Detailed Descriptions for Text Classification Applications

Gorka Artola*[1], German Rigau[1]

[1]University of the Basque Country (UPV/EHU), Faculty of Informatics, Manuel Lardizabal pasealekua, 1, 20018 Donostia-San Sebastián, Spain

Abstract

The development of effective domain specific text classification systems generally requires the availability of large amounts of high quality labeled domain data. In domains such as BioNLP, eHealth, NLP for Legal Purposes, NLP for Social Media and Journalism, etc., obtaining the needed volume of data manually-labeled by domain experts is not usually feasible or affordable. In this work we propose a new method for text classification based on the use of detailed class descriptions instead of using a large number of labeled instances for training the classifiers. Our method, experimentally tested on the classification of titles of scientific papers on the domain of the Sustainable Development Goals of the United Nations, consistently outperforms mainstream NLP classification approaches, radically faster and at a fraction of their cost due to it does not need a previous process of hand-labelling thousands of samples.

Keywords

text classification, class descriptions, sustainable development goals

1. Introduction

The dominant approach in the last few years for specific domain Natural Language Processing (NLP) applications is the use of transformer-based [1] general language models (PLMs) fine-tuned on domain specific and task relevant labeled data. To reach top performance, these downstream training processes easily require several thousands of high quality labeled samples in the shape of the targeted task, but for many real-world applications obtaining the minimum volume of data manually-labeled by domain experts is not affordable or even feasible. For these cases, there is a variety of zero-shot classification models and techniques [2] using the PLMs, but they offer worse results than the mentioned methods based on fine-tuning the models with large amounts of data.

In some cases, despite not having labeled samples, we have detailed descriptions of the main classes of the specific domain, usually crafted under consensus of relevant stakeholders. Our goal in this work is to explore the potential impact of using these detailed descriptions instead of the labeled samples for text classification tasks in specific application domains. Despite descriptions of classes and samples of the task to be performed may differ in both shape and domain, our main hypothesis is that the use of detailed descriptions of the classes will multiply the transfer-learning capacity of PLMs and noticeably improve the performance of NLP methods.

1.1. Description of the Task

The selected application domain for our experimentation is the 2030 Agenda for Sustainable Development of the United Nations (UN) with its 17 Goals (SDGs). Being a relatively recent concept, there is not an extensive corpus of data in the subject on which PLMs can be pre-trained or fine-tuned, but their official descriptions elaborated collaboratively by the members of the UN are detailed and public. The selected NLP task for this study is the multi-class classification of titles of scientific papers by SDGs as described in Figure 1: given a title of a scientific paper the method or system must select the SDGs most related to the paper.

As an example of the relevance of the work proposed in this paper, the generation of the hand-labeled AURORA dataset [3], used in this work to train the main baseline model for comparison purposes, has required the design of an international survey, the participation of 244 expert respondents from Europe and North America from October 2019 to January 2020 to gather the raw data, a post-processing phase to generate a labeled set of samples that ended in May 2020, and a multi-disciplinary work-team of 15 people for whole dataset generation process. The method proposed in this paper avoids all this trouble using instead just the SDG descriptions provided by the UN to train a competitive classifier immediately and for free.

1.2. Summary of Contributions

The main contributions of this work are:

- We propose the use of already existing or handcrafted detailed descriptions of the classes for multi-label sentence classification with PLMs as a better performing and more resource-efficient

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

gorka.artola@ehu.eus (G. Artola*); german.rigau@ehu.eus (G. Rigau)

*Corresponding author

Figure 1: SDG headlines and an example of the targeted task. In this example, the system should classify the title of the scientific paper in SDGs 1 "No poverty", 5 "Gender equality", 8 "Decent work and economic growth", and/or 10 "Reduced inequalities"

way than investing in manual-labelling of samples.

- We propose guidelines to decide between working in the generation of detailed descriptions or investing in hand-labeling samples, considering the availability or not of either detailed descriptiosn or labeled samples, and depending on this decision, to select the most appropriate multi-class classification technique with PLMs.
- We establish a new SOTA for classification of titles of scientific papers by SDGs.
- We publicly disclose the most relevant datasets and code used in our experimentation.

2. Related Work

PLMs, such as *BERT* [4] and *GPT* [5], have achieved state-of-art performance on many NLP tasks [6], and among them on multi-class text classification [7]. The research community has developed several lines of work to improve text classification in different data availability scenarios:

- When we have abundant unlabeled data related to the specific application domain but lack of labeled data **Weakly-supervised** techniques [8] show promising results. The most recent of them leverage the capacities of transformer-based PLMs, like LOTClass [9], which uses label names as initial keywords and augments the keywords with *BERT*'s MLM module to train classification models on unlabeled data, or FastClass by Xia et al. [10], that proposes the use of dense text representation techniques in semantic spaces.
- In the case we have large amounts of unlabeled data, but non related to the domain or the task, **Unsupervised text classification** techniques [11] show the capacity to improve text classification.
- When lacking of any data, PLMs allow the generation of improved semantically meaningful text representation models like Sentence-BERT [12], and the enunciation of the text classification task as a natural language inference (NLI) problem are the SOTA techniques [2]. Recently, Schopf et al. [13] proposed the combination of the embedding-based method Lbl2Vec and transformer-based PLMs to further improve their performance on unsupervised text classification.

