OCTOBER 2024
No 326

Authors
Steve Borchardt
Jean-Baptiste Jacouton
Michele Maroni
Luisa Marelli

Research papers

In Quest for Meaning: Towards a Common Understanding of the 2030 Agenda?

European Commission

ÉDITIONS AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

Agence française de développement

Papiers de recherche

Les *Papiers de Recherche de l'AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ils s'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les *Papiers de Recherche de l'AFD* n'en exclut aucune autre.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s) ou des institutions partenaires.

Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. The paper is therefore published under the sole responsibility of its author(s) or its partner institutions.

In Quest for Meaning: Towards a Common Understanding of the 2030 Agenda?

Auteurs / Authors

Steve BORCHARDT
Consultant to the European Commission, Joint Research Centre (JRC), Ispra, Italy & University of Bologna, Department of Agricultural and Food Sciences, Bologna, Italy

Jean-Baptiste JACOUTON
Agence française de développement

Michele MARONI
Consultant to the European Commission, Joint Research Centre (JRC), Ispra, Italy

Luisa MARELLI
European Commission, Joint Research Centre (JRC), Ispra, Italy

Abstract
Driven by recent advancements in data science, numerous initiatives seek to employ automated text analysis for tracking references to the Sustainable Development Goals (SDGs). This paper provides a thorough comparison of the *SDG Mapper* and the *SDG Prospector*, two text classification tools designed to detect the SDGs in text. The tools were tested on a set of European policy documents and World Bank project documents to evaluate their classification efficacy. The findings indicate that both tools demonstrate high convergence on SDGs related to health (3), economic growth (8), and climate action (13), but differ significantly in detecting goals like poverty reduction (1) and inequality (10). We conclude that with appropriate training, AI-based language models are more accurate than keyword approaches for SDG classification. The paper calls for systematic benchmarking exercises to foster accuracy and effectiveness of automated SDG classification solutions independently from their underlying methodology. We highlight a series of recommendations to strengthen the robustness and reliability of comparisons across SDG classification tools.

Acknowledgements
The authors warmly thank Régis Marodon (AFD), Meike Morren (University of Amsterdam), Abdulaziz Sadi-Cherif (AFD), Finn Woelm (Global Goals Directory), Chiara Gastaldi (JRC),Matteo Trane (JRC), Sophie Salomon (AFD) and Marilou Vincent (AFD) for their insightful comments and expertise.

Keywords
Sustainable Development Goals (SDGs), Artificial intelligence, Benchmarking

JEL Classification
C88, Q01

Original version
English

Accepted
July 2024

Résumé

Soutenues par les avancées récentes dans le domaine de la science des données, de nombreuses initiatives visent à exploiter l'analyse automatique de texte pour identifier les références aux Objectifs de développement durable (ODD). Cet article présente une analyse comparative de deux outils de classification, le *SDG Mapper* développé par la Commission Européenne, et le *SDG Prospector* développé par l'AFD. Afin de mesurer leur efficacité, les outils ont été évalués en utilisant un corpus de politiques européennes et des rapports de projets de la Banque mondiale. Les résultats démontrent une forte convergence entre les deux outils concernant les ODD relatifs à la santé (ODD 3), à la croissance économique (ODD 8) et à l'action climatique (ODD 13). Cependant, des divergences significatives apparaissent dans la détection d'objectifs tels que la réduction de la pauvreté (ODD 1) et l'inégalité (ODD 10). Nous concluons qu'avec un entraînement adéquat, les modèles de langage basés sur l'intelligence artificielle surpassent les approches par mots-clés en termes de précision pour la classification des ODD. L'article préconise la réalisation d'exercices systématiques d'évaluation comparative afin d'améliorer la précision et l'efficacité des solutions automatisées, indépendamment de leur méthodologie sous-jacente. Enfin, nous formulons une série de recommandations visant à renforcer la robustesse et la fiabilité des comparaisons entre les outils de classification des ODD.

Remerciements

Les auteurs tiennent à remercier Régis Marodon (AFD), Meike Morren (Université d'Amsterdam), Abdulaziz Sadi-Cherif (AFD), Finn Woelm (Global Goals Directory), Chiara Gastaldi (JRC), Matteo Trane (JRC), Sophie Salomon (AFD) et Marilou Vincent (AFD) pour leur relecture attentive.

Mots-clés

Objectifs de développement durable (ODD), Intelligence artificielle, *Benchmarking*

Classification JEL

C88, Q01

Version originale

Anglais

Acceptée

Juillet 2024

Contents

"The limits of my language mean the limits of my world."
Ludwig Wittgenstein, *Tractatus Logico-Philosophicus* (1922)

Introduction

Recent developments in Natural Language Processing (NLP) are revolutionizing knowledge management (Hu *et al*. 2023). The generation of Large Language Models, like ChatGPT, breaks down barriers between different types of languages (Naveed *et al*. 2023). In a few seconds, it is now possible to edit complex programming codes from a prompt written in vernacular language (Xu *et al*. 2022).

As a specific branch of NLP, classification involves the recognition and mapping of references to a specific topic within a text. This technique is useful for analyzing large corpuses of documents. Conceptually, the classification of Sustainable Development Goals (SDGs) is a particularly technical case. Adopted in 2015, the 2030 Agenda constitutes a common framework for approaching and implementing human development policies while respecting environmental boundaries. The 2030 Agenda is structured around 17 objectives, which are themselves broken down into 169 targets. In this regard, training an NLP model for SDG classification requires a detailed understanding of the specificities of each objective, as well as their interactions.

The SDGs are widely used by public and private organizations, as they constitute a shared approach to sustainable development. Over the last years, numerous SDG classification projects have emerged (Lafleur, 2023). The first applications have appeared within the United Nations ecosystem, to facilitate documents' classification with respect to the 17 Sustainable Development Goals (LaFleur, 2019; Joshi *et al*., 2020; Pukelis *et al*., 2020). Similarly, SDG classification is increasingly used in academia to facilitate SDG-related bibliographies (Elsevier, South African SDG Hub). SDG classification also finds significant resonance within international organizations in order to map public policies' alignment on the 2030 Agenda. As such, the OECD has developed a tool to analyze the SDG content of all projects submitted to the DAC (Pincet *et al*., 2019). Finally, the private sector is also developing tailored SDG classification to work with municipalities (Global Goals Directory), or to support sustainable finance objectives.

The proliferation of SDG classification solutions raises a question regarding the extent to which their results converge. Differences principally stem from varying methodological approaches. As it is relatively easy to implement, most classification tools rely on counting keywords' frequencies in a text approach. However, this approach, if not properly designed and revised by experts, is likely to generate biases because the same word does not have the same meaning in the context of the sentence in which it is used. Keyword approaches

may also struggle in detecting sparse, abstract or complex relevant information as their baseline ontology rarely includes comprehensive semantic universes, i.e. keywords and all their permutations. However, if access to the keywords and explanation on the detection methodology is provided, they maintain a desirable degree of transparency and explainability.

