different rating agencies, namely Refinitiv, Bloomberg and S&P. We collected the ESG scores from their respective platforms, with the data downloaded in May 2024. The timeframe for the Bloomberg ratings is from 2015 to 2022, providing us with 8 years of data. For S&P, we have ESG scores from 2014 to 2023, and for Refinitiv, from 2002 to 2023. However, we do not have the full timeframe for all companies in our dataset. A detailed overview of the dataset is provided in Tables 5.9, 5.11, and 5.13 in the appendix. It is important to note that for some companies, we have data for the entire time frame. However, in some cases, the ratings for certain reports are missing for certain years. From all rating agencies, we use annualized data for each ESG pillar E, S, and G. The scores range from 0 to 10 for Bloomberg and from 0 to 100 for Refinitiv and S&P. Additionally, for Refinitiv, we have data for subcategories within each pillar, but this data is only available from 2018 to 2022. To facilitate the results of our pVAR models, we have min-max scaled the rating scores between 0 and 1, as our sentiment scores range from -1 to 1.

5.2.2 Descriptive Statistics of the Dataset

In our study, we utilized reporting data from EuroSTOXX 50 companies. We downloaded all available reports from the Refinitiv Database for the period between January 1, 1999, and June 14, 2023. We identified and downloaded a total of 2,072 reports in PDF format across various reporting types. Of these, 14 were either empty or entirely corrupted. Additionally, 18 documents were not in English and were excluded as well. Consequently, 2,038 were machine-readable and free from

coding errors, which were used for our further analysis. Table 5.1 summarizes the key statistics of the companies in our dataset, including the total number of reports, average number of reports per year, number of different report types, and the number of years with at least one report. Of the 50 companies in the EuroSTOXX 50, the total number of reports ranges from 9 for Adyen NV to 80 for Intesa Sanpaolo SpA. While the average number of reports per year ranges between 2 and 3 for most companies, each company has at least 2 different report types available. Although most companies have reports spanning ten or more years, two companies have reports for only 4 years (Prosus NV) and 5 years (Adyen NV).

Figure 5.1 provides a chart of the number of reports over the years. The chart illustrates the distribution of different report types over time, highlighting the increase in reporting activity and diversity in recent years. We distinguish between eight different reporting types: Corporate Governance, ESG, Environment, Health and Safety Reports, Full Year, GRI Report, Remuneration Committee, Social Reports, and Sustainability Committee. One can clearly see that for the period between 1999 and 2010, there were not many different reporting types available in our dataset. Most reports from this period are annual reports, which we denote as Full Year. Notably, in 2005, there is a significant lack of reporting, with only 11 of the 50 companies having reports available in the Refinitiv database. Starting in 2011, there is a clear increase in ESG and Remuneration Committee reports, while Corporate Governance reports are only available between 2011 and 2014.

Table 5.1: Company Summary with Report Types (Part 1)

Company Name	Total Reports	Average Reports per Year	Number of Report Types	Years with Reports
ASML Holding NV	58	2.4	4	24
AXA SA	46	1.9	4	24
Adidas AG	46	2.0	5	23
Adyen NV	9	1.8	2	5
Airbus SE	59	2.8	4	21
Allianz SE	65	2.7	5	24
Anheuser-Busch Inbev SA	54	2.3	2	23
BASF SE	36	1.5	5	24
BNP Paribas SA	40	2.4	4	17
Banco Bilbao Vizcaya Argentaria SA	30	2.3	4	13
Banco Santander SA	31	2.6	6	12
Bayer AG	38	1.6	4	24
Bayerische Motoren Werke AG	44	1.9	5	23
CRH PLC	45	2.0	4	23
Danone SA	32	1.8	4	18
Deutsche Boerse AG	36	1.6	4	22
Deutsche Post AG	34	1.5	3	23
Deutsche Telekom AG	45	2.0	6	23
Enel SpA	58	2.5	5	23
Eni SpA	54	2.3	7	23
EssilorLuxottica SA	43	2.1	4	20
Flutter Entertainment PLC	29	1.3	3	22
Hermes International SCA	21	1.1	2	20
ING Groep NV	37	1.6	3	23
Iberdrola SA	37	2.8	4	13

Table 5.2: Company Summary with Report Types (Part 2)

Company Name	Total Reports	Average Reports per Year	Number of Report Types	Years with Reports
Industria de Diseno Textil SA	37	3.1	5	12
Infineon Technologies AG	41	1.8	3	23
Intesa Sanpaolo SpA	80	3.6	6	22
Kering SA	42	2.2	6	19
Koninklijke Ahold Delhaize NV	39	1.6	4	24
L'Air	38	1.9	2	20
L'Oreal SA	37	1.6	4	23
LVMH Moet Hennessy Louis Vuitton SE	53	2.9	3	18
Mercedes Benz Group AG	46	1.8	3	25
Muenchener Rueck	43	2.0	4	22
Nokia Oyj	53	2.2	5	24
Nordea Bank Abp	48	2.0	4	24
Pernod Ricard SA	28	1.9	2	15
Prosus NV	12	3.0	3	4
SAP SE	34	1.4	3	24
Safran SA	25	1.3	3	19
Sanofi SA	39	2.2	5	18
Schneider Electric SE	35	1.9	2	18
Siemens AG	43	1.8	4	24
Stellantis NV	27	2.5	5	11
TotalEnergies SE	36	1.8	5	20
UniCredit SpA	73	3.8	6	19
Vinci SA	31	1.9	2	16
Volkswagen AG	41	1.8	4	23
Vonovia SE	30	2.7	4	11

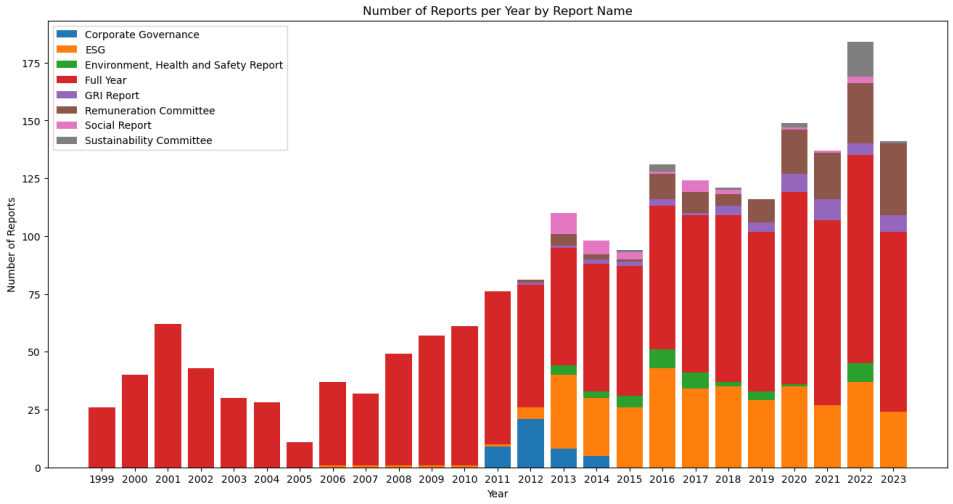


Figure 5.1: Number of Reports per Year by Report Name.

5.3 Methodology

5.3.1 Extracting text from PDFs

All of the reports utilized in our analysis are in PDF format. The conversion of textual information present within PDF documents into a format suitable for further NLP analysis is a more intricate process than that required for textual data stored in CSV or TXT files. In this paper, we apply a layout-parsing model, which is able to detect and extract actual text from PDF documents. In the context of our analysis, the text of the reports is included, while that of tables and graphs is deliberately omitted. This not only improves the quality of our data but also allows us to save computation time (Auzepy et al., 2023).

Our parsing model is based on Visual-Layout (VILA) groups introduced by Shen et al. (2021). VILA is able to convert textual data into groups of tokens (text lines or blocks) and to assign a layout tag to these tokens. There are several variants of VILA, called H-VILA (Visual Layout-guided Hierarchical Model) and I-VILA (Injecting Visual Layout Indicators). After several trials, we choose the H-VILA block variant trained on grotoap2 using the layoutLM model (Xu et al., 2020) as a base model since it delivered the best extraction and tokenization results. The output consists of the extracted text as groups of tokens together with the corresponding layout tags. Depending on the training set, the layout tags can be figures, body content, abstract and title. For our analysis, we keep the parts tagged as body content and abstract (Auzepy et al., 2023).

To implement this, we used several tools and models. We initialized the PDFEx-

tractor with "pdfplumber" for extracting text and images from PDFs. The EfficientDetLayoutModel from layoutparser was used for detecting the layout of the documents. We then employed the HierarchicalPDFPredictor from the VILA library, specifically the "allenai/hvila-block-layoutlm-finetuned-grotoap2" model. However, in some instances, these extraction methods failed, so we used a fallback option to retrieve as much textual information as possible. First, in a few cases, coding errors appeared. In these instances, we simply skipped the problematic byte and moved on to the next one. For more complex cases, we integrated an OCR agent using Tesseract. This involved converting the PDF documents to images using the pdf2image library, after which OCR was performed. The disadvantage of this approach is that detecting text blocks and layout in each page of the PDF using the VILA model was no longer possible. Despite this limitation, our comprehensive approach ensures that we accurately and efficiently extract the relevant textual data from PDF reports, thereby improving the overall quality and reliability of our analysis (Auzepy et al., 2023).

5.3.2 Few-shot SetFit Model

In recent years, few-shot models have gained popularity due to their improving performance. Unlike standard models, which require thousands of data samples for training, few-shot and zero-shot models are advantageous as they require minimal to no labeled data, making them cost-effective and time-efficient. In the NLP context, zero-shot models do not require any labeled data for prediction, relying solely on semantic understanding. Conversely, few-shot models need only a small set of labeled data to outperform some standard models that depend on thousands of labeled training sentences.

Our analysis employs SetFit (Sentence Transformer Fine-tuning) (Tunstall et al., 2022), an efficient, prompt-free few-shot model using sentence transformers (ST) available on Hugging Face (`https://huggingface.co/docs/setfit/index`). The authors demonstrate that with merely 8 labeled sentences per class, SetFit surpasses the performance of a standard fine-tuned RoBERTa large model trained on a full set of three thousand examples (Tunstall et al., 2022). We further validate these findings, showing that with significantly fewer labeled sentences, SetFit achieves superior performance on the same datasets compared to the ESG RoBERTa and DistilRoBERTa models, each trained with two thousand sentences (Schimanski et al., 2024).

SetFit offers several advantages over existing few-shot models. When using

RoBERTa as its base, SetFit outperforms smaller prompt-based models like GPT-3 and PET, although it does not surpass T-FEW (H. Liu et al., 2022). However, it is worth noting that SetFit is thirty times smaller than T-FEW, making it more compact while still delivering commendable performance without relying on prompts. This aspect is crucial as dependence on prompts, as seen with models like GPT-3, can lead to sensitive and unstable outcomes due to minor variations in wording (Tunstall et al., 2022).

The training of SetFit involves two key steps. Initially, the sentence transformers are trained in a Siamese manner on sentence pairs. Subsequently, a classifier head is trained using the encoded data from the first step. This bifurcation allows for a distinct separation between the ST fine-tuning phase and the classification head training phase, streamlining the process.