Focusing specifically on the use of descriptions of classes, there is also a body of research studying question answering task embodiment for text classification like the one proposed by Chai et al. [14]. These techniques in combination with strategies for the development of better class descriptions [15], label noise reduction methods [16], and the recent emergence of generative large language models (LLMs) [17] set the ground for future research in the use of descriptions for specific domain NLP applications.

Regarding the classification of scientific papers by SDGs, related literature describes several approaches grouped in two different working principles:

- Boolean query based approaches for information retrieval from databases like the ones developed

Table 1
Macro-averaged F1-score, tested on the test split of the Paper Titles Gold Dataset, of baselines and SDG-Descriptions based scientific paper classifiers

Model	Available labeled samples	F1-score
Baselines		
General PLM fine-tuned on labeled gold dataset	8,339	66.14%
Zero-shot classification with NLI-PLM on prompted keywords	17	61.20%
Description based models		
General PLM fine-tuned on SDG-Descriptions	447	64.56%
Zero-shot classification with NLI-PLM on prompted SDG-Titles	17	51.68%
Few-shot classification with NLI-PLM fine-tuned on SDG-Descriptions	447	**67.12%**

by Elservier [18] [19] [20], Digital Science [21], the University of Bergen [22], the University of Auckland [23] and the AURORA European University Alliance[*]. The most relevant among them is the AURORA SDG Queries v5 method [24] [25].
- NLP based methods like the AURORA-ML[*] method [26] [27]. This approach comprises 169 $Multilingual\ BERT$ [28] based models, one for each SDG target, fine-tuned on abstracts of papers obtained with the AURORA SDG Queries v5 method.

3. Experimental Setup

The following **Datasets** contain all the data used in our experimentation:

- The "SDG-Descriptions Dataset" comprising 447 sentences of different semantic natures (SDG-Headlines, SDG-Titles, SDG-Targets and SDG-Indicators) developed by the UN and published in a dedicated website[*] describing the 17 SDGs. Altogether, we name the samples of this dataset SDG-Descriptions. Considering we have 447 descriptive sentences of SDGs, we have built training dataset with 430 entailment samples and 7,152 contradiction samples.
- The "Paper Titles Gold Dataset" with 9,382 scientific paper titles labeled by experts. This dataset includes two families of samples that are disjoint, i.e., no paper title appears in both families:
  - "Positive samples" of titles labeled to one or more specific SDGs they are related to.
  - "Negative samples" of titles labeled to one or more specific SDGS they are not related to.

  This Gold Dataset is a subset of the AURORA dataset [3], elaborated surveying expert scientist, and that shows a human agreement level of 70.10% in this task. We have developed several splits of this dataset for training, development and evaluation purposes. The train-split contains more than 8,000 positive samples, and the test-split contains 2,086 labeled paper titles unevenly distributed by SDG but with the same amount of positive and negative samples for each one of them.

The **Classification Approaches** and **Models** we have experimented with are:

- Fine-tuning classifiers from general PLMs. After experimenting with different general PLMs we have selected $BART_{LARGE}$ [29] for its better results. We have developed different classifiers fine-tuning $BART_{LARGE}$ on different amounts of samples of the train-split of the Paper Titles Gold Dataset, on different amounts of samples of the SDG-Descriptions Dataset, and on the combinations of both of them.
- Zero-shot classification with NLI-PLMs. After experimenting with different NLI-PLMs and querying/prompting setups, we have obtained the best results querying $BART_{LARGE}\ MNLI$ [29] with either SDG-Headlines or SDG-Titles and prompting the queries with the expression "The subject is ".
- Few-shot classification. Building upon the previous approaches, we have developed a new method for multi-class text classification fine-tuning $BART_{LARGE}\ MNLI$ on pairs of SDG-Description sentence/SDG-Headline, and applying the resulting model for NLI based zero-shot classification of paper titles. For the initial fine-tuning we have built a training dataset with samples composed by pairs of sentences, being the first each one of the SDG description sentences and the second each one of the SDG-Headlines

[*] https://aurora-universities.eu/
[*] https://github.com/Aurora-Network-Global/TMD
[*] https://metadata.un.org/sdg/

prompted with the text "The Sustainable Development Goal is". This way we have generated 17 samples from each SDG description sentence, out of which the one pairing the sentence with its correspondent SDG-Headline is labeled as "entailment" and all the rest (16) as "contradiction". We generate a zero-shot classifier fine tuning the $BART_{LARGE}\ MNLI$ model with this dataset. The classification of each test sample is finally performed querying the model with the SDG-Headlines and prompting the queries with the expression "This is".

