SDGi Corpus: A Comprehensive Multilingual Dataset for Text Classification by Sustainable Development Goals

Mykola Skrynnyk[1,*,†], Gedion Disassa[1], Andrey Krachkov[1] and Janine DeVera[1]

[1]*United Nations Development Programme, New York, USA*

Abstract

We introduce SDGi Corpus (SDG Integration Corpus), the most comprehensive multilingual collection of texts labelled by Sustainable Development Goals (SDGs) to date. SDGi Corpus is a text dataset for multi-label classification that contains over 7,000 examples in English, French and Spanish. Leveraging years of SDG reporting on the international and subnational levels, we hand-picked texts from Voluntary National Reviews (VNRs) and Voluntary Local Reviews (VLRs) from more than 180 countries to create an inclusive dataset that provides both focused and broad perspectives on the SDGs. This paper reports on the dataset creation effort and use cases we envision for it. We also establish baselines for text classification by SDG using traditional machine learning (ML) and deep learning (DL) approaches. These illustrate the opportunities and challenges of using this dataset for social good. The dataset is available on Hugging Face as UNDP/sdgi-corpus.

Keywords

Dataset, Text Classification, Sustainable Development Goals, Supervised Machine Learning

1. Introduction

Adopted by the United Nations General Assembly in 2015, the 2030 Agenda for Sustainable Development establishes 17 Sustainable Development Goals (SDGs) to guide international development efforts. The SDGs are high-level objectives that focus on a range of core social, economic and environmental issues [1]. Recently, the 17 Goals have *de facto* become a global benchmark of human progress adopted by non-governmental actors outside the United Nations System. The SDGs have been increasingly utilised in private sector sustainability reports [1], scholarly publications mapping [2] and university ranking systems [3, 4].

As government agencies, academic institutions, businesses and non-governmental organisations are all looking for ways to link their work to the 2030 Agenda, the demand for text classification by SDGs continues to grow. The ability to accurately classify text by SDG would enable more efficient resource allocation, improve monitoring and evaluation efforts as well

2nd Symposium on NLP for Social Good 2024, April 25-26, 2024, Liverpool, the UK

*Corresponding author.

†Main author.

mykola.skrynnyk@undp.org (M. Skrynnyk); gedion.disassa@undp.org (G. Disassa); andrey.krachkov@undp.org (A. Krachkov); janine.devera@undp.org (J. DeVera)

0009-0003-2661-9303 (M. Skrynnyk); 0009-0000-8552-154X (G. Disassa)

[1]For details about the Agenda 2030 and the SDGs framework, visit the websites of the United Nations Department of Economic and Social Affairs or United Nations Development Programme.

as increase transparency in reporting. The amount of unstructured information of analytical interest is vast, highlighting the need for an automated solution for SDG classification. However, the efforts to develop a viable solution have been scattered, and no single benchmark for testing SDG classifiers has been adopted by the community thus far.

Against this backdrop, we introduce SDGi Corpus, the most comprehensive multilingual collection of texts labelled by SDGs to date. The dataset is designed to fill the gap in the availability of SDG-labelled data, facilitate the development of SDG classification systems and enable a consistent comparison among them. While we envision this dataset to be used as a standard benchmark for multi-label (and multi-class) text classification by SDG, we also encourage researchers and practitioners to use it for topic modelling, text mining and quantitative text analysis more broadly. The dataset includes rich metadata that allows to slice and dice data to answer various SDG-related questions.

Our contribution is two-fold. First and foremost, we introduce the dataset to encourage the ML community in developing state-of-the-art (SOTA) solutions for SDG classification. Secondly, we report preliminary results for a series of supervised and zero-shot ML experiments, illustrating the non-triviality of the SDG classification task in general as well as challenges of overfitting and out-of-domain generalisation for our dataset in particular.

The remainder of this paper is structured as follows. Section 2 discusses existing research on SDG classification. Section 3 introduces SDGi Corpus, describing the data sources and curation process as well as providing key descriptive statistics. In section 4, we report preliminary results for a series of experiments using the dataset. The final section 5 is used to draw conclusions and discuss avenues for future improvements.