In the first stage, a contrastive training approach commonly employed in image similarity detection (Koch et al., 2015) is adopted to address the challenge of limited training data in few-shot scenarios (Tunstall et al., 2022). This approach utilizes a small set of K labeled examples, denoted as $D = \{(x_i, y_i)\}$, where x_i represents sentences and y_i their corresponding class labels. For each class label $c \in C$, a set of R positive triplets, $T_p^c = \{(x_i, x_j, 1)\}$, is generated, where x_i and x_j are randomly selected sentence pairs from the same class ($y_i = y_j = c$). Similarly, a set of R negative triplets, $T_n^c = \{(x_i, x_j, 0)\}$, is formed, where x_i are sentences from class c and x_j are sentences from different classes ($y_i = c, y_j \neq c$). The contrastive fine-tuning dataset T is then assembled by merging these positive and negative triplets across all classes: $T = \{(T_p^0, T_n^0), (T_p^1, T_n^1), \ldots, (T_p^{|C|}, T_n^{|C|})\}$, where $|C|$ denotes the number of class labels, $|T| = 2R|C|$ represents the total number of pairs in T, and R is a hyperparameter set to 20 in all evaluations as per Tunstall et al. (2022).

This contrastive training strategy effectively enlarges the training dataset in few-shot scenarios. Given a small number of labeled examples K for a binary classification task, the potential size of the contrastive fine-tuning set T is derived from the total number of unique sentence pairs that can be generated, amounting to $K(K-1)/2$, which significantly exceeds the original count of K samples (Tunstall et al., 2022).

In the subsequent stage, the fine-tuned sentence transformer (ST) encodes the original labeled training data $\{x_i\}$, producing a single sentence embedding per sample, denoted as $Emb^{x_i} = ST(x_i)$, where $ST()$ symbolizes the fine-tuned STs from the first step. These embeddings, alongside their corresponding class labels, form the training set for the classification head, $T^{CH} = \{(Emb^{x_i}, y_i)\}$, where $|T^{CH}| = |D|$ build the training set for the text classification step. A logistic regression model

serves as the classification head throughout this model (Tunstall et al., 2022).

To perform inference with the trained model, the pre-fine-tuned sentence transformer (ST) first encodes an unseen input sentence, denoted as x_i, generating a sentence embedding. Following this, the classification head, which was trained in the preceding step, determines the class prediction for the input sentence based on its embedding. This process is formally represented as $x_i^{pred} = CH(ST(x_i))$, where CH represents the function used by the classification head to predict the class (Tunstall et al., 2022).

The SetFit model described above performs a standard text classification task. However, in this paper, we will use the ABSASetFit model, which classifies both an entity and the corresponding sentiment in a sentence. The modifications to the standard SetFit model will be explained in the following two sections.

5.3.3 Aspect-based Sentiment Analysis

In our analysis, we employed ABSA to extract ESG-related entities and their corresponding polarity. This methodology, augmented with insights from various studies (Zhang et al., 2018; Saeidi et al., 2016; Jo & Oh, 2011; Pontiki et al., 2015, 2016) and methodologies from Tunstall et al. (2022), helps us generate ESG criteria-specific sentiment time series. Furthermore, these time series are employed in our analysis of their relationship with ESG ratings. Compared to vanilla sentiment analysis, ABSA can extract a text's sentiment regarding a specific entity, such as a person, location, company, and more (B. Liu, 2020).

ABSA is typically employed by businesses to gain insights into customer sentiment regarding specific aspects of products or services. Nevertheless, this form of enhanced sentiment analysis can also be beneficial for other domains, as any entity can be extracted from texts. In this study, we focus on entities related to ESG issues and the tones in which they are discussed—namely, positive, negative, or neutral.

5.3.4 Combining SetFit and ABSA

The SetFit model is tailored for ABSA tasks in a specialized variant, SetFitABSA, which is accessible via Hugging Face (`https://huggingface.co/docs/setfit/how_to/absa`). As ABSA models in particular demand a substantial quantity of labeled data, requiring annotators to identify both the entity in question and its sentiment within the training sentences, this process is particularly labor-intensive; therefore, a few-shot ABSA model like SetFitABSA can substantially reduce the effort and time required for annotation (Laperdon et al., 2023). In particular, ABSA

models that are particularly traditional in nature require a substantial volume of labelled data, necessitating that annotators identify not only the entity in question but also its sentiment polarity within the training sentences. The labeling task is notably labor-intensive, and thus a few-shot ABSA model, such as SetFitABSA, can substantially reduce the effort and time required for annotation (Laperdon et al., 2023).

The core architecture of the SetFit model is retained, as outlined in section 5.3.2. In the ABSA framework, the few-shot model is deployed in two of three stages. Concisely, the process unfolds as follows:

1. Aspect Candidate Extraction: In the first stage, the candidate aspect or entity is extracted from the sentence. The 'SpaCy' library is used to tokenize the sentences and extract all nouns or noun compounds. Not all extracted nouns are actual aspects; therefore, they are referred to as aspect candidates (Laperdon et al., 2023).

2. Aspect Classification: In the second stage, a SetFit model determines whether the extracted candidate qualifies as an aspect. Training samples containing examples of aspect/non-aspect labels are needed for this step. Aspect candidates are merged with the entire training sentence to create a training instance following this template: `aspect_candidate:training_sentence` (Laperdon et al., 2023). For example, given the sentence "Waiters aren't friendly but the cream pasta is out of this world," assuming the nouns 'Waiters' and 'cream pasta' are aspects and 'world' is not an aspect, the templates would be:
   'Waiters: Waiters aren't friendly but the cream pasta is out of this world." with the label 1,
   "cream pasta: Waiters aren't friendly but the cream pasta is out of this world." with the label 1, and
   "world: Waiters aren't friendly but the cream pasta is out of this world." with the label 0.
   By training on such sentences, the model learns which aspect candidates are aspects or non-aspects (Laperdon et al., 2023).

3. Sentiment Classification: In the final stage, another instance of the SetFit model classifies the sentiment associated with the aspect. Training is similar to the aspect classification stage, but instead of a binary label, the label is one of three possible polarities: 'POS' for positive, 'NEG' for negative, and 'NEU' for neutral. Here, non-aspects are not included since only aspects are associated

with polarities (Laperdon et al., 2023).

This streamlined approach to ABSA using SetFitABSA represents a significant advancement, leveraging the few-shot learning capabilities of SetFit to efficiently process and analyze sentiment with minimal labeled data. The authors claim that SetFitABSA, when applied to the SemEval14 ABSA Datasets 'Laptop14' and 'Restaurant14', SetFitABSA performs with a low number of training samples remarkably better than T5 (Raffel et al., 2023), despite being two times smaller, and better than GPT2-medium (Radford et al., 2019), even though being three times smaller. SetFitABSA even perfoms better than the 64 times bigger Llama2 (Touvron et al., 2023) when compared on equal training sample size (Laperdon et al., 2023).

5.4 Model Training

5.4.1 General Description of Model Training

In this study, we leverage our model to extract and analyze ESG aspects, along with their corresponding sentiment, from sentences within corporate reports. This approach offers a refinement over the methodology described by Schimanski et al. (2024), who categorize content strictly under the broad labels of 'E' (Environmental), 'S' (Social), and 'G' (Governance). Unlike these models, our technique seeks to uncover more nuanced ESG entities, subsequently mapping these to predefined ESG subcategories for a more granular analysis.

To construct our training sample, we employed sentences from the datasets provided by Schimanski et al. (2024). The original work involved labeling approximately 2,000 sentences per model, assigning a '1' to sentences addressing the respective ESG aspect (e.g., 'E' for Environmental) and '0' otherwise. It's noteworthy that while a sentence labeled '0' in the 'E' dataset might not discuss environmental issues, it could still pertain to social or governance themes. Each of the three models developed by Schimanski et al. (2024) focuses on a specific ESG aspect, yet their datasets largely comprise identical sentences.

Our methodology involved a more selective approach, utilizing roughly 100 sentences from each of Schimanski et al. (2024)'s datasets. For example, from the dataset designated for the 'E' aspect, approximately 100 sentences were selected that were marked '1,' indicating a focus on environmental concerns. In total, our environment training set comprises 105 unique sentences, our social training set 95 unique sentences, and our governance training set 102 unique sentences. The

discrepancies arose because we determined that some of the sentences were not suitable for governance labeling and that some sentences for 'E' and 'S' were not included in the initial 100 sentences that were deemed important. The labeling process entailed the identification of the specific ESG aspect and sentiment within each sentence. This process was conducted by a team of three, comprising the authors of this paper, allowing for an initial intimate understanding and subsequent refinement of the labeling.

Our labeling process diverges from conventional methods by not starting with a predefined set of labels, as the relevant entities can vary significantly across sentences. Initially, one team member labeled the dataset to identify potential entities. In the second phase, a second reviewer examined the dataset for inconsistencies in labeling, which were then discussed and resolved collaboratively. Finally, a third reviewer, provided with a manual for labeling, compared their independent analysis with the previously labeled dataset to highlight and discuss discrepancies.

5.4.2 ESG Subcategories, Entities and Sentiment

To achieve a more detailed analysis, which identifies potential underrepresentation of ESG categories in reports, we employ predefined subcategories. These subcategories are detailed in Tables 5.6, 5.7 and 5.8 which outlines the ESG subcategories utilized in our labeling process. Although a variety of definitions exist, they generally converge on the same fundamental subcategories, exhibiting minimal variation. The subcategories utilized in this study are derived from those identified in Boffo & Patalano (2020) and represent a synthesis of the categorizations proposed by Thomson Reuters, MSCI, and Bloomberg. During the labeling process, these subcategories guided our selection of relevant entities from each category. The table also includes examples of entities from our training set.

In assessing the sentiment of sentences, context played a crucial role in determining the appropriate polarity. Sentences were deemed positive if they indicated that a company was taking steps to enhance its performance concerning ESG criteria. On their own, most sentences may appear neutral. For instance, a statement about a company offering employee training would typically be considered neutral. However, within the context of ESG implementation, such an initiative is regarded positively, leading us to label these sentences as positive. Conversely, sentences merely stating the existence of certain criteria without indicating the company's adherence were labeled neutral. Sentences that mentioned a company's failure to implement or improve upon ESG-related practices were labeled negative. Given the tendency of

companies to enhance their image in such reports, the majority of sentences were labeled positive, resulting in imbalanced classes. This imbalance might adversely affect the model's ability to predict sentiment accurately, a challenge that is elaborated upon in section 5.4.4.

5.4.3 Training Process

Table 5.3: Training Results for Entity and Sentiment

Metric	Entitiy	Sentiment
Accuracy	0.9173	0.7917
Precision	0.9204	0.8096
Recall	0.9173	0.7917
F1 Score	0.9184	0.7999

A single base model was trained on our NVIDIA RTX A5000 GPU. While other models may exhibit slight improvements, a comprehensive five-fold cross-validation would be necessary to ascertain whether another model could be deemed significantly superior to the base model employed. The base model selected was 'sentence-transformers/paraphrase-mpnet-base-v2,' chosen due to its relatively fast training time of approximately five hours and ten minutes on the GPU. Additionally, the model exhibited satisfactory performance. However, larger models, which could potentially yield better results, cannot be trained on the GPU due to memory limitations. Further research may be warranted to investigate the potential for enhancing the model by selecting a more optimal base model.

A grid search was deemed unnecessary due to the impracticality of the extensive training process. Default parameters were used, consisting of one set for fine-tuning the sentence transformer and one for the classification head. For the sentence transformer, we used a batch size of 16, one epoch, and a body learning rate of 2e-05. For the classification head, we used a batch size of 2, 16 epochs, and a body learning rate of 1e-05. The default head learning rate for the entire model was set, and the CosineSimilarityLoss function was chosen for the entire model's loss function.

