

How Is NLP Addressing the 17 UN Sustainability Goals? A Challenge Set of Social Good Paper Classification and Information Extraction

Fernando Gonzalez^{1,*} Zhijing Jin^{1,2,*} Jad Beydoun³ Bernhard Schölkopf^{1,2}

Tom Hope⁵ Rada Mihalcea³ and Mrinmaya Sachan¹

¹ETH Zürich, ²Max Planck Institute, ³University of Michigan,

⁴Hebrew University of Jerusalem, ⁵AI2

{fgonzalez, jinzhi, msachan}@ethz.ch, bs@tue.mpg.de,

{jadb, mihalcea}@umich.edu, tomh@allenai.org

Abstract

NLP for social good (NLP4SG) is becoming an increasingly important dimension to look at in NLP research. A vast number of NLP applications can be seen across various use cases and many academic researchers want to produce work with a positive social impact by addressing tasks such as hate speech detection, gender bias mitigation, and biomedical name entity recognition. However, good intentions need to be accompanied by a wise allocation of efforts. To provide researchers with the necessary information on the amount of attention social issues receive from the NLP4SG community, we introduce an NLP4SG scientific dataset with 3 subtasks that help to identify and characterize NLP4SG papers. We apply state-of-the-art Machine Learning (ML) models to solve each of these subtasks and identify the papers that address a social problem, describe them by obtaining the UN Sustainable Development Goals (SDG) to which they are associated, the task they solve, and the methods they use. Finally, we apply our models to the entire ACL anthology dataset and create a visualization workspace that allows researchers to get an overview of the NLP4SG field.¹

1 Introduction

Much of the organizational work in the community of natural language processing (NLP) is done by volunteers. For many communities, there are yearly workshops, paper lists, and various efforts to collect useful resources to share with researchers.

However, a blocker in such a process is that often the organization requires lots of work, which risks the sustainability of the community in the long run. Non-trivial efforts are required to maintain a list of resources and keep them up to date, inspirational for future research in this direction, and useful so

* Equal contribution.

¹Our code and data are available at <https://github.com/causalNLP/nlp4sg>.

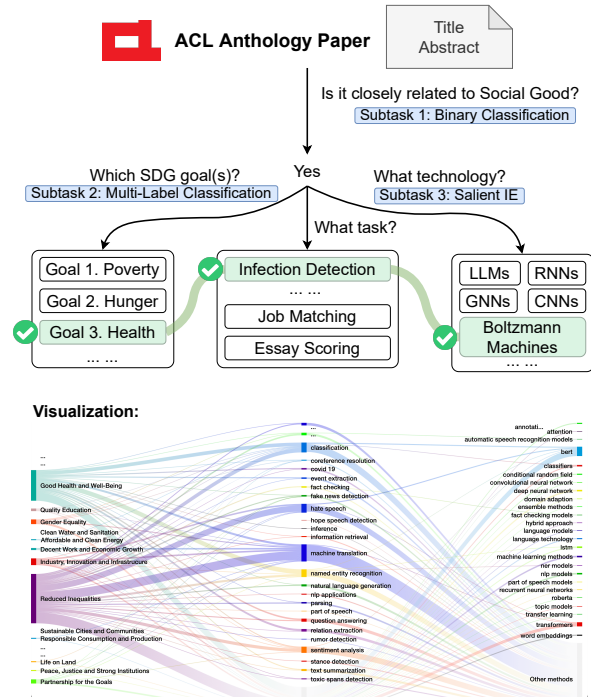


Figure 1: Overview of this paper.

that outsiders can quickly grasp a general picture of the research outputs in this domain, and find good ways to support or collaborate.

To provide a general picture of the NLP4SG domain, it is necessary to identify the most important characteristics of the scientific contributions, i.e., papers. One way to define these characteristics is by answering the questions: (1) which social problem does this paper address? (2) which aspect of this problem does it solve?, and (3) how does it solve this problem? The last two are usually defined as tasks and methods in the ML community, and some tools have been built to extract this type of information from papers. Regarding the first question, in 2015 the UN created a committee to define the most important objectives to achieve prosperity while protecting the planet, resulting in

the 17 Sustainable Development Goals (SDGs).² Based on these categories, we define the social problem that NLP papers solve.

Our goal is to provide the community with tools to get an overview of NLP4SG’s efforts. In this way, the collection of resources would represent fewer manual efforts, and it would be easier for researchers to find inspiration and contribution opportunities in the NLP4SG field. We make the following key contributions:

- An NLP4SG scientific dataset NLP4SGPAPERS that evaluates 3 social good (SG) subtasks and a series of baseline models for each one of them. The dataset provides a high-level summary of the distribution of NLP4SG work, and the subtasks work as a motivation for researchers to improve the benchmarks and thus create a more accurate summary.
- A visualization workspace in which more than 8,000 papers of ACL are shown after having been processed by our social good ML pipeline. The workspace provides an easy way to analyze NLP contributions to the UN SDGs and serves as inspiration for exploring innovative combinations of NLP tasks and methods to solve social problems.

2 NLP4SGPAPERS Dataset

We introduce the NLP4SGPAPERS challenge set which serves for two goals: to provide a high-level summary of the distribution of NLP4SG work for researchers, fund granters, and the general public; and to inspire future researchers what the neglected social good areas are, and encourage proposals of novel tasks.

To achieve the above two goals, we construct a pipeline (Figure 1) for NLP4SGPAPERS: (1) first, we identify whether a paper is closely related to social good or not; (2) then, for each identified NLP4SG paper, we further classify it into the UN SDG(s) that it contributes to; (3) finally, to gain further insights into the actual research trend, we identify more fine-grained information in the NLP4SG papers, such as the tasks and methods. In the following, we propose three corresponding subtasks for each step in the pipeline. For each subtask, we

²<https://www.un.org/sustainabledevelopment>

introduce task definition, our annotation process, and an analysis of our collected dataset.

2.1 Data Source

We use data from the ACL Anthology³, an official repository of NLP papers. The data was downloaded on May 27, 2022 and it contains 76,229 papers. Then we randomly sample 5,000 papers, for which we parse the title and abstracts provided by the ACL Anthology.⁴ For papers that ACL Anthology does not provide abstracts, we download their PDFs and parse the abstracts by the PDF-to-JSON converter `doc2json` by Semantic Scholar.⁵

Note that there are other sources of NLP papers, such as the arXiv category `cs.CL`, but the scope of our dataset mainly focuses on ACL Anthology papers, since most of the papers have peer-reviewed quality, and it is a commonly recognized and up-to-date source of NLP papers. We welcome future work to adapt our models and demo on other sources of papers.