2. Related Work

The body of works that explore text classification by SDG is relatively small and can be characterised by two distinct approaches: keywords-based and ML-based[2]. Thus, a number of initiatives directed their efforts to creating curated lists of keywords or queries to be able to link texts to SDGs in a rules-based fashion. Most popular of these are SIRIS [5], Elsevier [6], Aurora [7] and Auckland [8] systems. These systems differ in matching logic[3] and SDG coverage, with the original Elsevier and Auckland systems excluding SDG 17. These approaches mostly consider SDG classification from bibliometric or scientometric perspectives, although LinkedSDG [9] and SDG Mapper [10] adopted similar methodology in a policy context. While keyword-based systems were instrumental in popularising SDG classification, their systematic evaluations revealed issues with both accuracy and robustness [11, 12].

Attempts to classify text by SDG using ML algorithms adopted methodologies as diverse as semantic search [13, 14], topic modelling [15, 16] and text classification [17, 18, 19, 20, 21]. Despite a variety of methods used by ML-based approaches, one recurrent theme in the literature is the lack of readily available SDG-labelled datasets. Most studies using supervised learning

[2]We use these terms to refer to the way systems can be applied to classify text by SDG and not to the way they were derived. For example, SIRIS [5] system was enriched by matching terms using word embeddings. This however does not change the fact that the final system is keyword-based.

[3]Queries allow for a more complex keyword-matching logic but they are fundamentally rules-based classification systems with all the associated limitations.

are based on different sets of data collected by their respective authors. Examples include training an XGBoost model on a combination of 200 SDG-specific reports, SDG descriptions and a few hundred manually-labelled project descriptions [17], fine-tuning BERT on a collection of several hundred news, articles and policy briefs from IISD "SDG Knowledge Hub" website [4] [18]; fine-tuning DistilRoBERTa using a set of more than 8,000 paragraphs manually-labelled by unidentified experts [19]. Several studies have also relied on keyword-based systems to label data for supervised learning [20, 21]. Some of these studies have resulted in open- or closed-source products for SDG classification, including SDG Prospector based on [19] and text2sdg package in R based on [21].

To the best of our knowledge, there has been only one focused effort to create a public SDG-labelled dataset, namely OSDG Community Dataset (OSDG-CD) [22]. OSDG-CD is an English-only collection of passages from policy documents and reports labelled with respect to SDGs by online volunteers. The original release of the dataset included examples for the first 16 SDGs only but examples for SDG 17 have been added recently. At the time of writing, the dataset includes over 40,000 examples evaluated by over 1400 volunteers.

OSDG-CD has several major limitations that render it inappropriate as an SDG classification benchmark. Firstly, SDG labels are assigned by online volunteers and not domain experts. This can lead to systematic biases that cannot be mitigated by the fact that the same example is evaluated by multiple volunteers. Secondly, each example is evaluated by several volunteers with respect to one SDG only, i.e., every volunteer is presented with a binary question to determine if an example is relevant to a pre-defined SDG. Consequently, each text in the dataset has only one associated label, which reduces the SDG task to a multi-class classification problem. Finally, a significant number of examples have been judged not relevant by the majority of labellers, making the effective size of the dataset much smaller.

3. SDGi Corpus

To enable a consistent comparison among existing solutions and facilitate the development of SOTA models for SDG classification, we introduce SDGi Corpus, the most comprehensive multilingual collection of texts labelled by SDG to date. This section explains the origin of the data, details the data curation process and provides descriptive statistics of the dataset.

3.1. Data Sources

Although readily available datasets for SDG classification are few, textual information linked to SDGs is abundant in reporting documents such as Voluntary National Reviews (VNRs) and Voluntary Local Reviews (VLRs). VNRs are voluntary, multi-stakeholder and government-led reports submitted to the High-level Political Forum on Sustainable Development (HLPF). The rationale behind VNRs is to enable countries to share their experiences in the implementation of the 2030 Agenda. In a similar vein, VLRs are subnational reviews from local and regional governments that are increasingly engaged in the SDG implementation process. Unlike VNRs,

[4]Note that this website is managed by an independent think tank and is not part of the United Nations System.