To overcome these limitations, another category of SDG classification tools relies on language models (Guisiano et Chiky (2020), Jacouton *et al.* (2022)). This new generation of artificial intelligence models uses deep neural networks. They are pre-trained on very large volumes of texts coming mainly from the web, such as English Wikipedia (Liu *et al.*, 2019). The models are then fine-tuned with SDG-related texts to allow for a more thorough understanding of sentences and their context[1]. However, these models are harder to implement as their classifying capacity heavily depends on the quality of the learning base. The latter requires a large number of parameters and strong computing power (Bender *et al.*, 2021). Furthermore, their power and flexibility bear trade-offs with transparency and explainability of the predictions.

Another fundamental issue relates to classification tools' degree of precision. How to ensure that the classification is correct and that detected topics reflect the actual meaning of the text? The concept of sustainable development is particularly subject to interpretations. For example, Berg *et al.* (2022) show that in the financial sector, the correlation between different ESG assessments is low. As such, differences in SDG classifications should be carefully analyzed. Our underlying objective is not to provide an illusory "correct" interpretation of the SDGs, but rather to highlight possible biases in classification tools. This would contribute to transparency objectives regarding the use of automatic solutions and could help improving the quality of SDG classification.

The aim of this paper is to provide a comparison between two classification tools. The *SDG Mapper* was developed by the Joint Research Centre (JRC) of the European Commission to better understand the semantic interplay between European Union policies and the 2030 Agenda (Borchardt *et al.*, 2023). Simultaneously, the Agence française de développement developed the *SDG Prospector* to analyze public development banks' mandates (Jacouton *et al.*, 2022). Interestingly enough, the *Mapper* and the *Prospector* use different methods to classify the SDGs in written documents. The former applies a rule-based keyword approach, while the *SDG Prospector* is based on a large language model that has been trained on an expert-labelled learning base. Table 1 describes the main characteristics of each tool.

The remainder is structured as follows. Section 2 describes the dataset and the methodological choices to conduct the cross-comparison. Section 3 presents the classification results obtained with *SDG Mapper* and *Prospector* and explores

[1] Methodological choices to build such learning base are discussed in Jacouton *et al.* (2022).

possible reasons for areas of divergence through detailed case studies. Finally, Section 4 highlights possible research areas to sustain SDG classification cross-comparison, and Section 5 concludes.

Table 1: An overview of the two SDG classifiers' characteristics

SDG Mapper	SDG Prospector
• **Methodology**: Rules-based keyword approach (semantics) • **Keyword base**: 3,398 keywords (230 goal-level, 3,168 target-level) • **Precision**: Goal and target levels • **Language of application**: Multiple via eTranslation conversion to english • **Publications**: Borchardt *et al.* (2023) • **Website**: https://knowsdgs.jrc.ec.europa.eu/sdgmapper	• **Methodology**: Large Language Model (Distilled RoBERTa) • **Learning base**: 8,500+ paragraphs collected and labelled manually • **Precision**: Goal and target levels • **Language of application**: English only • **Publications**: Jacouton, Marodon et Laulanié (2022) • **Website**: https://sdgprospector.org

1. Comparing approaches to SDG classification

1.1 – Consolidating the benchmarking corpus

A robust comparative analysis requires a high quality and heterogeneous benchmarking dataset to account for different use cases of SDG classification, and to ensure a meaningful comparison of the classification outputs. To build a heterogeneous benchmarking dataset, two types of documents were considered: EU policy documents and project documentation of World Bank projects. Both types of documents differ in their structure, style and terminology.

A set of 155 EU policy documents corresponding to 89 policy initiatives was retrieved from EUR-Lex[2] and official European Commission websites. The sample covers a wide range of policy domains (e.g. poverty reduction, climate change adaptation, occupational safety and health, sustainable economic development, etc.) with different types of policy documents (preparatory documents like Communications or Staff Working Documents as well as legal acts like Directives or Regulations) of varying document length. To leverage a dataset of manually reviewed and SDG-classified documents for the tool comparison (Miola *et al.* 2019), we selected policy initiatives spanning over the Commission period 2014-2019..

For the set of World Bank project documents, 1,134 documents corresponding to 561 projects were randomly selected within the World Bank's portfolio[3]. The selected projects span across a broad range of sectoral investments (infrastructure, water, energy), as well as transversal issues which relate to global goods (e.g. decarbonization, COVID-19 recovery plans, governance). For each project, we collected three types of documents. The Project Information Document, and the Project Appraisal Document provide a detailed description of each project, including the economic, social, political and environmental context in which they were financed. Besides, we collected the Environmental and Social Commitment Plans where available. For the analysis, these three types of documents were considered together to apprehend classification results at the project level. This is more consistent as all three documents refer to the same project.

1.2 – Processing of mapping results

The *SDG Mapper* and *SDG Prospector* utilize different classification approaches that further determine the level of granularity for the comparison. The *SDG Mapper* provides keyword frequencies directly aggregated at the document-level. This limits the comparative

[2] See: https://eur-lex.europa.eu/homepage.html?locale=en

[3] See: https://projects.worldbank.org/en/projects-operations/projects-home

exercise as classification results can only be compared at the document-level and not at a more granular level (e.g. paragraphs). In this perspective, the outputs from the *SDG Prospector* that are provided at the paragraph-level needed to be aggregated to the document-level by summing up the counts of paragraphs classified with a specific SDG.

For this exercise, we chose to focus our comparative analysis at the goal-level and to leave comparisons of target classification results for further research. As such, outputs coming from *SDG Mapper* which provided counts for both goal- and target-related keywords needed to be aggregated to the goal-level by summing up the corresponding keyword frequencies. Due to the complex nature and specific role of SDG 17 "Partnerships" within the 2030 Agenda (Le Blanc, 2015), this SDG was not considered for the comparison.

Both, keyword frequencies and number of SDG-related snippets were converted to percentages within each document to facilitate harmonization and further processing of the classification results. Using relative values instead of absolute frequencies allows controlling for documents' length. To retain the most pertinent classification results and avoid potential inflation of SDG counts, we discarded SDGs whose detection was below 5% of total SDG references in each document. This threshold was defined after robustness checks presented below to ensure that our comparison would focus only on significant SDGs while keeping enough observations. In a final step, percentages were converted to ranks to allow for a better comparison of results.

The processing of both classification outcomes allowed for direct rank comparison between the results by subtracting one dataframe from the other to obtain rank differences. Mean absolute rank differences as well as standard deviation of absolute rank differences were calculated to assess convergence between both tools. Further comparisons between ranks were calculated to assess the level of convergence on a broader scale, for instance by verifying the presence of one tool's main classified SDG (highest frequencies in S*DG Mapper*, highest number of classified paragraphs by *SDG Prospector*) in the other tool's top-3 classified SDGs. This convergence check was then repeated for each SDG to get a more granular overview on the convergence between the tools for specific SDGs. Results were further contextualized with descriptive statistics on the results of each tool. Based on the outcomes of the rank comparison, several case studies from both datasets were selected for a deeper, more qualitative assessment of the results by looking at both -cases with high convergence as well as low convergence and to further validate the classification outcomes. Case studies were manually classified by the authors with respect to the SDGs addressed within each document to establish a concordant baseline. Results are presented in the next Section.

