United Nations | Department of Economic and Social Affairs

# Using large language models to help train machine learning SDG classifiers

DESA WORKING PAPER NO. 180

Marcelo T. LaFleur

NOVEMBER 2023

**UNITED NATIONS**
**Department of Economic and Social Affairs**
UN Secretariat, 405 East 42nd Street
New York, N.Y. 10017, USA
e-mail: undesa@un.org
www.www.un.org/desa

# Using large language models to help train machine learning SDG classifiers

DESA WORKING PAPER NO. 180

**ST/ESA/2022/DWP/180**

**Abstract:** This paper proposes the use of synthetic training data generated by large language models to improve machine learning SDG classifiers. It shows that supplementing existing training data with synthetic data produced by the ChatGPT tool improves the performance of the SDGClassy classifier. This addition of synthetic data is especially useful in building SDG classifiers given the limited availability of properly labeled data and the complex, interconnected nature of the SDGs. Synthetic data thus enables more effective machine-learning applications in this context.

**JEL Classification:** O0 General Economic Development; O20 General Development Policy and Planning; C88 Other Computer Software.

**Sustainable Development Goals:** All.

**Keywords:** Sustainable Development Goals (SDGs); Machine learning; Generative AI models; ChatGPT; SDG classification; Topic models.

# Table of Contents

# 1. Introduction

The Sustainable Development Goals (SDGs) encompass a broad range of interconnected social, economic, and environmental challenges. A vast and rapidly growing body of work aims to advance and evaluate progress toward these 17 goals, to identify ongoing challenges and inform policymaking and related actions. Machine learning methods for automatically classifying text according to the SDGs can help organize and analyze this diverse body of work. However, the multi-dimensional nature of these goals, combined with their interconnectedness, presents unique challenges for machine learning classification.

One of the key barriers to applying machine learning for SDG classification is the scarcity of labeled data with which to train a model. Each of the SDGs encompasses multiple themes and concepts and traditional data collection and labeling methods struggle to capture the complex relationships within and between each SDG. As a result, the process of manually categorizing text according to the 17 SDGs is labor-intensive, subjective, and difficult to scale.

To address this limitation, this paper proposes the use of synthetic data generation using the generative AI model ChatGPT to improve the performance of machine learning SDG classifiers. ChatGPT, a model trained on a large and diverse dataset, has a remarkable ability to generate coherent and contextually relevant text that can mimic human language. It is posited that ChatGPT can generate synthetic text that represents the scope and interconnected nature of each SDG. This paper explores this ability and evaluates the performance gains achieved by supplementing existing labeled data with synthetic data generated by ChatGPT.

This paper contributes a novel methodology for generating synthetic data to mitigate the lack of sufficient labeled data for complex classification tasks in the domain of SDG classification. The results of this work could help pave the way for improved monitoring and analysis of progress toward achieving the SDGs.

# 2. Challenges with SDG Classification

Machine learning classifiers require large training datasets that can provide the algorithm with examples of the "right answers." Such a labeled training dataset is lacking for use in SDG classification. Creating this dataset requires that a significant number of texts are manually categorized according to the SDGs, optimally with a weight for each of the 17 goals. This process is necessarily subjective, prone to inconsistencies, resource-intensive, and difficult to scale.

Developing guidelines and training annotators to achieve agreement on how to consistently label data according to the 17 SDGs is challenging given the complexity and interconnectedness of the concepts. Each SDG encompasses a broad range of social, economic, and environmental themes, captured in over 200 targets and indicators. Progress in some is deeply dependent on progress in others. For instance, achieving zero hunger (SDG 2) is tied to access to healthcare (SDG 3), education (SDG 4), sustainable energy (SDG 7) and sustained economic growth (SDG 8). Similar interconnections exist between all the other SDGs and subjectivity and inconsistency in the labels are therefore hard to avoid. While it may be possible to identify the SDG corresponding to a single sentence, quantifying the degree to which a longer text addresses each of these deeply intertwined goals involves a high degree of subjectivity and is difficult to achieve consistently at the level of detail required for a well-performing classifier.

This problem is compounded by the volume of data needed to represent the full scope of the 17 SDGs. Datasets used for machine learning algorithms can range from thousands to millions of observations.[1] The costs involved in manually labeling this volume of data to the level of detail necessary are substantial. Some approaches are focused on methodically building a training dataset through crowd-sourcing the training classification. Others rely on keywords in abstracts, or a selection of SDG-specific journals as sources (OSDG, UNDP IICPSD SDG AI Lab, and PPMI, 2021).

Early attempts to classify documents in the SDG space were limited by the lack of large datasets (LaFleur, 2019; LaFleur and Kim, 2020; Le Blanc, Freire and Vierros, 2017; Le Blanc, 2015; CDP Subgroup on voluntary national reviews, 2019). More recent initiatives have made important methodological advances, but agreement on an appropriate training dataset remains elusive (Table 1).

Two of these classifiers use methodological approaches that greatly reduce the need for a labeled training dataset. UN DESA's "LinkedSDG" tool relies on an "SDG ontology" that maps the relationships between individual SDGs, targets, and indicators to terms from the UNBIS Thesaurus, and employs natural language processing to extract terms from unstructured text. By using the interrelationships in the UNBIS Thesaurus, LinkedSDG associates documents with specific goals, targets, and indicators (W3C,

1 Wikipedia has a helpful list of some of the more popular machine-learning training datasets and their sizes: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research.