Considering that a paper title may be related to several SDGs, our **Metrics** on the experiments consider true positives (TP) the right predictions on positive samples, false positives (FP) the wrong predictions on negative samples and true negatives (TN) the right predictions on negative samples. The **Prediction Criterion** used in this analysis of the results is Topk-3, i.e., the top 3 scores given by the models for each tested sample are considered predictions for all considerations.

The current SOTA for the studied task and domain is the top macro averaged F1-score of 55% offered by the AURORA-ML method referenced in section 2. In the experimentation we have observed that the F1-score registered in a vanilla fine-tuning of $BART_{LARGE}$ on the full train split of the "Paper Titles Gold Dataset" goes above 60%. Therefore, we have considered this vanilla approach our **Baseline** or the analysis of the impact of the use of SDG-Descriptions. In the zero-shot approach the considered baseline is the direct use of $BART_{LARGE}\ MNLI$ with a collection of keywords, namely SDG-Subjects, also enunciated by the UN and related to the SDGs that we have not considered part of the SDG-Descriptions because they are not shaped as the descriptive sentences we intend to study.

These choices are the result of an extensive experimentation process comprising different PLMs and meta-parameters looking for the best performing ones.

4. Results

Table 1 shows a comparison between the best macro-averaged F1-scores obtained with our description-based models and the baselines. **Our few-shot classification method using 447 publicly available SDG-Descriptions overcomes the general baseline** trained with over 8,000 hand-labeled samples. On the other hand, our zero-shot classification using SDG-Descriptions lags far behind the zero-shot baseline

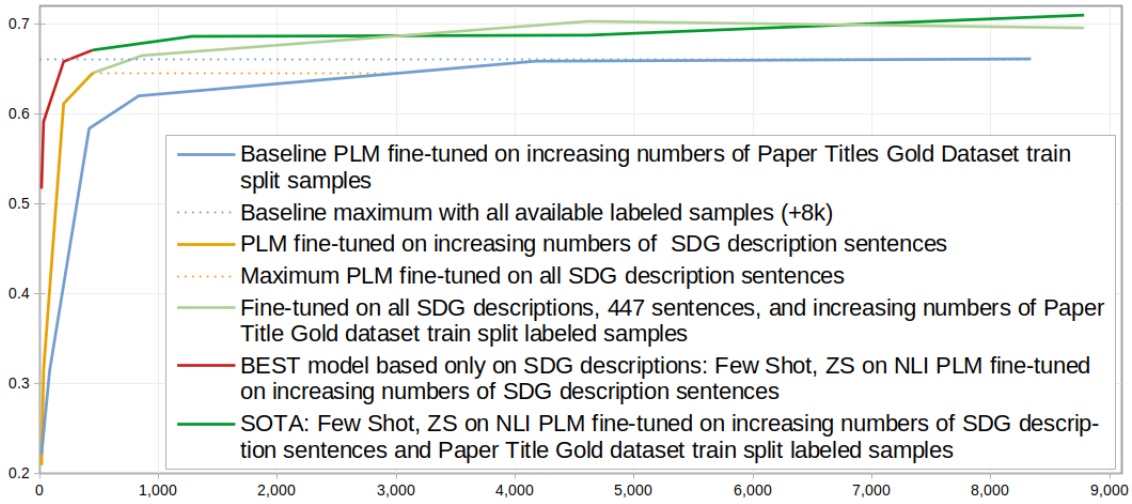
For the analysis of these results we will consider the following two scenarios of data availability:

- A "Labeled samples available" scenario, in which different amounts of labeled samples are available. In our study we simulate this scenario fine-tuning the base models with different numbers of labeled samples (75%, 50%, 25%, 10%, 5% and 1%) of the train split of the "Paper Titles Gold Dataset".
- A "Class descriptions available" scenario, in which different amounts and types of description sentences of the classes are available. We simulate this in our study by splitting the SDG-Descriptions Dataset in sub-sets of SDG-Headlines, SDG-Titles, SDG-Targets and SDG-Indicators and fine-tuning the base models in accumulative combinations of them.

Considering these simulations of scenarios, we have studied how the baseline and the different SDG-Description based models evolve with increasing numbers of available samples and descriptive sentences for training. In Figure 2 we can observe that the general baseline (blue line) requires almost 3,000 labeled samples to overcome our most simple model exclusively trained with 447 description sentences (yellow line). Furthermore, if we continue training our description based models with increasing numbers of hand-labeled samples, we can observe that the obtained fine-tuned classifier (light-green line), beats the top F1-score of the general baseline (+8000 labeled samples) with only around 500 labeled samples additional to the SDG-Descriptions. Furthermore, our few-shot classification model defines a new estate of art for classification of scientific papers by SDGs when using all SDG-Descriptions and hand-labeled samples. At this point, the peak measured F1-score is 71.01%, slightly over the human agreement level of 70.10% observed in the AURORA dataset.