As a sampling method, oversampling was used to ensure an even number of positive and negative sentence pairs until every sentence pair had been drawn. This methodology ensures that all sentence pairs are included at least once, thereby preventing an imbalance of positive and negative pairs. Given that our polarity

Table 5.4: Performance Comparison of E, S and G Models

Model	**Accuracy**		**F1 Score**		**Precision**		**Recall**	
SetFit-Model E	0.9950	±	0.9952	±	0.9957	±	0.9945	±
	0.0050		0.0048		0.0043		0.0050	
EnvRoBERTa	0.9565	±	0.9319	±	0.9330	±	0.9331	±
	0.0098		0.0140		0.0399		0.0314	
SetFit-Model S	0.9600	±	0.9543	±	0.9566	±	0.9600	±
	0.0292		0.0341		0.0329		0.0292	
SocRoBERTa	0.9341	±	0.9190	±	0.9035	±	0.9366	±
	0.0140		0.0179		0.0345		0.0292	
SetFit-Model G	0.9750	±	0.9738	±	0.9747	±	0.9750	±
	0.0194		0.0207		0.0205		0.0194	
GovRoBERTa	0.8961	±	0.7848	±	0.8562	±	0.7252	±
	0.0113		0.0262		0.0184		0.0378	

data is imbalanced, this approach is beneficial to our model, as oversampling serves to balance the training data and improve the model's performance. To assess the performance of the models, an 80% training set and 20% test set split was employed. The structure of the training and test sets can be found in Table 5.5 in the Appendix.

5.4.4 Training Results

In evaluating our performance, it is essential to distinguish between two key aspects: entity accuracy and sentiment accuracy. Entity accuracy pertains to the ability of the spaCy model to correctly classify aspect candidate spans as either true entities or non-entities. Sentiment accuracy, on the other hand, concerns the model's capacity to correctly categorize only the filtered aspect candidate spans into their respective classes. With default parameters, the entity prediction accuracy is 91.73%, while the sentiment prediction accuracy is 78.13%. The F1 scores for the entity and sentiment predictions can be found in Table 5.3. Although our model is more complex than the three distinct E, S, and G models presented in Schimanski et al. (2024), our performance is comparable.

To compare the SetFit model with the models trained by Schimanski et al. (2024), we trained three distinct SetFit models on a considerably smaller training set. We utilized the initial 200 sentences from Schimanski et al. (2024), implementing a 5-fold cross-validation procedure. Our training set comprised 160 sentences, with 40 sentences in our test set. Our model consistently demonstrated superior performance relative to the three models presented in Schimanski et al. (2024). The results are presented in Table 5.4. For a meaningful comparison with our ABSA model results, the same base model was employed. Despite Schimanski et al. (2024) arguing that the extensive and necessary nature of pretraining is a limitation in the ESG framework, our model, which was not pretrained on our subdomains, still demonstrated superior performance.

5.4.5 Panel VAR estimation

To estimate the pVAR, we employed our sentiment scores and the ESG scores for the main pillars E, S, and G. Additionally, we attempted to estimate more granular pVARs at a subcategory level; however, we only had data from Refinitiv for the years 2018 until 2022. Unfortunately, the available data for this period was insufficient for a reliable pVAR estimation. As a model, we estimated panel VARs with fixed effects and a System Generalized Method of Moments (GMM) approach. The System GMM approach enhances efficiency and addresses potential endogeneity issues.

For each ESG dimension (E, S, G), the dependent variables included the respective ESG score and the net sentiment score. A one-lag approach was employed for the dependent variables, based on the assumption that a company report in a specific year would not have a long-lasting impact on the ESG scores, nor vice versa. To mitigate potential issues associated with serial correlation in the transformed error terms, which could result in less efficient and reliable estimates, and to address potential non-stationarity, we employed forward orthogonal deviations. A two-step estimator was used to obtain robust standard errors. To avoid the proliferation of instruments and the resulting overfitting, we opted to collapse the instruments. The models were estimated separately for each ESG dimension (E, S, G) and each rating agency (Bloomberg, Refinitiv, S&P) using the panelvar package in R (Sigmund & Ferstl, 2021).

For each ESG dimension (E, S, G), the pVAR model without exogenous variables and predetermined variables can be represented as follows:

$$\boldsymbol{y}_{i,t} = \boldsymbol{\mu}_i + \sum_{l=1}^{p} \boldsymbol{A}_l \boldsymbol{y}_{i,t-l} + \boldsymbol{\epsilon}_{i,t}$$

Where:

- $\boldsymbol{y}_{i,t}$ is the vector of endogenous variables for company i at time t. In our context, $\boldsymbol{y}_{i,t}$ includes the ESG score and the sentiment score:

$$\boldsymbol{y}_{i,t} = \begin{bmatrix} \text{ESG}_{i,t} \\ \text{Sentiment}_{i,t} \end{bmatrix}$$

- $\boldsymbol{\mu}_i$ represents the individual fixed effects for company i.
- p is the number of lags of the endogenous variables.
- $\boldsymbol{A}_l$ is the matrix of coefficients for the l-th lag of the endogenous variables.
- $\boldsymbol{\epsilon}_{i,t}$ is the vector of idiosyncratic error terms.

Since we used forward orthogonal deviations, the model becomes:

$$\tilde{\boldsymbol{y}}_{i,t} = \boldsymbol{\mu}_i + \sum_{l=1}^{p} \boldsymbol{A}_l \tilde{\boldsymbol{y}}_{i,t-l} + \boldsymbol{\epsilon}_{i,t}$$

The transformed variables, denoted by $\tilde{\boldsymbol{y}}_{i,t}$, represent a deviation from the average of future observations for the same individual. For our specific case with one lag ($p = 1$) and the System GMM approach, the model can be simplified to:

$$\tilde{\boldsymbol{y}}_{i,t} = \boldsymbol{\mu}_i + \boldsymbol{A}_1 \tilde{\boldsymbol{y}}_{i,t-1} + \boldsymbol{\epsilon}_{i,t}$$

In matrix form, considering $\boldsymbol{y}_{i,t} = [\text{ESG}_{i,t}, \text{Sentiment}_{i,t}]^\top$, the model for each ESG dimension (E, S, G) for each rating agency (Bloomberg, Refinitiv, S&P) is:

$$\begin{bmatrix} \tilde{\text{ESG}}_{i,t} \\ \tilde{\text{Sentiment}}_{i,t} \end{bmatrix} = \boldsymbol{\mu}_i + \boldsymbol{A}_1 \begin{bmatrix} \tilde{\text{ESG}}_{i,t-1} \\ \tilde{\text{Sentiment}}_{i,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{\text{ESG},i,t} \\ \epsilon_{\text{Sentiment},i,t} \end{bmatrix}$$

In contrast to the fixed effects model proposed in Schimanski et al. (2024), this equation is designed to capture the dynamic interactions between ESG scores and sentiment scores over time for each company in the dataset. It should be noted, however, that no exogenous variables were included, such as company fundamentals like the current ratio or revenues. Moreover, the data utilized in this study was limited to that of the EUROSTOXX50, whereas Schimanski et al. (2024) employed data from the EUROSTOXX600.

5.5 Results

5.5.1 Descriptive Results

The output of our model comprises ESG-related entities and their corresponding sentiment for each report under consideration. This detailed data can be invaluable for an analyst focusing on specific companies. However, while a detailed examination of each company is possible, it would provide too much information for the scope of this work. Therefore, we have summarized the results for all reports and years under consideration simultaneously. Figure 5.2 illustrates the relative importance and sentiment of all ESG-related entities across all reports, years, and ESG categories. The entities displayed in the word cloud represent a summary of issues from reports spanning from 1999 to 2023. Each word cloud displays the 100 most frequently occurring words. The size of the font indicates the frequency of occurrence of the corresponding word in a given report. The average sentiment score of the overall word cloud is 0.5642. Green shades indicate a positive tone, red shades indicate a negative tone, and shades of yellow represent words discussed with neutral sentiment, lying between the most positively and most negatively discussed terms.

It appears that "board" and "employee" are the most pivotal words, with "employee" being discussed in a more favorable manner on average and "board" in a more unfavorable manner. While not as crucial, terms such as "stakeholder", "culture", and "community" are frequently discussed in a positive light. These terms indicate the ESG topics that companies claim to be improving or are already on a positive trajectory. Some of the less frequently but negatively discussed topics include "depreciation", "volatility", and "consolidated financial statement".

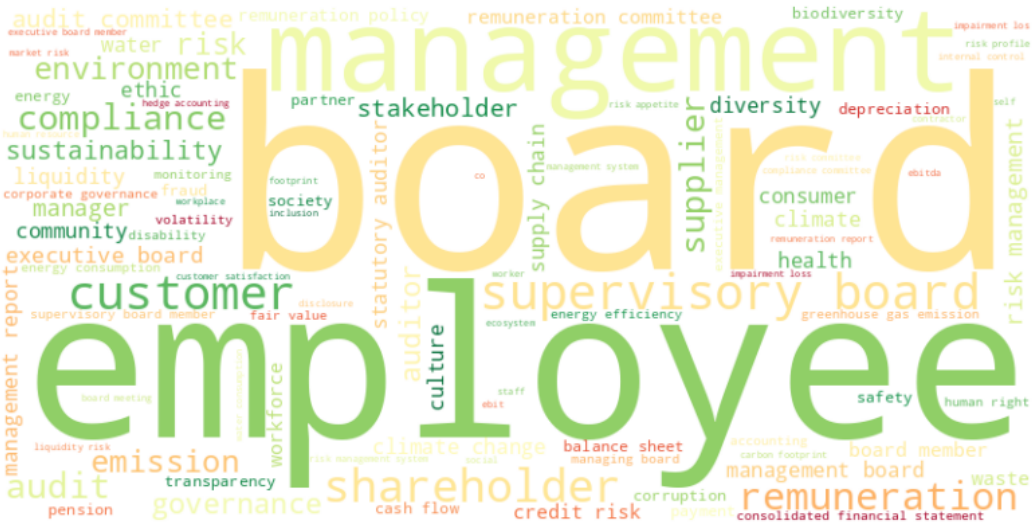


Figure 5.2: The mean sentiment for the top 100 words in all reports across all ESG pillars.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

Moreover, a word cloud was generated for the top 100 words in each of the three pillars. A comparison of the three pillars reveals that the social pillar is the most positive, with an average sentiment score of 0.7025. The governance pillar appears to be more negative, with an average sentiment score of 0.47081. In contrast, the environment pillar has a more positive tone, with an average sentiment score of 0.55072, though this is less positive than that of the social pillar. In general, companies tend to report the most negative experiences with governance-related issues. However, the subjective nature of the reporting does not permit the assumption that they experience greater difficulties with governance issues than with social issues. Nevertheless, an analysis of the sentiment scores on their own reveals the key areas of concern for companies in the ESG context.

Upon examining the environment pillar in Figure 5.3, it becomes evident that the term "sustainability" is a significant topic with a positive tone. The term "environment" is also frequently mentioned, although with a less positive tone. The term "emission" appears to be a source of greater concern, potentially leading to difficulties for the companies in question. While not frequently discussed, other environmental issues, such as weather conditions, weather events, oil prices, gas prices, and droughts, also appear to be significant concerns for companies. In contrast, topics such as recycling, energy efficiency, and ecosystems are associated with a more positive tone.

In conclusion, weather-related events and fluctuations in prices present significant challenges for companies within the EUROSTOXX50. In contrast, sustainability issues such as recycling, the carbon footprint, and energy efficiency appear to be less pressing concerns.

