Unsupervised Sustainability Report Labeling Based on the Integration of the GRI and SDG Standards

Seyed Alireza Mousavian Anaraki, Danilo Croce and Roberto Basili
Department of Enterprise Engineering
University of Rome, Tor Vergata
Via del Politecnico 1, 00133, Rome, Italy
seyedalireza.mousaviananaraki@students.uniroma2.eu
{croce,basili}@info.uniroma2.it

Abstract

Sustainability reports are key instruments for communicating corporate impact, but their unstructured format and varied content pose challenges for large-scale analysis. This paper presents an unsupervised method to annotate paragraphs from sustainability reports against both the Global Reporting Initiative (GRI) and Sustainable Development Goals (SDG) standards. The approach combines structured metadata from GRI content indexes, official GRI–SDG mappings, and text semantic similarity models to produce weakly supervised annotations at scale. To evaluate the quality of these annotations, we train a multi-label classifier on the automatically labeled data and evaluate it on the trusted OSDG Community Dataset. The results show that our method yields meaningful labels and improves classification performance when combined with human-annotated data. Although preliminary, this work offers a foundation for scalable sustainability analysis and opens future directions toward assessing the credibility and depth of corporate sustainability claims.

1 Introduction

Sustainability reporting is increasingly critical as organizations around the world address urgent global challenges such as climate change. Clear and standardized disclosures on how companies contribute to sustainability goals help stakeholders assess corporate impacts and inform responsible investments. Among several recognized sustainability frameworks, two have emerged as particularly influential: the United Nations Sustainable Development Goals (SDGs) (UN, 2015), which establish high-level sustainability targets, and the Global Reporting Initiative (GRI)[1], which provides detailed disclosure guidelines for organizations.

With sustainability gaining increasing global attention (from climate change to poverty and gender equality), it is becoming essential to understand how companies report and communicate their actions in these areas. However, sustainability reports are typically extensive, complex, and unstructured documents, making manual annotation and information extraction challenging, costly, and error-prone. Despite the significance of climate-focused analysis, the use of NLP to systematically evaluate how companies align their reports with global climate-related sustainability frameworks remains significantly underexplored. For example, consider the following excerpt from Pfizer's recent sustainability report: "*Pfizer was one of the initial signatories to the U.S. Department of Health and Human Services (HHS) climate pledge. The pledge, launched in 2022, calls on stakeholders in the U.S. healthcare system (including hospitals, health systems, payers, suppliers, and pharmaceutical companies) to reduce GHG emissions and build a more climate-resilient healthcare infrastructure. By signing, we committed to reduce GHG emissions, publicly report our progress, and develop a climate resiliency plan.*"

This paragraph explicitly addresses climate action and emission reductions. Using our proposed method, the paragraph can be automatically linked to specific SDG and GRI categories because the text clearly refers to underlying principles and requirements of these frameworks. In particular, it can be annotated as follows:

- **SDG 13** (CLIMATE): "*Take urgent action to combat climate change and its impacts.*"

- **GRI 305** (EMISSIONS), specifically disclosure **GRI 305-5**: "*Reduction of GHG emissions.*"

This automated labeling approach bridges the gap between structured sustainability frameworks and unstructured corporate reports, facilitating a

[1] https://www.globalreporting.org/standards/

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 151–162
July 31, 2025 ©2025 Association for Computational Linguistics

large-scale, efficient, and systematic analysis of climate-related disclosures.

In this work, we propose an unsupervised annotation pipeline designed to simplify the identification and annotation of paragraphs within lengthy sustainability reports. Given the significant costs and limitations associated with manual annotation (including expense, time consumption, and potential interannotator disagreements), our goal is to reduce the manual effort by automatically suggesting high-confidence annotations, which can subsequently be verified by domain experts.

We leverage structured human-generated metadata known as GRI content indexes, as in (Nechaev and Hain, 2023). These indices, typically included in sustainability reports, explicitly link GRI standards (covering various environmental, economic, and social impacts) to specific report pages. By focusing on these pages, we narrow the search space, ensuring that we analyze only the potentially relevant sections of otherwise lengthy and complex documents. Given these candidate pages, we use established mappings between GRI codes and SDGs to further constrain the possible annotations. Such mappings significantly reduce ambiguity by limiting the combinatorial explosion that arises from jointly considering 33 GRI codes and 17 SDGs.

Finally, we apply semantic similarity methods based on a pre-trained sentence encoder, such as (Devlin et al., 2019) or (Reimers and Gurevych, 2019), to compare each paragraph against textual definitions of the corresponding GRI disclosures and SDG targets. This step allows us to rank and select the most plausible annotation pairs (GRI, SDG) for each paragraph, resulting in a high-confidence annotated dataset.

Evaluating the quality of such an unsupervised annotation pipeline directly can be challenging. Thus, we propose an indirect evaluation method: training a supervised model on our automatically annotated dataset and testing its performance on an existing benchmark, i.e., the OSDG Community Dataset, presented in (Pukelis et al., 2022). Our hypothesis is that if the addition of our dataset, albeit automatically generated, improves the classification performance, then the generated annotations must contain valuable information[2]. Our preliminary experimental results show that combining automatically labeled (auto-labeled) data with existing annotated datasets improves the classification accuracy, especially in complex texts, e.g., controversial cases.

In the remainder of this paper, Section 2 reviews the related literature. Section 3 details the annotation methodology. Section 4 describes our experimental evaluation and results, and Section 5 presents conclusions and future research directions.

2 Related Work

Previous research relevant to our work can be broadly categorized into three main areas: (1) sustainability reporting frameworks, (2) challenges in annotating sustainability reports, and (3) automated annotation approaches based on NLP.