Up to now we have studied the results as a whole, but the task includes 17 different classes that may behave differently. Table 2 shows the detailed global and per SDG results of the test performed with our Few-Shot approach on the Titles-Test split of the Paper Titles-Gold dataset. SDG 6 "Clean water and sanitation" and SDG 17 "Partnership for the goals" show the worst results. The model has been trained with 21 sentences describing SDG 6 and 46 sentences of SDG 17, similar or higher than the number of sentences used to train other much better performing SDGs like SDG 7 "Affordable and clean energy" (13 sentences, F1-score 83.93%) or SDG 3 "Good health and well-being" (42 sentences, F1-score 67.7%). This suggests that there is no clear correlation or proportion between the number of sentences included in the description and the performance of the model, and that the reasons for a better classification may relay on other features probably related to the semantics of the description sentences and the sentences to be classified. The study of the features that make a description good for this classification approach are lines for further research.

Figure 2: Macro-averaged F1-score curves by number of samples, on the test split of the Paper Titles Gold Dataset, of baselines and SDG description based scientific paper classifiers



5. Error Analysis

We focus the error analysis on the results of the few-shot model, the best performing model among those that use exclusively SDG-Descriptions for training and classification. Table 3 summarizes how the model gives right, wrong or inconclusive predictions. More than 90% of the good predictions are obtained with the first (Top 1 - 74.00%) and second (Top 2 - 17.79%) highest scores. The average scoring pattern gives a relatively high value at Top 1 (0.55-0.77) and drops significantly at every next prediction, scoring in the range 0.03-0.15 at Top 2 and 0.002-0.02 at Top 3. Nevertheless, the highest average Top 2 (0.1514) and Top 3 (0.0214) scores correspond to right predictions obtained at second and third guesses. Both right and wrong predictions at Top 1 score on average around 0.77, meaning that the model is particularly mistaken in the wrong predictions.

Deepening one step further, the test dataset has a particular set of samples; those that have been labeled positively or negatively by more than one expert. They could be referred as "strong true samples" or "strong false samples" if they are either positive or negative labeled samples. In opposition, we call the test samples labeled by a single person as "weak true samples" or "weak false samples". Table 4 shows the results of the test on these particular samples. The model has been able to classify correctly all the strong true samples with a particularly high average Top 1 score of over 0.88, but at the same time has classified incorrectly around 30% (12) of the strong false samples.

Overall, the behaviour of the model seems to follow what could be expected by common sense on the scenario we are working on, considering (i) that each tested paper title is most likely related with several SDGs but not in the same extent, (ii) that the label given to each test sample is not necessarily the one of the SDG they are most related to, and (iii) that the more SDGs a paper title is related to, the lower score it will give at each one of them individually. Coherently, the scores observed in the "no coincidence" predictions of the positive samples happen to be ones with lowest scores. Also, the low average scores registered in the false positive coincidences in Top 2 and Top 3 can be explained as those debatable cases that even with human observers reduce the agreement level to the previously mentioned 70.1%. Nevertheless, the following results appear to be relevant failures of the model worth to be analysed in detail:

- The 63.02% of false positive coincidences with the highest scores (0.7795) at Top 1 prediction.
- The 12 bad predictions or false positive coincidences on strong false samples.
- The high rate of undetected true positives on SDG 6 and SDG 17.

Appendix A shows several examples of these failures. Regarding the false positives in weak false samples, the wrong guesses are absolutely arguable and may fall in the side of the measured roughly 30% of human disagreement level, with the exception of the example of SDG 17 "Partnership for the goals" with the title "Tuple-based

Table 2
Testing of the few-shot model: number of sentences in the SDG-Descriptions dataset used for training the model, number of positive (True) and negative (False) samples in the Titles-Test split of the Paper Titles-Gold dataset, number of true positive (TP) and false positive (FP) predictions on this test dataset, obtained recall (r), precision (p), F1-score (f1) and accuracy (acc), and global micro and macro averages for the same metrics

SDG	Sentences in SDG descriptions	True test samples	False test samples	TP	FP	r	p	f1	acc
1	23	2	2	2	0	100.00%	100.00%	100.00%	100.00%
2	23	30	30	25	23	83.33%	52.08%	64.10%	53.33%
3	42	313	304	283	240	90.42%	54.11%	67.70%	56.24%
4	23	87	87	73	42	83.91%	63.48%	72.28%	67.82%
5	25	94	96	63	25	67.02%	71.59%	69.23%	70.53%
6	21	62	62	19	20	30.65%	48.72%	*37.62%*	49.19%
7	13	63	69	47	2	74.60%	95.92%	83.93%	86.36%
8	31	17	17	13	10	76.47%	56.52%	65.00%	58.82%
9	22	65	65	34	16	52.31%	68.00%	59.13%	63.85%
10	23	31	31	17	9	54.84%	65.38%	59.65%	62.90%
11	27	57	57	50	28	87.72%	64.10%	74.07%	69.30%
12	26	48	49	40	31	83.33%	56.34%	67.23%	59.79%
13	15	36	36	24	13	66.67%	64.86%	65.75%	65.28%
14	22	17	17	13	1	76.47%	92.86%	83.87%	85.29%
15	28	77	77	60	35	77.92%	63.16%	69.77%	66.23%
16	37	40	40	33	28	82.50%	54.10%	65.35%	56.25%
17	46	29	29	8	7	27.59%	53.33%	*36.36%*	51.72%
Total	**447**	**1,068**	**1,068**	**804**	**530**				
				Micro avg.		75.28%	60.27%	66.94%	62.83%
				Macro avg.		71.51%	66.15%	67.12%	66.05%