We split the data into the train ($n=2,500$), development ($n=500$), and test set ($n=2,000$). The general statistics of the splits are shown in Table 1.

Preprocessing. We remove entries that are not papers, by a set of hand-crafted rules such as removing entries whose title starts with “Book Reviews: ...” or ends with “... Workshop.” We also remove the special symbols “{” and “}” in titles such as “NoPropaganda at SemEval-2020 Task 11: A Borrowed Approach to Sequence Tagging and Text Classification.” Our preprocessing codes are available at our code repository.

2.2 Subtask 1: Binary Classification of Social Good Papers

	# All Papers	# SG Papers	# Sents	# Tokens	Vocab
Total Data	5,000	624	40,525	659,975	45,454
Train	2,500	297	20,305	333,094	29,563
Dev	500	63	4,112	66,435	10,640
Test	2,000	264	16,108	260,446	24,899

Table 1: Statistics of SG binary classification task.

Task Definition. Some scientific papers justify their contribution with technical reasons, and others

³<https://aclanthology.org>

⁴<https://aclanthology.org/anthology+abstracts.bib.gz>

⁵<https://github.com/allenai/s2orc-doc2json>

find motivation in social problems. The first task is to identify which are the papers that directly address a social problem, using for that the title and the abstract of the paper.

The US has published standards for AI regulation: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

Annotation Process. We ask an NLP4SG research expert to identify if a paper directly addresses a social problem. We provide them with a set of guidelines on what an NLP4SG paper is. The guidelines are divided into 2 parts. The first part is about the definition of social good, we compile a list of social good topics, namely the 17 UN SDGs plus well-known digital problems, e.g., hate speech, fake news, and privacy. Based on that list, we create a list of keywords for each topic. In the second part of the guidelines, we define what means to directly address a social problem. We rely on the 4-stage classification proposed by Jin et al. (2021) and define “directly” as stages 3 and 4 which correspond to applicable tools and deployed applications. The diagram in Figure 10 illustrates the annotation rules we use.

Exploratory Data Analysis. As shown in Table 1 our dataset has 12.48% of SG papers. From year 2000 the proportion of NLP4SG papers steadily increased until 2022 when it dramatically increased as can be seen in Figure 2. The snapshot of ACL Anthology data we are working with has a lower proportion of papers in 2022.

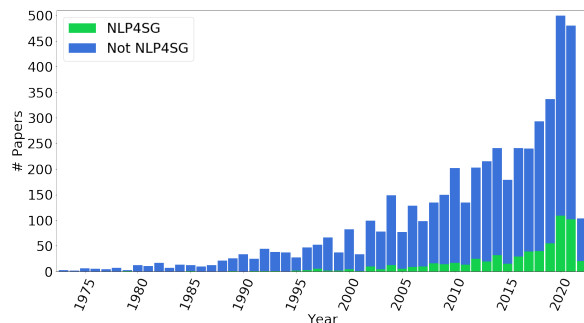


Figure 2: Distribution of SG papers in NLP4SGPAPERS

2.3 Subtask 2: Multi-Label Classification to Match with UN SDG Goals

Task Definition. The second subtask is to identify what type of social problem is being addressed by the papers by classifying them into one or more

of the 17 UN SDGs. We treat this task as a multi-label classification one because some papers contribute to more than one SDG or their contribution is on the borderline of 2 SDGs.

UN SDG	Original	Upsampled
Goal 1. No Poverty	—	—
Goal 2. Zero Hunger	—	0.14%
Goal 3. Good Health and Well-Being	32.42%	28.61%
Goal 4. Quality Education	16.21%	14.31%
Goal 5. Gender Equality	2.57%	11.53%
Goal 6. Clean Water and Sanitation	—	0.28%
Goal 7. Affordable and Clean Energy	—	—
Goal 8. Decent Work and Economic Growth	8.19%	7.64%
Goal 9. Industry, Innovation and Infrastructure	16.21%	14.17%
Goal 10. Reduced Inequalities	5.14%	4.72%
Goal 11. Sustainable Cities and Communities	0.16%	0.42%
Goal 12. Responsible Consumption and Production	0.16%	0.14%
Goal 13. Climate Action	0.16%	2.78%
Goal 14. Life Below Water	—	0.14%
Goal 15. Life on Land	—	—
Goal 16. Peace, Justice and Strong Institutions	30.66%	28.19%
Goal 17. Partnership for the Goals	2.57%	2.50%

Table 2: UN SDGs and their frequencies (%) in the dataset.

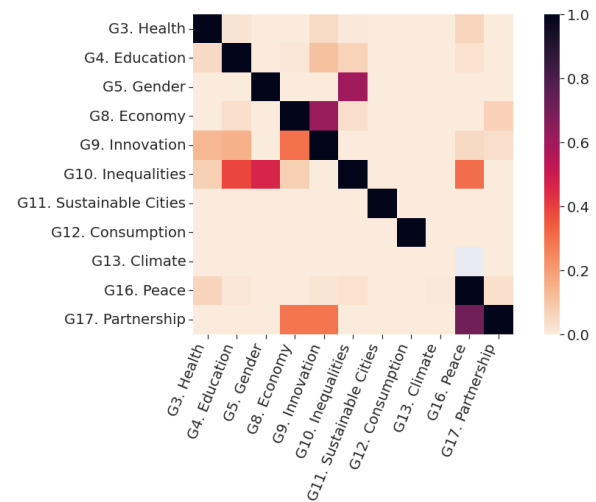


Figure 3: Heatmap of co-occurrences of UN SDGs on papers.

Annotation Process. We provide the annotator with a description of each one of the UN SDGs. We ask them to select up to 3 goals for papers that they marked as SG.

Upsampling of low occurrence classes In Table 2 we show that some UN SDG classes represent a very low proportion of the observations, some of them even 0%. On one hand, this reflects the small attention these social problems are getting from the NLP4SG community. However, for the purpose of getting a good estimate of the performance of the model in all the classes, we up-sample the low

occurrence classes, that being classes that have less than 5% of the observations. We do this by getting the papers that match one or more of the social good keywords from our list. Those additional papers are manually annotated. The new proportions are also shown in Table 2. After upsampling the total number of papers is 5,097, and the number of SG papers is 720.

Data Analysis. On average, there are 1.14 goals per paper. In Figure 3 we show a heatmap of co-occurrences. The goals with most co-occurrences are Goal 9 Industry, Innovation, and Infrastructure and Goal 8 Decent Work and Economic Growth with 18.