Sustainability Reporting Frameworks. Sustainability reporting has become increasingly standardized through widely adopted frameworks such as the Sustainable Development Goals (SDGs) of the United Nations (UN, 2015) and the Global Reporting Initiative (GRI)[3]. Additionally, the Environmental, Social, and Governance (ESG) framework influences access to corporate financing (Zou et al., 2025). The SDGs define 17 general goals and 169 specific targets to guide global development efforts through 2030 (Smith et al., 2021). Each target is accompanied by indicators to monitor progress. For example, SDG 13-CLIMATE includes goals such as **13.1**, which aims to "*strengthen resilience and adaptive capacity to climate-related hazards and natural disasters in all countries*" (see Appendix A). GRI, first established in 1997, provides a complementary framework for revealing the impacts of sustainability in the economic, environmental, and social domains (Ngee et al., 2024). It defines structured standards and disclosures (some required, others recommended) that help organizations systematically report relevant actions and outcomes. The Action Platform Reporting on the SDGs[4], in collaboration with GRI, has developed guidance to help companies integrate the SDGs effectively into their reporting processes. This database enables businesses to identify relevant disclosures that reflect their contributions to achieving the SDGs. Given the strong relationship between SDGs and GRIs (where SDGs represent strategic goals that can be

[2] Although our method is capable of generating both SDG and GRI labels for individual paragraphs, this paper primarily evaluates the quality of SDG annotations. The assessment of GRI labels, and the development of joint evaluation protocols, remain open directions for future work.

[3] https://www.globalreporting.org/standards/

[4] https://www.globalreporting.org/reporting-support/goals-and-targets-database/

mapped to more granular GRI codes and disclosures), a linking database connects SDG targets to GRI subcodes at different levels. For example, target **13.1** of SDG 13-CLIMATE is connected to GRI 101-BIODIVERSITY, GRI 201-ECONOMIC PERFORMANCE, GRI 302-ENERGY, and GRI 305-EMISSIONS at the code level. Each of these GRI codes encompasses multiple disclosures and subcodes that further refine their alignment with SDG 13. For example, GRI 302-4: REDUCTION OF ENERGY CONSUMPTION is one of such disclosures related to GRI 302-ENERGY and is related to SDG 13 (more details in Appendix C).

Challenges in Annotating Sustainability Reports. Sustainability reports are essential to understand corporate strategies and impacts, but their length, unstructured format, and use of technical language make automated analysis difficult (Kang and Kim, 2022). At the same time, institutions and companies are increasingly expected to align their disclosures with global frameworks such as the SDGs and GRI (Angin et al., 2022). Manual annotation efforts have been used to assess alignment with the 2030 Agenda (Calabrese et al., 2022, 2021), but are labor-intensive, subjective, and difficult to scale. To address this, researchers have turned to Natural Language Processing (NLP) to automate the extraction, classification, and evaluation of sustainability content. These methods improve scalability and reduce bias, and recent work has explored the potential of Large Language Models (LLMs) to support such tasks.

Automated NLP-based Annotation Approaches. Recent progress in NLP has allowed the development of scalable methods for automating the annotation of sustainability content. These approaches support information extraction at multiple levels (ranging from keywords to entire pages) and often rely on semantic similarity between report content and textual definitions from frameworks such as GRI (Ngee et al., 2024; Gutierrez-Bustamante and Espinosa-Leal, 2022) and SDGs (Kang and Kim, 2022), and ESG topics (Morio et al., 2024; Bronzini et al., 2024). Supervised learning remains the dominant strategy, with models trained on labeled data to recognize SDG (Angin et al., 2022; Li and Rockinger, 2024), GRI (Polignano et al., 2022; Hillebrand et al., 2023), and ESG mentions (Ghosh and Naskar, 2022; Koloski et al., 2022; Schimanski et al., 2024). However, these methods are limited by the high cost and low scalability of manual annotations. More lightweight solutions have emerged, such as using SDG icons in reports as weak labels to train multilabel classifiers (Jakob et al., 2024), or using AutoGluon to automatically tag open resources with SDG labels, as explored in the OSDG initiative (Yao et al., 2024). Large Language Models (LLMs) have further increased automation potential. However, as noted in (Ngee et al., 2024), they often do not capture the contextual depth and nuance required for an accurate evaluation of ESG. Human expertise remains essential, particularly for fine-grained judgment in overlapping sustainability domains.

We propose an unsupervised method that aims to complement expert analysis and support supervised approaches. By integrating semantic similarity scoring with structured information, such as GRI content indexes and SDG-GRI mappings, our method identifies relevant paragraph-label pairs without requiring manual annotation. These weakly labeled outputs can be used to initialize supervised models or to highlight content for expert review, thereby enhancing scalability while maintaining interpretability.

3 Bootstrapping GRI-SDG Annotations via Report Structure and Semantic Similarity

We propose an unsupervised approach for automatically annotating paragraphs from sustainability reports with relevant GRI and SDG labels. Our method takes advantage of the structure provided by human-generated GRI indexes, known mappings between GRI codes and SDGs, and embedding-based semantic similarity.

Formal Problem Definition. Let us denote with $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ a set of sustainability reports provided as unstructured PDF documents. Each document $d \in \mathcal{D}$ contains textual content segmented into paragraphs and a table of the GRI content index that lists the GRI codes covered by specific ranges of pages.

Let $\mathcal{G}$ represent the set of all the possible 33 GRI codes, and let $\mathcal{S}$ represent the set of all 17 SDGs (more details in Appendices A and B).

An official semantic mapping between GRI and SDG codes is available as a function $\mathcal{M} : \mathcal{G} \to 2^{\mathcal{S}}$. Our task is to generate annotated paragraphs for each p occurring in some $d \in \mathcal{D}$, in form of triples (p, g, s) consisting of:

- a textual paragraph p occurring in $d \in \mathcal{D}$,

- the GRI label $g \in \mathcal{G}$ that characterizes p possibly identified by the GRI content index table of the document d,
- an SDG label $s \in \mathcal{S}$ associated to the GRI label g by the mapping $\mathcal{M}$, i.e. $s \in \mathcal{M}(g)$.