2. Results

2.1 – Descriptive analysis

Each tool provided extensive results for both datasets (see table 2 below).

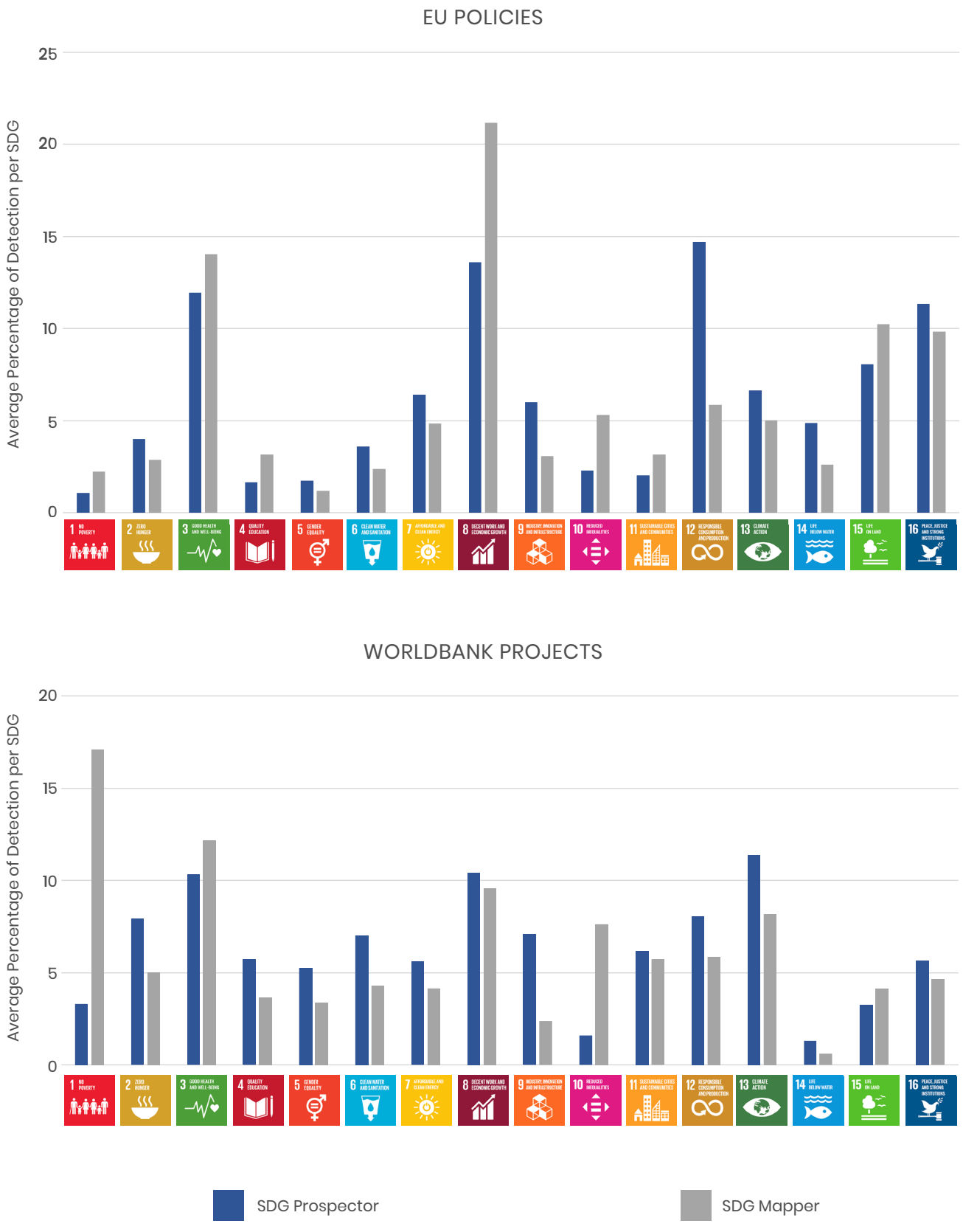
Table 2: SDG classification outcomes per type of document

	SDG MAPPER	SDG PROSPECTOR
WORLD BANK PROJECTS	119,194 references to the SDGs	84,942 paragraphs associated with at least one SDG
EU POLICIES	13,693 references to the SDGs	15,361 paragraphs associated with an SDG

The figure 1 shows a further breakdown of those detections and considers the average converted percentages of an SDG per tool and dataset to facilitate the comparison of classification outcomes.

This breakdown indicates notable differences between the tools in the detection of various SDGs. Notably, both tools converge in frequently classifying SDGs 3 (Good Health and Well-being) and 8 (Decent Work and Economic Growth) across both datasets. This indicates a shared sensitivity towards or focus on these SDGs in varying contexts. Additionally, SDG 13 (Climate Action) is another area where both tools demonstrate high detection rates within the World Bank dataset. However, the analysis also uncovers notable disparities in the detection of certain SDGs. A significant divergence is observed in the detection of SDG 1 (No Poverty), where *SDG Mapper* shows notably higher detections in the World Bank dataset compared to *SDG Prospector*. For SDG 10 (Reduced Inequalities), *SDG Mapper* again demonstrates higher detections in both datasets, indicating a greater sensitivity in classifying this goal. On the contrary, SDG 12 (Sustainable production and consumption) has higher detection rates across both datasets (in particular within EU policies) for *SDG Prospector*, as does SDG 14 (Life below water). These findings underscore that while *SDG Prospector* and *SDG Mapper* have areas of analytical synergy, they also possess distinct detection characteristics.

Figure 1: Detection frequency by SDG.



Based on the classification differences, classification results coming from *SDG Mapper* were further contextualized and analyzed to assess a potential over-detection for certain SDGs. Detected keywords and their frequencies were scanned for keywords associated with SDG 1 and 10, looking into their distribution to identify keywords potentially causing inflationary counts for those SDGs (see figure 2).

Figure 2: *SDG Mapper*'s detected keywords for SDG 1 and SDG 10.

Figure 2 shows the top-5 detected keywords for SDG 1 and 10, divided by the two datasets for a more nuanced overview. Keyword frequencies are much higher for World Bank projects as the average text length as well as the number of documents is significantly higher. This is also reflected in the classification results of *SDG Prospector* showing significantly higher numbers of SDG-relevant paragraphs identified in the World Bank dataset compared to EU policies. Diving deeper into the detected keywords of each SDG, keywords detected for SDG 1 within World Bank projects, show that "social risk" linking to target 1.3 on implementing appropriate social protection systems constitutes more than 50% of detected keyword frequencies for SDG 1's top-5 detected keywords (and beyond). Looking at a sample of documents with high frequencies for this keyword reveals certain types of World Bank project documents (e.g. Additional Financing Appraisal Environmental and Social Review Summary) that mention social risk in their document structure and hence automatically lead to higher detections (the document requires to report on assessment and Management of Environmental and Social Risks and Impacts). For EU policies, detected keywords for SDG 1 appear to be more balanced. Detected keywords of SDG 10 also indicate higher detection

rates driven by individual keywords in both datasets, but with emphasis placed on different topics (e.g. migration in World Bank projects and discrimination in EU policies).