2.4 Subtask 3: Salient Information Extraction (IE)

Task Definition The third task is scientific IE. It has been explored already by multiple works (Luan et al., 2018; Gábor et al., 2018; Augenstein et al., 2017; Jain et al., 2020), and some datasets that address paragraph and document-level scientific IE have been released. Unlike those approaches, we do not have span-level annotations, but entities obtained via abstractive summarization. The task also has similarities with keyword extraction (Bennani-Smires et al., 2018; Sun et al., 2020). However, it has the additional component of extracting different types of keywords (methods and tasks).

Title: Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections

Abstract:

Detecting healthcare-associated infections poses a major challenge in healthcare. Using natural language processing and machine learning applied on electronic patient records is one approach that has been shown to work. However the results indicate that there was room for improvement and therefore we have applied deep learning methods. Specifically we implemented

a network of stacked sparse auto encoders and

a network of stacked restricted Boltzmann machines .

Our best results were obtained using the stacked restricted Boltzmann machines with a precision of 0.79 and a recall of 0.88.

Task: Detecting healthcare-associated infections

Method: Boltzmann machines, auto encoders

Annotation We present the title and abstract of social good papers to the annotator and ask them to write down a list of the main tasks and methods addressed by each paper. In Table 3 we show an example of the type of annotations we collect.

Data Analysis. For this subtask, we have 100 annotated SG papers. An average of 1.21 tasks and 1.69 methods per paper, and a total of 116 unique tasks and 159 unique methods, which means we have very few repeated tasks and methods. In the dataset, we have elements that are the same but written slightly different, which explains this behavior.

3 Models for Subtask 1

We explore the use of multiple deep-learning models for subtask 1. We use GPT-3 (Brown et al., 2020) in a zero-shot and few-shot settings and also fine-tune BERT-base models (Devlin et al., 2019). Since we have a big set of unlabeled data from the ACL anthology, we also perform experiments using distant supervision data in order to augment the available training data and aim for better performances.

3.1 Data augmentation

We include other sources of gold data + weakly supervised data in our training set. The additional gold data comes from a reading list of 213 NLP4SG papers, and a previously annotated dataset of 534 ACL 2020 papers with both positive and negative examples. We consider 3 ways to obtain the weakly supervised data:

- We create a list of “social good keywords”. If the title or abstract of the paper contains one of those keywords, the paper is marked as SG.
- We create a list of “social good workshops”. We consider that a paper is an NLP4SG paper if it is part of one of the workshops. The list of workshops was extracted from the ACL website.⁶
- We measure the cosine similarity between the text embeddings of the title + abstract of the paper and the text embeddings of the description of each one of the UN SDGs and took the most similar UN SDG. We consider a paper as SG if its text similarity is among the

Table 3: Example annotation of subtask 3.

⁶<https://aclanthology.org/volumes/>

	Class = Positive			Class = Negative		
	F1	P	R	F1	P	R
Workshop Name Match	11.73	70.37	6.40	93.88	88.76	99.64
Title Keyword Match	68.76	91.11	55.22	96.71	94.27	99.27
Title+Abstract Keyword Match	61.71	51.58	76.77	93.36	96.65	90.29
SG similarities	8.08	47.06	4.42	92.45	86.57	99.20
Title keyword+similarities	75.06	86.98	66.01	95.22	92.78	97.79

Table 4: Quality evaluation of the distant supervision heuristics. We use the training data to evaluate all heuristics and rank them by quality.

top 1% most similar papers in the unlabeled data from the ACL anthology. We consider that a paper is not a NLP4SG one if the similarity is in the bottom 50%. For the text embeddings, we use a general-purpose sentence similarity model, all-distilroberta-v1⁷ which is a pre-trained distilroberta-base (Sanh et al., 2019) model fine-tuned on a 1B sentence pairs dataset.

We measure the performance of these methods using the training set as an evaluation set as shown in Table 4. Title keyword+similarities implies selecting positive examples with: keywords in the title, most similar papers to the description of SDGs; then selecting negative examples with the least similar papers. The rest of the papers are labeled as undefined and therefore not used for the evaluation of the data augmentation technique.

3.2 Models

SciBERT We fine-tune SciBERT (Beltagy et al., 2019) on the text classification task to detect if a paper is SG or not by freezing the embedding layer and the first 11 encoder layers and learn the parameters of the 12th encoder layer, the pooler, and the classifier layer. The details about training time, GPU used, and hyperparameter tuning can be found in the Appendix B.1.

GPT-3 For the second approach, we design a prompt and ask GPT-3 in a zero-shot setting if a paper is NLP4SG or not. The prompt contains 3 elements:

- Introduction to the task: We use the sentence “There is an NLP paper with the title and abstract:”
- Content of the paper: The title and abstract concatenated

⁷<https://huggingface.co/sentence-transformers/all-distilroberta-v1>

- Final question: We use the sentence “Is this paper contributing to the UN Sustainable Development Goals? Answer yes or no.”

We also perform experiments using few-shot settings, i.e., 2,4,8,16, and 32 shots, and report the results for the setting with the best performance in the dev set. For those experiments, we only use the title of the paper and the same proportion of positive and negative examples.

3.3 Results Analysis

In Table 5 we compare the performance of the models under different settings and different ways to build the training set. The SciBERT fine-tuned classifiers achieve better performance than GPT-3. In particular, the model using gold data+ title keyword matching + SDG similarity achieves the highest F1-score.

Explainability In order to understand what kind of information makes the model decide if a paper is SG or not, we use LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016). LIME explains the predictions of black box models by creating local interpretable models around the observation that it tries to explain. An example of a true positive (TP) prediction of the model is shown in figure Figure 4. The model uses the words “counseling” and “psychosocial” as positive evidence that the paper is NLP4SG and the word “qualitative” as negative evidence. We also use LIME to explain a true negative (TN), false positive (FP), and false negative (FN) observation. The details can be found in the Appendix C.1.

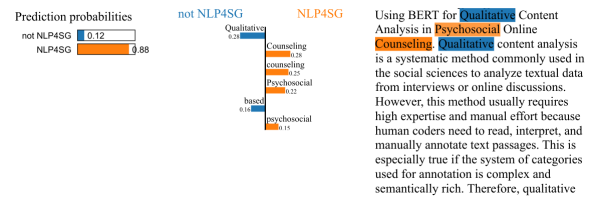


Figure 4: LIME output for a TP example

Common Errors We analyze the type of errors that the model makes most frequently. We divide it into 2 parts, the analysis of false positives and false negatives.