In order to generate the set of triples (p, g, s) useful for training a GRI and SDG classifier, different steps are needed to select representative and meaningful examples p regarding natural language descriptions of categories g and s.

Step 1: Extraction of paragraphs. Given the input set of all sustainability reports $\mathcal{D}$, the textual content of each document $d \in \mathcal{D}$ corresponds to $\mathcal{P}_d$, which is the set of paragraphs[5] in d. The extraction procedure of documents $\mathcal{D}$ results in the collection of paragraphs $\mathcal{P}_\mathcal{D} = \bigcup_{d \in \mathcal{D}} \mathcal{P}_d$, which includes cleaned and segmented paragraph-level texts able to trigger the annotation stage according to GRI and SDG labels, $\mathcal{G}$ and $\mathcal{S}$, respectively.

Step 2: Initializing GRI Annotations. Sustainability reports d typically include a structured *GRI content index*, which is a table summarizing the sections of the report that address specific GRI disclosure standards. Rows in this table link a GRI standard g (identified by its unique code) to one or more specific pages in d. It precisely indicates on which pages of d information relevant to g is discussed. Formally, the GRI content index allows one to define the set of GRI codes g and pages $\Pi(d, g)$ where g is discussed, i.e.

$$\Pi(d, g) = \{\pi |\ g \text{ is discussed at page } \pi \text{ of } d\}$$

As this information is provided by the authors, we use it to generate a set of GRI labels, named **candidates**, for each paragraph $p \in \mathcal{P}_d$ included on page π. Formally, for each $p \in \mathcal{P}_d$ we define the set of candidate GRI labels as $\text{CAN}(p) = \{(p, g) \mid \exists \pi \in \Pi(d, g) \text{ and } p \text{ is a paragraph in } \pi\}$. The set of candidates $\text{CAN}(p)$ includes potential GRI labels according to the GRI index table of d.

This automated labeling provides an initial "approximate" annotation, as it is based on the human-curated structure (GRI Index) of each report. However, index tables, although manually prepared, are page-based. This implies that not all paragraphs on the page may reflect the table entries, so false positives may well exist. Moreover, further relevant GRI-related textual content can be found beyond the sections mentioned in the GRI Index table, as false negatives.

To address these limitations, for each paragraph in which we have identified GRI candidates, we also define **alternative labeling**. Formally, for each $p \in \mathcal{P}_d$ we define the set of alternative labels as $\text{ALT}(p) = \{(p, g') \mid \exists \pi \in \Pi(d, g) \wedge p \text{ is a paragraph in } \pi \wedge g' \neq g\}$. In other words, $\text{ALT}(p) = \{(p, g) \mid \forall g \in \mathcal{G} \text{ such that } (p, g) \notin \text{CAN}(p)\}$.

In this way, GRI codes that are not in the list of candidates are also retained as additional potential GRI categories relevant for paragraphs, although not explicitly mentioned in the GRI index. As a result, triples

$$\big(p, \text{CAN}(p), \text{ALT}(p)\big)$$

can be obtained for every paragraph $p \in d$.

Step 3: SDG annotation using GRI-SDG mapping. To extend GRI labels to suitable SDG codes, we use GRI-SDG mapping $\mathcal{M} : \mathcal{G} \to 2^\mathcal{S}$ (more details in Appendix C). Specifically, each paragraph p previously annotated in $\text{CAN}(p)$ and $\text{ALT}(p)$ is augmented with SDG labels derived by mapping its GRI codes. Given a paragraph p, we derive $\text{CAN}^+(p) = \{(p, g, s) \mid g \in \text{CAN}(p) \wedge s \in \mathcal{M}(g)\}$, which suggests the set of all SDG labels s that are compatible with GRI candidates for p. Similarly, the alternatives $\text{ALT}(p)$ are extended. A second extended set $\text{ALT}^+(p)$ is obtained as $\text{ALT}^+(p) = \{(p, g, s) \mid g \in \text{ALT}(p) \wedge s \in \mathcal{M}(g)\}$.

The result of this step is the enriched set of annotations for each paragraph p, i.e.,

$$\big(p, \text{CAN}^+(p), \text{ALT}^+(p)\big)$$

while $\text{CAN}^+(p)$ and $\text{ALT}^+(p)$ suggest the GRI and SDG labels for p.

Step 4: Embedding-based similarity scoring. The candidate and alternative annotations are still ambiguous, and several multiple interpretations (i.e., different triples) are likely to be obtained for a paragraph p. Notice that each label, among GRI codes or SDG codes, is described by one or more textual descriptions that we call disclosures R^{GRI}, for GRI, as well as target descriptions T^{SDG}, for SDG (more details in Appendices A and B).

[5] Paragraph boundaries are identified by extracting and cleaning text blocks from PDF files using PyMuPDF (https://pymupdf.readthedocs.io/) which effectively preserves the original layout. We define a paragraph as a continuous block of text containing at least 20 words, after removing duplicate or abnormally short entries.