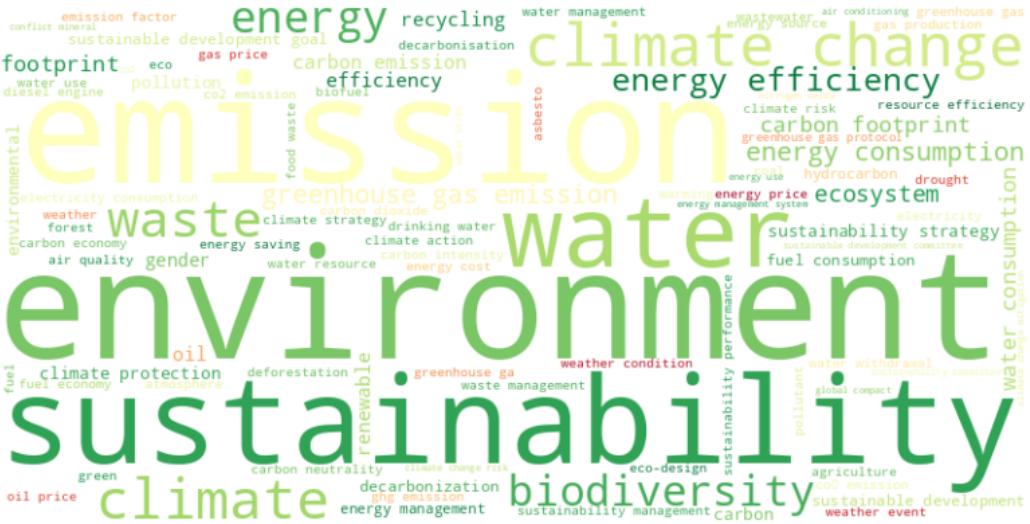


Figure 5.3: The mean sentiment for the top 100 words in all reports for the environment pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

As previously stated, the social setting is characterized by a positive tone. The color green is frequently associated with words in Figure 5.4. This is particularly evident in the instances of the most discussed words "employee" and "customer". Conversely, the words "community", "culture", and "diversity" appear to be the most positively mentioned in the social setting. The social pillar presents a challenge for companies in the areas of pension, employee benefits, and fair value. In general, companies tend to report more on the positive aspects and achievements related to social issues in the ESG context.

The wordcloud representing the governance pillar, as shown in Figure 5.5, is predominantly composed of words in shades of yellow and orange, which collectively convey a more negative tone. The most frequently discussed topics in the context of governance are "management", "shareholder", "board", and "supervisory board". These topics are associated with a more negative tone. It appears that the board and remuneration settings are a more negative issue for companies in terms of governance. The annotated words that are more positive in tone are "stakeholder", "transparency", and "compliance". These words are more closely related to the

Figure 5.4: The mean sentiment for the top 100 words in all reports for the social pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

concept of business transparency.

While an analysis of individual ESG-related entities can provide detailed insights into companies' reporting on ESG topics, displaying changes in tone over time for each entity is not feasible. Consequently, to analyze ESG issues in general, including a time component, the entities were grouped into ESG subcategories as previously mentioned in Section 5.4.2. Figure 5.6 depicts the average sentiment scores per year per ESG subcategory for all EUROSTOXX50 companies. A darker shade of red indicates a more negative tone, while a darker shade of green indicates a more positive tone. It is important to note that all sentiment scores over the years are positive. Consequently, our findings indicate a tendency towards more positive reporting on ESG-related issues in general, with individual reports exhibiting a negative sentiment score. However, when considering the overall tone across all companies and years, a positive sentiment is evident. As our model is trained using an oversampling method to prevent class imbalance, we still observe a greater tendency towards positivity in our training sample. This is likely because reports are, on average, more positive in nature. This is a logical conclusion, as companies appear to prioritize reporting on more positive achievements and topics, which could be perceived as greenwashing (Bingler et al., 2022). Nevertheless, to substantiate these assertions, further analysis is required, which will be briefly discussed in the Conclusion section.

Prior to 2005, there appeared to be a greater range of tones in reporting, with

Figure 5.5: The mean sentiment for the top 100 words in all reports for the governance pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

particularly positive coverage of "Energy Use" in 1999, "Human Rights" and "Labor Practices" in 2000, and "Community Engagement and Impact" in 2002. In contrast, more negative reporting was observed for "Executive Compensation" in 1999 and "Climate Change Policies and Carbon Footprint" in 2005. However, given the paucity of reports in the early years and the fact that special ESG reports were not introduced until 2004, the results from this period may be subject to bias.

Upon examination of the data from 2005 onwards, it becomes evident that certain subsections are more positively discussed than others. In general, the subcategories "Community Engagement and Impact", "Customer Satisfaction and Data Protection", "Employee Relations and Diversity", and "Sustainability and Environmental Stewardship" appear to have been discussed in the most positive manner over time. The sentiment expressed towards "Energy Use" is consistently positive, with the exception of the years 2004 to 2006 and 2023. On the less favorable end of the spectrum are the categories "Board Composition and Structure", "Climate Change Policies and Carbon Footprint", "Executive Compensation", and "Shareholder Rights". The remaining categories exhibit a more diverse range of results.

The scores generated for all companies in a given year can also be created for a specific company or a specific report. Consequently, this tool allows for the generation of alternative ESG scores. While such scores are inherently subjective and reflect the perspectives of the companies themselves, they can be used as an adjunct to

traditional scores, which are subject to similar limitations. This aligns with the findings of Berg et al. (2022), who posited that aggregate scores may be unreliable. Additionally, the correlation coefficients between the Bloomberg, Refinitiv, and S&P's scores are relatively low, illustrating the variability between the scores of the rating agencies and the questionable transparency of their methodologies.

Despite the subjectivity of our scores, they can provide valuable insights into a company's strengths and weaknesses on ESG topics. Furthermore, changes in score values can be interpreted as improvements or declines in certain ESG areas within a company.

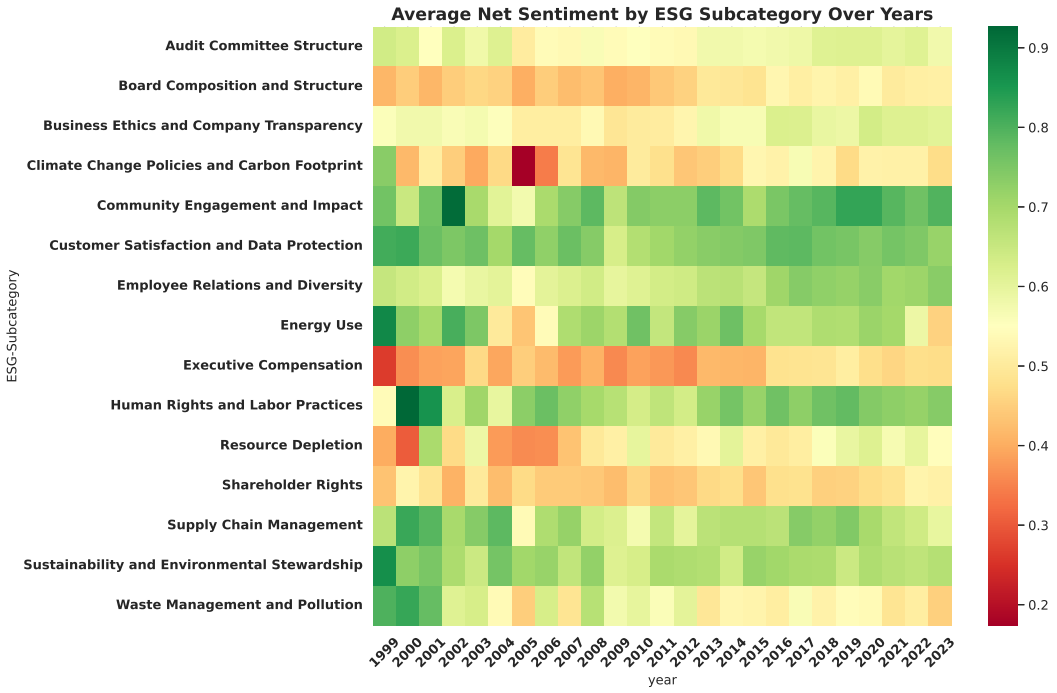


Figure 5.6: Average Sentiment for each ESG subcategory over time.
The legend ranges from a more negative sentiment in blur to a more positive senitment in red.

5.5.2 Panel VAR model results

To interpret the results of our pVAR models, we employ generalized orthogonal response functions. The results can be found in Figure 5.8 in the appendix. In all models, except for the model for governance with S & P data, where we find a positive effect in the first period for both sentiment on score and score on sentiment, there are no significant reactions from the ESG scores to a shock in sentiment, nor vice versa. This may be attributed to model misspecification due to a lack of exogenous variables. Alternatively, this could be viewed as another point of criticism regarding ESG scores from rating agencies. In future research, it would be beneficial to include

company fundamentals to improve the predictive value of the pVAR models, though this would also increase model complexity.

Due to the criticism outlined in the literature (Berg et al., 2022), we also calculated the correlations between the ESG scores from all agencies. Figure 5.7 illustrates the results, demonstrating a lack of correlation between the E scores, S scores, and G scores across the agencies. This aligns with the general criticism of these scores and shows significant variation between ESG scores, raising questions about their reliability. Consequently, the lack of significance in the relationship between the scores and our sentiment scores does not necessarily imply a deficiency in the quality of our sentiment scores. It is possible that the rating scores do not adequately capture the subjective views of the companies regarding ESG issues.

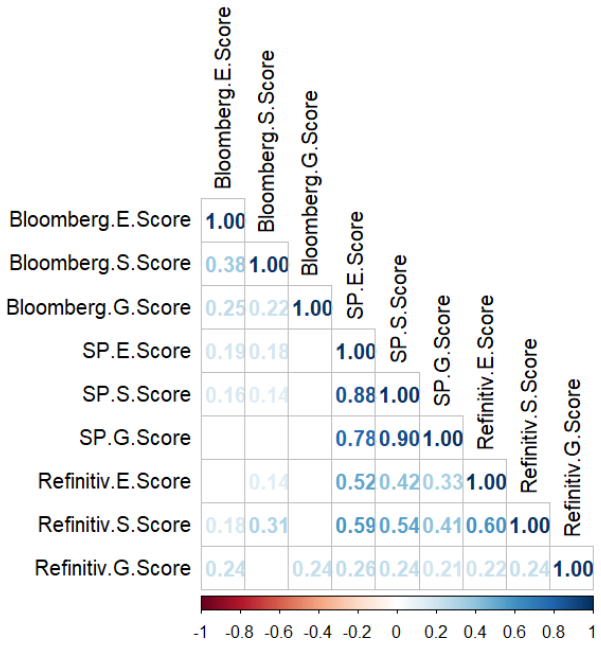


Figure 5.7: Correlation Analysis between Ratings from Refinitiv, S&P, and Bloomberg.

The figure displays the Pearson correlation coefficients between the ratings obtained from Refinitiv, S&P, and Bloomberg. Significant correlations are shown with their respective values, while non-significant correlations are blanked out.

Consequently, our scores may be employed as an alternative measure, albeit subjective, in conjunction with the ESG scores provided by rating agencies. The aggregation of our scores into ESG subcategories provides a more detailed view of ESG issues. Alternatively, companies can be analyzed on an entity-by-entity basis, without the necessity of manually reading their reports. Furthermore, stakeholders may utilize reports from a single company and employ the provided tool to identify the issues that are most negatively discussed in the company report. This information

can then be subjected to further investigation in greater detail. To provide a more objective measure, the model could be trained and applied to news data (Fischbach et al., 2023). This approach would diversify the sources of ESG information, potentially reducing bias and offering a broader perspective on ESG performance.

5.6 Conclusion

In this study, we developed an ABSA model that serves as a systematic tool for extracting entities related to ESG issues and their associated sentiment from company reports. Unlike previous studies that have employed NLP on company reports (Schimanski et al., 2024), our model provides more detailed insights into a company's perspective on ESG issues. While Schimanski et al. (2024) adopted a quantitative approach by examining the frequency of mentions of E, S, and G in company reports, our approach is more qualitative, incorporating the tone of reporting. This method provides deeper insights into a company's performance in relation to ESG matters beyond mere quantity. Furthermore, we employed a distinct time series analysis methodology. By estimating a pVAR, we examined the bidirectional effects of the variables, considering that ESG scores could influence the tone of reporting and vice versa. Despite this comprehensive approach, no significant relationship, except for one model, was found between the ESG scores provided by rating agencies and the sentiment scores. This result may be attributed to the controversy surrounding the quality of ESG scores, as discussed in Berg et al. (2022). Our correlation analysis between the ESG scores revealed a lack of consistency among the rating agencies, indicating inconsistencies in their methodologies. Consequently, there appears to be no predictive capacity of ESG sentiment in company reports on ESG scores from rating agencies.