For the case of false negatives, we use the UN SDG annotations that are already part of the dataset. In this way we were able to identify that even when

Model	Input Text	Data + Distant Supervision				Performance			
		Our Annot.	+ Prev. Annot.	+ Title Keyword	+ SDG Sim.	F1	Acc	P	R
Random 10:90	N/A					7.81	78.75	9.14	6.82
Random 50:50	N/A					16.29	46.55	10.27	39.39
GPT3-Instruct (0-Shot)	Title					28.83	80.25	27.49	30.30
GPT3-Instruct (16-shot)	Title	✓				40.72	86.75	49.73	34.47
SciBERT	Title&Abstract	✓				69.49	92.80	78.47	62.36
SciBERT	Title&Abstract	✓	✓	✓		74.16	94.25	90.66	62.74
SciBERT	Title&Abstract	✓	✓	✓	✓	75.26	94.15	84.76	67.68

Table 5: Performance of different models on the binary classification of social good papers.

peace (29.16%) and health (26.51%) are the most frequent categories in the test set. Education with 20.83% of the NLP4SG papers in the test set has a proportion of 30.33% among the false negatives. Which indicates that there is room for improvement in the identification of education papers.

Regarding false positives, we analyze the frequency of the words in false positive papers compared to the frequency in the complete test set. One interesting case is the word “social” that appears in 14.29% of the FP papers, but only in 3.65% of the papers in the test set. In particular, this word is a good indicator that a paper is NLP4SG given that it appears in 18.28% of the TP, but it is also used in papers that analyze social media not necessarily with a focus in positive social impact.

4 Models for Subtask 2

We use 3 models to address the SDGs classification task. As we mention in 2.3 we do not have data for some classes in the entire NLP4SGPAPERS set. Moreover, in our training set, we only have 297 positive examples for a classification task with 17 classes. For that reason, we decided to focus our experiments on zero-shot approaches.

4.1 Models

Model 1: GPT-3 Instruct We design a prompt to get which UN goal a paper contributes to. Since GPT-3 already has background knowledge of which are the UN SDGs, and what is the purpose of each one of them we do not need to include that explanation in the prompt. The prompt has the following components:

- Introduction to the task: We use the sentence “There is an NLP paper with the title and abstract:”
- Content of the paper: The title and abstract concatenated.

- Final question: We use the sentence “Which of the UN goals does this paper directly contribute to? Provide the goal numbers and names.” as well as that same question but in a singular form, i.e., asking for the goal number and name.

We parse the response, allowing multiple goals for each paper.

Model 2: Pretrained MNLI classifier We use a pretrained MNLI classifier (bart-large-mnli) as a zero-shot text classifier, an approach proposed by Yin et al. (2019). We get the probability of entailment/contradiction of the content of the paper with the description of each one of the UN goals obtained from the UN website. We consider that a paper contributes to a UN SDG if the probability of entailment is 0.5 or higher. If there is no SDG with a probability higher than 0.5 we take the SDG with the highest probability.

4.2 Results Analysis

We show the multi-label classification performance by SDG of GPT-3 and the Zero-shot classifier in Table 6. GPT-3 achieves an F1 score above 65 in the 3 most common UN SDGs when using the prompt that asks for a single UN SDG. We show the details of its performance in Table 7 and binary confusion matrices in Figure 5. When using the prompt in plural form, we notice that in some answers, GPT-3 lists all the goal names and numbers, which negatively impacts the precision. We perform a visual inspection of the responses and found that GPT-3 gives an accurate reasoning explanation of why a paper is contributing to a SDG, so we decide to further analyze GPT-3 responses in the next section.

4.3 Analyzing LLMs Understanding

We design a prompt where, in addition to asking which SDG the paper addresses, we ask about the

reasoning behind the answer. GPT-3’s responses typically contained an explanation linking some sentence in the paper to how this contribution could help the specific SDG. For the responses that correctly classified the UN SDG, we analyze 2 aspects of the answer:

- Is the “impact chain” reasoning complete or too abrupt because of missing any intermediate steps? We assign 0 if the reasoning chain is too abrupt, 1 if the reasoning is complete and 0.5 if it is in a mid-point.
- Does it miss any obvious other impact chains?. We assign 0 if it misses an obvious impact chain, 1 if it doesn’t and 0.5 if it misses a not obvious impact chain.

We analyze 200 positive scenarios. In 33.3% of the cases, GPT-3 correctly identified the SDG. For those correctly classified cases, the average response for the first aspect was 0.776 which means that around 77.6% of the answers are considered as well reasoned. 18% of the answers miss an obvious impact chain. We show some example responses in the Appendix C.2.

Model	F1	PaM	ExM	P	R
Zero-shot classifier	28.38	39.44	10.69	46.06	35.79
GPT3 (plural)	54.55	67.50	39.86	54.36	61.06
GPT3 (singular)	65.04	70.69	62.08	77.91	61.55

Table 6: Multi-label UN SDG classification performance (weighted average) in terms of F1, exact match (ExM), partial match (PaM), precision and recall on 720 SG papers.

5 Models for Subtask 3

After identifying NLP4SG papers, we characterize them, that is, extract the tasks, and methods mentioned on them. We test 3 approaches, since the task is very similar to NER + coreference resolution we perform experiments with two scientific information extraction models, and as a third approach, we design a prompt for a Large Language Model. Due to the format of the annotations, we do not evaluate on the span level, we use a similarity metric instead. We describe our evaluation metric in 5.1.

5.1 Models

1) PURE As we mentioned in 2.4 the task is very similar to scientific NER, for our first approach

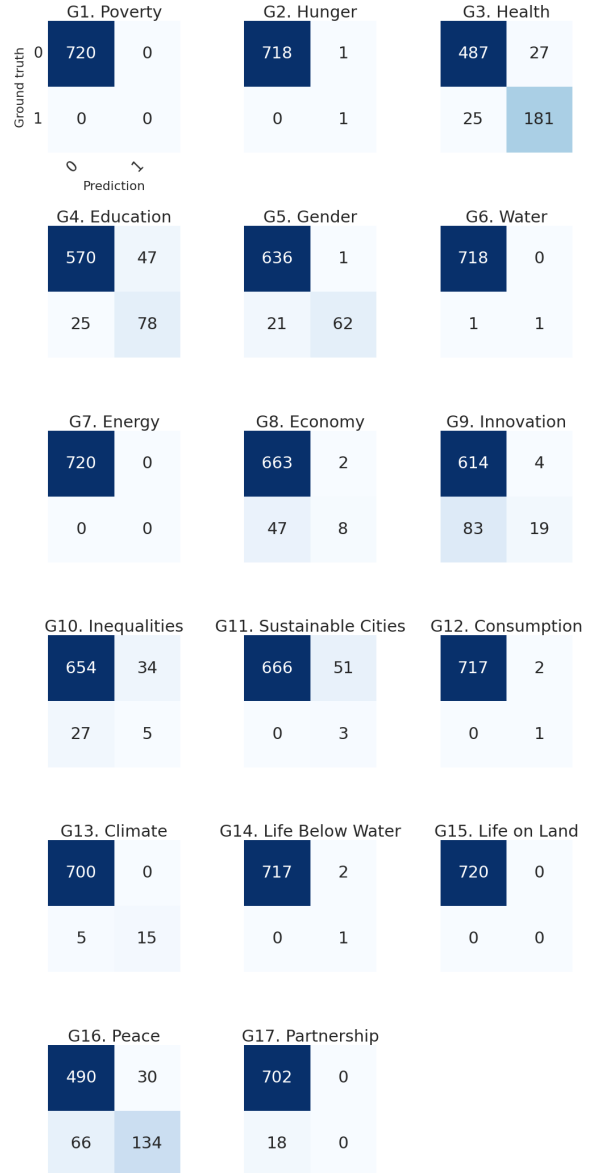


Figure 5: Confusion matrix of GPT3 goal multi-label classification.