(a) Raw example (b) Edited example with redacted content

Figure 1: An illustration of raw and edited examples for SDG 14 from 2023 Amsterdam VLR.

VLRs were not envisioned in the 2030 Agenda but emerged as a popular means of communication about SDG localisation.

To construct SDGi Corpus, we web-scraped more than 350 VNRs and close to 200 VLRs, all in PDF format, from the United Nations[5]. Our dataset includes extracts from reviews published before December 2023, covering 8 years of SDG reporting in total. While we collected all available reports, our initial release only includes examples from reports written in English, French or Spanish, which together constitute over 90% of the original document set [6].

Being review documents, VNRs are typically structured in such a way that each chapter discusses progress in achieving a specific SDG. However, not all countries cover all SDGs [7]. VLRs are similar reports produced by subnational governments and are therefore more likely to focus on SDGs 9 "Industry, innovation and infrastructure" and 11 "Sustainable cities and communities" than other SDGs. It is also common to find case studies or examples of projects in VLRs that contribute to one or more SDGs. The next subsection describes how we leveraged the structure of the documents to extract examples for SDGi Corpus.

3.2. Document Curation

To create SDGi Corpus, we manually analysed each document, searching and extracting specific parts clearly linked to SDGs. Our curation process can be summarised in 4 steps as follows:

1. Manually examine a given document to identify SDG-labelled content.
2. Extract pages containing relevant content to SDG-specific folders.
3. Edit extracted pages to redact irrelevant content before and after the relevant content.
4. For content linked to multiple SDGs, fill out a metadata sheet.

While creating examples, we made as few assumptions and judgements about the data as possible. Our goal was not to label examples but to extract content labelled by the authors of

[5]VNRs and VLRs are available on HLPF and UN DESA websites respectivelly.
[6]We will explore the possibility of extending SDGi Corpus to include data in all the 6 official languages of the United Nations in the future updates to the dataset.
[7]For instance, land-locked countries commonly exclude SDG 14 "Life Below Water" from their VNRs.

VNRs and VLRs who are assumed to be domain experts. To identify relevant content, we relied on visual and textual clues that indicated SDG relevance, such as chapter and section titles and SDG icons. SDG-labelled content varies greatly in size and relevance. Some examples are chapters from VNRs that span dozens of pages whereas others are short paragraphs discussing an SDG-related project or initiative.

The PDF pages containing relevant content were extracted into one of the 17 SDG-specific folders or a dedicated folder for multi-labelled examples. Whenever pages contained clearly separable content, they were split further into multiple examples. However, a large number of long sections were left intact to preserve cohesion, while statistical tables and appendices were excluded altogether.

The next step was to edit every example to mask noisy or irrelevant content. Since examples come from PDFs, the structure and layout of pages differ greatly. For all examples, contents from the sections that precede and follow the relevant part were masked out. For a small portion of the data, especially shorter examples of up to 4 pages, we also made a reasonable effort to redact other irrelevant or noisy content within the section itself, including headers, footers, tables, figures, image credits etc. Figure 1 illustrates this masking process with an example for SDG 14[8]. Finally, we extracted text from PDFs using pypdfium2 package in Python.

Overall, our dataset represents data extracted from a variety of real-world layouts, which results in a fair degree of noise in the extracted texts. Given the public nature and intended use of the reviews, VNRs and VLRs are extremely unlikely to contain any sensitive Personally Identifiable Information (PII).

3.3. Train/Test Split

One peculiarity of the dataset is that single-labelled examples may give away their true label in their content. For instance, many sections on SDG 2 start with text containing "Zero hunger" or will contain "Zero hunger", "SDG 2" or "Goal 2" in the header of every page. This makes it easy to train a seemingly accurate classifier that just badly overfits the data. We consider this as both a limitation of the dataset and technical challenge for the ML community in that training a generalisable classifier becomes non-trivial, despite training data containing abundant information about different aspects of all 17 SDGs.

