Beyond Good Intentions: Reporting the Research Landscape of NLP for Social Good

Fernando Gonzalez[1,*] **Zhijing Jin**[1,2,*] **Bernhard Schölkopf**[1,2]
Tom Hope[4,5] **Mrinmaya Sachan**[1,†] and **Rada Mihalcea**[3,†]
[1]ETH Zürich [2]MPI for Intelligent Systems [3]University of Michigan
[4]The Hebrew University of Jerusalem [5]Allen Institute for AI (AI2)
{fgonzalez,jinzhi,msachan}@ethz.ch bs@tue.mpg.de
tomh@allenai.org mihalcea@umich.edu

Abstract

With the recent advances in natural language processing (NLP), a vast number of applications have emerged across various use cases. Among the plethora of NLP applications, many academic researchers are motivated to do work that has a positive social impact, in line with the recent initiatives of NLP for Social Good (NLP4SG). However, it is not always obvious to researchers how their research efforts tackle today's big social problems. Thus, in this paper, we introduce NLP4SGPAPERS, a scientific dataset with three associated tasks that can help identify NLP4SG papers and characterize the NLP4SG landscape by: (1) identifying the papers that address a social problem, (2) mapping them to the corresponding UN Sustainable Development Goals (SDGs), and (3) identifying the task they solve and the methods they use. Using state-of-the-art NLP models, we address each of these tasks and use them on the entire ACL Anthology, resulting in a visualization workspace that gives researchers a comprehensive overview of the field of NLP4SG.[1]

1 Introduction

With the rapid advancement of natural language processing (NLP) technology (Radford et al., 2018; Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022), there has been a growing interest in using NLP for applications with social impact in recent years (Wu et al., 2020; de Mello et al., 2019; Jin and Mihalcea, 2023). This effort has been condensed into the initiative of "NLP for Social Good" (NLP4SG) (Jin et al., 2021), which began with an overall theoretical framework (Jin et al., 2021) in 2021, and is reaching broader community impact through various workshops at NLP conferences (Field et al., 2021; Biester et al., 2022).

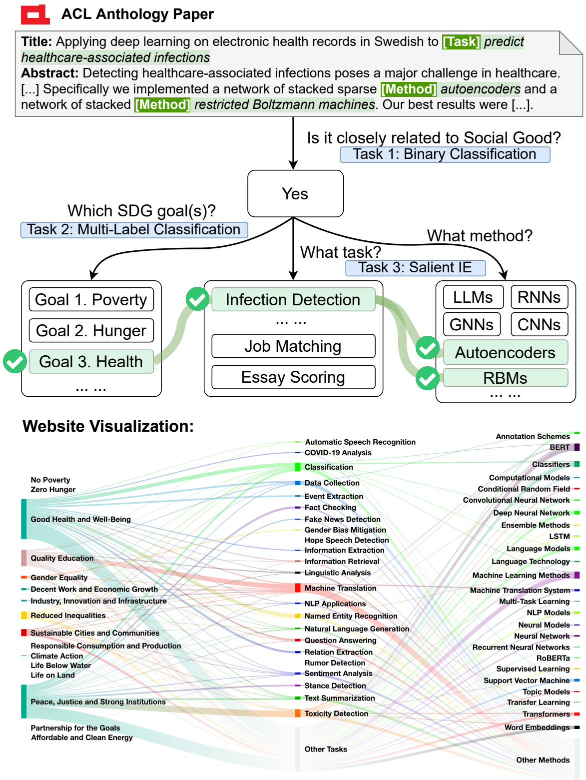


Figure 1: To generate the NLP4SG progress report, we provide a *PaperAnalyzer*, which identifies social good papers (Task 1), classifies the relevant UN SDGs (Task 2), and analyzes salient scientific terms (Task 3). We process all ACL Anthology papers to create a Sankey diagram (bottom) on our website.