we use a model trained on NER. PURE (Zhong and Chen, 2021) is a model for end-to-end relation extraction. It is trained on 3 datasets, we use the model trained on SciERC, which is a collection of scientific abstracts.

2) SciREX SciREX (Jain et al., 2020) is a document level IE dataset. In Jain et al. (2020), they present a model that identifies mentions, their saliency, and their coreference links. The reasons behind choosing SciREX are the coreference and saliency identification modules which in combination with NER can potentially detect the main tasks and methods as in our annotated dataset. We also

UN SDG	GPT-3			
	F1	P	R	Freq.
G1. Poverty	–	–	–	0.00%
G2. Hunger	66.67	50.00	100.00	0.12%
G3. Health	87.44	87.02	87.86	24.91%
G4. Education	68.42	62.40	75.73	12.45%
G5. Gender	84.93	98.41	74.70	10.04%
G6. Water	66.67	100.00	50.00	0.24%
G7. Energy	–	–	–	0.00%
G8. Economy	24.62	80.00	14.55	6.65%
G9. Innovation	30.40	82.61	18.63	12.33%
G10. Inequalities	14.08	12.82	15.62	3.87%
G11. Sustainable Cities	10.53	5.56	100.00	0.36%
G12. Consumption	50.00	33.33	100.00	0.12%
G13. Climate	85.71	100.00	75.00	2.42%
G14. Life Below Water	50.00	33.33	100.00	0.12%
G15. Life on Land	–	–	–	0.00%
G16. Peace	73.63	81.71	67.00	24.18%
G17. Partnership	0.00	0.00	0.00	2.18%
Weighted Average	65.04	77.91	61.55	100%

Table 7: GPT-3 multi-label classification performance on each individual SDG.

report the performance without saliency detection.

3) GPT-3 We use GPT-3 in a zero-shot setting in order to extract the relevant information. We use a different prompt for each of the 2 extraction tasks (tasks and methods). The prompt consists of 3 elements:

- Instruction: Here we tell GPT-3 what to do in natural language form.
- Paper content: We concatenate the title and abstract of the paper.
- Final question: We present the question again to reinforce the explanation.

The exact prompts can be found in the Appendix B.1.

Evaluation metric To assess how well our models extract information from the papers, we evaluate the precision, recall, and f1 score per paper and then report the average over all the test set. Instead of considering exact matches we make use of the ratio score in the library `thefuzz`⁸ which is defined by

$$\text{ratio} = \text{int}(\text{round}((2.0 * E/T) * 100)), \quad (1)$$

where T is the total number of characters in both strings, and E is the number of matches in the two

strings. The resulting score is in the range $[0,100]$, we consider a match if the ratio score is greater than 75.

An annotated paper p_j contains N tasks $[t_1, t_2, \dots, t_n]$ and M methods $[m_1, m_2, \dots, m_m]$. For each paper, our model predicts a set of P coreference task groups, $[\hat{g}_1, \hat{g}_2, \dots, \hat{g}_p]$ where each group can have 1 or more (P_g) mentions of the task $\hat{g}_1 = [\hat{g}_1^1, \hat{g}_1^2, \dots, \hat{g}_1^{P_{g1}}]$. We measure the precision, recall, and f1 score between the set of annotated tasks and the set of predicted task groups in the following way: We take each task in the gold set and compare it with each element of every coreference group. The matched coreference group is the one with the element with the highest similarity score > 75 , so a gold task can only match with 1 coreference group. If there is no element that meet these criteria, we say that the gold task does not have a match. Moreover, a coreference group can match with only one gold task. The precision is computed as $\# \text{matches} / \# \text{predicted_groups}$ and the recall $\# \text{matches} / \# \text{gold_elements}$. Finally, we average the precision, recall and f1 over all the papers in the test set. We define the partial match (PaM) metric as the proportion of observations that have at least 1 similarity match.

5.2 Result Analysis

We show the performance of task extraction in Table 8, and method extraction in Table 9. We report additional experiments where we combine SciREX with GPT-3. For those experiments, we mainly use SciREX and just fill the observations without prediction with GPT3’s prediction. This event happens when the paper mentions methods or tasks implicitly, since GPT-3 is a generative model, it is able to generate a prediction that is not in the text. The best performing model is SciREX+GPT3. As shown in Table 8 and Table 9, we achieve a good recall performance. However, the precision is low. This has to do with the systematic difference between our annotations and the models’ predictions, i.e., the list of predicted entities is usually larger than the list of gold entities. The test set has 1.21 tasks and 1.69 methods in average per paper. The predictions have an average of 2.61 tasks and 3.0 methods per observation. We tried to mitigate this error with the saliency module of SciREX, but the performance did not improve for the method extraction task.

⁸<https://github.com/seatgeek/thefuzz>

Model	F1	PaM	P	R
GPT3	27.56	51.46	20.67	48.06
PURE (SciERC)	36.03	57.28	29.53	55.34
SciREX	52.88	84.47	44.19	81.23
SciREX+salient	53.08	76.70	50.26	71.84
SciREX+GPT3	52.88	84.47	44.19	81.23
SciREX+salient+GPT3	53.08	76.70	50.26	71.84

Table 8: Test set performance task extraction. For task extraction, the results of SciREX+GPT3 are the same as SciREX alone because there are no missing values. The evaluation metric is described in Section 5.1

Model	F1	PaM	P	R
GPT3	33.47	57.28	28.62	47.04
PURE (SciERC)	34.50	59.22	32.38	45.61
SciREX	42.13	67.96	39.42	55.11
SciREX+salient	36.56	62.14	33.30	51.55
SciREX+GPT3	44.08	71.84	40.71	59.00
SciREX+salient+GPT3	38.50	66.02	34.60	55.44

Table 9: Test set performance method extraction. The evaluation metric is described in 5.1

6 Demo Website Visualization

We build a visualization workspace that provides a general overview of the NLP4SG research status. We show papers from the ACL anthology and their main features obtained by our models. In this section, we describe the components of our visualization workspace.