The ABSA model may prove to be a valuable resource for stakeholders seeking insights into a company's stance on ESG issues. The tool can provide a comprehensive understanding of ESG topics and their associated tone at the entity level. This can be achieved by examining detailed topics or aggregating data at a higher level, focusing on specific ESG subcategories or even on the E, S, and G pillar levels. The tool can be applied to specific reports or all reports over time from a company to identify which ESG matters might be problematic or unproblematic at specific points in time. As company reports are inherently subjective, future research could apply our tool to news data, as suggested by Fischbach et al. (2023), to obtain a more objective measure. Given the ongoing debate regarding the quality of ESG scores from rating

agencies (Schimanski et al., 2024; Berg et al., 2022), our tool offers an alternative measure that aligns with the companies' own views on ESG matters. Furthermore, our findings indicate that the SetFit few-shot model with standard parameters yields superior outcomes on the same dataset, despite using only one-tenth of the training data.

Nevertheless, our analysis is subject to several limitations. In our time series analysis, we did not incorporate any exogenous variables, such as company fundamentals, which could have enhanced the analysis. The absence of a correlation between the scores may be attributed to this limitation, suggesting an opportunity for future research to include such variables. Additionally, the available data did not permit a time series analysis on an ESG subcategory basis. With more than five years of data from Refinitiv matched to our ESG subcategories, future research could explore relationships at a more granular level. Regarding the training process of our model, there is considerable scope for improvement. The ABSA model currently employs standard parameters. While a grid search could potentially improve the model, it requires significant computational resources, and the sustainability of such extensive training must be considered. Furthermore, comparing different base models, especially larger ones, may yield more favorable outcomes. Future research could also incorporate additional variables beyond those provided by ESG scores from rating agencies to assess the quality of our sentiment scores. A larger sample size, such as that from the EUROSTOXX600 or S&P 500, could facilitate the development of more robust time series models and provide insights into industry-specific ESG issues.

D Appendix to Chapter 5

Table 5.5: Distribution of Sentiment Labels in Datasets

Dataset	Set	Positive	Neutral	Negative	Total
Environment	Training	86	30	21	137
	Test	14	9	7	30
Sum		100	39	28	167
Social	Training	122	6	3	131
	Test	27	1	0	28
Sum		149	7	3	159
Governance	Training	119	23	4	146
	Test	33	3	2	38
Sum		152	26	6	184

Table 5.6: ESG Subcategories and Definitions - Environmental

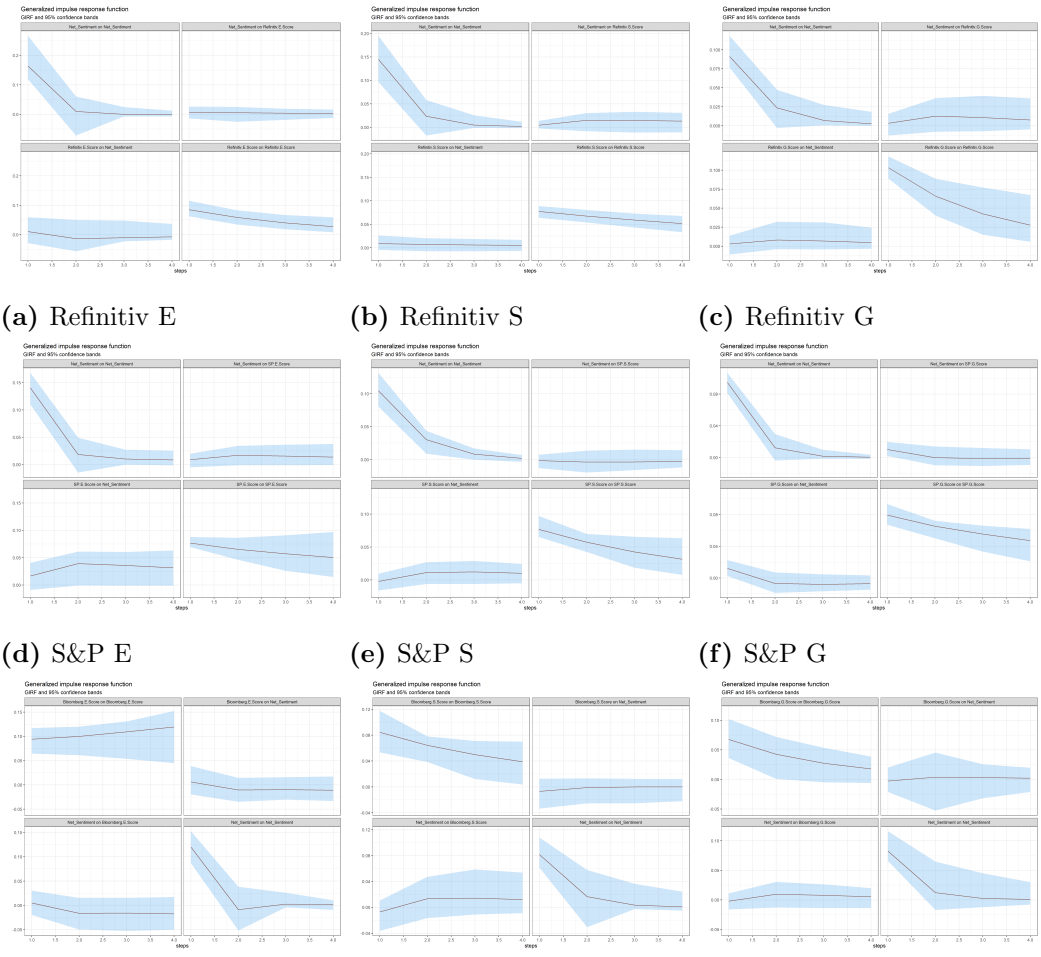
ESG Category	**Subcategory Definition**	**Example Entities**
Environmental	Climate Change Policies and Carbon Footprint: Measures the company's contribution to climate change through greenhouse gas emissions and carbon footprint management.	Greenhouse Gas Emission, Carbon Emission, Decarbonization, Greenhouse Gas, Climate Risk
	Energy Use: Assesses the company's energy efficiency and renewable energy usage.	Energy Efficiency, Energy, Renewable, Energy Sector, Fuel Efficiency
	Waste Management and Pollution: Evaluates waste management practices, pollution prevention, and handling of toxic emissions.	Recycling, Carbon Dioxide, Waste Management, Food Waste, Air Pollution
	Resource Depletion: Considers the company's use of resources, such as water and raw materials, and its impact on biodiversity.	Fuel Economy, Natural Resource, Forest, Resource Management, Coal
	Sustainability and Environmental Stewardship: Looks at the company's overall commitment to environmental sustainability practices.	Environment, Sustainability, Climate, Environmental, Sustainable Development

Table 5.7: ESG Subcategories and Definitions - Social

ESG Category	**Subcategory Definition**	**Example Entities**
Social	Employee Relations and Diversity: Involves employee treatment, diversity, labor standards, and fair wages.	Employee, Diversity, Health, Woman, Culture
	Customer Satisfaction and Data Protection: Focuses on product quality, customer service, data security, and privacy.	Customer, Fair Value, Customer Satisfaction, Cybersecurity
	Community Engagement and Impact: Looks at how the company contributes to the communities in which it operates, including charitable efforts and community service.	Community, Society, Social, Corporate Social Responsibility, Citizen
	Human Rights and Labor Practices: Assesses the company's adherence to fair labor practices, human rights, and avoiding exploitation.	Human Right, Discrimination, Harassment, Refugee, Child Labor
	Supply Chain Management: Evaluates the social aspects of the supply chain, including labor practices and human rights of suppliers.	Global Supply Chain, Supply Chain, Supplier, Contractor

Table 5.8: ESG Subcategories and Definitions - Governance

ESG Category	**Subcategory Definition**	**Example Entities**
Governance	Board Composition and Structure: Analyzes the diversity, independence, and expertise of board members.	Board, Management, Executive Board, Supervisory, Top Management
	Executive Compensation: Looks at how executives are compensated and whether it aligns with the company's long-term goals and shareholders' interests.	Renumeration, Management Renumeration, Supervisory Board Renumeration, Board Renumeration, Cash Renumeration
	Audit Committee Structure: Evaluates the quality and independence of internal audits and controls.	Audit, Compliance, Auditor, Tax, Accounting
	Business Ethics and Company Transparency: Considers ethical business practices, transparency in reporting, and avoiding conflicts of interest.	Risk Management, Fraud, Crisis Management, Business Ethics, Credit Risk Management
	Shareholder Rights: Examines the rights of shareholders and how well the company listens to and integrates their feedback.	Shareholder, Minority Shareholder, Ordinary Shareholder, Minority Interest, Shareholder Renumeration



(a) Refinitiv E **(b)** Refinitiv S **(c)** Refinitiv G

(d) S&P E **(e)** S&P S **(f)** S&P G

(g) Bloomberg E **(h)** Bloomberg S **(i)** Bloomberg G

Figure 5.8: Generalized Impulse Response Function (GIRF) of pVAR Model with 95% confidence bands.

The figure displays the Generalized Impulse Response Function (GIRF) of the pVAR model with 95% confidence bands for ratings obtained from Refinitiv, S&P, and Bloomberg. The confidence bands are based on 1000 bootstrap samples.

Table 5.9: Summary of ESG scores from S&P (Part 1).

Company Name	**From Year**	**To Year**	**Count of Years**
adidas ag	2014	2023	10
adyen nv	2019	2023	5
airbus se	2014	2023	10
allianz se	2014	2023	9
anheuserbusch inbev	2014	2023	10
asml holding	2014	2023	9
axa sa	2014	2023	9
banco bilbao	2014	2023	10
banco santander	2014	2023	9
basf se	2014	2023	10
bayer ag	2014	2023	10
bayerische motoren	2014	2023	10
bnp paribas	2014	2023	9
crh plc	2014	2023	9
danone sa	2014	2023	10
deutsche boerse	2014	2022	9
deutsche post	2014	2023	8
deutsche telekom	2014	2023	8
enel spa	2014	2023	9
eni spa	2014	2023	9
essilorluxottica sa	2014	2023	10
flutter entertainment	2016	2023	6
hermes international	2014	2023	10
iberdrola sa	2014	2023	9

Table 5.10: Summary of ESG scores from S&P (Part 2).

Company Name	From Year	To Year	Count of Years
industria de	2014	2023	8
infineon technologies	2014	2022	9
ing groep	2014	2023	8
intesa sanpaolo	2014	2023	9
kering sa	2014	2023	9
koninklijke ahold	2014	2023	10
lair liquide	2014	2023	10
loreal sa	2014	2023	10
lvmh moet	2014	2023	9
mercedes benz	2014	2023	9
muenchener rueckversicherungs	2014	2023	9
nokia oyj	2014	2023	10
nordea bank	2014	2023	9
pernod ricard	2014	2022	8
prosus nv	2020	2022	2
safran sa	2014	2023	10
sanofi sa	2014	2023	9
sap se	2014	2023	8
schneider electric	2014	2023	8
siemens ag	2014	2023	9
stellantis nv	2015	2023	9
totalenergies se	2014	2023	8
unicredit spa	2014	2023	9
vinci sa	2014	2023	9
volkswagen ag	2014	2023	9
vonovia se	2015	2023	9

Table 5.11: Summary of ESG ratings from Bloomberg (Part 1).