Table 3
Testing of the few-shot model: share (%) of positions of predictions matching the label of the test samples for true positive (TP) and false positive (FP) predictions, and average prediction scores registered in each position (average score of matching positions in bold) for TP, FP and predictions not matching neither positive nor negative samples

	Position of prediction		
	Top 1	**Top 2**	**Top 3**
Positive samples			
Prediction-test label coincidence	74.00%	17.79%	8.21%
Average prediction scores			
TP coincidence in Top 1	**0.7741**	0.0528	0.0040
TP coincidence in Top 2	0.6667	**0.1514**	0.0117
TP coincidence in Top 3	0.5558	0.0975	**0.0214**
No coincidence	0.6313	0.0655	0.0097
Negative samples			
Prediction-test label coincidence	63.02%	23.58%	13.40%
Average prediction scores			
FP coincidence in Top 1	**0.7795**	0.0393	0.0038
FP coincidence in Top 2	0.6379	**0.0782**	0.0027
FP coincidence in Top 3	0.6087	0.0751	**0.0207**
No coincidence	0.6175	0.0703	0.0134

semantic and structural mapping for sustainable interoperability" not objectively relatable with this SDG. When it comes to the false positives related to strong false samples, that have happened exclusively for samples of the SDG 3 "Good health and well-being", we can observe several possible reasons for the failures like:

- Debatable or arguable labelling.
- A possible tendency of the model to relate tobacco with health (SDG 3), and a tendency of experts not to do it when the paper titles refer to its economic dimensions.
- A difficulty of the model to distinguish between

Table 4
Testing of the few-shot model: number of samples and average Topk-3 predictions scores for good and bad predictions on positive (strong true) and negative (strong false) samples labeled coincidentally by more than one expert

	Number of samples	Average prediction scores Top 1	Top 2	Top 3
Strong True samples	**59**			
Good prediction	59	**0.8816**	0.0092	0.0020
Bad prediction	0	-	-	-
Strong False Samples	**38**			
Good prediction	26	0.6390	0.0726	0.0034
Bad prediction	12	**0.8030**	0.0223	0.0046

Figure 3: Testing of the few-shot model: co-occurrence matrix, considering the positive samples of the test dataset

SDG		Prediction																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Gold-dataset label	1	25%		13%	13%							13%	13%			25%		
	2	4%	32%	19%			4%		10%		1%		17%	1%		12%		
	3	7%	4%	31%	2%	3%	3%	0%	16%	0%	2%	2%	14%	2%	1%	7%	3%	1%
	4	10%	0%	3%	32%	1%	1%		15%	0%	4%	1%	13%	0%	0%	5%	6%	7%
	5	5%		7%	4%	28%			10%	2%	7%	1%	6%	2%		4%	18%	4%
	6	4%	4%	4%			21%	1%	5%	2%		2%	10%	9%	7%	27%		2%
	7	4%	1%	1%			1%	27%	14%	6%	1%	2%	18%	7%		11%	3%	5%
	8	5%			2%				30%	7%	16%	2%	14%	2%		5%	5%	12%
	9	2%	1%	1%				4%	9%	24%	1%	9%	17%	1%	2%	14%	8%	8%
	10	11%		3%	6%	2%			5%		26%	2%	5%	2%	2%	9%	17%	14%
	11	12%		3%		1%	2%	3%	6%	7%	2%	28%	12%	4%	2%	8%	6%	5%
	12	2%	5%	1%			2%	12%	8%	10%	2%	12%	31%	2%	1%	12%		2%
	13	15%	4%	3%			4%	1%	4%	1%		8%	7%	18%	1%	24%	1%	8%
	14	14%	3%	3%			14%		6%				17%	9%	3%	29%	3%	
	15	23%	10%	3%	1%		6%	1%	3%			6%	15%	3%	3%	23%	1%	5%
	16	14%	2%	2%	3%	3%			4%		18%	2%	4%			8%	28%	11%
	17	5%							5%	3%	13%	3%	3%	13%	5%	25%	10%	18%

animal health and human health.