TABLE 1

## List of recent SDG classification initiatives

| NAME | ORGANIZATION | URL |
|---|---|---|
| **CountryRisk.io** | CountryRisk.io | https://www.countryrisk.io/ |
| **EUR-SDG-Mapper** | Erasmus University Rotterdam & Dialogic | https://github.com/dialogicnl/eur-sdg |
| **Gemeinschaftswerk Nachhaltigkeit** | German Council for Sustainable Development: (together with Exxeta and GFA) | https://gemeinschaftswerk-nachhaltigkeit.de/en |
| **Global Goals Directory** | Global Goals Directory (2030 Ecosystems UG) | https://globalgoals.directory/ |
| **LinkedSDGs** | UN DESA | https://linkedsdg.officialstatistics.org/#/ |
| **SDGClassy** | UN DESA | https://github.com/SeaCelo/SDGclassy |
| **OSDG.ai** | OSDG | https://www.osdg.ai/ |
| **SDG Classifier** | Athena Research & Innovation Center | https://explore.openaire.eu/sdgs |
| **SDG Classifier** | Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) | https://huggingface.co/spaces/GIZ/sdg_classification |
| **SDG Mapper** | European Commission Joint Research Center (JRC) | https://knowsdgs.jrc.ec.europa.eu/sdgmapper |
| **SDG Pathfinder** | OECD | https://sdg-pathfinder.org/ |
| **SDG Prospector** | Agence Française de Développement | https://sdgprospector.org/ |
| **SDG Research Dashboard** | Aurora Alliance | https://aurora-universities.eu/sdg-research/ |
| **SDG Research Mapping Initiative** | Elsevier | https://www.elsevier.com/about/partnerships/sdg-research-mapping-initiative |
| **South African SDG Hub** | South African SDG Hub | https://sasdghub.up.ac.za/ |

**Source:** https://globalgoals-directory.notion.site/List-of-SDG-Classification-Initiatives-7ab6bdc16e8e484793c9903e853ff0d5.

n/d; UN DESA, 2019). The result of this structured approach is the ability to effectively determine the relatedness of different items to the SDGs, helping to link unstructured documents to SDG concepts.[2]

"SDGClassy" uses a relatively small collection of 17 representative texts as the training dataset and trains a topic model that can identify the differences between them. SDGClassy currently relies on a survey of existing documents that are reasonably focused on each SDG by design, accurately capturing the breadth, depth, and intricacy of each SDG. Using this data, the SDGClassy classifier leverages the ability of LDA algorithms to create probabilistic models capable of differentiating among the desired classification groups (Blei, 2012; LaFleur, 2019).[3] This selected group of documents can be used as models of how the SDGs are discussed in published documents rather than in theory. A classification is possible by comparing the similarity of any document to these examples. The training dataset includes a total of 68,647 words from the following sources:

The text for the UN webpage that describes each SDG. For example, the representative text for SDG 1 includes the text for the webpage found at www.un.org/sustainabledevelopment/poverty/.

The relevant section of the Secretary-General's annual report "Progress towards the Sustainable Development Goals" for 2016, 2017, and 2018 (for example, http://undocs.org/E/2018/64).

The relevant sections of the "Special Edition of the Sustainable Development Goals Progress

2 While the LinkedSDG tool avoids the need to manually tag each document with a specific SDG, it does require expert intervention to link each goal, target, and indicator to specific terms in the UNBIS Thesaurus.

3 The training dataset is processed by removing words and terms that may confuse the classifier and are not meaningful to a classification. For example, the term "United Nations" and regional terms such as "Africa" are excluded as they are very prominent in the training data and are not associated with any of the concepts of the 17 SDGs. The list of excluded words is available from the SDGClassy repository: https://github.com/SeaCelo/SDGclassy.

Report," available at https://sdgs.un.org/documents/special-edition-progress-towards-sustainable-25359.

The full text of all the targets and indicators for each SDG, available at https://unstats.un.org/sdgs/indicators/indicators-list/.

The approach used by SDGClassy offers great flexibility since it mostly relies on identifying examples of documents that discuss each SDG. However, much like all other classifiers, the SDGClassy tool cannot be expected to give accurate results when applied to a text with a very different vocabulary. To improve its applicability to a wider range of document types, the classifier can be trained on a broader range of representative examples (e.g., analytical reports, speeches, academic writing, and news items). But this brings us back to the limitations of existing training data: how to identify appropriate examples for each of the SDGs? Existing classification systems depend on problematic training data and so cannot be used to identify representative documents. A new approach must be used.

Rather than trying to identify more representative texts for each SDG, it is possible to instead use the generative abilities of large language models to produce the desired representative data. With carefully designed prompts, AI tools can generate large quantities of text that reflect how each SDG is discussed in the billions of documents used to train these models. Such synthetic data generation using AI models could therefore provide the training data needed to improve the performance of machine learning classifiers.