Company Name	**From Year**	**To Year**	**Count of Years**
adidas ag	2015	2022	8
adyen nv	2018	2022	5
airbus se	2015	2022	8
allianz se	2015	2022	8
anheuserbusch inbev	2015	2022	8
asml holding	2015	2022	8
axa sa	2015	2022	8
banco bilbao	2015	2022	8
banco santander	2015	2022	8
basf se	2015	2022	8
bayer ag	2015	2022	8
bayerische motoren	2015	2022	8
bnp paribas	2015	2022	8
crh plc	2015	2022	8
danone sa	2015	2022	8
deutsche boerse	2015	2022	8
deutsche post	2015	2022	8
deutsche telekom	2015	2022	8
enel spa	2015	2022	8
eni spa	2015	2022	8
essilorluxottica sa	2015	2022	8
flutter entertainment	2021	2022	2
hermes international	2015	2022	8
iberdrola sa	2015	2022	8

Table 5.12: Summary of ESG ratings from Bloomberg (Part 2).

Company Name	From Year	To Year	Count of Years
industria de	2015	2022	8
infineon technologies	2015	2022	8
ing groep	2015	2022	8
intesa sanpaolo	2015	2022	8
kering sa	2015	2022	8
koninklijke ahold	2016	2022	7
lair liquide	2015	2021	7
loreal sa	2015	2022	8
lvmh moet	2015	2022	8
mercedes benz	2015	2022	8
muenchener rueckversicherungs	2015	2022	8
nokia oyj	2015	2022	8
nordea bank	2015	2022	8
pernod ricard	2015	2022	8
prosus nv	2020	2022	3
safran sa	2015	2022	8
sanofi sa	2015	2022	8
sap se	2015	2022	8
schneider electric	2015	2022	8
siemens ag	2015	2022	8
stellantis nv	2015	2022	8
totalenergies se	2015	2022	8
unicredit spa	2015	2022	8
vinci sa	2015	2022	8
volkswagen ag	2015	2022	8
vonovia se	2015	2022	8

Table 5.13: Summary of ESG ratings from Refinitiv (Part 1).

Company Name	**From Year**	**To Year**	**Count of Years**
adidas ag	2002	2022	21
adyen nv	2018	2022	5
airbus se	2002	2022	21
allianz se	2002	2023	22
anheuserbusch inbev	2002	2022	21
asml holding	2002	2022	21
axa sa	2002	2022	21
banco bilbao	2002	2022	21
banco santander	2002	2022	21
basf se	2002	2022	21
bayer ag	2002	2022	21
bayerische motoren	2005	2022	18
bnp paribas	2002	2022	21
crh plc	2005	2022	18
danone sa	2005	2022	18
deutsche boerse	2002	2022	21
deutsche post	2005	2022	18
deutsche telekom	2002	2022	21
enel spa	2002	2022	21
eni spa	2002	2022	21
essilorluxottica sa	2002	2022	21
flutter entertainment	2005	2022	18
hermes international	2005	2022	18
iberdrola sa	2002	2022	21

Table 5.14: Summary of ESG ratings from Refinitiv (Part 2).

Company Name	From Year	To Year	Count of Years
industria de	2002	2023	22
infineon technologies	2002	2023	22
ing groep	2002	2022	21
intesa sanpaolo	2002	2022	21
kering sa	2002	2022	21
koninklijke ahold	2002	2023	21
lair liquide	2005	2022	18
loreal sa	2002	2022	21
lvmh moet	2002	2022	21
mercedes benz	2002	2022	21
muenchener rueckversicherungs	2002	2022	21
nokia oyj	2002	2022	21
nordea bank	2005	2023	19
pernod ricard	2002	2022	21
prosus nv	2020	2023	4
safran sa	2002	2022	21
sanofi sa	2002	2022	21
sap se	2002	2022	21
schneider electric	2002	2022	21
siemens ag	2002	2023	22
stellantis nv	2002	2022	21
totalenergies se	2002	2022	21
unicredit spa	2007	2022	16
vinci sa	2002	2022	21
volkswagen ag	2002	2022	21
vonovia se	2015	2022	8

Part IV

Conclusion

Chapter 6

Conclusion

This PhD thesis is comprised of four principal parts. The first part serves as an introduction, while the final part provides a conclusion. Part two encompasses papers that examine textual content within the context of sustainable development, with a particular focus on the Sustainable Development Goals (SDGs). The third part of the thesis is devoted to the analysis of sustainability in company reports. In both parts, the focus is on the creation of indices that can be used either standalone of or in conjunction with other key data, or on the analysis of the distribution of discourse on environmental, social, and governance topics. The following presents a brief summary of the findings from the aforementioned studies, along with a discussion of potential limitations and avenues for future research.

The first paper, presented with the title *Reading Between the Lines: The Intersection of Research Attention and Sustainable Development Goals*, introduces a Research Attention Index, which gauges research attention by calculating the quantity of each SDG discussed in the scientific literature. We employ a zero-shot model for text classification, where the output is a probability between 0 and 100 percent. The Research Attention Index is comprised of outputs for each of the 17 SDGs. Additionally, the index is compared to the official SDG scores provided by the United Nations, revealing a complex non-linear relationship between them. This study is limited by the interpretation of the index as a relative measure of research attention and the inability to provide an absolute measure of research attention. Moreover, the SDG scores are a composite of indices derived from different time periods, which precludes the possibility of conducting a time series analysis. The data employed in this study were drawn exclusively from the Web of Science. However, the inclusion of data from other sources could be considered in future research. Furthermore, the incorporation of additional key data could facilitate a more comprehensive comparison of our index

with other indices. Another limitation of this study is that potential confounding variables were not accounted for. These variables could include variables that might influence the relationship between the index and the SDG scores.

The second paper in part two is entitled: *Finding common development paths in voluntary national reviews reporting on sustainable development goals using aspect-based sentiment analysis*. In this paper, we develop a sentiment index by analyzing the tone in Voluntary National Reviews (VNRs) using an aspect-based sentiment analysis model. For each SDG, an index is created, ranging from -1 to 1. A value of -1 represents the most negative sentiment, while a value of 1 represents the most positive sentiment. Additionally, we compare our sentiment indices with the official SDG scores provided by the UN. A positive correlation was observed between the index and the SDG score for SDG 2 (Zero Hunger) and SDG 11 (Sustainable Cities and Communities), while a negative correlation was found for SDG 5 (Gender Equality). In the case of the remaining SDGs, no relationship was identified. This indicates that the tone of the VNRs may not accurately reflect the actual progress being made towards achieving the SDGs. Moreover, we cluster the reporting countries by their reporting tone towards all SDGs and identify clusters of countries that may benefit from collaboration. Nevertheless, the disparity in reporting sentiment is more pronounced among LDCs, EMs, and FMs than in developed countries. However, it should be noted that the study also has some potential limitations. Some reports are not published in English, which necessitates the translation of those reports. This process may result in a reduction in the quality of the data. Furthermore, there is a discrepancy in the reports from the same country over different periods. This discrepancy may be attributed to external events, such as the impact of the COVID-19 pandemic. Unfortunately, it is not possible to conduct a time series analysis due to the lack of quality SDG scores as data. Future research could employ different or more sophisticated natural language processing techniques than those proposed in this study to further investigate the sensitivity of such valuable reports. Another potential research gap is the analysis of reports at a more local level, given that this study is based on reports provided at the country level. Moreover, there is potential for further analysis to identify the reasons behind the discrepancies observed between the sentiment scores.

The third paper presented in this thesis belongs to part three, which focuses on environmental, social, and governance topics in company reports. In this paper, we analyze company reports published by banks that have committed to publish in accordance with the TCFD guidelines. We employ a zero-shot model for text classifi-

cation, and our findings indicate that climate-related reporting increased following the introduction of the guidelines. Nevertheless, we find that not all guidelines are addressed in the same proportion. Consequently, some recommendations appear to be not met, which might be indicative of selective reporting. The limitations of this study include the possibility that an increase in reporting does not necessarily lead to actions by banks. Future research could investigate the factors that lead to an increase in climate-related reporting. As we have identified instances where banks have not addressed all of the topics included in the TCFD recommendations, future research could investigate the motivations behind this selective reporting. Additionally, we have identified weaknesses in the TCFD recommendations, such as the use of vague wage formulations and the inclusion of overlapping topics. Future research could investigate how climate-related disclosure recommendations could improve the quality of reporting.

The final paper in this thesis also forms part of the third part of this thesis. The data analyzed in this paper comprises company reports published by EUROSTOXX50 companies. An ABSA model is employed to generate a sentiment index for ESG subcategories. The results from our model can be accessed in two ways: on a very fine-grained entity level or grouped to 15 ESG subcategories. The indices in question provide a subjective measure of the companies' views on ESG matters, and may therefore serve as an alternative to ESG scores provided by rating agencies. Furthermore, we employ pVARs to assess the relationship between our indices and ESG ratings. Our findings indicate that there is no discernible correlation between any of the sentiment indices and ESG ratings. However, this lack of relationship may also be attributed to the quality of the ESG ratings themselves, a topic that has been extensively discussed in the literature (Berg et al., 2022; Schimanski et al., 2024).

In conclusion, this doctoral dissertation proposes the use of text-based indicators as an alternative to official scores in the context of the SDGs and ESG indices. Such indicators facilitate the systematic extraction of information from available text that offers additional valuable insights. Moreover, the indices endeavor to quantify qualitative information, and the methodologies employed in this thesis can be applied to other sustainability contexts. As the field of NLP continues to evolve, future research may enhance the quality of the indicators developed in this thesis by employing more sophisticated models. Furthermore, the collection of more accurate data will permit the comparison of these indices with other matrices. Moreover, the models employed in this thesis can be applied to other forms of textual data, thus facilitating research based on different text data, such as news or social media data.

Bibliography

Abidoye, B., Felix, J., Kapto, S., & Patterson, L. (2021). Leaving no one behind: Impact of covid-19 on the sustainable development goals (SDGs). *New York, NY and Denver CO: United Nations Development Programme and Frederick S. Pardee Center for International Futures*. Retrieved from `https://sdgintegration.undp.org/sites/default/files/Leaving_No_One_Behind,_COVID_impact_on_the_SDGs_second_flagship.pdf` (Last accessed June 25, 2024)

Allen, C., Metternicht, G., Wiedmann, T., & Pedercini, M. (2019). Greater gains for australia by tackling all SDGs but the last steps will be the most challenging. *Nature Sustainability*, *2*(11), 1041–1050. doi: 10.1038/s41893-019-0409-9

Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angin, P., & Dikmener, G. (2022). A RoBERTa approach for automated processing of sustainability reports. *Sustainability*, *14*(23). doi: 10.3390/su142316139

Anselmi, D., D'Adamo, I., Gastaldi, M., & Lombardi, G. V. (2023). A comparison of economic, environmental and social performance of european countries: A sustainable development goal index. *Environment, Development and Sustainability*, 1–25. doi: 10.1007/s10668-023-03496-3

Auzepy, A., Tönjes, E., Lenz, D., & Funk, C. (2023). Evaluating TCFD reporting — a new application of zero-shot analysis to climate-related financial disclosures. *PLOS ONE*, *18*(11), e0288052. doi: 10.1371/journal.pone.0288052

Bali Swain, R., & Ranganathan, S. (2021). Modeling interlinkages between sustainable development goals using network analysis. *World Development*, *138*, 105136. doi: 10.1016/j.worlddev.2020.105136

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370. doi: 10.1037/1089-2680.5.4.323

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72. doi: 10.1162/tacl_a_00254

Bellantuono, L., Monaco, A., Amoroso, N., Aquaro, V., Lombardi, A., Tangaro, S., & Bellotti, R. (2022). Sustainable development goals: conceptualization, communication and achievement synergies in a complex network framework. *Applied Network Science*, *7*(1), 14. doi: 10.1007/s11625-022-01093-3