6.1 NLP4SG vs UN SDG nlp4sg.vercel.app/social_needs

In order to better allocate NLP research efforts to social issues, it is important to look at the alignment of three elements.

1. The importance of each SDG. Yang et al. (2020) conducted a worldwide survey of sustainability researchers and measure how experts rank the SDGs. They computed a priority score, which we use as an importance metric.
2. The intention of NLP researchers. For this, we use a survey of NLP researchers conducted on ACL 2021. The question presented on the survey is: what social good applications should we pursue?
3. The actual effort made by NLP researchers to address each social problem.

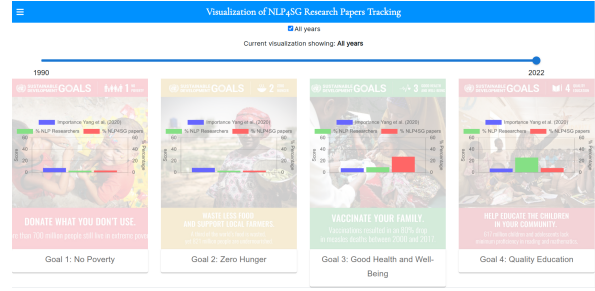


Figure 6: Website visualization NLP4SG vs importance UN SDGs

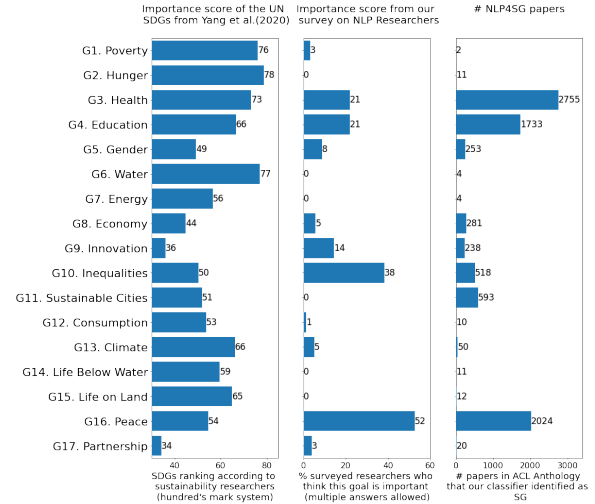


Figure 7: SDGs importance vs actual efforts

In Figure 6 we illustrate our first visualization component which shows an overview of the evolution of NLP4SG research through the years. It shows what proportion of NLP papers address each SDG, and it allows us to compare the three elements stated above and see how aligned they are. In Figure 7 we present the distribution of the three elements side by side.

6.2 NLP4SG Sankey Diagram nlp4sg.vercel.app/sankey

This component is the main part of the workspace, and has two elements:

Sankey diagram The main element is a sankey diagram that shows how much effort has been put into each SDG and how they are being addressed by the NLP community, that is, which tasks are being solved, and the methods used to solve them. The diagram has 3 columns, the left column is the result of subtask 2 and shows the number of papers that address each goal. The central and right columns correspond to the results of subtask 3. The

central column shows the most common tasks in NLP4SG papers, e.g., machine translation, name entity recognition, and the right column presents the methods used to solve them, e.g., BERT, GPT-3. The links between columns represent the papers that solve a specific goal through a certain task using a certain method. With this, the researchers can gain insight into, for a specific goal, which tasks are addressed more often and which have not been explored as much, as well as the methods commonly used to solve them, thus consider new ways to contribute to the solution of social problems. The diagram is illustrated in Figure 1.

NLP4SG papers list The second element is a list of papers. When the user clicks on a link in the sankey diagram the list shows the papers that correspond to that specific goal, task, or method. We show the title and abstract of the papers and highlight the text spans that contain the task and methods. This function is helpful in NLP4SG literature review process, since it facilitates the exploration of papers using the key elements of the abstract. We show an example of this visualization in Figure 8.

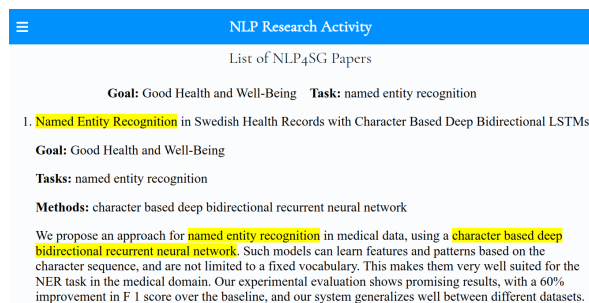


Figure 8: NLP4SG papers list after clicking on Health and NER.

Topic Cluster nlp4sg.vercel.app/papers

6.3 Organizations Sankey Diagram

nlp4sg.vercel.app/sankey_org

To have an impact on society, researchers usually partner with organizations that contribute with resources to reach shared goals. Our dashboard contains a visualization that helps to identify which organizations are associated with which type of papers, that is, which UN SDG they solve. We present the relations that exist between goals and organizations in the form of a Sankey diagram in

which the users can also click on the links to see the title and abstract of the papers. In this way, researchers can look for collaborations with organizations that usually support a specific research path.

7 Related Work

NLP4SG Previous works analyze the role and status of AI4SG and NLP4SG. Shi et al. (2020) create a survey of AI4SG, they analyze the application domains of the papers and the AI techniques used. They perform a case study for eight application domains of AI4SG. Cows et al. (2021) introduce a database of AI4SG initiatives and analyze 108 projects that address one or more of the SDGs. Yeh et al. (2021) introduce a collection of datasets and benchmarks related to the UN SDGs. They focus on encouraging the development of machine learning models for each individual UN SDG. Fortuna et al. (2021) provides an analysis of the research tendencies on NLP4SG. They identify explicit NLP4SG papers with a set of keywords and then analyze the proportion of NLP4SG papers in different areas and venues. We focus on the domain of NLP4SG, characterize the papers not only by the social problem they address, but by the tasks and methods they use and present a visualization workspace that let users explore the NLP4SG field interactively.

Paper Classification and IE visualization: Lahav et al. (2021); Hope et al. (2020) present visualization workspaces that allow researchers to explore topics in scientific literature related to COVID-19. (Lahav et al., 2021) extract research challenges and directions from the abstract of COVID-19 related papers. Hope et al. (2020) present a system where researchers can explore authors, topics, and associations between COVID-19 related concepts. Our visualization rather focuses on tasks, methods and social problems. Our main visualization provides a general picture of the contributions of NLP to social problems.