In the case of the undetected positive samples of SDG 6 "Clean water and sanitation" and SDG 17 "Partnership for the goals", all cases appear to be very debatable. An explanation may be that in these cases the titles of the papers do not describe properly the contents of the paper, or even may be misleading, but the experts have labeled the papers not by their title but by their content. For instance, the paper titled "Local renewable energy cooperatives: revolution in disguise?", may be related with the SDG 6 "Clean water and sanitation", but the title itself suggests it may be more related to SDG 7 "Affordable and clean energy" as the model predicts, or the paper titled "Sustainability of small water supplies: Lessons from a Brazilian program (SESP/FSESP)" may of course be related to SDG 17 "Partnership for the goals" but the title suggests it may be mainly related to SDG 6 "Clean water and sanitation" as the model predicts.

These phenomena are most likely related to the evident overlaps that exist between the SDGs. Figures 3 and 4 depict the co-occurrence and confusion matrices of the test. The co-occurrence matrix plots all Topk-3 predictions of the model on the positive samples of the test dataset. Generally the model predicts more frequently the right SDG, but we can also observe that SDG 15 "Life on land" is remarkably more predicted than the other SDGs, followed by SDG 12 "Responsible consumption and production", SDG 1 "No poverty" and SDG 8 "Decent work and economic growth". SDG 15 "Life on land" is even more prevalent than the labeled SDG in the case of SDG 6 "Clean water and sanitation", SDG 13 "Climate action", SDG 14 "Life below the water" and SDG 17 "Partnership for the goals". The confusion matrix plots whe wrong predictions of the model. In this case most fre-

Figure 4: Testing of the few-shot model: confusion matrix, considering only wrong predictions on positive samples of test dataset

SDG		Prediction																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Gold dataset test sample label	1																	
	2	20%		13%	7%	7%			7%			7%	20%			20%		
	3	11%	1%		10%	11%	1%	3%	9%	2%	9%	6%	12%	3%	1%	4%	16%	
	4	14%	2%	14%		7%		2%	17%		7%	7%	10%			5%	10%	5%
	5	12%	2%	16%	3%		13%		12%	2%	6%	1%	11%	1%		8%	4%	9%
	6	2%		1%	2%			16%	16%	19%	2%	4%	18%	4%	1%	2%	5%	9%
	7	6%	6%	4%	4%		4%		4%	4%	8%	10%	27%	6%		6%	4%	4%
	8			8%	8%		8%			8%	8%	17%	17%			8%	8%	8%
	9	3%	5%	4%	4%		1%	8%	14%		5%	8%	27%	3%		8%	2%	8%
	10	14%		12%	5%	12%			12%	5%		7%	7%	7%		2%	10%	7%
	11	24%	5%	5%			19%	10%	14%				14%	5%		5%		
	12	4%	4%	17%			13%		8%	4%		8%			4%	29%		8%
	13		8%				8%	14%	11%	8%		3%	22%		3%	17%	3%	3%
	14	17%							8%			8%	17%	8%		25%	8%	8%
	15	14%	14%	8%		6%	8%		12%	2%	4%	6%	22%		4%			2%
	16	10%		5%		5%				10%	10%	10%	10%	19%		19%		5%
	17	14%	2%	10%	5%		6%	3%	13%	3%	5%	10%	19%	3%		6%	2%	

quently mistaken prediction is for SDG 12 "Responsible consumption and production", followed in this case by the same SDGs outstanding in the co-occurrence matrix.

According to these results, the SDGs that most overlap with the rest are the SDG 15 "Life on land", the SDG 12 "Responsible consumption and production", SDG 1 "No poverty" and SDG 8 "Decent work and economic growth". This may mean that these SDGs are the ones that most diversely may impact the UN 2030 Agenda for Sustainable Development, what could be an excellent bonus insight offered by the model, but once again, this may be related only to the different quality of the descriptions of each SDG, and for sure a question worth to be further studied.

6. Conclusions and future work

This work offers initial experimental evidences that using detailed descriptions of the main classes that shape an specific domain has the potential to benefit Text Classification. All the experiments reported have been developed classifying automatically scientific papers to UN SDGs.

The use of class descriptions may reduce significantly or even eliminate the need to develop hand-labeled samples for training NLP models, reducing drastically the development cost. Depending on the availability on descriptions of classes we recommend:

- In scenarios with detailed descriptions available: Few shot approaches, fine tuning NLI-PLM for zero-shot exclusively with class descriptions, have the potential to outperform conventional classifiers fine-tuned on PLMs with thousands of hand labeled samples.
- In the case of both detailed descriptions and labeled samples available: Conventional PLM classifiers fine-tuned with a combination of class descriptions and labeled samples have the potential to reduce the need of labeling by an order of magnitude, being able to establish a new SOTA in our case study.
- On the contrary, on a pure zero-shot approach, in cases with only a single keyword or description sentence available per class, the classical prompted keyword classification seems to be better than any similar description sentence based classifier.