To partially mitigate this issue, we deviate from the standard practice of creating train/test splits randomly, adopting an adversarial approach instead. In so doing, we first fit a shallow neural network to classify examples using embeddings from OpenAI's *text-embedding-ada-002*. Then, we calculate cross-entropy loss for every example in the dataset. The loss values are used as weights for sampling. We sample 20% of the examples into the test set, stratifying by language, text size bucket and a binary variable indicating whether an example has one label or more.

As the result, the train and test sets follow a noticeably different distribution of labels, with SDGs 5, 6 and 7 being underrepresented and SDGs 8, 10 and 11 being overrepresented in the test set when compared to the train set, as seen in Figure 2. At the same time, the test set contains more challenging examples that simpler models struggle with.

[8]The same raw PDF can be used to create examples for SDG 15 and 16 by masking other parts of the page.

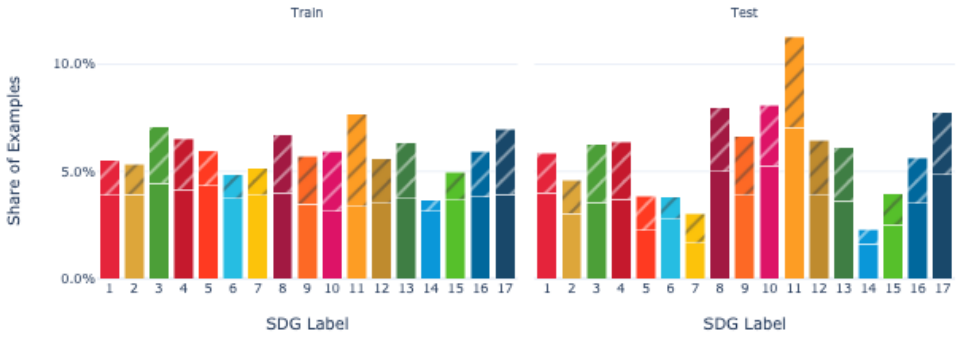

Figure 2: Distribution of labels in SDGi Corpus by train/test split. Solid parts show the labels that come from single-labelled examples while shaded parts reflect contributions of multi-labelled examples. The height of every bar corresponds to the share of examples that have that bar's label in the respective split.

Table 1
Average Number of Tokens in SDGi Corpus

Size / Language	English (EN)	Spanish (ES)	French (FR)	All (XX)
Short (S)	190	158	144	180
Medium (M)	852	557	673	797
Long (L)	2,678	3,367	2,665	2,787
All (X)	1,306	1,480	1,700	1,382

3.4. Descriptive Statistics

In total, SDGi Corpus contains 7,350 examples, with 5,880 (80%) and 1,470 (20%) examples in the train and test sets respectively. The dataset is dominated by examples in English (71.9%), followed by Spanish (15.9%) and French (12.2%). To allow different yet consistent uses of the dataset, we include a grouping variable for text size buckets. Examples shorter than 512 tokens[9] are considered short (S), between 512 and 2048 are medium (M) and longer than 2048 tokens are long (L)[10]. There are 1,662 short, 3,013 medium and 2,675 long examples. Most examples (89.1%) in the dataset have just one label, but those that are multi-labelled have anywhere between 2 to 17 labels. The average text length in the corpus is 1,382 tokens, although this slightly varies from language to language. Admittedly, this is considerably longer than the expect input length of most transformer models. We refer the reader to Table 1 for more details.

[9]Token counts here and in Table 1 are based on OpenAI's *cl100k_base* tokenisation after replacing all numbers with *"NUM"* placeholder value.
[10]When referring to all sizes, we use "X". Similarly, when referring to all languages, "XX" designation is used.

4. Experiments

In this section, we present preliminary results for multi-label classification using SDGi Corpus[11]. We test several popular modelling approaches, combining sparse and neural representations with support vector machines (SVMs), feedforward neural networks (FFNs) and graph neural networks (GNNs) [23]. In addition, a GPT architecture [24] is employed for zero-shot learning.