However, most of the efforts so far have consisted of disparate community discussions, opinion pieces, and projects, yet without a unifying community-based insight. This is partly due to a lack of understanding of how NLP research efforts align with social good implications. We argue that it is critical to have community reports

[*] Equal contribution.
[†] Equal supervision.
[1] We open-source our entire project:
- Website: nlp4sg.vercel.app
- Data: huggingface/feradauto/NLP4SGPapers
- Code: github.com/feradauto/nlp4sg

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 415–438
December 6-10, 2023 ©2023 Association for Computational Linguistics

based on *solid empirical evidence*, similar to the United Nations' (UN) annual compilation of the report on Sustainable Development Goals (SDGs) (e.g., United Nations, 2021, 2022). Such progress reports for NLP4SG can help calibrate intuitions with facts and (1) inspire the NLP community to understand how well NLP research address social good; (2) raise awareness of under- or unaddressed areas; (3) serve as a reference for researchers who start in the field when deciding on what topics to take on; and (4) foster better connections between the research community and social science experts, allowing for more impactful research and feedback to improve NLP4SG.

To this end, we design a new suite of tasks for analysing NLP4SG papers, spanning from identification of NLP4SG papers, classifying them into SDGs, to identifying the task and methods used in them. We create NLP4SGPAPERS, a scientific dataset of 5,000 papers annotated with the above three tasks, and use the dataset to build an "NLP for NLP" *PaperAnalyzer* system (Figure 1) that applies the state-of-the-art NLP models to the NLP papers.

Using our *PaperAnalyzer* system, we parse the entire database of 76K ACL Anthology papers, and find the following trends: (1) 13.07% of the papers are related to social good; (2) healthcare and education are popular goals among NLP4SG papers, but goals such as poverty and hunger are largely unaddressed; (3) most common tasks addressed by the community in NLP4SG are machine translation, text classification, and toxicity detection, and although the most common models are large language models (LLMs) (e.g., Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019), these innovations have not been deployed in all tasks.

In summary, our contributions are as follows:

1. We introduce a suite of novel tasks to analyze NLP4SG papers, and support it with NLP4SGPAPERS, a scientific dataset annotated with 5,000 NLP papers.
2. We provide our *PaperAnalyzer* system, which consists of LLM-based models on the three tasks.
3. We analyze the entire ACL Anthology to understand key trends in NLP4SG from 1980.
4. We open-source our NLP tools, and build a publicly accessible website to visualize the analysis results, making the progress tracking accessible for anyone interested in NLP4SG.

2 Goals of the NLP4SG Report

We begin by listing the key questions that drive our NLP4SG progress report and the desired properties of a good NLP4SG report, and highlight what purpose the report will serve. Finally, we introduce our proposed NLP-for-NLP pipeline.

Key Questions. When envisioning the NLP4SG progress report, we find the following questions crucial to address:

- Q1 **Overall estimate.** How large is the proportion of NLP4SG papers among all NLP papers? How does this proportion change over the years?
- Q2 **Social good aspects.** Among NLP4SG papers, what specific social good aspects do they tend to address?
- Q3 **Technical solutions.** For each social good goal, how are the papers addressing them? For example, what tasks do people formulate, and what models do people use? Is there an overlooked space?

Properties of a Good Progress Report. After settling in the key questions that the report should aim to answer, we propose several criteria for a good automation of the NLP4SG progress report. A good report should (**P1**) comprehensively address the questions; (**P2**) base its answers on representative, large data; (**P3**) require little repetitive manual effort, ideally largely automated; (**P4**) be transparent and open-sourced so that it is trustworthy; (**P5**) be reader-friendly for both technical and non-technical audiences; and (**P6**) be easy to access for all parties, including but not limited to researchers, social good organizations such as NGOs, funding agencies, and the general public.