# 3. Generating Synthetic Training Data for SDG Classification

Generative AI tools like OpenAI's ChatGPT can produce large volumes of text based on user-provided prompts, offering a high level of control over the generated content. Having been trained on vast and diverse corpora of text data, these models can generate coherent and contextually relevant text spanning a wide array of topics. The quality of these models has increased significantly over the last five years and well-designed prompts yield high-quality, reliable results.

A large language model such as ChatGPT is fundamentally trying to produce a "reasonable continuation" of whatever text it already has (Wolfram, 2023). Essentially, the model continuously answers the question "Given the text so far, what should the next word be?" In this way, the model generates text that one might expect from the examples found in the billions of web pages and documents used to train the model. How well the model achieves this has been improving in line with two factors: the quantity of training data that is used and the number of parameters, or model size (OpenAI, 2023). While the quantity of training data is not published (GPT-3 was said to be trained on 40 Gigabytes of text), the number of parameters is documented. OpenAI's GPT-1 model used 117 million parameters, while GPT-2 used 1.5 billion.[4] GPT-3, released in February 2021, marked a watershed in model performance and was trained with 175 billion parameters. The most recent version, GPT-4, is another leap in performance and is estimated to have been trained on 100 trillion parameters (Geyer, 2023).

The methodology and the performance of these large language models make it possible to leverage them to generate large quantities of texts that are representative of each SDG and to build a robust training dataset. Although the models are not designed to provide factually correct results, they can provide representative vocabulary and semantic structures that reflect how the SDGs are discussed in the billions of published documents.

## Prompt design strategy and methodology

Large language models must be given clear instructions that elicit relevant responses.[5] The proposed methodology for generating synthetic datasets for SDG classification consists of a careful approach to formulating the prompts that rely on two principles. First, the prompts must be broad and contextual, rather than narrow and isolated, to reflect the interconnected nature of the SDGs. This means that prompts should request examples about the overall SDG, rather than a specific aspect such as a policy or a target. This encourages the generation of text that not

4 Parameters are the values that a neural network tries to optimize during training for the task at hand.

5 There is a growing scholarship on how to optimize these instructions. The Prompt Engineering Guide project tracks academic work on the impact of prompt design on output (https://www.promptingguide.ai/papers).

Table 2

**Variations in the prompts used to generate representative texts for each of the 17 SDGs**

| PUBLICATION TYPE | STYLE VARIATION | AREAS OF FOCUS |
|---|---|---|
| **Newspaper article** | Various newspapers | Global |
| **Op-ed** | Various authors | Middle-income countries |
| **Magazine article** | Various magazines | Least developed countries, small island developing states, and landlocked developing countries. |
| **Academic article** | Various journals | |
| **Essay** | Various speakers | |
| **Speech** | Various audiences | |
| **Analytical report** | Various institutions | |

only focuses on the primary themes of each SDG but also explores areas that overlap with other goals. It is this holistic approach that allows us to capture the complex and interconnected nature of the SDGs.

Second, there must be sufficient but consistent variety in the data generated by the prompts to give the classifier a broad set of examples. This is not a trivial task since simply asking for multiple examples of the same prompt results in repetitive results and there are limited ways that one can ask for an explanation of a particular SDG. The variety must also be consistent for all 17 SDGs to avoid introducing a bias to the classifier. The strategy used allows for a degree of variation in the publication type, publication style, and area of focus of the texts. This same variation is used for each of the 17 SDGs (Table 2).

Various newspaper styles are important to broadly emulate the style of mainstream news coverage. The styles included are The New York Times, the Financial Times, USA Today, Bloomberg, and the Wall Street Journal. The different styles of these news outlets enable the models to generate texts that not only cover the SDGs but also incorporate the unique writing style, tone, and perspective of each publication. This aids in creating a more diverse dataset and training the classifier to handle different styles and tones of writing.

Different styles of opinions (op-eds) were included to emulate texts that possess a more subjective, forceful, informed, and focused style than is produced in normal news writing. The prompts used for the dataset include opinions from Nicholas Kristof, Paul Krugman, and Mark Whitehouse. The specific authors were selected based on a cursory search of SDG-related opinion pieces. Future work should expand the list of authors. Different styles of magazine articles were included to emulate in-depth features and analyses. The prompts include The Economist, the Harvard Business Review, Time Magazine, Forbes, and Newsweek.

Variation in academic articles is important to mimic scholarly discourse from journals and produce texts that are dense, nuanced, and heavily informed by research. The prompts include the Journal of International Development, the Journal of Development Economics, the Review of Development Economics, the Journal of Economic Literature, the World Development Journal, and the World Bank Economic Review. College-level essays were also included as an academic style.

Different styles of analytical reports are created to emulate research based on systematic, data-driven examinations to inform policymaking as is produced by researchers from several United Nations Departments, Divisions, Regional Commissions, Funds, and Programs. The styles included are UN DESA, EAPD, DPIDG, DISD, DSDG, FSDO, UNDP, UNCTAD, ECA, ECE, ECLAC, ESCAP, and ESCWA. These prompts are vital for generating synthetic text that mirrors

Table 3

**Selected examples of prompts used to generate representative texts for each of the 17 SDGs**

| |
|---|
| Draft a long newspaper article about the SDG in the style of the Financial Times |
| Draft a long academic article about the SDG in the style of the Journal of Development Economics |
| Draft a long article about the SDG in the style of The Economist |
| Draft a long fictional detailed analytical report on the current SDG in the style of UN DESA targeting an expert audience |
| Draft a long fictional detailed analytical report on the current SDG in the style of UNDP targeting an expert audience and with a focus on middle-income countries |
| Draft a long fictional detailed analytical report on the current SDG in the style of ECLAC targeting an expert audience and with a focus on least developed countries, small island developing states, and landlocked developing countries. |

the detailed, technical language and in-depth analysis that are characteristic of such reports. The prompts also included styles of speeches from leading figures in development discourse to generate text that is structured and replete with rhetorical devices and persuasive language, allowing the classifier to categorize spoken content effectively.