Using encoding neural methods, we can obtain for each disclosure $r \in R^{GRI}$ or target $t \in T^{SDG}$ unique embedding vectors $\mathbf{r}$ and $\mathbf{t}$, respectively. Notice that paragraphs p and textual descriptions of GRI disclosures r and SDG targets t can all be encoded in vector embeddings $\mathbf{p}$, $\mathbf{r}$, and $\mathbf{t}$ through the application of pre-trained encoder models, such as `all-mpnet-base-v2`, (Song et al., 2020), adopted in this work. Thus, a measure of text similarity is achieved through cosine similarity among vector pairs: it measures the semantic alignment between paragraph content and label definitions. Formally, evaluating the alignment of a paragraph p with a given GRI code g, given that

$$\sigma(p, r) = \frac{\mathbf{p} \cdot \mathbf{r}}{||\mathbf{p}||\,||\mathbf{r}||}$$

and given all disclosures $r^g \in R^{GRI}$ textually describing g, is thus achieved through

$$\sigma_{pg} = \sigma(p, g) = \max_{r^g \in R^{GRI}} \sigma(p, r^g)$$

Similarly, for each SDG code s, and its descriptions $t^s \in T^{SDG}$

$$\sigma_{ps} = \sigma(p, s) = \max_{t^s \in T^{SDG}} \sigma(p, t^s)$$

Notice that similarity scores are normalized and this allows a fair comparison between multiple alternative codes g and s in $(p, g, s) \in \text{CAN}^+(p) \cup \text{ALT}^+(p)$. The selection of the most relevant labels for each paragraph may therefore require a disambiguation step, modeled as a function acting on the similarity ranking in sets $\text{CAN}^+(p)$ and $\text{ALT}^+(p)$.

Step 5: Disambiguation of p for label assignment. The final annotation for each paragraph is obtained by selecting the best pair of labels (g^*, s^*) that can explain the relationship of p with the GRI and SDG standards. A good pair should satisfy the following properties:

- Both g^* and s^* should exhibit a high similarity with respect to p
- g^* and s^* should satisfy the mapping $\mathcal{M}$, i.e. $s^* \in \mathcal{M}(g^*)$

Notice that both properties may not be satisfied in cases where alternative codes in $\text{ALT}^+(p)$ receive larger similarities than candidate codes in $\text{CAN}^+(p)$: in these cases, GRI codes not explicitly mentioned in the GRI Index Table show higher similarity scores with a paragraph p. Thus, an adversarial comparison between $\text{ALT}^+(p)$ and $\text{CAN}^+(p)$ is needed. Given any triple $(p, g, s) \in \text{CAN}^+(p) \cup \text{ALT}^+(p)$, the quantity that appears to be maximized to fulfill the above properties is

$$\omega_{pgs} = \omega(p, g, s) = \sigma(p, g) \cdot \sigma(p, s)$$

The selection considers two major cases during the population of the final annotated dataset. Given a paragraph p:

- If the similarity score for the best triple $(p, g, s) \in \text{CAN}^+(p)$ is equal or higher than that of an **alternative** $(p, g', s') \in \text{ALT}^+(p)$, the paragraph is annotated with the *candidate* $(p, g, s) \in \text{CAN}^+(p)$ as it maximizes ω_{pgs}
- On the contrary, when no such triple $(p, g, s) \in \text{CAN}^+(p)$ can be found, as a better alternative $(p, g', s') \in \text{ALT}^+(p)$ exists with
$$\max_{(p,g',s') \in \text{ALT}^+(p)} \omega_{pg's'} > \max_{(p,g,s) \in \text{CAN}^+(p)} \omega_{pgs},$$
both the best triples are accepted as valid annotations.

For each paragraph p, given the best candidate

$$(p, g, s)^* = \operatorname*{arg\,max}_{(p,g,s) \in \text{CAN}^+(p)} \omega_{pgs}$$

and the best alternative label

$$(p, g', s')^* = \operatorname*{arg\,max}_{(p,g,s) \in \text{ALT}^+(p)} \omega_{pg's'}$$

thus the labeling of p is thus computed as:

$$lab(p) = \begin{cases} \{(p, g, s)^*\} \text{ if } \omega_{(p,g,s)^*} \geq \omega_{(p,g',s')^*} \\ \{(p, g, s)^*, (p, g', s')^*\} \text{ otherwise} \end{cases}$$

As a result, given all the paragraphs p in some $\mathcal{P}_d$ for $d \in \mathcal{D}$ then the overall training set can be defined as

$$\bigcup_{p \in \mathcal{P}_\mathcal{D}} lab(p)$$

.

The resulting training set is made of all triples (p, g, s) for which both d and p exist such that

$$d \in \mathcal{D} \wedge p \in \mathcal{P}_d \wedge (p, g, s) \in lab(p)$$

This approach aims to balance automated semantic analysis with structured metadata, addressing key limitations of both fully manual annotation (such as scalability and cost) and purely automatic approaches, which may struggle to capture nuanced context. Our method thus represents a pragmatic solution to assist human annotators rather than replacing their expertise entirely.

4 Experimental Evaluation

In this section, we evaluate the effectiveness of the weakly supervised dataset produced by our annotation pipeline. Since direct validation of unsupervised labels is challenging, we adopt an indirect strategy: we train a classifier on our automatically labeled data and test it on a manually annotated benchmark. The underlying assumption is that if our annotations capture meaningful SDG-related semantics, they should improve performance on the downstream classification task.

Experimental Setup. The proposed methodology was applied to a dataset consisting of 30 sustainability reports published in 2023. These reports span 10 industry sectors (Energy, Health Care, Mining, Food and Beverages, Chemicals, High-tech, Land and Soil, Manufacturing, Services, and Textile) with three reports selected per sector. In total, the corpus includes 3,663 pages. After preprocessing and paragraph segmentation (see Step 1 in Section 3), we extracted 19,133 paragraphs. Of these, 10,303 paragraphs fall within the page ranges indicated by GRI content indexes (Step 2), and are retained for candidate annotation. For SDG enrichment, we use the official GRI-SDG mapping from the Action Platform Reporting on SDGs[6]. The mapping includes all 33 GRI codes and all 17 SDGs, resulting in 89 distinct $(g, s) \in \mathcal{G} \times \mathcal{S}$ pairs used to guide the annotation process.

We use the `all-mpnet-base-v2` pre-trained model from the sentence-transformers library[7] to compute paragraph and label embeddings. This model, based on MPNet (Song et al., 2020), has been fine-tuned for sentence-level tasks such as semantic similarity and sentence matching. It combines strengths from both BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), and has demonstrated strong performance in sustainability-related annotation tasks (Ngee et al., 2024).