Beltagy, I., Cohan, A., & Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *arXiv:1903.10676[Preprint]*. Retrieved from `http://arxiv.org/abs/1903.10676` (Last accessed Feb. 28, 2022)

Bennich, T., Weitz, N., & Carlsen, H. (2020). Deciphering the scientific literature on sdg interactions: A review and reading guide. *The Science of the Total Environment*, *728*, 138405. doi: 10.1016/j.scitotenv.2020.138405

Berg, F., Koelbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, *26*(6), 1315–1344. doi: 10.1093/rof/rfac033

Beyene, W., De Greiff, K., Delis, M., & Ongena, S. (2021). Too-big-to-strand: Bond to bank substitution in the transition to a low-carbon economy. *SSRN Electronic Journal*. Retrieved from `https://ssrn.com/abstract=3960296` (Last accessed June 25, 2024)

Beyene, W., Ongena, S., & Delis, M. (2022). *Financial institutions' exposures to fossil fuel assets: An assessment of financial stability concerns in the short term and in the long run, and possible solution.* Retrieved from `https://www.europarl.europa.eu/RegData/etudes/STUD/2022/699532/IPOL_STU(2022)699532_EN.pdf` (Last accessed May 22, 2023)

Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, *47*, 102776. doi: 10.1016/j.frl.2022.102776

Boffo, R., & Patalano, R. (2020). *ESG investing: Practices, progress and challenges.* Paris. Retrieved from `https://www.oecd.org/finance/ESG-Investing-Practices-Progress-and-Challenges.pdf` (Last accessed June 25, 2024)

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, *142*(2), 517–549. doi: 10.1016/j.jfineco.2021.05.008

Brulhardt, M. (2021). *Where should i eat after the pandemic?: Decision making with aspect-based sentiment analysis using transformers.* Retrieved from `https://github.com/mwbrulhardt/yelp-absa` (Last accessed June 25, 2024)

CDSB. (2019). *TCFD good practice handbook 1st edition.* Retrieved from `https://www.cdsb.net/sites/default/files/tcfd_good_practice_handbook_web_a4.pdf` (Last accessed May 22, 2023)

CDSB. (2021). *TCFD good practice handbook 2nd edition.* Retrieved from `https://www.cdsb.net/sites/default/files/tcfd_good_practice_handbook_v5_pages.pdf` (Last accessed May 22, 2023)

Chang, I.-C., Yu, T.-K., Chang, Y.-J., & Yu, T.-Y. (2021). Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals. *Sustainability*, *13*(19), 10856. doi: 10.3390/su131910856

Claessens, S., & Yurtoglu, B. B. (2013). Corporate governance in emerging markets: A survey. *Emerging Markets Review*, *15*, 1–33. doi: 10.1016/j.ememar.2012.03.002

D'Adamo, I., Gastaldi, M., Imbriani, C., & Morone, P. (2021). Assessing regional performance for the sustainable development goals in italy. *Scientific Reports*, *11*(1), 24117. doi: 10.1038/s41598-021-03635-8

Davison, J. (2020). *Zero-shot learning in modern NLP.* Retrieved from `https://joeddav.github.io/blog/2020/05/29/ZSL.html` (Last accessed May 22, 2023)

Decouttere, C., De Boeck, K., & Vandaele, N. (2021). Advancing sustainable development goals through immunization: a literature review. *Globalization and Health*, *17*(1), 95. Retrieved from `10.1186/s12992-021-00745-w`

Demaria, S., & Rigot, S. (2020). Corporate environmental reporting: Are french firms compliant with the task force on climate financial disclosures. *Business Strategy and the Environment*, *30*(1), 721–738. doi: 10.1002/bse.2651

Department for Business, E. . I. S. (2021). *Consultation response: Mandatory climate-related financial disclosures by publicly quoted companies, large private companies, and llps.* Retrieved from `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1029354/tcfd-consultation-government-response.pdf` (Last accessed May 22, 2023)

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805[Preprint]*. Retrieved from `http://arxiv.org/abs/1810.04805` (Last accessed Feb. 28, 2022)

Dias, J., Salgado, E., Barbosa, S., Alvarenga, A., & Lira, J. (2017). Assessment of the sustainability of countries at worldwide. *Journal of Management and Sustainability*, *7*, 51. doi: 10.5539/jms.v7n4p51

Diaz-Sarachaga, J. M., Jato-Espino, D., & Castro-Fresno, D. (2018). Is the sustainable development goals (SDG) index an adequate framework to measure the progress of the 2030 agenda? *Sustainable Development*, *26*(6), 663-671. doi: 10.1002/sd.1735

Ding, D., Liu, B., & Chang, M. (2022). Carbon emissions and tcfd aligned climate-related information disclosures. *Journal of Business Ethics*, *182*, 967–1001. doi: 10.1007/s10551-022-05292-x

Fischbach, J., Adam, M., Dzhagatspanyan, V., Mendez, D., Frattini, J., Kosenkov, O., & Elahidoost, P. (2023). Automatic esg assessment of companies by mining and evaluating media coverage data: Nlp approach and tool. In *2023 ieee international conference on big data (bigdata)* (p. 2823-2830). doi: 10.1109/BigData59044.2023.10386488

Friederich, D., Kaack, L. H., Luccioni, A., & Steffen, B. (2021). Automated identification of climate risk disclosures in annual corporate reports. *arXiv:2108.01415 [Preprint]*. Retrieved from `https://arxiv.org/abs/2108.01415` (Last accessed June 25, 2024)

FSB. (2015, November 9). *Proposal for a disclosure task force on climate-related risks.* Retrieved from `https://www.fsb.org/2015/11/disclosure-task-force-on-climate-related-risks-2/` (Last accessed May 22, 2023)

Fuso Nerini, F., Sovacool, B., Hughes, N., Cozzi, L., Cosgrave, E., Howells, M., ... Milligan, B. (2019). Connecting climate action with other sustainable development goals. *Nature Sustainability*, *2*(8), 674–680. doi: 10.1038/s41893-019-0334-y

Föhr, T. L., Schreyer, M., Juppe, T. A., & Marten, K.-U. (2023). Assuring sustainable futures: Auditing sustainability reports using ai foundation models. *SSRN*

Electronic Journal. Retrieved from `https://ssrn.com/abstract=4502549` (Last accessed 25 June, 2024)

Gibson, R., Glossner, S., Krueger, P., Matos, P., & Steffen, T. (2022). Do responsible investors invest responsibly? *Review of Finance*, *26*(6), 1389–1432. doi: 10.1093/rof/rfac064

Gillan, S. L., Koch, A., & Starks, L. T. (2021). Firms and social responsibility: A review of ESG and CSR research in corporate finance. *Journal of Corporate Finance*, *66*, 101889. doi: 10.1016/j.jcorpfin.2021.101889

Goloshchapova, I., Poon, S.-H., Pritchard, M., & Reed, P. (2019). Corporate social responsibility reports: topic analysis and big data approach. *The European Journal of Finance*, *25*(17), 1637-1654. doi: 10.1080/1351847X.2019.1572637

Hackl, A. (2018). Mobility equity in a globalized world: Reducing inequalities in the sustainable development agenda. *World Development*, *112*, 150-162. doi: 10.1016/j.worlddev.2018.08.005

Hegre, H., Petrova, K., & von Uexkull, N. (2020). Synergies and trade-offs in reaching the sustainable development goals. *Sustainability*, *12*(20), 8729. doi: 10.3390/su12208729

Jain, Y., Gupta, S., Yalciner, S., Joglekar, Y. N., Khetan, P., & Zhang, Q. (2023). Overcoming complexity in esg investing: The role of generative ai integration in identifying contextual esg factors. *SSRN Electronic Journal*. Retrieved from `https://ssrn.com/abstract=4495647`

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 151–160). Portland, Oregon, USA. doi: 10.18653/v1/P11-1016

Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth acm international conference on web search and data mining* (pp. 815–824). New York, NY, USA. doi: 10.1145/1935826.1935932

Jung, J., Herbohn, K., & Clarkson, P. (2018). Carbon risk, carbon risk awareness and the cost of debt. *Journal of Business Ethics*, *150*, 1151–1171. doi: 10.1007/s10551-016-3207-6

Kearney, C. (2012). Emerging markets research: Trends, issues and future directions. *Emerging Markets Review*, *13*(2), 159–183. doi: 10.1016/j.ememar.2012.01.003

King, D. A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316. doi: 10.1038/430311a

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *Icml deep learning workshop* (Vol. 2). Retrieved from `https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf` (Last accessed June 25, 2024)

Kravets, O., & Sandikci, O. (2014). Competently ordinary: New middle-class consumers in the emerging markets. *Journal of Marketing*, *78*(4), 125–140. doi: 10.1509/jm.12.0190

Kroll, C., Warchold, A., & Pradhan, P. (2019). Sustainable development goals (SDGs): Are we successful in turning trade-offs into synergies? *Palgrave Communications*, *5*(1). doi: 10.1057/s41599-019-0335-5

Krueger, P., Sautner, Z., & Starks, L. T. (2020). The importance of climate risks for institutional investors. *The Review of Financial Studies*, *33*(3), 1067–1111. doi: 10.1093/rfs/hhz137

Laperdon, R., Aarsen, T., Tunstall, L., Korat, D., Pereg, O., & Wasserblat, M. (2023). *Setfitabsa: Few-shot aspect based sentiment analysis using setfit.* Retrieved from `https://huggingface.co/blog/setfit-absa` (Last accessed June 07, 2024)

Le Blanc, D. (2015). Towards integration at last? the sustainable development goals as a network of targets. *Sustainable Development*, *23*(3), 176–187. doi: 10.1002/sd.1582

Lee, H., Lee, S. H., Lee, K. R., & Kim, J. H. (2023). ESG discourse analysis through bertopic: comparing news articles and academic papers. *Computers, Materials & Continua*, *75*(3), 6023–6037. doi: 10.32604/cmc.2023.039104

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2019). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv:1910.13461[Preprint]*. Retrieved from `http://arxiv.org/abs/1910.13461` (Last accessed May 22, 2023)

Li, X., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting bert for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th workshop on noisy user-generated text (w-nut 2019)* (pp. 34–41). doi: 10.18653/v1/D19-5505

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv:2205.05638 [Preprint].* Retrieved from `https://arxiv.org/abs/2205.05638` (Last accessed June 25, 2024)

Luccioni, A., Baylor, E., & Duchene, N. (2020). Analyzing sustainability reports using natural language processing. *arXiv:2011.08073 [Preprint].* Retrieved from `https://arxiv.org/abs/2011.08073` (Last accessed June 25, 2024)

Ma, E. (2019). *Nlp augmentation.* Retrieved from `https://github.com/makcedward/nlpaug` (Last accessed June 25, 2024)

Mehra, S., Louka, R., & Zhang, Y. (2022). ESGBERT: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv:2203.16788 [Preprint].* Retrieved from `https://arxiv.org/abs/2203.16788` (Last accessed May 22, 2023)

Moinuddin, M. (Ed.). (2017). *Sustainable development goals interlinkages and network analysis: A practical tool for SDG integration and policy coherence.* Institute for Global Environmental Strategies. Retrieved from `https://www.greenpolicyplatform.org/research/sustainable-development-goals-interlinkages-and-network-analysis-practical-tool-sdg` (Last accessed June 25, 2024)

MSCI Inc. (2021). *MSCI global market accessibility review.* Retrieved from `https://www.msci.com/documents/1296102/1330218/MSCI_2021_Global_Market_Accessibility_Review_Report.pdf/d88d8bc0-a882-58c7-35f0-bef191e0ebe2` (Last accessed June 25, 2024)