Prompts about analytical reports also vary according to geographic and economic contexts: middle-income countries, least-developed countries, small island developing states, and landlocked developing countries. This allows for the generation of text that considers the SDGs within different socio-economic and geographic contexts, enhancing the breadth and depth of our synthetic dataset.

Finally, before submitting each of the unique prompts to the ChatGPT API, the model is given a common preamble instruction. This preamble asks the model to provide responses based on a common set of assumptions and serves to focus the responses on a common audience type. The preamble used to generate the training data is:

"You are a knowledgeable assistant. You are an expert on the Sustainable development goals, including how they are discussed in the annual report of the secretary-general titled "progress towards the sustainable development goals," and in speeches and reports from the United Nations. You are familiar with the work of the following organizations: UN DESA (and its divisions EAPD, DPIDG, DISD, DSDG, FSDO), UNDP, UNCTAD), and the five UN regional commissions. You will respond to the following prompt with a long response with as much detail as possible."

For each of the 17 SDGs, the combination of style variations resulted in 62 unique prompts, and a total of 1,054 texts were synthetically created using ChatGPT's public API. An illustrative selection of the wording used in each prompt is shown in Table 3. The 62 prompts used in this paper are meant as an initial list to validate this approach and are not intended to represent the optimal set of prompts for SDG classification.

Adjustments can likely be made to the synthetic data by changing the choice of publications, authors, and institutions in a way that both improves classifier's performance and reduces any bias. This could be done by, for instance, randomizing the selection of publication and author styles and studying if the classification results are sensitive to different prompt choices. Such "prompt engineering" – improving text inputs for better communication with advanced language models – is a fruitful area of research in the use of large language models and the World Economic Forum named it the top "job of the future" (World Economic Forum, 2023). This paper does not explore how such prompt engineering can impact the training dataset and the results of the classification model and this is left for future research. Instead, the guiding focus of generating prompts is to achieve a consistent and balanced training dataset that minimizes the sources of classifier bias, as explained in LaFleur (2019).

# 4. Experimental Results and Evaluation

The first step to evaluate the performance of the classifier is to determine how well it differentiates among the 17 SDGs using the training dataset. Like the original SDGClassy classifier, the new "SDGClassy+" does an excellent job of differentiating among the 17 goals in the training sample (Table 4). In this matrix, the diagonal terms represent the correct classification for each SDG and the percentage indicates the degree of confidence.

Table 4

**Correspondence between estimated topics and SDG-representative texts using SDGClassy+**

| | ESTIMATED SDG USING SDGCLASSY+ | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TRAINING DATA | SDG 1 | SDG 2 | SDG 3 | SDG 4 | SDG 5 | SDG 6 | SDG 7 | SDG 8 | SDG 9 | SDG 10 | SDG 11 | SDG 12 | SDG 13 | SDG 14 | SDG 15 | SDG 16 | SDG 17 |
| SDG 1 | 98% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 2% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 2 | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 3 | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 4 | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 5 | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 6 | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 7 | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 8 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 9 | 0% | 0% | 0% | 0% | 0% | 0% | 3% | 0% | 97% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 10 | 2% | 0% | 0% | 0% | 3% | 0% | 0% | 0% | 0% | 95% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 11 | 0% | 0% | 0% | 0% | 0% | 1% | 0% | 0% | 0% | 0% | 99% | 0% | 0% | 0% | 0% | 0% | 0% |
| SDG 12 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% |
| SDG 13 | 0% | 0% | 0% | 0% | 0% | 0% | 3% | 0% | 0% | 0% | 0% | 0% | 97% | 0% | 0% | 0% | 0% |
| SDG 14 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 98% | 1% | 0% | 0% |
| SDG 15 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 1% | 1% | 99% | 0% | 0% |
| SDG 16 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% |
| SDG 17 | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% |

**Note:** The percentage in each cell represents the association between the model and the training dataset. This is computed as the optimal sampling distribution that minimizes the distance between the data and the model. See Blei (2012) for a detailed description. Cell shading corresponds to its proportion in each SDG (row). Totals may not add to 100% due to rounding.