To further analyze possible implications emerging from the different datasets, the table below provides a general overview on the average number of detected SDGs per tool and dataset considering different thresholds for retaining SDGs above certain detection percentages.

Table 3: Average number of detected SDGs

DATASET	TOOL	AVERAGE NUMBER OF SDGS (ALL VALUES)	AVERAGE NUMBER OF SDGS (VALUES ABOVE 5%)	AVERAGE NUMBER OF SDGS (VALUES ABOVE 10%)
World Bank Goal-level	SDG Mapper	12.6	4.5	2.0
	SDG Prospector	12.0	4.8	2.6
EU Policies Goal-level	SDG Mapper	6.2	3.2	2.2
	SDG Prospector	6.5	3.4	2.5

When considering all SDG detections, the average number of SDGs per document is substantially higher compared to filtering out detected SDGs below 5% or 10%. This might allude to an overdetection of relevant SDGs in both tools, potentially linked to the tools' classification approaches (e.g. 1 SDG-related keywords in *SDG Mapper* will indicate an SDG link as well as 1 paragraph classified as relevant to an SDG in *SDG Prospector*). The table also underlines differences between the two datasets, most likely stemming from the longer text lengths in the World Bank projects dataset. However, the average number of SDGs per document becomes very similar in both datasets when filtering out detected SDGs below 10%. This suggests that a comparison between these tools should focus on the top-detected SDGs rather than all detections as the classification outcomes in both tools might carry substantial noise in the detected SDGs. Therefore, the comparison of ranked SDGs focused on the top-3 detected SDGs of each tool exclusively. The table below provides a first overview on the tools' convergence by looking at the presence of a tool's top-detected goal within the other tool's top-3 detected goals for each document, showing that in the majority of cases top-detected goals are present in both tools across both datasets.

Table 4: Convergence between top-detected SDGs.

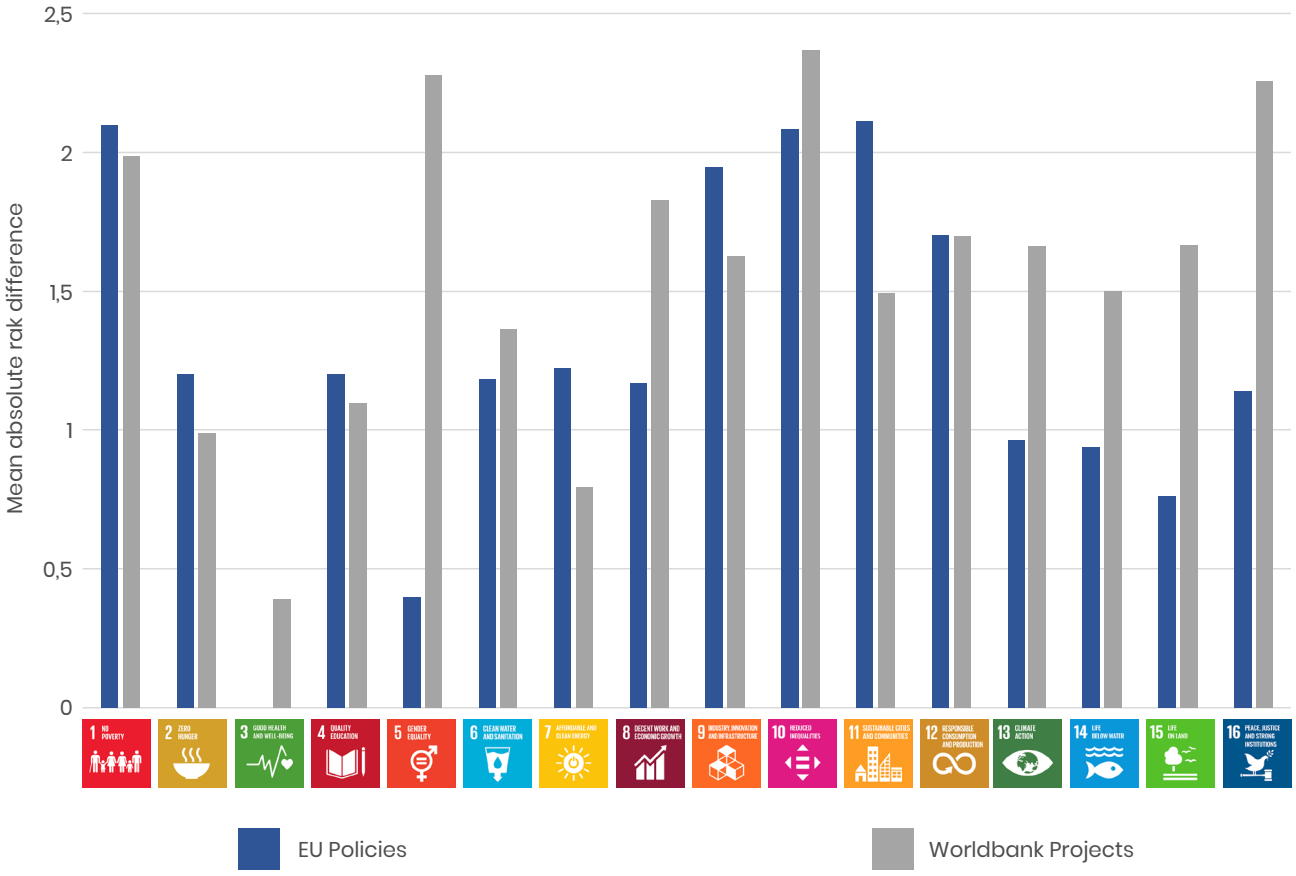
SDG MAPPER TOP-1 IN SDG PROSPECTOR TOP-3			
World Bank Projects Dataset		EU Policies Dataset	
Present	296 (80%)	Present	129 (85%)
Not Present	76 (20%)	Not Present	22 (15%)
SDG PROSPECTOR TOP-1 IN SDG MAPPER TOP-3			
World Bank Projects Dataset		EU Policies Dataset	
Present	318 (85%)	Present	131 (87%)
Not Present	54 (15%)	Not Present	20 (13%)

However, looking at actual rank differences indicates more granular disparities between the tools. While for EU policies, the average absolute rank difference per document is 4.1, for the larger dataset of World Bank projects, the classification results have a mean absolute rank difference of 7.1, likely due to the greater amount of text data leading to more detections in both tools and therefore, increasing the possibility for greater rank differences. This is further underlined when looking at the mean rank differences per SDG (see figure 3).