We obtain a total of $10,303$ annotated paragraphs, each labeled with one or more (GRI, SDG) pairs. The distribution of labels in the 17 SDGs is markedly uneven. SDG 2 - HUNGER (about *Achieve food security and sustainable agriculture*) has the fewest annotations (128), while SDG 8 - ECONOMY (about *Promote sustainable economic growth and decent work*) appears most frequently (3,857). On average, each SDG is associated with 894 annotated paragraphs, with a standard deviation of 976, reflecting significant variability in the frequency of the label.

To evaluate the quality of our automatically labeled data, we rely on the Open Source SDG (OSDG) Community Dataset[8]. In this dataset, each paragraph is associated with a single SDG and validated through binary judgments (agree/disagree) by multiple annotators. While this one-label-per-paragraph setting is a simplification (since many texts plausibly relate to multiple SDGs), it reflects the task's inherent ambiguity. Each example is also accompanied by an agreement score that indicates how consistent the annotators were in accepting the proposed SDG.

We use this agreement score to build two sets of evaluations of increasing complexity:

- **Simple test set:** examples in full agreement (agreement $= 1$).
- **Complex test set:** examples with partial consensus (agreement $\in [0.7, 1]$).

Lower-agreement examples (below 0.7) are excluded from the test due to their inherent ambiguity and noise. For training, we consider four configurations to assess the utility of our auto-labeled data and its interaction with OSDG:

- **Auto-labeled:** our dataset of $10{,}303$ annotated paragraphs.
- **OSDG (100% agreement):** a high-confidence subset of $11{,}938$ examples.
- **OSDG (full):** the full OSDG training set, including all paragraphs with agreement ≥ 0.1 (28,478 examples).
- **Combined:** the union of our auto-labeled data and the full OSDG set (38,781 examples).

For the evaluation, we fine-tuned a custom BERT-based classifier (`bert-base-cased`) for multi-label classification[9] as such encoder based classifiers provide a strong baseline and have shown robust performance on SDG, GRI, and ESG label prediction in prior work (Angin et al., 2022;

[6] https://www.globalreporting.org/reporting-support/goals-and-targets-database/

[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[8] https://github.com/osdg-ai/osdg-data

[9] Training was performed using an effective batch size of 16 (with gradient accumulation every 4 mini-batches of size 4), a learning rate of 2×10^{-5}, weight decay of 0.1, and a warmup ratio of 0.1. The model was trained for 5 epochs using the AdamW optimizer and a linear learning rate scheduler.

Polignano et al., 2022; Hillebrand et al., 2023). A key challenge arises from the difference in annotation schemes: the OSDG dataset provides only a single SDG code per paragraph, whereas our auto-labeled dataset may associate a paragraph with multiple SDG codes, an approach we argue to be more realistic. To accommodate this, the model was trained in a multi-label setting, allowing it to output potentially more than one SDG per instance. However, during the evaluation, only the label with the highest predicted probability was selected to remain consistent with the single-label format of the OSDG test set. Model performance was reported in terms of accuracy, calculated as the percentage of paragraphs that were correctly reassigned to their original SDG label.

Results. Table 1 reports preliminary results on the OSDG benchmark, comparing the classification accuracy in both simple and complex test subsets in different training configurations. As expected, the lowest performance is observed when training the model solely on our auto-labeled data, which reaches 0.762 accuracy on the simple set and 0.737 on the complex one. This outcome is understandable given the domain shift between datasets: our data come from corporate sustainability reports, while the OSDG dataset is composed primarily of general policy texts from international organizations, mainly the United Nations. Despite this thematic and stylistic gap, achieving an accuracy of up to 76% in a 17-class setting without human supervision remains a promising and nontrivial result. Using the high-confidence portion of OSDG, where all annotators agreed on the label, yields 0.919 on the simple test set and 0.899 on the complex one. Interestingly, expanding the training data to include all OSDG samples in agreement ≥ 0.1 (OSDG full) slightly reduces performance on the simple test set (0.917), but improves accuracy on the complex one (0.907), likely due to increased exposure to more nuanced and ambiguous examples. Crucially, the best results are obtained by combining our auto-labeled data with the entire OSDG dataset, reaching 0.924 on the simple test set and 0.910 on the complex set. This suggests that our automatically generated annotations contribute positively to model generalization, despite being derived from a different textual domain. Rather than introducing noise, they provide complementary information that enriches the training data. In general, these findings support the use of unsupervised annotations as a valuable resource to scale up sustainability-related NLP applications.

Table 1: Accuracy of BERT-based classifier on OSDG simple (agreement $= 1$) and complex (agreement $\in [0.7, 1]$) test sets under different training setups.

Training Data	**Simple Test Set**	**Complex Test Set**
Auto-labeled	0.762	0.737
OSDG (100% agreement)	0.919	0.899
OSDG (full)	0.917	0.907
Auto-label. + OSDG (full)	**0.924**	**0.910**

To further investigate the domain distance between the two training corpora, we also tested the generalization ability of the BERT-based classifier in an inverse setting. Specifically, we trained a classifier only on OSDG (100% agreement) data and evaluated it on the 10% held-out portion of our auto-labeled company data (the same subset previously reserved for the classifier reported in the first row of Table 1). In this setting, using a single-label prediction against our multi-label ground truth (i.e., considering the prediction correct if it matched any of the gold labels), the model achieved only 40% accuracy. This low result highlights the significant dissimilarity between the two domains, likely driven by differences in textual genre, style, and content focus. In contrast, our model trained on company data appears more robust when evaluated in an out-of-domain setting (row 1 in Table 1).