Table 5

### WESS 2013 chapter titles and equivalent SDGs

| WESS 2013 CHAPTERS | EQUIVALENT SDGS |
|---|---|
| **1. Global trends and challenges to sustainable development post-2015** | Goal 13: Climate Action |
| **2. Strategies for development and transformation** | Goal 12: Responsible Consumption and Production |
| **3. Towards sustainable cities** | Goal 11: Sustainable Cities and Communities |
| **4. Ensuring food and nutrition security** | Goal 2: Zero Hunger |
| **5. The energy transformation challenge** | Goal 7: Affordable and Clean Energy |

To evaluate the performance of the classifier out-of-sample, the analysis follows the same procedure and the same report as in the original paper describing the SDGClassy methodology (LaFleur, 2019). The results of the classifier are evaluated using a complex report where the expected results are reasonably obvious (though the specific weightings are not knowable). The ability of the classifier to differentiate between SDG 16 and 17 is also evaluated, something that was identified as a weakness in the original SDGClassy classifier. Finally, the augmented classifier is applied to several editions of a complex multithemed report to evaluate its ability to correctly identify multiple themes.

## Testing the classification with complex reports: WESS and WSR

To compare the performance, both versions of the classifier are used to compute SDG Scores for the 2013 World Economic and Social Survey titled "Sustainable Development Challenges."

This report is useful because it is a long-form publication that touches on several SDGs. The five chapters of the 2013 WESS are a good indication of the five main SDG themes that are the focus of the report and against which the SDG classifiers are tested (Table 5).

The findings from the two classifiers are consistent with the expected results: the top five SDGs identified by both classifiers are a good reflection of the actual contents of the 2013 WESS. Notably, SDGClassy+ more strongly identifies the five SDGs, as shown by a higher sum of adjusted SDG scores for the top 5 SDGs (Table 6).[6]

We also reviewed how well the classifier captures the nuances of the World Social Report (WSR). The WSR advances discussion and policy analysis of socio-economic matters by identifying emerging social trends of international concern and relationships among major development issues. In the execution of this mission, the report addresses how thematic questions such

Table 6

### Comparing the results of SDGClassy and SDGClassy+ on the contents of WESS 2013

| SDGCLASSY | | SDGCLASSY+ | |
|---|---|---|---|
| **SDG** | **ADJUSTED SCORE** | **SDG** | **ADJUSTED SCORE** |
| Goal 2: Zero Hunger | 16.4% | Goal 2: Zero Hunger | 26.1% |
| Goal 11: Sustainable Cities and Communities | 14.9% | Goal 11: Sustainable Cities and Communities | 16.5% |
| Goal 12: Responsible Consumption and Production | 13.0% | Goal 7: Affordable and Clean Energy | 15.6% |
| Goal 13: Climate Action | 12.3% | Goal 12: Responsible Consumption and Production | 10.5% |
| Goal 7: Affordable and Clean Energy | 10.4% | Goal 13: Climate Action | 9.9% |
| **Sum of Top 5 scores** | **66.9%** | **Sum of Top 5 scores** | **78.6%** |

6 The adjusted scores are computed as a percentage of the total sum of the 17 scores. The sum of the 17 adjusted SDG scores is 100%.

Table 7

**World Social Report 2005-2018 classification using SDGClassy+**

| YEAR | TITLE | SDG 1 | SDG 2 | SDG 3 | SDG 4 | SDG 5 | SDG 6 | SDG 7 | SDG 8 | SDG 9 | SDG 10 | SDG 11 | SDG 12 | SDG 13 | SDG 14 | SDG 15 | SDG 16 | SDG 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2005 | The Inequality Predicament | 11.6% | 3.7% | 15.9% | 6.5% | 12.2% | 0.0% | 0.0% | 25.9% | 0.0% | 15.3% | 0.0% | 1.3% | 0.0% | 0.0% | 0.4% | 7.1% | 0.0% |
| 2007 | The Employment Imperative | 7.2% | 2.2% | 4.7% | 8.2% | 11.3% | 0.0% | 0.0% | 58.4% | 0.0% | 7.9% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| 2010 | Rethinking Poverty | 39.8% | 7.1% | 7.6% | 6.0% | 5.9% | 0.1% | 0.0% | 20.1% | 0.0% | 12.5% | 0.5% | 0.0% | 0.0% | 0.0% | 0.0% | 0.5% | 0.1% |
| 2011 | The Global Social Crisis | 13.1% | 26.2% | 8.7% | 6.8% | 2.9% | 0.4% | 0.3% | 36.7% | 0.0% | 4.2% | 0.0% | 0.3% | 0.1% | 0.4% | 0.0% | 0.0% | 0.0% |
| 2013 | Inequality Matters | 11.0% | 2.0% | 11.5% | 19.7% | 6.8% | 0.0% | 0.0% | 17.1% | 0.0% | 27.4% | 4.4% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| 2016 | Leaving no one Behind: The Imperative of Inclusive Development | 12.7% | 0.0% | 5.9% | 14.5% | 22.4% | 0.0% | 0.0% | 16.2% | 0.0% | 18.4% | 0.5% | 0.0% | 0.0% | 0.0% | 0.0% | 9.3% | 0.0% |
| 2018 | Promoting Inclusion Through Social Protection | 38.3% | 0.0% | 12.6% | 11.0% | 9.0% | 0.0% | 0.0% | 16.6% | 0.0% | 11.3% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 1.2% | 0.0% |

**Note:** Row totals add to 100%. Cell shading corresponds to its proportion in each SDG (row).

as aging and equality matter for poverty (SDG 1), employment (SDG 8), inequality (SDG 10), and social issues such as hunger (SDG 2), health (SDG 3), education (SDG 4), and gender (SDG 5).