In the World Bank dataset, rank differences are higher (greater 1.5) for 9 out of 17 SDGs, while in the EU policies dataset rank differences are higher (greater 1.5) for 5 SDGs and mostly only marginally higher. High rank differences for both datasets occur for SDG 1 and SDG 10 which might be influenced by the high keyword detections in *SDG Mapper* for those goals. Within the EU policies dataset, the average rank difference for SDG 3 on health is 0 and remains also relatively low for the World Bank dataset, illustrating overall convergence between the tools in classifying this goal. Other SDGs with relatively low mean rank differences across both datasets are SDG 2, 4 as well as 6 and 7. The average rank differences per SDG as well as the average absolute rank differences per document underline that - despite some areas of convergence - the tools demonstrate substantial differences in the SDG classification for certain SDGs.

To gain a deeper understanding of those classification differences, a few documents were selected as case studies to further contextualize and validate the obtained classification results.

Figure 3: Significant differences in identifying multidimensional inequalities and poverty.



2.2 – Case studies

To complete the analysis, we present case studies to inform possible sources of discrepancies between the *Mapper* and the *Prospector*. We selected a set of 20 documents among World Bank's project descriptions and European Commission's policies. For each document, we relied on expert-knowledge to identify the main SDGs which were the most likely to be detected. To limit subjective biases, each document was reviewed by three separate analysts. In case of discrepancy, consensus was reached among the analysts by jointly reviewing the documents. Among these examples, we identified three possible outcomes in the results. First, we qualify quasi-perfect convergence whenever both tools identify the same main SDGs but their rank differs. The second outcome describes situations when both tools identify SDGs as one of the key topics, whereas this SDG was not identified by the analysts. This second case is described as a 'false positive', under the assumption that human analysis was correct. Finally, we identify projects in which there is a high discrepancy across both tools. The case studies below illustrate each of the aforementioned scenarios.

2.2.1 – Case study No1: Quasi-perfect convergence

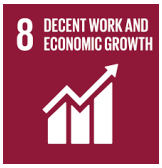
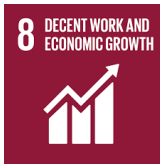
The number of documents for which there is perfect convergence between the *Mapper* and the *Prospector* across the 3 first SDGs is rare (only 3 occurrences out of the 561 projects from the World Bank). However, we identify projects for which both tools identified the same main SDGs, but their rank varies across the tools.

To illustrate this case, we rely on a communication document from the European Commission regarding the Eco-innovation Action Plan[4]. The document details the opportunities of putting in place a European action plan in order to stimulate productivity, sustain growth and job creation, especially for Small and Medium Enterprises (SMEs), while safeguarding the environment. As detailed in Table 5, the document echoes 3 main SDGs, namely SDG 9 "Industry, Innovation and Infrastructure", SDG 12 "Responsible Consumption and Production" and SDG 8 "Decent work and economic growth".

In accordance with the manual analysis, both the *SDG Mapper* and the *SDG Prospector* identified the three main topics in the document, which signifies a high level of convergence. However, we note slight differences between the two approaches. First, the aggregated frequencies for the 3 SDGs are stronger for the *Prospector* (77%) than for the *Mapper* (54%). This result is consistent with Table 3, which shows that on average, the *SDG Mapper* tends to identify a higher number of SDGs in analyzed documents. In this sense, the *SDG Mapper* may have a larger coverage of the SDGs contained in a text but when it comes to defining the main topics of the document, its results are less distinct than the *Prospector*'s. Second, we observe that the *Prospector* and the *Mapper* do not yield similar rankings across the 3 main SDGs. In this example, the *Mapper* emphasizes the importance of the eco-innovation plan in terms of growth and job creation, while the *Prospector* highlights the fact that the plan fosters sustainable production practices. In practice, this difference does not raise any issue for interpreting the content of the document as human analysts themselves may not be able to rank SDGs with certainty. This result supports the idea that when comparing SDG-classification tools, more emphasis should be put on the nature of detected SDGs rather than on their exact ranking.

[4] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A52011DC0899

Table 5: Case study no1: Innovation for a sustainable Future
The Eco-innovation Action Plan

RANK	EXPECTED SDGS BY THE ANALYSTS	SDG MAPPER		SDG PROSPECTOR	
		DETECTED SDG	FREQUENCY	DETECTED SDG	FREQUENCY
#1			27%		28%
#2			14%		26%
#3			13%		23%

2.2.2 – Case study N°2: The false positive

Interestingly enough, we observe cases in which both classification tools highlight SDGs which were not identified by human analysis. To illustrate this case, we rely on a World Bank project related to the management of coastal natural resources.

To start with, we note that the *SDG Mapper* makes the case for a strong prevalence of SDG 1 in this project while neither the *Prospector* nor the human-based analysis identified this SDG. This result is consistent with the descriptive statistics of Section 3.1. which shows that the *SDG Mapper* tends to over-identify SDG 1 compared to the *Prospector*, especially in World Bank project documents. While the *SDG Prospector* seems consistent with the analysts regarding the prevalence of SDG 14 "Life below Water", it failed to identify SDG 15 "Life on Land" contrarily to the *Mapper*. This is also consistent with the fact that, on average, the *Prospector* tends to under-identify this SDG compared to the *Mapper*.

Both tools identified SDG 8 "Decent work and economic growth" as one of the main SDGs targeted by this project, while analysts did not stress this particular topic. Two explanations can support this result. First it could be the case that both tools outperformed human analysis in identifying SDG 8. Undoubtedly, SDG 8 is relevant for this project, but it does not constitute the main objective of the project as detailed by the Project Development Objectives (PDO) in the document:

> ***"The proposed project development objective and global environment objective is to improve the capacity of the Project Countries to manage the transboundary natural resources of the Gulf of Fonseca, including for climate change adaptation."***

Moreover, while project documents show that the project can contribute to improving livelihoods of local populations, it does not include result indicators regarding the creation of jobs or economic value added. For these reasons, we lean towards the second explanation described as a 'false positive' situation. According to this explanation, both tools over-identified SDG 8 as one of the main objectives of the project. It should be noted that this result does not imply a structural weakness of both classification tools. Indeed, it could be the case that the authors of the analyzed document particularly stressed the role of this project in terms of job creation throughout the text, while the objectives stated in the section "Project Development Objectives" of the document are not explicit in this regard. Moreover, the analyzed documents include sections dedicated to the economic context of the project. It could be the case that the *Prospector* and the *Mapper* retrieved information from these sections and labeled them as references to SDG 8. To avoid such biases, it is important to isolate the parts of the text which are the most relevant to the comparative exercise, in particular when documents are only a few pages long.