Error Analysis. A qualitative analysis of misclassifications reveals that many errors are not due to clear model linguistic failures but rather stem from semantic overlaps between SDG categories. In several cases, the predicted label (though incorrect according to the gold annotation) is still plausible and semantically coherent. Let us consider some errors of the classifier trained on the combined dataset.

The assumed correct label for the paragraph "*Over time, personalized technology will supersede one-size-fits-all models of education... mobile apps now make it easy for teachers to administer quizzes...*" is SDG 9 - Industry, which pertains to *Build resilient infrastructure and foster innovation*. However, the predicted label is SDG 4 - Education, which aligns with *Ensure inclusive, quality education for all*, which also seems to be meaningful and basically correct.

Furthermore, most of the misclassified samples were related to SDG 8 - ECONOMY (about *Promote sustainable economic growth and decent work*) and SDG 10 - INEQUALITY (about *Reduce inequality within and among countries*). This can be attributed to the close relationship between these two SDGs, which cover broad and general concepts, leading to ambiguous text samples.

For example, consider the text: "*The chapter explores the relationship between own-use production work and household income... and derives measures of inequality.*" The actual label for this sample is SDG 8 - ECONOMY, but the predicted label was SDG 10 - INEQUALITY. Both SDGs are related to the ambiguous nature of sentences, demonstrating how the overlap of themes between these SDGs contributed to misclassifications.

As an example of how the enriched dataset, with our annotated data, is helpful, we can consider this ambiguous text: "*Highly qualified teachers address gender-specific attitudes... studies show gender differences in competencies...*" The actual label is SDG 5 - GENDER, which pertains to *Achieve gender equality and empower women*. Based on the classifier trained only on the OSDG material, the model predicted SDG 4 - EDUCATION, which aligns with *Ensure inclusive, quality education for all*. However, when using the enhanced dataset, the model predicted the correct label, demonstrating the effectiveness of the enriched dataset in handling ambiguous cases.

The role of the enriched dataset in predicting the correct labels for climate-related samples has also shown improvement. For example, consider the following text: "*Between 2002 and 2008, carbon footprint of Dutch holidaymakers increased by 16.8%... 9% of national emissions.*" The actual label is SDG 13 - CLIMATE, which pertains to *Take urgent action to combat climate change and its impacts*. Based on the classifier trained only on OSDG material, the model predicted SDG 12 - CONSUMPTION, which aligns with *Ensure sustainable consumption and production patterns*. However, when using the enhanced dataset, the model predicted the correct label, demonstrating the effectiveness of the enriched dataset in climate-related texts. These observations suggest that some errors are attributable to inherent subjectivity and overlap among SDG definitions, rather than to annotation noise or model limitations. They also motivate future directions toward more nuanced evaluation strategies, including multi-label classification and uncertainty-aware models.

5 Conclusion and Future Work

This work introduces an unsupervised pipeline for annotating sustainability report paragraphs with both GRI and SDG labels. By combining structured metadata from GRI content indexes, official SDG-GRI mappings, and semantic similarity scoring via sentence embeddings, we create a weakly supervised dataset that can be used to support downstream classification tasks. A preliminary evaluation using the OSDG benchmark demonstrates that our auto-labeled data (despite being generated without human supervision) contributes positively when combined with high-quality manually annotated datasets. These early results suggest that our annotations capture meaningful sustainability semantics and can complement existing resources.

This paper represents the first step in a wider ongoing effort. Although our current evaluation focuses on SDG labels and uses an indirect performance-based validation, further work is needed to better assess the quality and reliability of the SDG annotations- ideally through targeted human evaluations and to benchmark our fine-tuned classifier, trained on automatically labeled data, against established methods in the literature (Angin et al., 2022; Polignano et al., 2022; Yao et al., 2024). In addition, our method currently generates GRI labels, but these have not yet been empirically validated; assessing their accuracy is a priority for future iterations. More importantly, our long-term goal extends beyond surface-level classification. We aim to deepen the analysis of sustainability discourse within reports by examining the nature of reported content (distinguishing between fact-based, temporally grounded disclosures and vague, qualitative claims). This is particularly relevant in detecting subtle forms of greenwashing. Future work will focus on moving beyond topic classification to assess the substance and credibility of reported content, to identify meaningful disclosures as opposed to vague or unsubstantiated claims, and extend the scope to include ESG reporting.

Acknowledgments

We thank Armando Calabrese, Roberta Costa and Luigi Tiburzi for their valuable advice, insightful discussions on sustainability reporting, and for generously sharing documents that inspired and enabled this initial experimental study.

Limitations

This work represents a first step toward an automated integrated GRI-SDG annotation, with several limitations. Our evaluation relies on indirect metrics and single-label test data, which may not fully reflect the multi-label nature of real sustainability disclosures. GRI annotations are currently not evaluated because of the lack of ground-truth data. Moreover, while our method identifies relevant topics, it does not assess the quality or truthfulness of the claims, which is crucial to distinguish meaningful reporting from generic or greenwashed content.

In the paragraph extraction process, overlapping disclosures are not explicitly handled, each paragraph is treated as an independent unit. We acknowledge that this may lead to content overlap across segments and plan to explore more refined segmentation strategies in future work.

A limitation of our approach is the reliance on company-provided GRI content indexes as a primary source of structured metadata. Since these tables are compiled by the reporting organizations themselves, they may reflect selective transparency or introduce bias in the disclosure of sustainability topics. This means that some relevant information may be omitted or presented in a favorable light, potentially impacting the objectivity of our initial candidate labels. To mitigate this, we complement index-based labels with alternative GRI codes inferred through semantic similarity, broadening the annotation framework and reducing dependence on potentially biased or incomplete company disclosures. However, residual bias cannot be entirely excluded, and human expert review remains essential for high-stakes applications.

References

Merih Angin, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp, Mert Atay, Pelin Angin, and Gökhan Dikmener. 2022. A roberta approach for automated processing of sustainability reports. *Sustainability*, 14(23):16139.

Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or gold? deriving structured insights from sustainability reports via large language models. *EPJ Data Science*, 13(1):41.

Armando Calabrese, Roberta Costa, Massimo Gastaldi, Nathan Levialdi Ghiron, and Roberth Andres Villazon Montalvan. 2021. Implications for sustainable development goals: A framework to assess company disclosure in sustainability reporting. *Journal of Cleaner Production*, 319:128624.

Armando Calabrese, Roberta Costa, Nathan Levialdi Ghiron, Luigi Tiburzi, and Roberth Andres Villazon Montalvan. 2022. Is the private sector becoming cleaner? assessing the firms' contribution to the 2030 agenda. *Journal of Cleaner Production*, 363:132324.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.

Sohom Ghosh and Sudip Kumar Naskar. 2022. Ranking environment, social and governance related concepts and assessing sustainability aspect of financial texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 243–249.

Marcelo Gutierrez-Bustamante and Leonardo Espinosa-Leal. 2022. Natural language processing methods for scoring sustainability reports—a study of nordic listed companies. *Sustainability*, 14(15):9165.

Lars Hillebrand, Maren Pielka, David Leonhard, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Milad Morad, Christian Temath, Thiago Bell, and 1 others. 2023. sustain. ai: a recommender system to analyze sustainability reports. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 412–416.

Charlott Jakob, Vera Schmitt, Salar Mohtaj, and Sebastian Möller. 2024. Classifying sustainability reports using companies self-assessments. In *Future of Information and Communication Conference*, pages 547–557. Springer.

Hyewon Kang and Jinho Kim. 2022. Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods. *Applied Sciences*, 12(11):5614.

Boshko Koloski, Syrielle Montariol, Matthew Purver, and Senja Pollak. 2022. Knowledge informed sustainability detection from short financial texts. In *Proceedings of the fourth workshop on financial technology and natural language processing (FinNLP)*, pages 228–234.

Yao Li and Michael Rockinger. 2024. Unfolding the transitions in sustainability reporting. *Sustainability*, 16(2):809.

Gaku Morio, Soh Young In, Jungah Yoon, Harri Rowlands, and Christopher Manning. 2024. Reportparse: A unified nlp tool for extracting document structure and semantics of corporate sustainability reporting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8749–8753.

Ivan Nechaev and Daniel S Hain. 2023. Social impacts reflected in csr reports: Method of extraction and link to firms innovation capacity. *Journal of Cleaner Production*, 429:139256.

Hui Qian Ngee, Asha Ganesh, Muhammad Aizat Noor Azmi, Tiong Yew Tang, Muaadh Mukred, Fathey Mohammed, and Adi Affandi Bin Ahmad. 2024. Environmental. social and governance (esg) scores automation in global reporting initiative (gri) with natural language processing. In *2024 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS)*, pages 1–7. IEEE.

Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. An nlp approach for the analysis of global reporting initiative indexes from corporate sustainability reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 1–8.

Lukas Pukelis, Nuria Bautista-Puig, Gustė Statulevičiūtė, Vilius Stančiauskas, Gokhan Dikmener, and Dina Akylbekova. 2022. Osdg 2.0: a multilingual tool for classifying text data by un sustainable development goals (sdgs). *Preprint*, arXiv:2211.11252.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.

T. B. Smith, R. Vacca, L. Mantegazza, and I. Capua. 2021. Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals. *Sci. Rep.*, 11(1):22427.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

UN. 2015. Transforming our world: The 2030 agenda for sustainable development. *New York: United Nations, Department of Economic and Social Affairs*, 1:41.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Rui Yao, Meilin Tian, Chi-Un Lei, and Dickson KW Chiu. 2024. Assigning multiple labels of sustainable development goals to open educational resources for sustainability education. *Education and Information Technologies*, 29(14):18477–18499.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. Esgreveal: An llm-based approach for extracting structured data from esg reports. *Journal of Cleaner Production*, 489:144572.

A Sustainable Development Goals(SDGs)

The United Nations Sustainable Development Goals (UN SDGs)[10] constitute a universal call to action to end poverty, protect the planet, and ensure prosperity for all as part of a new sustainable development agenda. The SDGs framework comprises 17 overarching Global Goals, each of which is further specified by a set of 169 detailed targets (see Table 2 for a complete list of the goals). Among these, SDG 13-CLIMATE is of particular importance, as it aims to "*Take urgent action to combat climate change and its impacts*". Given its relevance, in this work we provide the full list of targets associated with SDG 13 in Table 3. These target descriptions serve as concrete examples of the input material utilized in our annotation approach, illustrating the level of textual granularity considered when linking sustainability report content to SDG objectives.

B Global Reporting Initiative(GRI)

In general, GRI comprises UNIVERSAL STANDARDS (Codes 1–3) and three sets of topic-specific standards: ECONOMIC (Codes 201–207), ENVIRONMENTAL (Codes 301–308), and SOCIAL (Codes 401–419). According to the latest published version (February 5, 2024), a new independent standard, BIODIVERSITY 2024 (Code 101), has been introduced. Additionally, GRI 307 and GRI 419 have been withdrawn and replaced by Disclosure 2-27, while the content of GRI 412 has been integrated into the Universal Standards. Table 4 presents the GRI codes (GRI labels for our approach) and their descriptions without considering GRI 1 and 3 pertain to FOUNDATION and MATERIAL TOPICS.