Missions. This report aims to serve several purposes. First, we want to convey an overall positive and encouraging message that there is a growing and diverse community of researchers who are interested in NLP4SG. Second, we want to improve the cohesion of this community, by making the information on the existing research landscape easier to access, so that people can build on each other's research. Third, we aim to inspire more people to develop novel tasks and use

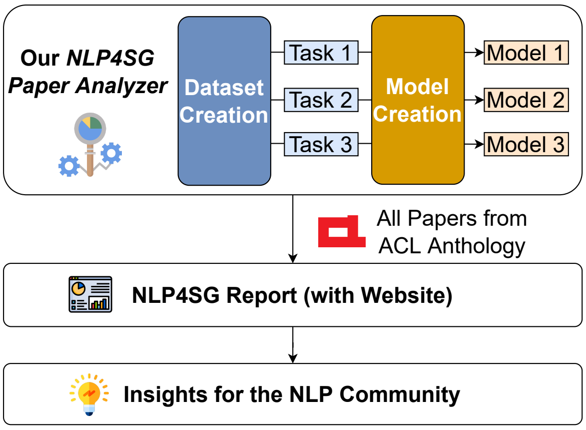


Figure 2: Our pipeline to automate the NLP4SG report.

cases to fill in the under-addressed social impact areas. Finally, we want to make the current progress transparent and public to all interested parties, especially non-research parties, including organizations such as NGOs working on NLP4SG, decision-makers, and the general public. By doing this, we can better connect the research community with all other related communities, not only to broadcast the impact of NLP4SG to the broader community, but also to help improve NLP research with external feedback, and formulate research questions more aligned with real-world problems.

Our Proposed NLP-for-NLP pipeline. We develop an NLP-for-NLP pipeline to automatically analyze papers for the NLP4SG report, as shown in Figure 2. As part of this pipeline, we develop a *PaperAnalyzer* system consisting of various models trained to answer the three questions (Q1-Q3).

The pipeline consists of the following steps: (1) We annotate various datasets that can help us answer the three questions (meeting P1 and P2). (2) We use this data to train and evaluated our *PaperAnalyzer* models. (3) Next, we run our NLP4SG *PaperAnalyzer* (which meets P3) on a large repository of NLP papers, and compose a comprehensive NLP4SG progress report, together with an open-source GitHub release of our code (which meets P4) and an interactive interface whose screenshots are in Appendix A (which meets P5 and P6). (4) Finally, we distill some key findings for the NLP community, to pave the way for more cooperative, evidence-guided community efforts to fuel NLP4SG.

3 NLP4SGPAPERS Dataset

We build a NLP4SGPAPERS dataset to address the following three tasks as illustrated in Figure 1:

- **Task 1:** Binary classification of NLP4SG papers (for Q1).
- **Task 2:** Multi-label multi-class categorization of NLP4SG papers into the 17 UN SDGs (for Q2).
- **Task 3:** Identification of salient scientific terms (tasks and methods) in NLP4SG papers (for Q3).

Next, we introduce the data collection process, and then analyze the data statistics for each task.

3.1 Dataset Collection

Data Source and Preprocessing. We identify the ACL Anthology as an ideal data source. We obtain 76,229 papers from ACL Anthology that were published by May 27, 2022.[2] We randomly sample 5,000 papers uniformly over the entire ACL Anthology and parse their titles and abstracts. See Appendices B.1 and B.2 for discussions about the data sources and our data preprocessing details. We split the dataset into 2,500 train samples, 500 development samples, and 2,000 test samples. Statistics of the data are shown in Table 1.

Task Formulation. Let us denote our dataset as $\boldsymbol{D} := \{\boldsymbol{x}_i\}_{i=1}^N$ consisting of N papers, each paper $\boldsymbol{x} := (\boldsymbol{t}, \boldsymbol{a})$ contains a title $\boldsymbol{t}$ and abstract $\boldsymbol{a}$.

Task 1 is a binary classification task $f_1 : \mathcal{X} \rightarrow \mathcal{Y}$, which maps each paper to a binary label space $\mathcal{Y} = \{0, 1\}$ representing whether a paper address a social good task ($y = 1$) or not ($y = 0$).

Task 2 is a multi-label multi-class classification task $f_2 : \mathcal{X}' \rightarrow \mathcal{Z}$, which maps NLP4SG papers $\mathcal{X}' = \{\boldsymbol{x} | f_1(\boldsymbol{x}) = 1, \boldsymbol{x} \in \boldsymbol{D}\}$ to a space $\mathcal{Z}$, which is the power set of all 17 UN SDGs.