8 Discussion

8.1 Limitations and Future Directions

A limitation to obtaining an accurate picture of the state of the NLP4SG research is the performance of the models. The models allow us to present a general overview of which social problems are being

addressed and how. However, there are some papers that may have been misclassified, which could cause some confusion to users of the visualization tool. As future work, we want to explore new models for each subtask and explore other sources of NLP papers, such as the arXiv category cs.CL or the NLP dataset compiled by [Mohammad \(2020\)](#).

9 Conclusion

We introduce a new NLP4SG dataset and proposed three subtasks that together help to get an overview of the papers that address social problems within the NLP scientific community. We explore different models to solve the 3 subtasks and create a visualization workspace through which the efforts that the NLP community has made to solve social problems can be explored.

Acknowledgments

We thank Flavio Schneider for deploying the basic structure of the website based on which we built the visualizations.

This material is based in part upon works supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by the Precision Health Initiative at the University of Michigan; by the John Templeton Foundation (grant #61156); by a Responsible AI grant by the Haslerstiftung; and an ETH Grant (ETH-19 21-1). Zhijing Jin is supported by PhD fellowships from the Future of Life Institute and Open Philanthropy. We also thank OpenAI for granting Zhijing free access to their API of GPT series through the Researcher Access Program.

References

- Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. [SemEval 2017 task 10: SciencE - extracting keyphrases and relations from scientific publications](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics. 4
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics. 5
- Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. [Simple unsupervised keyphrase extraction using sentence embeddings](#). 4
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Matheus Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc. 4
- Josh Cows, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. [A definition, benchmark and database of ai for social good initiatives](#). *Nature Machine Intelligence*, 3:111–115. 10
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 4
- Paula Fortuna, Laura Pérez-Mayos, Ahmed AbuRa’ed, Juan Soler-Company, and Leo Wanner. 2021. [Cartography of natural language processing for social good \(NLP4SG\): Searching for definitions, statistics and white spots](#). In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 19–26, Online. Association for Computational Linguistics. 10
- Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. [SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers](#). In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics. 4
- Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel Weld, Marti Hearst, and Jevin West. 2020. [SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 135–143, Online. Association for Computational Linguistics. 10

- Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. [SciREX: A challenge dataset for document-level information extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics. 4, 7
- Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. [How good is NLP? a sober look at NLP tasks through the lens of social impact](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics. 3
- Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S. Weld, and Tom Hope. 2021. [A search engine for discovery of scientific challenges and directions](#). 10
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. [Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics. 4
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc. 14
- Saif M. Mohammad. 2020. [NLP scholar: A dataset for examining the state of NLP research](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 868–877, Marseille, France. European Language Resources Association. 11
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 14
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. 5
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108. 5
- Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. [Artificial intelligence for social good: A survey](#). *CoRR*, abs/2001.01818. 10
- Yi Sun, Hangping Qiu, Yu Zheng, Zhongwei Wang, and Chaoran Zhang. 2020. [Sifrank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model](#). *IEEE Access*, 8:10896–10906. 4
- Siqi Yang, Wenwu Zhao, Yanxu Liu, Francesco Cherubini, Bojie Fu, and Paulo Pereira. 2020. [Prioritizing sustainable development goals and linking them to ecosystem services: A global expert's knowledge evaluation](#). *Geography and Sustainability*, 1(4):321–330. 9
- Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B. Lobell, and Stefano Ermon. 2021. [Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 10
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [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)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics. 6
- Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics. 7

A Additional dataset details

In Figure 9 we show the co-occurrence matrix with the absolute number of co-occurrences.

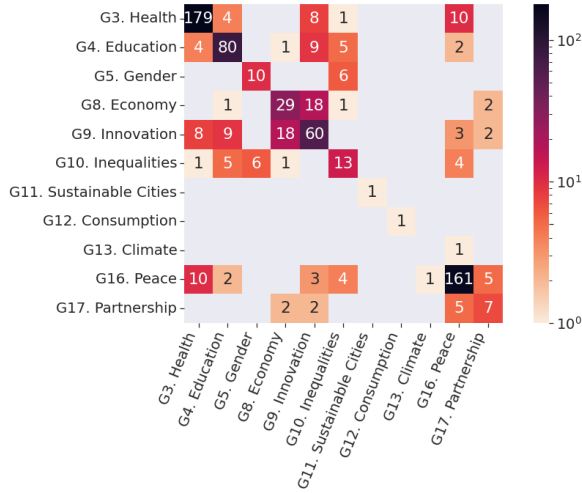


Figure 9: Heatmap of co-occurrences of UN SDGs on papers.

Figure 10 shows the decision diagram we use for the social good NLP papers annotation.

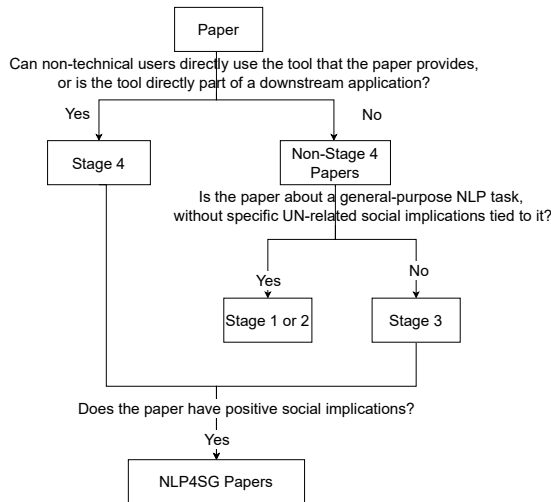


Figure 10: NLP4SG decision diagram used for annotations.

B Experimental Details

B.1 Implementation Details

SG classifier We fine-tune SciBERT for classification by adding dropout and a single linear layer. We use one GPU model NVIDIA Quadro RTX 6000 with 24 GiB. We train the classifier for 25 epochs using a batch size of 32. The training time

for our best performing model is approximately 2 hours. We manually grid search with different learning rates (5e-03, 5e-04, 5e-05, 5e-06) and different scheduler types (“constant”, “linear”, “cosine with restarts”) to make a total of 12 runs and chose the best hyperparameters based on f1 score on the dev set. The chosen learning rate is 5e-05 and a linear scheduler.

We tried with a similar setting using BERT-base and BERT-large obtaining lower performances.

GPT Implementation We use the OpenAI API⁹ to access GPT. We use the largest engine “davinci-text-002.” We keep default values of the API, and set the temperature to zero and maximum number of tokens of the response as 50.

Information extraction For SciREX and PURE we use one GPU model NVIDIA Quadro RTX 6000 with 24 GiB. It takes around 1hr to extract the information from 9k social good papers.