The SDGClassy+ classifier correctly captures the nuance in the content of each WSR (Table 7). The 2007 report, for instance, is about employment, which is reflected in the strong weight of SDG 8. The 2010 report is focused on poverty. The 2011 report is titled "The Global Social Crisis," but the main chapters discuss jobs (SDG 8), income and poverty (SDG 1), and food access (SDG 2). The 2018 report discusses the broader concept of social protection, but it has a heavy focus on its role in preventing and reducing poverty (SDG 1) and promoting broader social development (SDGs 3, 4, 5, 8, and 10).

## SDGClassy+ is better at classifying SDGs 16 and 17

The new training dataset also improves the ability of the classifier to differentiate between SDG 16 and SDG 17, a problem that was observed when classifying reports dealing with governance. This issue was evident when analyzing UN DESA's World Public Sector Reports (WPSR), which are analytical reports on how to foster "effective, efficient, transparent, accountable, inclusive and innovative public governance, administration and services for sustainable development."[7] The original SDGClassy often confused the language of SDG 16 "Peace, justice and strong institutions" with the language of SDG 17 "Partnerships for the goals" (Table 8). This was mostly because discussions about both

7 https://publicadministration.un.org/en/About-Us/Who-We-Are

Table 8

## Average classification of UN DESA publications by type using SDGClassy

| SDG | CDP POLICY NOTE | DESA WORKING PAPER | GSDR | NDS POLICY NOTE | DESA POLICY NOTE | OTHER REPORT | RWSS | WESS | WPSR | WYR |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.90% | 7.20% | 8.50% | 5.60% | 4.40% | 4.70% | 20.60% | 5.00% | 5.00% | 4.30% |
| 2 | 5.10% | 7.90% | 1.10% | 10.00% | 10.90% | 1.30% | 5.60% | 9.30% | 1.60% | 0.60% |
| 3 | 4.00% | 1.60% | 0.90% | 0.50% | 3.60% | 3.90% | 3.80% | 2.50% | 0.90% | 5.00% |
| 4 | 1.10% | 2.40% | 3.50% | 1.90% | 0.10% | 1.60% | 7.00% | 1.90% | 2.30% | 15.40% |
| 5 | 1.40% | 2.20% | 3.70% | 1.10% | 0.30% | 14.10% | 6.40% | 1.00% | 5.60% | 8.60% |
| 6 | 2.30% | 2.90% | 5.60% | 1.50% | 2.80% | 1.80% | 0.50% | 1.90% | 5.00% | 1.00% |
| 7 | 3.40% | 1.30% | 2.50% | 1.10% | 9.60% | 0.00% | 0.00% | 3.90% | 0.20% | 0.30% |
| 8 | 6.70% | 12.00% | 1.30% | 10.50% | 5.80% | 7.30% | 22.20% | 8.60% | 3.50% | 30.40% |
| 9 | 5.70% | 7.20% | 10.40% | 10.00% | 5.70% | 4.90% | 2.70% | 8.60% | 3.80% | 3.00% |
| 10 | 11.20% | 17.30% | 3.50% | 14.40% | 6.20% | 8.10% | 15.90% | 13.70% | 7.10% | 6.90% |
| 11 | 0.80% | 1.80% | 6.00% | 1.60% | 3.30% | 11.50% | 1.90% | 2.50% | 3.80% | 1.20% |
| 12 | 3.80% | 4.70% | 6.00% | 7.10% | 2.60% | 1.40% | 0.90% | 5.00% | 2.40% | 1.80% |
| 13 | 19.70% | 9.00% | 11.30% | 6.60% | 20.00% | 9.80% | 3.10% | 13.60% | 11.20% | 9.80% |
| 14 | 1.30% | 1.70% | 10.70% | 1.50% | 3.50% | 0.50% | 0.10% | 1.60% | 0.80% | 0.40% |
| 15 | 1.40% | 1.80% | 4.60% | 1.20% | 5.60% | 1.20% | 0.30% | 2.40% | 1.40% | 0.70% |
| 16 | 5.40% | 4.20% | 4.20% | 5.80% | 1.50% | 4.80% | 5.00% | 3.70% | 20.40% | 5.20% |
| 17 | 22.70% | 14.70% | 16.20% | 19.80% | 14.10% | 23.30% | 4.00% | 14.90% | 25.00% | 5.40% |