4.1. Data Preparation

For sparse representations, we apply standard preprocessing such as lowercasing, removing numbers and punctuation. We also remove stop words using a combined list of English, French and Spanish stop words but do not lemmatise or stem tokens. Texts are vectorised using Term Frequency - Inverse Document Frequency (TF-IDF). Models using this type of features are denoted with BOW. Neural representations are derived from *text-embedding-ada-002* and are 1536-dimensional. Models based on embeddings are prefixed with "Ada".

For GNN, we transform texts into graphs. Our approach is similar to [25]. We treat unique tokens as nodes and connect tokens that co-occur within a window of size 2 with directed edges. Only 30,000 most common tokens are kept in the vocabulary. For each graph, up to 10,000 nodes and 10,000 edges with the highest weight are used.

4.2. Models

SVM. Support vector machines used to be SOTA models for text (and document) classification [26, 27]. It is a fast and simple architecture that has a proven track record of robust performance when combined with sparse word representations. Our experiments are based on linear SVMs.

FNN. We use a shallow neural network with one hidden layer and 100 neurons, ReLU activation and Adam optimiser. To reduce overfitting, early stopping mechanism is used. We combine SVM and FNN with both sparse representations and embeddings from OpenAI.

GNN. Graph neural networks are flexible architectures that have only recently been applied to text with some success [25]. We use a GNN model based on GraphSAGE architecture [28]. The model consists of an embedding layer, which embeds every node (token) into a 128-dimensional embedding, and two SAGEConv layers that propagate information from the neighbouring nodes. The first layer outputs 256-dimensional vector, while the second one outputs 17-dimensional vector. The final prediction is done by reading out the representations for all nodes and averaging them. Depending on the vocabulary size, our models have between 1 to 4 million parameters.

GPT. GPT-family models are currently among the most powerful ones when it comes to generative AI. We use GPT-3.5 Turbo (16k), relying on its zero-shot learning capabilities to classify text by SDG without any training data. We manually experiment with several prompt designs to maximise the performance.

We optimise the SVM and FNN models by grid search, using various settings for regularisation term as well as tf-idf transformation, if applicable. For the GNN, we manually test several settings of hyperparameters, preferring a model with a smaller size. When training the GNN, we use 20% of the training data for validation and early stopping.

[11]The codebase for the experiments is available in UNDP-Data/dsc-sdgi-corpus repository on GitHub.

Table 2
Macro-average F_1 scores for multi-label classification on SDGi Corpus (test set) and out-of-domain data

Train Set (Size)	Test Set (Size)	BOW SVM	Ada SVM	BOW FFN	Ada FFN	GNN	GPT-3.5 Turbo 16k (Zero-shot)
sdgi-s-en (965)	sdgi-s-en (241)	39.40	48.60	39.84	**52.96**	43.12	15.13
sdgi-m-en (1,860)	sdgi-m-en (466)	57.34	53.39	63.15	**70.41**	59.49	26.89
sdgi-l-en (1,400)	sdgi-l-en (350)	79.45	23.00	71.23	64.54	**80.62**	24.54
sdgi-x-en (4,225)	sdgi-x-en (1,057)	60.23	53.23	**67.85**	64.06	65.47	23.89
sdgi-x-fr (720)	sdgi-x-fr (180)	47.81	54.24	58.60	61.14	**78.05**	32.42
sdgi-x-es (935)	sdgi-x-es (233)	55.30	42.82	41.65	51.97	**57.90**	22.31
sdgi-x-xx (5,880)	sdgi-x-xx (1,470)	63.85	54.47	**66.34**	66.30	65.41	24.63
sdgi-x-en (4,225)	IISD News (724)	40.77	**48.27**	20.67	40.81	26.12	12.26
sdgi-x-xx (5,880)	IISD News (724)	40.00	**47.70**	22.62	40.11	28.94	12.26

4.3. Results

The main results for the experiments are shown in Table 2. The test set performance varies greatly depending on the subset of data. But even the highest scores are relatively low, with no model exceeding 81.0 macro-average F_1 score. This illustrates the overfitting problem we indicated earlier. Secondly, the Ada-based models perform better on shorter text, especially when coupled with an FFN. The best score on the full dataset is achieved by BOW FFN, but ADA FFN and GNN are only slightly behind. Admittedly, the GNN model significantly outperforms others for Spanish and French texts. It may be more robust when training on small training sets due to the way the graph representation is constructed. GPT-3.5 performance is subpar but our experiments showed that it tends to greatly overassign labels to texts, calling for a more careful prompt engineering or even few-short learning to constrain this behaviour.