Table 6: Case study no2: Gulf of Fonseca Transboundary Management of Coastal Natural Resources (World Bank - P176323)

RANK	EXPECTED SDGS BY THE ANALYSTS	SDG MAPPER		SDG PROSPECTOR	
		DETECTED SDG	FREQUENCY	DETECTED SDG	FREQUENCY
#1			30%		26%
#2			8%		14%
#3			9%		13%

2.2.3 – Case study No3: Discrepancies

Finally, we identify cases in which there is a relatively strong degree of discrepancy between the results of the *Prospector* and the *Mapper*. To illustrate such cases, we focus on a European regulation on classification, labeling and packaging of chemical substances and mixtures[5].

From the table below, we note that the *SDG Mapper* and the *SDG Prospector* both identified SDG 3 "Good health and well-being" and SDG 12 "Responsible Consumption and Production" as prevalent SDGs in the document. Both tools show that the SDG detections in this document are concentrated as the 3 most prevalent SDGs totalize 94% and 85% of the SDG detections respectively. However, the tools diverge regarding the weights they assign to each SDG: SDG 3 is particularly prevalent according to the *Mapper* (71%) but appears lower according to the *Prospector*. To go further, we present the results in absolute value:

Table 7: Case study no3: Regulations on classification, labeling and packaging of substances and mixtures (EU - PD10035)

RANK	EXPECTED SDGS BY THE ANALYSTS	SDG MAPPER		SDG PROSPECTOR	
		DETECTED SDG	FREQUENCY	DETECTED SDG	FREQUENCY
#1			71%		50%
#2			12%		24%
#3	n.a.		11%		11%

When looking at absolute frequencies of detection, we note that the tools differ greatly. First, the *SDG Prospector* detected 9 times more SDG references than the *SDG Mapper*. However, these figures do not say much about the quality of the detection, and it may be the case that the *Prospector* (*Mapper*) over (under) detected SDG references. Consequently, we

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008R1272

observe that even though SDG 3 occupies a strong part of the SDG detections according to the *Mapper*, the *Prospector* identified more than 50 more occurrences than the *Mapper*. Interestingly enough, some SDGs were strongly emphasized by the *Prospector* but yielded no occurrence from the *Mapper*. This is the case for SDG 7 "Affordable and Clean Energy" and SDG 6 "Clean Water and Sanitation". The prevalence of SDG 7 can be explained by the fact that the document mentions materials that are derived from petroleum e.g. "This note applies only to certain complex oil-derived substances", while the learning base of the *SDG Prospector* includes fossil energy-related snippets. In this example, the *Prospector* may have over-identified SDG 7. On the other hand, the document includes several references to water pollution which are consistent with target 6.3. of the 2030 Agenda[6]. For example:

> ***"Classification of substances and mixtures for environmental hazards requires the identification of the hazards they present to the aquatic environment. The aquatic environment is considered in terms of the aquatic organisms that live in the water, and the aquatic ecosystem of which they are part."***

While the example presented above is related to target 6.3., it does not describe an active contribution towards the achievement of this target. As such, one could argue that this paragraph should not be identified as a relevant reference to the 2030 Agenda. Yet, the fact that the *Mapper* did not label this type of paragraphs illustrates one of the main limitations of keyword-driven approaches. In the absence of certain keywords in the classifier's ontology, the tool is unable to link texts with possibly relevant topics. In the case of the *Mapper*, the list of keywords for SDG 6 only includes one possibility using the word "aquatic" in the expression "pollution of the aquatic environment". This may explain why the *SDG Mapper* did not label this type of snippets as related to SDG 6.

[6] "By 2030, improve water quality by reducing pollution, eliminating dumping and minimizing release of hazardous chemicals and materials, halving the proportion of untreated wastewater and substantially increasing recycling and safe reuse globally"

Table 8: Absolute discrepancies

Frequency of detection in the text (1,355 pages)

	SDG MAPPER	SDG PROSPECTOR
SDG 1	1	0
SDG 2	1	2
SDG 3	132	185
SDG 4	0	1
SDG 5	0	1
SDG 6	0	54
SDG 7	0	413
SDG 8	0	3
SDG 9	22	3
SDG 10	7	0
SDG 11	0	0
SDG 12	20	852
SDG 13	0	180
SDG 14	0	0
SDG 15	2	0
SDG 16	1	14
TOTAL	**186**	**1,708**

3. Deliverables and discussions

Building upon the results of the comparative analysis and the case studies presented above, this section offers an overall discussion on the determinants which explain the discrepancies, methodological limits, and possible convergence between SDG classification tools.

3.1 – Main results and explained discrepancies

This working paper presents the outcomes of a comparison between two SDG classification tools which apply different approaches to identify SDGs in text. Since the adoption of the 2030 Agenda ten years ago, progress in natural language processing and artificial intelligence have fueled innovative approaches to SDG classification. Automating SDG analyses can support thorough understanding of sustainability and limit biases from human interpretation. Yet, the multitude of classification tools yielding different results may create misunderstanding and raise concerns regarding the quality and efficiency of SDG-related NLP. For this reason, it is necessary to better understand observed discrepancies between classification tools to increase confidence and ownership over these tools.

Comparing SDG classification results between *SDG Prospector* and *SDG Mapper* showed general convergence for the top-detected SDGs across both datasets with slightly higher convergence in the EU policies dataset. Document length seems to have an effect on the number of detected SDGs, but this effect is resolved when filtering out detected SDGs below certain detection percentages. When considering the actual ranking of SDGs however, differences become more apparent and seem to be more pertinent for some SDGs than others.

Observed differences in SDG classification stem from the different methodological approaches between both tools. As the methodologies substantially differ in nature, results comparison could only be conducted at the document level which ultimately hampered a more granular comparison of classification results (e.g. on paragraph level). In the case of *SDG Mapper*, the definition and quality of SDG-specific keywords might impact the detection of certain SDGs and can potentially cause over- or underdetection of those goals. A semantic keyword approach as utilized in *SDG Mapper* may further be prone to inflationary keyword counts based on different document structures (e.g. the case of SDG 1 in the World Bank projects). On the other hand, the *SDG Prospector* allows for a large understanding of SDG-related paragraphs. In this respect, we conclude that the AI-based language model used by the *SDG Prospector* allows more accurate classification results than a keyword approach. It should be noted that, when using language models, the quality of the learning base for each SDG has implications for the quality of classification results.

3.2 – Recommendations to build a benchmarking dataset

To facilitate systematic comparisons and to test the validity of different approaches (Wulff *et al*., 2023), we identify several guidelines to build a robust benchmarking dataset.

To schematize, 3 types of classification tools exist: one category is dedicated to analyze policy documents (e.g., Linked SDGs, *SDG Mapper*, OSDG to a certain extent…). Academic researchers have also developed their own classification tools to track SDG references in research papers (e.g., Aurora, Elsevier). Finally, tools like the *SDG Prospector* (Global Goals Directory, text2sdgs) were created to inform business decisions and initiatives by analyzing annual reports and project documentation. To ensure interoperability between different classification needs, the benchmarking dataset should include maximum diversity to test the tools on different types of documents, including policies, project documentation, annual reports, strategies, research papers, news article, tweets, and social media. At the same time, certain SDG classification approaches developed for certain contexts that perform well for their specific use case should be subject to benchmarks specifically designed for their respective domain to get a thorough understanding of the quality of classification results. For example, classification tools, which are specialized in analyzing policy documents, may strive for highest accuracy on policy documents, regardless of their performance on other types of texts.