**Note:** Column totals add to 100%. Cell shading corresponds to its proportion in each SDG (column).

Table 9

## Average classification of UN DESA publications by type using SDGClassy+

| SDG | CDP POLICY NOTE | DESA WORKING PAPER | GSDR | NDS POLICY NOTE | DESA POLICY NOTE | OTHER REPORT | RWSS | WESS | WPSR | WYR |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.20% | 10.10% | 6.30% | 7.70% | 4.30% | 3.90% | 18.80% | 5.90% | 3.00% | 2.60% |
| 2 | 5.30% | 5.80% | 3.60% | 6.30% | 12.10% | 1.40% | 5.80% | 9.70% | 0.80% | 1.20% |
| 3 | 12.10% | 6.40% | 6.80% | 5.00% | 7.70% | 9.80% | 9.50% | 9.20% | 7.70% | 8.00% |
| 4 | 3.60% | 4.10% | 5.80% | 3.90% | 0.40% | 4.50% | 10.50% | 3.80% | 6.00% | 26.50% |
| 5 | 3.00% | 3.30% | 6.20% | 2.30% | 0.40% | 24.60% | 10.10% | 1.60% | 10.90% | 17.00% |
| 6 | 0.50% | 2.20% | 7.40% | 1.70% | 3.20% | 1.30% | 0.10% | 1.70% | 2.50% | 0.80% |
| 7 | 7.10% | 3.20% | 4.20% | 5.50% | 15.30% | 0.40% | 0.00% | 9.00% | 0.20% | 0.80% |
| 8 | 8.90% | 21.10% | 2.80% | 19.40% | 8.10% | 9.60% | 27.60% | 14.50% | 4.20% | 27.40% |
| 9 | 2.60% | 4.40% | 8.60% | 9.70% | 2.00% | 3.60% | 0.00% | 5.50% | 3.90% | 0.10% |
| 10 | 7.70% | 12.30% | 2.60% | 9.30% | 1.70% | 3.40% | 13.80% | 8.00% | 3.60% | 1.30% |
| 11 | 0.20% | 1.20% | 6.80% | 1.10% | 4.70% | 18.30% | 0.80% | 2.20% | 4.50% | 0.70% |
| 12 | 2.10% | 4.60% | 4.50% | 5.10% | 2.50% | 0.40% | 0.20% | 4.20% | 2.40% | 0.90% |
| 13 | 17.80% | 6.10% | 5.10% | 2.20% | 19.10% | 1.70% | 0.00% | 10.80% | 2.20% | 5.30% |
| 14 | 0.60% | 1.60% | 12.20% | 2.10% | 3.60% | 0.30% | 0.10% | 1.10% | 0.70% | 0.40% |
| 15 | 1.50% | 1.50% | 8.30% | 0.50% | 7.80% | 0.90% | 0.10% | 2.80% | 1.10% | 0.80% |
| 16 | 5.80% | 3.80% | 4.20% | 7.40% | 1.60% | 4.70% | 2.50% | 1.70% | 36.10% | 5.20% |
| 17 | 13.10% | 8.30% | 4.30% | 10.70% | 5.60% | 11.10% | 0.00% | 8.30% | 10.20% | 1.10% |

**Note:** Column totals add to 100%. Cell shading corresponds to its proportion in each SDG (column).

SDGs often involve statements about agreements, governance, and institutions. The additional training data from ChatGPT used for the new SDGClassy+ greatly improves the ability of the classifier to differentiate between these two SDGs (Table 9).

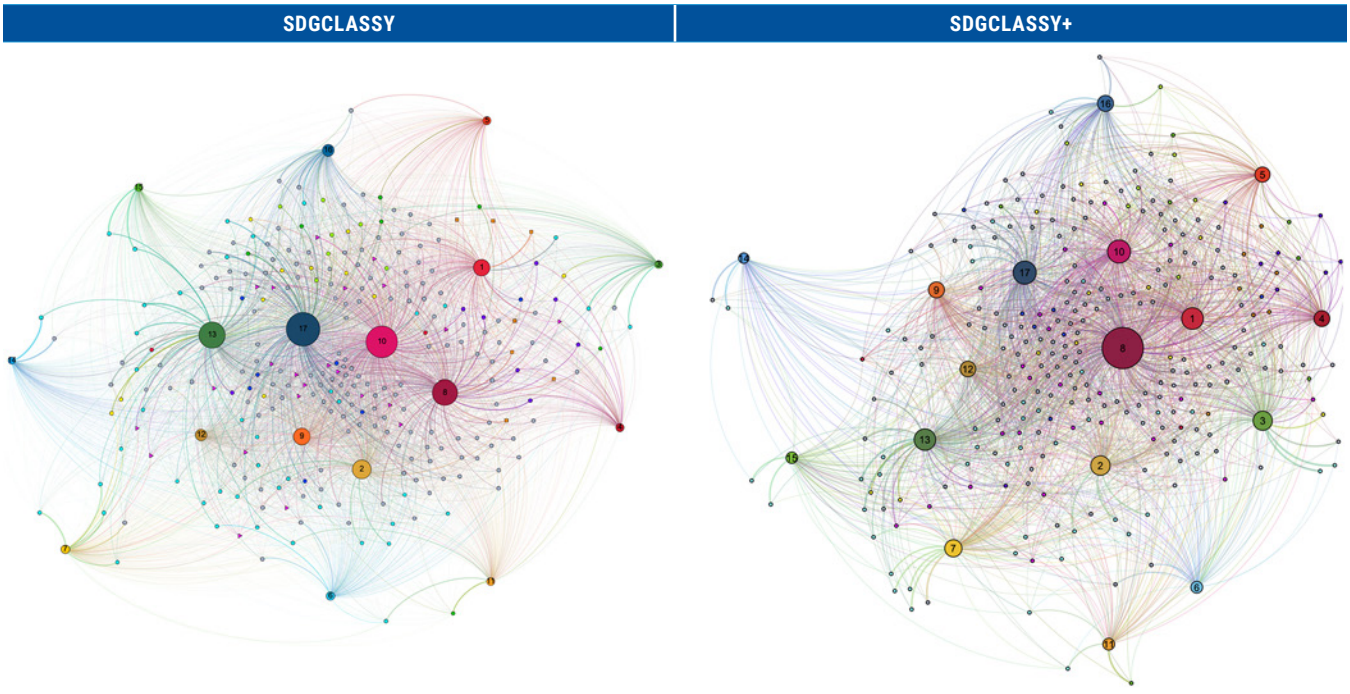
## Updated map of SDG connections using SDGClassy+