To provide a detailed breakdown of the disclosure identification requirements related to GRI 302

[10] https://sdgs.un.org/

Table 2: Abbreviation and Descriptions of SDGs

No.	Abbreviation	Description
1	POVERTY	End poverty in all forms
2	HUNGER	Achieve food security and sustainable agriculture
3	HEALTH	Ensure healthy lives and well-being for all
4	EDUCATION	Ensure inclusive, quality education for all
5	GENDER	Achieve gender equality and empower women
6	SANITATION	Ensure sustainable water and sanitation
7	ENERGY	Ensure access to sustainable energy
8	ECONOMY	Promote sustainable economic growth and decent work
9	INDUSTRY	Build resilient infrastructure and foster innovation
10	INEQUALITY	Reduce inequality within and among countries
11	SETTLEMENTS	Make cities sustainable and resilient
12	CONSUMPTION	Ensure sustainable consumption and production
13	CLIMATE	Take action against climate change
14	AQUATIC	Protect oceans and marine resources
15	TERRESTRIAL	Sustainably manage forests and biodiversity
16	PEACE	Promote peace, justice, and strong institutions
17	PARTNERSHIPS	Strengthen global partnerships for development

Table 3: Target Descriptions of SDG 13 "*Take urgent action to combat climate change and its impacts*"

No.	Description
13.1	*Strengthen resilience and adaptive capacity to climate-related hazards and natural disasters in all countries.*
13.2	*Integrate climate change measures into national policies, strategies and planning.*
13.3	*Improve education, awareness-raising and human and institutional capacity on climate change mitigation, adaptation, impact reduction and early warning.*
13.a	*Implement the commitment undertaken by developed-country parties to the United Nations Framework Convention on Climate Change to a goal of mobilizing jointly $100 billion annually by 2020 from all sources to address the needs of developing countries in the context of meaningful mitigation actions and transparency on implementation and fully operationalize the Green Climate Fund through its capitalization as soon as possible.*
13.b	*Promote mechanisms for raising capacity for effective climate change-related planning and management in least developed countries and small island developing States, including focusing on women, youth, and local and marginalized communities.*

Table 4: GRI Codes Descriptions

Code	Description
201	ECONOMIC PERFORMANCE
202	MARKET PRESENCE
203	INDIRECT ECONOMIC IMPACTS
204	PROCUREMENT PRACTICES
205	ANTI-CORRUPTION
206	ANTI-COMPETITIVE BEHAVIOR
207	TAX
301	MATERIALS
302	ENERGY
303	WATER AND EFFLUENTS
304	BIODIVERSITY
305	EMISSIONS
306	WASTE
308	SUPPLIER ENVIRONMENTAL ASSESSMENT
401	EMPLOYMENT
402	LABOR/MANAGEMENT RELATIONS
403	OCCUPATIONAL HEALTH AND SAFETY
404	TRAINING AND EDUCATION
405	DIVERSITY AND EQUAL OPPORTUNITY
406	NON-DISCRIMINATION
407	FREEDOM OF ASSOCIATION AND COLLECTIVE BARGAINING
408	CHILD LABOR
409	FORCED OR COMPULSORY LABOR
410	SECURITY PRACTICES
411	RIGHTS OF INDIGENOUS PEOPLES
413	LOCAL COMMUNITIES
414	SUPPLIER SOCIAL ASSESSMENT
415	PUBLIC POLICY
416	CUSTOMER HEALTH AND SAFETY
417	MARKETING AND LABELING
418	CUSTOMER PRIVACY
101	BIODIVERSITY
2	GENERAL DISCLOSURES

- ENERGY as an example illustrating the input materials used in our approach, we present the requirements descriptions for GRI 302-4: REDUCTION OF ENERGY CONSUMPTION, one of its subcodes or disclosures, as follows:

- *The reporting organization shall report the following information: a. Amount of reductions in energy consumption achieved as a direct result of conservation and efficiency initiatives, in joules or multiples.*

- *The reporting organization shall report the following information: b. Types of energy included in the reductions: whether fuel, electricity, heating, cooling, steam, or all.*

- *The reporting organization shall report the following information: c. Basis for calculating reductions in energy consumption, such as the base year or baseline, including the rationale for choosing it.*

- *The reporting organization shall report the*

Table 5: Connections between SDG 13-CLIMATE, Targets, GRI Codes, and Disclosures

GRI Code	GRI Sub-code (Disclosure Identifier)
GRI 101	GRI 101-2: Management of biodiversity impacts
GRI 201	GRI 201-2: Financial implications and other risks due to climate change
GRI 302	GRI 302-1: Energy consumption within the organization GRI 302-2: Energy consumption outside of the organization GRI 302-3: Energy intensity GRI 302-4: Reduction of energy consumption GRI 302-5: Reductions in energy requirements of products and services
GRI 305	GRI 305-1: Direct (Scope 1) GHG emissions GRI 305-2: Energy indirect (Scope 2) GHG emissions GRI 305-3: Other indirect (Scope 3) GHG emissions GRI 305-4: GHG emissions intensity GRI 305-5: Reduction of GHG emissions

following information: d. Standards, methodologies, assumptions, and/or calculation tools used.

- *Compilation requirements 2.7 When compiling the information specified in Disclosure 302-4, the reporting organization shall: 2.7.1 exclude reductions resulting from reduced production capacity or outsourcing; 2.7.2 describe whether energy reduction is estimated, modeled, or sourced from direct measurements. If estimation or modeling is used, the organization shall disclose the methods used.*

C GRI-SDG Linking Dataset

Table 5 presents an excerpt from the official SDG–GRI mapping[11] that forms the basis for our annotation strategy. This table demonstrates how SDG 13 - CLIMATE, together with its individual targets, is systematically connected to relevant GRI codes and their respective disclosure requirements. By providing a concrete example of these structured relationships, the table illustrates the type of cross-framework linkages that our annotation pipeline exploits to assign both SDG and GRI labels to sustainability report content. Such mappings are fundamental to reducing ambiguity and ensuring consistency when annotating real-world documents. These explicit associations between SDG targets and GRI disclosures guide the automated annotation process, enabling more interpretable and explainable results when analyzing unstructured sustainability reporting.

[11] https://www.globalreporting.org/reporting-support/goals-and-targets-database/