A similar concern relates to the length of texts composing the benchmarking dataset. As shown in Section 3, benchmarking results often depend on the length of analyzed texts. Snippets are useful to get a thorough understanding of sources for discrepancies between two tools. Comparing classification tools based on snippets can be particularly useful to assess their sensitivity to the presence/absence of certain keywords. On the other hand, analyzing longer text reveals useful information to appraise varying aggregation methods, whether they depend on keyword frequencies or on sliding windows.

Ideally, the benchmarking dataset should allow comparison over the whole scope of 2030 Agenda, at the goal and target level. Especially, the target level is particularly relevant for the analysis of operational documents. Furthermore, several tools do not allow for the classification of SDG 17 as it is considered a transversal objective based on broad targets. Digging in the classification of SDG 17-related text raises interesting methodological issues. Conceptually, it necessitates a clarification of concepts regarding notions such as partnerships, technical assistance, policy coherence, etc. As such, it questions our own understanding of language as a prerequisite for knowledge (Russell, 1950). Furthermore, it raises questions regarding classification tools' abilities to differentiate between operational targets (e.g. improve water quality) and what refers to means of achievement (e.g. strengthen the participation of local communities in improving water quality). This is an area where further research is needed.

Finally, the benchmarking dataset should ideally be constructed based on coordinated expert labelling to ensure such a benchmarking dataset matches the necessary diversity and quality. However, a benchmarking exercise can never be exhaustive: not all the SDGs (targets) can be covered with the same precision. If experts can devise rules to guide classification, then it should be possible to improve AI-classification. In this respect, the dataset should be available open-source and subject to ongoing refinements. This recommendation is in fact similar to the strategy which should be employed to fine tune a language-model. Ongoing efforts, as marked by the SDG Classification Expert Group[7] are steps in the right direction towards such a diverse, well-balanced, open-source and high-quality benchmarking dataset[8].

3.3 – Methodological limitations

To conduct this comparative analysis, we were constrained by the fact that the *Prospector* yields result at paragraph level while the *Mapper* provides results at document level. Even though the resolution of different tools might be larger, the document level is the lowest level at which we could make comparisons with our tools at the moment.

Due to the lack of a high-quality SDG-classification benchmarking dataset, the comparison between the tools was mainly focused on assessing the level of convergence of both tools on the two selected datasets. A common baseline has only been established for a selection of cases due to limited time and resources. Other datasets attempting to provide high-quality SDG-classification data like the OSDG-dataset[9] were not suitable for this analysis because of the structure of the dataset and the way the SDG-classification tasks were framed (asking community members to determine if a paragraph relates to a certain SDG but not determining which are the most relevant SDGs).

Eventually, we note that a paragraph does not need to mention an explicit advancement towards a given SDG to be labeled SDG-related by the *Mapper* or the *Prospector*. In other words, they are not able to differentiate between positive (e.g., “our project contributes to poverty reduction”), and negative contributions (e.g., “our project does not aim to combat poverty”). This broader definition may translate into over-detection of certain SDGs. Applications of sentiment analysis to SDG classification is beyond the scope of this report and left for further research.

These methodological limitations raise debates regarding the capacity of automatic solutions to fully harness the complexity of sustainability issues. Natural Language Processing,

[7] https://sdg-ai.org/

[8] https://github.com/SDGClassification/benchmark

[9] https://www.osdg.ai/

and more generally AI-based solutions relating to text, offer tremendous opportunities. They build bridges across different types of language and allow performing tasks in a snapshot, providing extremely capable approaches towards overcoming various communication hurdles. The advent of LLMs, with their ability to elaborate context and process vast datasets, suggest new paths toward more promising classification methodologies. In an age where LLMs exhibit a promising capacity for nuanced elaboration and support for decision-making, one must ponder whether these advanced computational models could redefine the standards for accuracy and reliability in SDG classification, thereby assessing the consistency of human analytical capacity in this domain. In essence, this analysis not only highlights the critical need for high-quality benchmarking datasets and the potential pitfalls of an oversaturated toolbox of classification methodologies but also suggests the need for a critical re-examination about the objective capacity of human analysts to conduct SDG classification consistently. It demands a forward-looking perspective on how we approach, develop, and evaluate SDG classification tools in alignment with the evolving landscape of artificial intelligence.

3.4 – Policy implications

Improving sustainability-linked classification can have several policy implications. For example, a common criticism is often addressed to extra-financial reporting as financial actors and firms use disparate and apply non-comparable methods to quantify their contributions, or negative externalities, to environmental and social objectives. Unlike traditional financial ratings, which extensively rely on subjective interpretation (UNDP, 2023), the use of automatic methods could minimize analytical biases and facilitate impact monitoring. Oyewole *et al.* (2024) provide a good review on current practices, and limitations, regarding the integration of AI-based solutions in sustainable finance. Robust data frameworks stand out as a key prerequisite to unleash the full potential of AI.

The 2030 Agenda has the ambition to constitute a common compass towards sustainable development worldwide. However, the Sustainable Development Goals undergo criticism by practitioners and in the academic literature: the SDGs are sometimes perceived as a set of objectives made up of too many targets (169) and indicators (232) leading to high degrees of complexity in operationalizing and implementing them. Biermann *et al.* (2022) assessed the political impact of the SDGs concluding that impacts have been mostly normative and discursive by establishing a common language that is shaping the way in which we talk about sustainable development on a global scale. Some authors (Mélonio & Tremel, 2021) also note tensions between the different goals, in particular the ability to reconcile environmental objectives with the growth objectives of SDG 8. Eventually, the United Nations (2024) observe that countries are off track on the 2030 Agenda as poverty

and climate objectives will not be met by 2030. If the SDGs were revised after 2030, it would necessitate updating classification tools to be consistent with the new Agenda as well as existing reporting and monitoring activities and capacities. Yet, investing in SDG classification today allows us to explore conceptual and methodological issues which can inform the design and application of NLP methods to other sustainability frameworks.

Conclusion

The comparative analysis of the *SDG Mapper* and the *SDG Prospector* underscores the pivotal role that high-quality benchmarking datasets play in the realm of SDG classification. Establishing a robust baseline is essential for evaluating the efficacy of novel classification methodologies against existing frameworks. Both tools demonstrated high convergence in detecting SDGs related to health, economic growth, and climate action, but significant divergences were observed in identifying goals such as poverty reduction and inequality. Our comparison further shows that the top sustainable development goal identified by either tool is highly consistent (80%) with the SDGs identified by the other. This finding highlights that, despite methodological differences, there is a significant overlap in the most critical areas of sustainable development detected by both tools. However, the proliferation of SDG classification tools raises significant concerns regarding their overall utility. Paradoxically, the abundance of such tools may be detrimental, sowing confusion and misunderstanding among users due to the lack of standardization and clarity in their application. In conclusion, while AI-based language models hold promise for more precise SDG classification, we stress the necessity of systematic benchmarking to enhance the robustness of automated SDG classification tools. Addressing methodological disparities, such as the differing levels of granularity in classification results, is crucial for refining these tools. A diverse and high-quality benchmarking dataset, encompassing various document types and lengths, would help in facilitating meaningful comparisons and improve classification accuracy. Future research should also explore the integration of sentiment analysis and the refinement of classification methodologies to further enhance their effectiveness. This will ensure that automated solutions can reliably support sustainable development efforts by providing accurate, transparent, and efficient SDG classification across diverse textual corpora.