We also report generalisation performance on IISD News, an out-of-domain dataset. This is a collection of English-only news articles from IISD used in [18]. We find generalisation performance lacking for all models but much less so for the Ada-based SVM.

5. Conclusion

We introduced SDGi Corpus, the most comprehensive multilingual collection of texts labelled by SDGs to date. The dataset contains over 7,000 examples in English, French and Spanish, hand-picked from VNRs and VLRs. We provided an overview of the data creation effort and set baselines for multi-label classification on different subsets of the dataset using SVMs, FFNs, GNNs and GPT-3.5 Turbo. Our experiments demonstrate that there is no "one-size-fits-all" solution to SDG classification and different models are better on different tasks. Overfitting remains a major challenge but overcoming it will lead to more robust models. Our future efforts will be directed to adding examples in more languages as well as extending SDGi Corpus with examples from other types of documents. We encourage researchers and practitioners to use SDGi Corpus for developing novel SDG classification systems for social good.

References

[1] United Nations, The Sustainable Development Goals Report 2023: Special Edition, Technical Report, United Nations, New York, 2023. URL: https://unstats.un.org/sdgs/report/2023/.

[2] Elsevier, The Power of Data to Advance the SDGs: Mapping research for the Sustainable Development Goals, Technical Report, Elsevier, 2020. URL: https://www.elsevier.com/connect/report-mapping-research-to-advance-the-sdgs.

[3] C. OCallaghan, QS World University Rankings: Sustainable Development Goals, 2021. URL: https://www.topuniversities.com/university-rankings/world-university-rankings/sustainable-development-goals.

[4] THE reporters, Impact Rankings 2023: methodology, 2023. URL: https://www.timeshighereducation.com/world-university-rankings/impact-rankings-2023-methodology.

[5] N. Duran-Silva, E. Fuster, F. A. Massucci, A. Quinquillà, A controlled vocabulary defining the semantic perimeter of Sustainable Development Goals, 2019. URL: https://zenodo.org/record/3567768. doi:10.5281/ZENODO.3567768.

[6] B. Jayabalasingham, Identifying research supporting the United Nations Sustainable Development Goals, 2019. URL: https://data.mendeley.com/datasets/87txkw7khs/1. doi:10.17632/87TXKW7KHS.1.

[7] M. Vanderfeesten, R. Otten, E. Spielberg, Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0.2, 2020. URL: https://zenodo.org/record/3817444. doi:10.5281/ZENODO.3817444.

[8] W. Wang, W. Kang, J. Mu, Mapping research to the Sustainable Development Goals (SDGs), preprint, In Review, 2023. URL: https://www.researchsquare.com/article/rs-2544385/v2. doi:10.21203/rs.3.rs-2544385/v2.

[9] UN DESA, LinkedSDG, 2019. URL: https://linkedsdg.officialstatistics.org.

[10] European Commission. Joint Research Centre., Mapping EU policies with the 2030 agenda and SDGs :fostering policy coherence through text based SDG mapping., Publications Office, LU, 2023. URL: https://data.europa.eu/doi/10.2760/87754.

[11] C. S. Armitage, M. Lorenz, S. Mikki, Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?, Quantitative Science Studies 1 (2020) 1092–1108. URL: https://direct.mit.edu/qss/article/1/3/1092-1108/96106. doi:10.1162/qss_a_00071.

[12] F. Schmidt, M. Vanderfeesten, Evaluation on accuracy of mapping science to the United Nations' Sustainable Development Goals (SDGs) of the Aurora SDG queries, Technical Report, [object Object], 2021. URL: https://zenodo.org/record/4964606. doi:10.5281/ZENODO.4964606, version Number: v1.0.2.