The use of class descriptions instead of labeled task samples may not require sophisticated NLP approaches, offering results comparable to human classification in the studied case, using only conventional and widely-used NLP methods and models.

To benefit from these advantages, the class descriptions may be written by non domain experts in plain non-scientific or technical language. In our experimental case, SDG-Descriptions are designed for the general public understanding in a public-policy style language while paper titles are written in a specialized scientific

language. Also, not all descriptive sentences offer the same improvement potential: single sentences describing the whole class (SDG titles) and collection of single sentences describing each one a particular relevant aspect of the class (SDG targets) contribute the most.

Finally, the results of this initial experimental study suggest the following future lines of research:

- Extending the study to further specific domain NLP applications to generate further evidence about the potential benefits of using class descriptions and grasp its limitations.
- Apply the use of class descriptions in methods more sophisticated than the conventional NLP approaches applied in this work to validate or refuse the hypothesis that advanced NLP techniques like generative LLMs and QA tasks may also benefit from them.
- Deep dive in what makes a description good for NLP applications and explore how advanced description development and improvement techniques can contribute.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[2] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, 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. 3914–3923. URL: https://aclanthology.org/D19-1404. doi:10.18653/v1/D19-1404.

[3] M. Vanderfeesten, E. Spielberg, Y. Gunes, Survey data of "Mapping Research Output to the Sustainable Development Goals (SDGs)", 2020. URL: https://doi.org/10.5281/zenodo.3813230. doi:10.5281/zenodo.3813230.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018.

[6] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, 2021. arXiv:2111.01243.

[7] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, L. He, A survey on text classification: From traditional to deep learning, ACM Transactions on Intelligent Systems and Technology (TIST) 13 (2022) 1 – 41.

[8] Y. Meng, J. Shen, C. Zhang, J. Han, Weakly-supervised neural text classification, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 983–992. URL: https://doi.org/10.1145/3269206.3271737. doi:10.1145/3269206.3271737.

[9] Y. Meng, Y. Zhang, J. Huang, C. Xiong, H. Ji, C. Zhang, J. Han, Text classification using label names only: A language model self-training approach, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9006–9017. URL: https://aclanthology.org/2020.emnlp-main.724. doi:10.18653/v1/2020.emnlp-main.724.

[10] T. Xia, Y. Wang, Y. Tian, Y. Chang, FastClass: A time-efficient approach to weakly-supervised text classification, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4746–4758. URL: https://aclanthology.org/2022.emnlp-main.313.

[11] M. Thangaraj, M. Sivakami, Text classification techniques: A literature review, Interdisciplinary Journal of Information, Knowledge, and Management 13 (2018) 117–135.

[12] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[13] T. Schopf, D. Braun, F. Matthes, Evaluating unsupervised text classification: Zero-shot and similarity-based approaches, 2023. arXiv:2211.16285.

[14] D. Chai, W. Wu, Q. Han, F. Wu, J. Li, Description based text classification with reinforcement

learning, in: International Conference on Machine Learning, 2020.
[15] X. Zhu, Z. Peng, J. Guo, S. Dietze, Generating effective label description for label-aware sentiment classification, Expert Syst. Appl. 213 (2023). URL: https://doi.org/10.1016/j.eswa.2022.119194. doi:10.1016/j.eswa.2022.119194.
[16] X. Gao, Z. Zhu, X. Chu, Y. Wang, W. Ruan, J. Zhao, Enhancing robust text classification via category description, 2022 IEEE International Conference on Data Mining (ICDM) (2022) 151–160.
[17] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023. arXiv:2303.18223.
[18] B. ayabalasingham, R. Boverhof, K. Agnew, L. Klein, Identifying research supporting the united nations sustainable development goals, 2019. URL: https://elsevier.digitalcommonsdata.com/datasets/87txkw7khs/1. doi:10.17632/87txkw7khs.1.
[19] M. Rivest, Y. Kashnitsky, A. Bédard-Vallée, D. Campbell, P. Khayat, I. Labrosse, S. Pinheiro, Henrique; Provençal, G. Roberge, C. James, Improving the scopus and aurora queries to identify research that supports the united nations sustainable development goals (sdgs) 2021, 2021. URL: https://elsevier.digitalcommonsdata.com/datasets/9sxdykm8s4/. doi:10.17632/9sxdykm8s4.4.
[20] G. Roberge, Y. Kashnitsky, C. James, Elsevier 2022 sustainable development goals (sdg) mapping, 2022. URL: https://elsevier.digitalcommonsdata.com/datasets/6bjy52jkm9. doi:10.17632/6bjy52jkm9.1.
[21] D. Science, J. Wastl, S. Porter, H. Draux, B. Fane, D. Hook, Contextualizing sustainable development research, 2020. URL: https://digitalscience.figshare.com/articles/report/Contextualizing_Sustainable_Development_Research/12200081. doi:10.6084/m9.figshare.12200081.v2.
[22] C. Armitage, M. Lorenz, S. Mikki, Replication data for: Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?, 2020. URL: https://doi.org/10.18710/98CMDR. doi:10.18710/98CMDR.
[23] W. Wang, W. Kang, J. Mu, Mapping research to the sustainable development goals (sdgs), 2023. URL: https://www.researchsquare.com/article/rs-2544385/v2. doi:10.21203/rs.3.rs-2544385/v2.
[24] M. Vanderfeesten, R. Otten, E. Spielberg, Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0.2, 2020. URL: https://doi.org/10.5281/zenodo.4883250. doi:10.5281/zenodo.4883250.
[25] F. Schmidt, M. Vanderfeesten, Evaluation on accuracy of mapping science to the United Nations' Sustainable Development Goals (SDGs) of the Aurora SDG queries, 2021. URL: https://doi.org/10.5281/zenodo.4964606. doi:10.5281/zenodo.4964606.
[26] M. Vanderfeesten, R. Jaworek, L. Keßler, AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs), 2022. URL: https://doi.org/10.5281/zenodo.6487606. doi:10.5281/zenodo.6487606.
[27] R. Jaworeck, SDG BERT - Multi-language Multi-label BERT model for classifying texts to Sustainable Development Goals (SDGs) based on Aurora SDG Query Model v5, 2022. URL: https://doi.org/10.5281/zenodo.7304547. doi:10.5281/zenodo.7304547.
[28] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