An updated map of 267 UN DESA publications using the augmented classifier gives a more precise view of how the 17 SDGs are interconnected in practice. Using the SDG scores computed by means of the classification algorithm as a measure of force, each publication "pulls" against each SDG node. Once all forces are in balance, a network diagram emerges showing the special distribution of the SDGs that emerge from the collection of published works (Figure 1).

As the figure shows, when compared to the results of the original SDGClassy, the augmented SDGClassy+ gives a slightly different understanding of how SDGs are connected. SDG 8, together with SDGs 1, 10, 17, and 13 are more central to the body of work analyzed here. The SDGs and the publications at the periphery of the space are those with weaker links, and therefore of less importance, to the corpus. These SDGs found in the periphery may well represent areas of opportunity for more specialized and focused policy research and advice.

Figure 1

**Network representation of how SDGs are connected through UN DESA publications using SDGClassy**




**Source:** Own calculations.

**Note:** Each dot represents an individual publication with links to each of the 17 SDGs based on the SDGClassy+ model (publication types have different colors and shapes). The SDGs are the larger numbered circles, sized according to their overall importance to the corpus. The SDG scores for each publication determine the length of each line. The SDGs at the center of each figure are the most important for the entire collection of DESA publications.

# 5. Discussion and Conclusion

This work proposed and evaluated the use of ChatGPT to generate synthetic data for improving machine learning SDG classifiers. The results indicate ChatGPT's suitability for generating data that are representative of the SDG concepts as discussed and understood in practice. This is a significant advantage in creating a model of what a given SDG is, rather than relying on a limited definition based on targets and indicators, for instance.

We demonstrated that by supplementing the dataset used to train an existing classifier with the synthetic samples generated by ChatGPT, model performance on a complex text classification task could be significantly improved. The augmented classifier, SDGClassy+, showed substantial improvements in distinguishing between and accurately classifying texts according to each SDG.

This approach works because it leverages generative AI models in the exact way that they are built: to create representative texts that mirror the language used in the millions of documents that discuss the SDGs. Further research applying and assessing this methodology on additional datasets and with other models could strengthen confidence in the broader utility and value of this approach. There are opportunities to refine how generated data is validated to ensure high-quality, representative samples. Combining multiple models has also been shown to enhance results (Hsu, LaFleur and Orazbek, 2022).

The results of the augmented SDGClassy+ demonstrate the value of generative AI for producing high-quality training data to advance machine learning on complex real-world problems. For tasks limited by scarce labeled data, synthetic data generation offers a promising solution to improving model performance, contributing to applications that propel social and scientific progress on a large scale. Although further improvements and evaluations are still needed, this work highlights the significant possibilities that exist for leveraging AI tools to better understand and guide research on sustainable development and other issues.

# References


Blei, David M. (2012). Probabilistic topic models. *Communications of the ACM*, vol. 55, No. 4. April.

CDP Subgroup on voluntary national reviews (2019). Voluntary National Reviews Reports – What do they (not) tell us? *CDP Background Paper*, vol. 49. July.

Geyer, Anne (2023). "GPT-4: All You Need to Know + Differences To GPT-3 & ChatGPT," *AX Semantics*.

Hsu, D. Frank, Marcelo T. LaFleur and Ilyas Orazbek (2022). Improving SDG Classification Precision Using Combinatorial Fusion. *Sensors*, vol. 22, No. 3. January.

LaFleur, Marcelo T. (2019). Art is long, life is short: An SDG Classification System for DESA Publications. *DESA Working Paper*, vol. 159. May.

LaFleur, Marcelo T. and Namsuk Kim (2020). What does the United Nations "say" about the global agenda? An exploration of trends using natural language processing for machine learning. *DESA Working Paper*, vol. 171. October.

Le Blanc, David (2015). Towards Integration at Last? The Sustainable Development Goals as a Network of Targets. *DESA Working Paper*, vol. 141. March.

Le Blanc, David, Clovis Freire and Marjo Vierros (2017). Mapping the linkages between oceans and other Sustainable Development Goals: A preliminary exploration. *DESA Working Paper*, vol. 149. February.

OpenAI (2023). GPT-4 Technical Report. arXiv.

OSDG, UNDP IICPSD SDG AI Lab, and PPMI (2021). "OSDG Community Dataset (OSDG-CD)."

UN DESA (2019). "LinkedSDGs." Available from https://linkedsdg.officialstatistics.org (accessed 22 December 2021).

W3C (n/d). "Semantic web." Available from https://www.w3.org/standards/semanticweb/.

Wolfram, Stephen (2023). "What Is ChatGPT Doing … and Why Does It Work?," *Stephen Wolfram Writings*.

World Economic Forum (2023). "3 new and emerging jobs you can get hired for this year," *Growth2023*.