B.2 GPT-3 prompts

We perform experiments with GPT-3 for the 3 sub-tasks. The prompts we use are as follows:

Few-shot SG binary classification

Are these papers contributing to the UN Sustainable Development Goals or helping society?

Paper: [Title + Abstract]

Answer: [Yes/No]

Paper: [Title + Abstract]

Answer: [Yes/No]

Paper: [Title + Abstract]

Answer:

[GPT-3 response]

SDG classification

There is an NLP paper with the title and abstract:

[Title + Abstract]

Which of the UN goals does this paper directly contribute to? Provide the goal number and name.

[GPT-3 response]

⁹<https://beta.openai.com/overview>

SDG classification explanation

There is an NLP paper with the title and abstract:

[Title + Abstract]

Is this paper contributing to the UN Sustainable Development Goals? Answer yes or no. If the answer is “yes”, mention which goal the paper is contributing to and in which way it is contributing to that goal.

[GPT-3 response]

IE tasks

Identify the NLP task(s) addressed in this paper. Answer in the format [task1,...,taskn,...]

[Title + Abstract]

The NLP tasks addressed in this paper are:

[GPT-3 response]

IE methods

Identify the NLP task(s) addressed in this paper. Answer in the format [task1,...,taskn,...]

[Title + Abstract]

What are the methods employed to solve the tasks? Don't include tasks in the answer. The NLP methods used in this paper are:

[GPT-3 response]

B.3 Tasks and methods clustering

C Additional results

C.1 Subtask 1

LIME In Figures 11 to 13 we show examples of LIME explanations. A particularly interesting example is Figure 12 where the word “suspicion” makes the model classify the paper as SG.

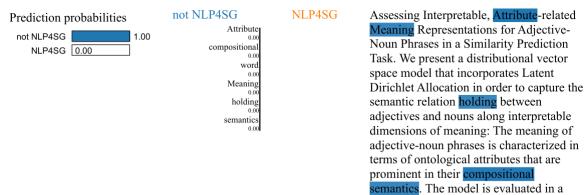


Figure 11: LIME output for a TN example

C.2 Subtask 2

Analyzing LLMs understanding In Table 10 we show an example where GPT-3 provides an

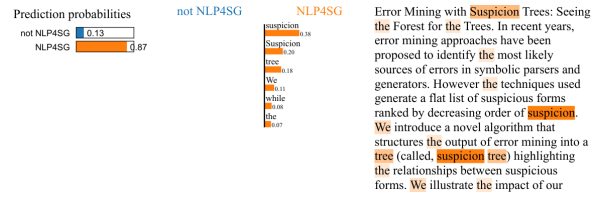


Figure 12: LIME output for a FP example

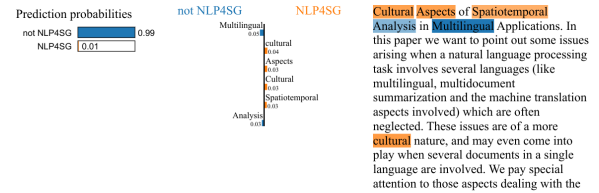


Figure 13: LIME output for a FN example

answer with intermediate steps for the reasons why the paper is contributing to the SDGs. In Table 11, GPT-3 doesn't provide a complete reasoning.

D Preprocessing for visualization

The output of subtask 3 is a list of methods and tasks names. However, these names are not normalized, that is, there are different spellings for the same method or task. To achieve a better visualization of the most frequent methods and tasks, we group similar elements together with the following procedure:

- We get a vector representation for each word with Word2Vec (Mikolov et al., 2013). We train Word2Vec with the abstract of all the ACL Anthology papers, we set the parameter window to 10 and vector size 100.
- We use the agglomerative clustering function in scikit-learn (Pedregosa et al., 2011) on the word vectors, with the parameter distance threshold set to 0.2.
- To double-check that the elements of the cluster are similar, we measure the ratio score similarity between each word of a group to its centroid and just keep in the cluster the words that have a similarity score greater than 75.
- We take that automatic clustering as a starting point and then manually check each one of the clusters, add new elements and create new clusters if necessary.

Title: Extracting Symptoms and their Status from Clinical Conversations

Abstract: This paper describes novel models tailored for a new application, that of extracting the symptoms mentioned in clinical conversations along with their status. Lack of any publicly available corpus in this privacy-sensitive domain led us to develop our own corpus, consisting of about 3K conversations annotated by professional medical scribes. We propose two novel deep learning approaches to infer the symptom names and their status: (1) a new hierarchical span-attribute tagging (SA-T) model, trained using curriculum learning, and (2) a variant of sequence-to-sequence model which decodes the symptoms and their status from a few speaker turns within a sliding window over the conversation. This task stems from a realistic application of assisting medical providers in capturing symptoms mentioned by patients from their clinical conversations. To reflect this application, we define multiple metrics. From inter-rater agreement, we find that the task is inherently difficult. We conduct comprehensive evaluations on several contrasting conditions and observe that the performance of the models range from an F-score of 0.5 to 0.8 depending on the condition. Our analysis not only reveals the inherent challenges of the task, but also provides useful directions to improve the models.

GPT3 response: Yes, this paper is contributing to the UN Sustainable Development Goal of Good Health and Well-Being. The paper describes models that are designed to extract symptoms from clinical conversations, which can assist medical providers in capturing symptoms mentioned by patients. This can ultimately lead to better diagnosis and treatment of diseases, and improve overall health and well-being.

Table 10: Example well reasoned answer.

Title: Extracting Symptoms and their Status from Clinical Conversations

Abstract: Lexically-Triggered Hidden Markov Models for Clinical Document Coding. The automatic coding of clinical documents is an important task for today's healthcare providers. Though it can be viewed as multi-label document classification, the coding problem has the interesting property that most code assignments can be supported by a single phrase found in the input document. We propose a Lexically-Triggered Hidden Markov Model (LT-HMM) that leverages these phrases to improve coding accuracy. The LT-HMM works in two stages: first, a lexical match is performed against a term dictionary to collect a set of candidate codes for a document. Next, a discriminative HMM selects the best subset of codes to assign to the document by tagging candidates as present or absent. By confirming codes proposed by a dictionary, the LT-HMM can share features across codes, enabling strong performance even on rare codes. In fact, we are able to recover codes that do not occur in the training set at all. Our approach achieves the best ever performance on the 2007 Medical NLP Challenge test set, with an F-measure of 89.84.

GPT3 response: Yes, this paper is contributing to the UN Sustainable Development Goal of good health and well-being. The paper is doing this by proposing a Lexically-Triggered Hidden Markov Model (LT-HMM) that leverages phrases to improve coding accuracy for clinical documents. This will ultimately help healthcare providers provide better care to patients.

Table 11: Example abrupt reasoning.