A. Examples of wrong predictions

Table 5
Testing of NLI-based few-shot approach with Topk-3 classification strategy: examples of wrong predictions

#	Paper title	Gold SDG		Predicted SGD		
False postives in weak false samples		**#**	**Headline**	**Top 1**	**Top 2**	**Top 3**
1	Food insecurity and effectiveness of behavioral interventions to reduce blood pressure , New York City, 2012-2013	2	Zero hunger	2	3	8
2	Global governance for facilitating access to medicines: Role of world health organization	3	Good health and wellbeing	3	12	8
3	Equipping Preservice Elementary Teachers for Data Use in the Classroom	4	Quality education	4	8	12
4	How to study varieties of opposition to gender+ equality in Europe?: Lessons from this book, conceptual building blocks, and puzzles to address	5	Gender equality	5	10	8
5	RETRACTED ARTICLE: Comparative advantage analysis for water utilization in Hubei province based on NRCA model	6	Clean water and sanitation	6	8	1
6	A study on factors affecting the youth employment rate: Focusing on data from 31 cities and counties in Gyeonggi-do, South Korea	8	Decent work and economic growth	8	11	9
7	Analysis of the inclusions in 38Si7 spring steel with fatigue failure	9	Industry, innovation and infrastructure	9	12	8
8	The development and transition of urban walking grey space in China, based on a unique model " Langpeng"	11	Sustainable cities and communities	11	1	12
9	Corporate sustainability in emerging markets: Insights from the practices reported by the Brazilian retailers	12	Responsble consumption and production	12	8	10
10	Sensitivity analysis with the regional climate model COSMO-CLM over the CORDEX-MENA domain	13	Climate action	13	15	8
11	Rainforest tourism, conservation and management: Challenges for sustainable development	15	Life on land	15	12	8
12	Capitalizing on Criminal Accomplices: Considering the Relationship between Co-offending and Illegal Earnings	16	Peace, justice and institutions	16	1	10
13	Tuple-based semantic and structural mapping for a sustainable interoperability	17	Partnership for the goals	17	12	9
False positives in strong false samples		**#**	**Headline**	**Top 1**	**Top 2**	**Top 3**
14	Nature, scope and use of economic evaluation of healthcare programmes: With special reference to Pakistan	3	Good health and wellbeing	3	15	12
15	Endovascular Aortic Repair for Thoracic Aortic Injuries	3	Good health and wellbeing	3	8	1
16	Comparison between online and offline price of tobacco products using novel datasets	3	Good health and wellbeing	3	12	8
17	An Assessment of the Forward-Looking Hypothesis of the Demand for Cigarettes	3	Good health and wellbeing	3	8	12
18	Mycobacterium marinum infection in fish and man: Epidemiology, pathophysiology and management; a review	3	Good health and wellbeing	14	3	6
Undetected SDG 6 and SDG 17 true positives		**#**	**Headline**	**Top 1**	**Top 2**	**Top 3**
19	An exploration of the boundaries of ‘community’ in community renewable energy projects: Navigating between motivations and context	6	Clean water and sanitation	7	11	10
20	Typology of future clean energy communities: An exploratory structure, opportunities, and challenges	6	Clean water and sanitation	7	8	12
21	A review of renewable energy investment in the BRICS countries: History, models, problems and solutions	17	Partnership for the goals	7	12	8
22	Sustainability of small water supplies: Lessons from a brazilian program (SESP/FSESP)	17	Partnership for the goals	6	10	8