References

Bender E.M., Gebru T., McMillan-Major A. & Shmitchell, S. (2021), On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? , in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623), ACM, New York, NY, USA, https://doi.org/10.1145/3442188.3445922

Berg F., Kölbel J.F., Rigobon R. (2022), Aggregate Confusion: The Divergence of ESG Ratings, *Review of Finance*, 1315-1344.

Biermann, F., Hickmann, T., Sénit, CA. *et al*. (2022), Scientific evidence on the political impact of the Sustainable Development Goals. *Nat Sustain* 5, 795–800. https://doi.org/10.1038/s41893-022-00909-5

Borchardt, S., Barbero Vignola, G., Buscaglia, D., Maroni, M. and Marelli, L. (2023), *Mapping EU Policies with the 2030 Agenda and SDGs*, EUR 31347 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-60474-7, doi:10.2760/87754, JRC130904.

Guisiano J. & Chiky R. (2021), *Automatic classification of multilabel texts related to Sustainable Development Goals*, TECHENV EGC2021, Montpellier, France. hal-03154261

Hu, X., Tian, Y., Nagato, K., Nakao, M., & Liu, A. (2023), Opportunities and challenges of ChatGPT for design knowledge management. ArXiv, abs/2304.02796. https://doi.org/10.48550/arXiv.2304.02796.

Jacouton JB., Marodon R., Laulanié A.(2022), *The Proof is in the Pudding - Revealing the SDGs with Artificial Intelligence*, AFD Research Papers N°262.

Joshi A., Gonzalez Morales L., Klarman S., Stellato A., Helton A., Lovell S. & Haczek A. (2020), *A Knowledge Organization System for the United Nations Sustainable Development Goals*.

LaFleur M.T. (2019), *Art is Long, Life is Short: an SDG Classification System for DESA Publications*, DESA Working Paper No.159.

LaFleur M.T. (2023), *Using large language models to help train machine learning SDG classifiers*, DESA WORKING PAPER NO. 180.

Le Blanc D. (2015), *Towards Integration at Last? The Sustainable Development Goals as a Network of Targets*, UN DESA Working Paper No. 14.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. & Stoyanov V. (2019), *RoBERTa:Facebook AI*, https://arxiv.org/abs/1907.11692.

Mélonio T., Tremel L. (2021), *Climat, Biodiversité, Inégalités… comment remettre les ODD sur les rails*, AFD Policy Paper N°7.

Miola A., Borchardt, S., Neher, F., Buscaglia, D. (2019), *Interlinkages and policy coherence for the Sustainable Development Goals implementation - An operational method to identify trade-offs and co-benefits in a systemic way*.

Naveed, H., Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023), *A Comprehensive Overview of Large Language Models*. ArXiv, abs/2307.06435. https://doi.org/10.48550/arXiv.2307.06435.

Oyewole A.T., Adeoye O.B., Addy W.A., Okoye C.C., Ofodile O.C., Ugochukwu C.E. (2024), Promoting sustainability in finance with AI: A review of current practices and future potential, *World Journal of Advanced Research and Reviews*, 21(03), 590–607.

Pincet A., Okabe S. & Pawelczyk M. (2019), *Linking Aid to the Sustainable Development Goals – A Machine Learning Approach*, OECD Working Papers.

Russell B. (1950), *An inquiry into Meaning and Truth*, Allen & Unwin, London.

Pukelis L., Bautista Puig N., Skrynik M. & Stanciauskas V. (2020), *Open-Source Approach to Classify Text Data by UN Sustainable Development Goals (SDGs)*, https://doi.org/10.48550/arXiv.2005.14569

United Nations (2024), *Inter-agency Task Force on Financing for Development, Financing for Sustainable Development Report 2024: Financing for Development at a Crossroads*. (New York: United Nations, 2024), available from:https://developmentfinance.un.org/fsdr2024.

United Nations Development Program (2023), *Reducing the Cost of Financing for Africa, The Role of Sovereign Credit Ratings* (New York: United Nations, 2023), available https://www.undp.org/

Wittgenstein L. (1922), *Tractatus Logico-Philosophicus*.

Wulff D.U., Meier D.S., Mata R. (2023), *Using novel data and ensemble models to improve automated labeling of Sustainable Development Goals*.

Xu F.F., Alon U., Neubig G., Hellendoorn V.J. (2022), A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (MAPS 2022)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3520312.353486

Glossary

CLASSIFICATION Classification is a type of NLP task where the goal is to analyze text and assign it to predefined categories or classes based on its features.

LARGE-LANGUAGE Model: "A Large Language Model (LLM) is an AI model trained on vast amounts of text data to understand and generate human language. It uses techniques like deep learning to perform tasks such as text completion, translation, summarization, and question answering based on patterns in the data." Definition obtained from Chat GPT.

LEARNING BASE Collection of input data which are used to train an AI-based model to replicate a task on new, unseen data.

NATURAL-LANGUAGE Processing: Among the many applications of artificial intelligence, text analysis is a matter of natural language processing (NLP). The aim is to model how humans understand and use language as a means of communication, so that the computer can perform tasks such as automatic translation, or the identification of specific themes addressed in a text.

SNIPPET In classification exercises, a snippet is usually defined as a short piece of text or data used as input to train or test a model. Here we use the term to qualify short paragraphs extracted from the analyzed documents.

Agence française
de développement
5, rue Roland Barthes
75012 Paris I France
www.afd.fr

What is AFD?
Éditions Agence française de développement publishes analysis and research on sustainable development issues. Conducted with numerous partners in the Global North and South, these publications contribute to a better understanding of the challenges faced by our planet and to the implementation of concerted actions within the framework of the Sustainable Development Goals.
With a catalogue of more than 1,000 titles and an average of 80 new publications published every year, Éditions Agence française de développement promotes the dissemination of knowledge and expertise, both in AFD's own publications and through key partnerships. Discover all our publications in open access at editions.afd.fr.
Towards a world in common.

Publication Director Rémy Rioux
Editor-in-Chief Thomas Melonio

Legal deposit 4rd quarter 2024
ISSN 2492 - 2846

Graphic design MeMo, Juliegilles, D. Cazeils
Layout PUB
Printed by the AFD reprography service

To browse our publications:
https://www.afd.fr/en/ressources-accueil