Mugellini, G., Villeneuve, J.-P., & Heide, M. (2021). Monitoring sustainable development goals and the quest for high-quality indicators: Learning from a practical evaluation of data on corruption. *Sustainable Development*, *29*(6), 1257–1275. doi: 10.1002/sd.2223

Murphy, E., Walsh, P. P., & Murphy, E. (2023). Nation-based peer assessment of europe's sustainable development goal performance. *PLOS ONE*, *18*(6), e0287771. doi: 10.1371/journal.pone.0287771

Nilsson, M., Chisholm, E., Griggs, D., Howden-Chapman, P., McCollum, D., & Messerli, P. e. a. (2018). Mapping interactions between the sustainable development goals: Lessons learned and ways forward. *Sustainability Science*, *13*(6), 1489–1503. doi: 10.1007/s11625-018-0604-z

OECD. (2002). *Foreign direct investment for development.* OECD Publishing. doi: 10.1787/9789264199286-en

OECD. (2021). *Development co-operation report 2021.* OECD Publishing. doi: 10.1787/ce08832f-en

Pedercini, M., Arquitt, S., Collste, D., & Herren, H. (2019). Harvesting synergy from sustainable development goal interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(46), 23021–23028. doi: 10.1073/pnas.1817276116

Pham-Truffert, M., Metz, F., Fischer, M., Rueff, H., & Messerli, P. (2020). Interactions among sustainable development goals: Knowledge for identifying multipliers and virtuous cycles. *Sustainable Development*, *28*(5), 1236–1250. doi: 10.1002/sd.2073

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., & AL-Smadi, M. e. a. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 19–30). San Diego, California. doi: 10.18653/v1/S16-1002

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)* (pp. 486–495). Denver, Colorado. doi: 10.18653/v1/S15-2082

Pradhan, P. (2023). A threefold approach to rescue the 2030 agenda from failing. *National Science Review*, *10*(7), 2095–5138. doi: 10.1093/nsr/nwad015

Pradhan, P., Costa, L., Rybski, D., Lucht, W., & Kropp, J. P. (2017). A systematic study of sustainable development goal (SDG) interactions. *Earth's Future*, *5*(11), 1169–1179. doi: 10.1002/2017EF000632

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9. Retrieved from `https://api.semanticscholar.org/CorpusID:160025533` (Last accessed June 25, 2024)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683 [Preprint]*. Retrieved from `https://arxiv.org/abs/1910.10683` (Last accessed June 25, 2024)

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296-320. doi: 10.1207/S15327957PSPR0504_2

Sachs, J., Kroll, C., Lafortune, G., Fuller, G., & Woelm, F. (2021). *Sustainable development report 2021*. Cambridge University Press. doi: 10.1017/9781009106559

Sachs, J., Kroll, C., Lafortune, G., Fuller, G., & Woelm, F. (2022). *Sustainable development report 2022*. Cambridge University Press. doi: 10.1017/9781009210058

Sachs, J., Schmidt-Traub, G., Kroll, C., Lafortune, G., & Fuller, G. (2019). *Sustainable development report 2019*. Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN). This work is licensed under a Creative Commons Attribution 4.0 International License.

Sachs, J., Schmidt-Traub, G., Kroll, C., Lafortune, G., Fuller, G., & Woelm, F. (2021). *Sustainable development report 2020: The sustainable development goals and covid-19 includes the SDG index and dashboards*. Cambridge University Press. doi: 10.1017/9781108992411

Sachs, J. D., Schmidt-Traub, G., Mazzucato, M., Messner, D., Nakicenovic, N., & Rockström, J. (2019). Six transformations to achieve the sustainable development goals. *Nature Sustainability*, *2*(9), 805–814. doi: 10.1038/s41893-019-0352-9

Saeidi, M., Bouchard, G., Liakata, M., & Riedel, S. (2016). Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. *arXiv:1610.03771 [Preprint]*. Retrieved from `https://arxiv.org/abs/1610.03771` (Last accesed Nov. 10 2023)

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in esg measurement: Using nlp to quantify environmental,

social, and governance communication. *Finance Research Letters*, *61*, 104979. doi: 10.1016/j.frl.2024.104979

Sebestyén, V., Domokos, E., & Abonyi, J. (2020). Focal points for sustainable development strategies-text mining-based comparative analysis of voluntary national reviews. *Journal of Environmental Management*, *263*, 110414. doi: 10.1016/j.jenvman.2020.110414

Shen, Z., Lo, K., Wang, L. L., Kuehl, B., Weld, D. S., & Downey, D. (2021). VILA: Improving structured content extraction from scientific pdfs using visual layout groups. *arXiv:2106.00676[Preprint]*. Retrieved from `https://arxiv.org/abs/2106.00676` (Last accessed Feb. 28, 2022)

Sigmund, M., & Ferstl, R. (2021). Panel vector autoregression in R with the package panelvar. *The Quarterly Review of Economics and Finance*, *80*, 693–720. doi: 10.1016/j.qref.2019.01.001

Smith, T. B., Vacca, R., Mantegazza, L., & Capua, I. (2021). Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals. *Scientific Reports*, *11*(1), 22427. doi: 10.1038/s41598-021-01801-6

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, *26*. Retrieved from `https://doi.org/10.48550/arXiv.1301.3666` (Last accessed Feb. 28, 2022)

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892. doi: 10.1073/pnas.1908369116

Subramaniam, N., Wahyuni, D., Cooper, B. J., Leung, P., & Wines, G. (2015). Integration of carbon risks and opportunities in enterprise risk management systems: evidence from australian firms. *Journal of Cleaner Production*, *96*, 407–417. doi: 10.1016/j.jclepro.2014.02.013

Sullivan, R., & Gouldson, A. (2012). Does voluntary carbon reporting meet investors' needs? *Journal of Cleaner Production*, *36*, 60–67. doi: 10.1016/j.jclepro.2012.02.020

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 conference*

of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 380–385). Minneapolis, Minnesota. doi: 10.18653/v1/N19-1035

TCFD. (2017a). *Final report: Recommendations of the task force on climate-related financial disclosures* (Tech. Rep. No. 11 (1)). Switzerland: TCFD. Retrieved from `https://assets.bbhub.io/company/sites/60/2020/10/FINAL-2017-TCFD-Report-11052018.pdf` (Last accessed June 25, 2024)

TCFD. (2017b). *Implementing the recommendations of the task force on climate-related financial disclosures.* Retrieved from `https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf` (Last accessed May 22, 2023)

The World Bank. (2020). *Capital markets.* Retrieved from `https://www.worldbank.org/en/topic/financialsector/brief/capital-markets` (Last accessed June 25, 2024)

Tkaczyk, D., Czeczko, A., Rusek, K., Bolikowski, L., & Bogacewicz, R. (2012). GROTOAP: Ground truth for open access publications. In *Proceedings of the 12th acm/ieee-cs joint conference on digital libraries* (pp. 381–382). doi: 10.1145/2232817.2232901

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288 [Preprint]*. Retrieved from `https://arxiv.org/abs/2307.09288` (Last accessed June 25, 2024)

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv:2209.11055 [Preprint]*. Retrieved from `https://arxiv.org/abs/2209.11055` (Last accessed June 25, 2024)

United Nations. (2004). *Who cares wins: Connecting financial markets to a changing world: Technical report* (Tech. Rep.). United Nations Global Compact. Retrieved from `https://www.unepfi.org/fileadmin/events/2004/stocks/who_cares_wins_global_compact_2004.pdf` (Last accessed June 25, 2024)

United Nations. (2016a). *A nexus approach for the SDGs.* Retrieved from `https://sdgtoolkit.org/tool/a-nexus-approach-for-the-sdgs-interlinkages-between-the-goals-and-targets/` (Last accessed Feb. 28, 2022)

United Nations. (2016b). *Transforming our world: The 2030 agenda for sustainable development.* Retrieved from `https://sdgs.un.org/2030agenda` (Last accessed Feb. 28, 2022)

United Nations. (2019). *Unlocking capital markets to finance the SDGs.* Retrieved from `https://www3.weforum.org/docs/WEF_Unlocking_Capital_Markets_to_Finance_the_SDGs_2019.pdf` (Last accessed June 25, 2024)

United Nations. (2021). *Sustainable development report 2021 supplementary online materials.* Retrieved from `https://dashboards.sdgindex.org/downloads` (Last accessed Feb. 28, 2022)

United Nations Department of Economic and Social Affairs. (2019). *Handbook for the preparation of voluntary national reviews, the 2019 edition.* Retrieved from `https://sustainabledevelopment.un.org/content/documents/20872VNR{_}hanbook{_}2019{_}Edition{_}v3.pdf` (Last accessed Feb. 28, 2022)

United Nations Economic and Social Council. (2016). *Progress towards the sustainable development goals: Report of the secretary-general.* Retrieved from `http://undocs.org/E/2016/75` (Last accessed Feb. 28, 2022)

United Nations Economic and Social Council. (2017). *Progress towards the sustainable development goals: Report of the secretary-general.* Retrieved from `http://undocs.org/E/2017/66` (Last accessed Feb. 28, 2022)

United Nations Economic and Social Council. (2018). *Progress towards the sustainable development goals: Report of the secretary-general.* Retrieved from `http://undocs.org/E/2018/64` (Last accessed Feb. 28, 2022)

United Nations Economic and Social Council. (2019). *Progress towards the sustainable development goals: Report of the secretary-general.* Retrieved from `http://undocs.org/E/2019/68` (Last accessed Feb. 28, 2022)

United Nations Economic and Social Council. (2020). *Progress towards the sustainable development goals: Report of the secretary-general.* Retrieved from `http://undocs.org/E/2020/57` (Last accessed Feb. 28, 2022)

United Nations Economic and Social Council. (2021). *Progress towards the sustainable development goals: Report of the secretary-general.* Retrieved from `http://undocs.org/E/2021/58` (Last accessed Feb. 28, 2022)

United Nations Global Compact. (2019). *SDG bonds — leveraging capital markets for the sdgs.* Retrieved from `https://www.unglobalcompact.org/library/5713` (Last accessed June 25, 2024)

United Nations IATF. (2022). *Financing for sustainable development report 2022.* United Nations. Retrieved from `https://www.un.org/ohrlls/sites/www.un.org.ohrlls/files/fsdr_2022.pdf` (Last accessed June 25, 2024)

United Nations Statistics Division. (2006). *Methodology - standard country or area codes for statistical use (M49).* Retrieved from `https://unstats.un.org/unsd/methodology/m49/` (Last accessed June 25, 2024)

Vo, D. T., & Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the 24th international conference on artificial intelligence (IJCAI'15)* (pp. 1347–1353). Retrieved from `https://www.ijcai.org/Proceedings/15/Papers/194.pdf` (Last accessed June 25, 2024)

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBert: A pretrained language model for climate-related text. *arXiv:2110.12010 [Preprint]*. Retrieved from `https://arxiv.org/abs/2110.12010` (Last accessed May 22, 2023)

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.emnlp-demos.6` (Last accessed Feb. 28, 2022)

Xiao, H., Liu, Y., & Ren, J. (2023). Synergies and trade-offs across sustainable development goals: A novel method incorporating indirect interactions analysis. *Sustainable Development*, *31*(2), 1135–1148. doi: 10.1002/sd.2446

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (Vol. 1, pp. 1192–1200). doi: 10.1145/3394486.3403172

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv:1909.00161 [Preprint]*.

Retrieved from `https://arxiv.org/abs/1909.00161` (Last accessed Feb. 28, 2022)

Zhang, L., Wang, S., & Liu, B. e. a. (2018). Deep learning for sentiment analysis: A survey. *arXiv:1801.07883 [Preprint]*. Retrieved from `https://arxiv.org/abs/1801.07883` (Last accesed Nov. 10 2023)