[13] J. Galsurkar, A. Vempaty, K. R. Varshney, L. Wu, M. Sushkov, M. Singh, D. Iyer, S. Kapto, I. Research, I. Watson, Semantic Searching for Efficient Assessment of Sustainable Development in National Plans, 2017. URL: https://api.semanticscholar.org/CorpusID:86858304.

[14] F. Sovrano, M. Palmirani, F. Vitali, Deep Learning Based Multi-Label Text Classification of UNGA Resolutions, in: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, 2020, pp. 686–695. URL: http://arxiv.org/abs/2004.03455. doi:10.1145/3428502.3428604, arXiv:2004.03455 [cs, stat].

[15] M. LaFleur, Art Is Long, Life Is Short: An SDG Classification System for DESA Publications, SSRN Electronic Journal (2019). URL: https://www.ssrn.com/abstract=3400135. doi:10.2139/ssrn.3400135.

[16] M. T. LaFleur, Using large language models to help train machine learning SDG classifiers, Working Paper 180, 2023. URL: https://desapublications.un.org/working-papers/using-large-language-models-help-train-machine-learning-sdg-classifiers.

[17] A. Pincet, S. Okabe, M. Pawelczyk, Linking Aid to the Sustainable Development Goals – a machine learning approach, OECD Development Co-operation Working Papers 52, 2019. URL: https://www.oecd-ilibrary.org/development/linking-aid-to-the-sustainable-development-goals-a-machine-learning-approach_4bdaeb8c-en. doi:10.1787/4bdaeb8c-en, series: OECD Development Co-operation Working Papers Volume: 52.

[18] J. E. Guisiano, R. Chiky, J. De Mello, SDG-Meter: A Deep Learning Based Tool for Automatic Text Classification of the Sustainable Development Goals, in: N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, E. Szczerbicki (Eds.), Intelligent Information and Database Systems, volume 13757, Springer International Publishing, Cham, 2022, pp. 259–271. URL: https://link.springer.com/10.1007/978-3-031-21743-2_21. doi:10.1007/978-3-031-21743-2_21, series Title: Lecture Notes in Computer Science.

[19] J.-B. Jacouton, R. Marodon, A. Laulanié, The Proof is in the Pudding. Revealing the SDGs with Artificial Intelligence, Research Document 262, Agence Française de Développement (AFD), 2022. URL: https://www.afd.fr/en/ressources/proof-pudding-revealing-sdgs-artificial-intelligence.

[20] M. Vanderfeesten, R. Jaworek, L. Keßler, AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs), Technical Report, Zenodo, 2022. URL: https://zenodo.org/record/5603019. doi:10.5281/ZENODO.5603019, version Number: 1.0.

[21] D. U. Wulff, D. S. Meier, R. Mata, Using novel data and ensemble models to improve automated labeling of Sustainable Development Goals, 2023. URL: http://arxiv.org/abs/2301.11353, arXiv:2301.11353 [cs].

[22] OSDG, UNDP IICPSD SDG AI Lab, PPMI, OSDG Community Dataset (OSDG-CD), 2021. URL: https://zenodo.org/record/5550238. doi:10.5281/ZENODO.5550238.

[23] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The Graph Neural Network Model, IEEE Transactions on Neural Networks 20 (2009) 61–80. doi:10.1109/TNN.2008.2005605.

[24] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, 2020. URL: http://arxiv.org/abs/2005.14165, arXiv:2005.14165 [cs].

[25] L. Huang, D. Ma, S. Li, X. Zhang, H. Wang, Text Level Graph Neural Network for Text Classification, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong,

China, 2019, pp. 3442–3448. URL: https://www.aclweb.org/anthology/D19-1345. doi:10.18653/v1/D19-1345.

[26] D. D. Lewis, Y. Yang, T. G. Rose, F. Li, RCV1: A New Benchmark Collection for Text Categorization Research, J. Mach. Learn. Res. 5 (2004) 361–397. Publisher: JMLR.org.

[27] S. Wang, C. Manning, Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 90–94. URL: https://aclanthology.org/P12-2018.

[28] W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 1025–1035. Event-place: Long Beach, California, USA.