In Task 3, we extract the main NLP tasks addressed and methods used in the NLP4SG papers, $f_3 : \mathcal{X}' \rightarrow \mathcal{W}_{\text{t}} \times \mathcal{W}_{\text{m}}$, where $\mathcal{W}_{\text{t}}$ and $\mathcal{W}_{\text{m}}$ represent the space of the power sets of the main tasks and methods in NLP4SG papers, respectively. Given our desiderata, it is better to make the extracted terms easy to use for summarizing trends in our

[2] https://aclanthology.org/anthology+abstracts.bib.gz

Sankey diagram (Figure 1). Inspired by the annotation scheme of Papers with Code,[3] we use task names such as "machine translation" and "toxicity detection"; and we use method names such as "BERT-based models," "long short-term memory networks (LSTMs)," and "support vector machines (SVMs)." We provide the detailed list in Appendix B.3.4. This allows us to use Task 3 to provide evidence that can help us answer high-level questions such as what NLP4SG work uses what type of technology, as shown in Figure 1.

Data Annotation. The data associated with each task was annotated by two NLP researchers who are proficient in English. We provide detailed annotation guidelines for each task in Appendix B.3. Briefly, the inter-annotator agreements are 92.93% Cohen's kappa for Task 1, 88.67% Cohen's kappa for Task 2, and achieve a high similarity score of 84.63% BERTScore (Zhang et al., 2020) between the annotator's responses for Task 3. We resolved cases of disagreement by first asking them to review potential oversights, and then clarifying the annotation principles for samples that are ambiguous. With these clarifications, the annotators converged to an agreement through re-iteration and discussions.

3.2 Dataset Characteristics

Data Overview for Task 1. We show the overall statistics in Table 1. We can observe that 11.84% of the papers are related to social good (592 papers of our sample of 5000).

	# All Papers	# SG Papers	# Sents	# Tokens	Vocab
Total	5,000	592 (11.84%)	40,533	656,939	45,088
Train	2,500	284	20,305	330,618	29,274
Dev	500	57	4,112	65,705	10,511
Test	2,000	251	16,116	260,616	24,913

Table 1: Statistics for Task 1. We show the number of social good-related papers (# SG Papers), along with the general statistics, including the total number of papers (# All Papers), sentences (# Sents), tokens (# Tokens), and vocabulary size (Vocab).

Data Overview for Task 2. For Task 2, we first show in Table 2 the distribution of all 17 UN SDGs among all the 592 social good-related papers. The distribution is strongly skewed, with some goals covering the majority of the data. For example, the top three goals – G3 Health (34.12%), G16 Peace (31.93%), and G4 Education (16.22%) – cover a total of 82.27% of the NLP4SG papers. Many goals rarely occur or are even not present in the dataset, such as G1 Poverty and G2 Hunger. This distribution suggests either a lack of attention from the NLP4SG community to some of these UN goals or challenges in addressing these goals through NLP. Additional analyses are presented in Section 5.

To enrich the data for low- or zero-occurrence classes, we perform upsampling to increase the number of samples corresponding to all the classes that constitute less than 5% of our dataset. Specifically, we take the unannotated part of the ACL Anthology, run the best model for Task 1 to identify social good papers, and use SDG-specific keywords to obtain highly possible paper candidates for the under-addressed SDGs. Then, we manually check the SDGs that these papers address. By this procedure, we enrich the original data by 167 more papers, which are more distributed over the low- or zero-occurrence classes. Details of the upsampling procedure are in Appendix B.3.3.

We report the statistics of the upsampled dataset in Table 2 and more in Appendix B.3.3. Since each paper can correspond to several UN SDGs (on average, 1.10 goals per paper), we plot the heatmap of co-occurrences of the SDGs in Figure 3. As we can see, some of the top correlations are seen between G2 Hunger and G8 Economy, as well as between G5 Gender and G10 Inequalities.

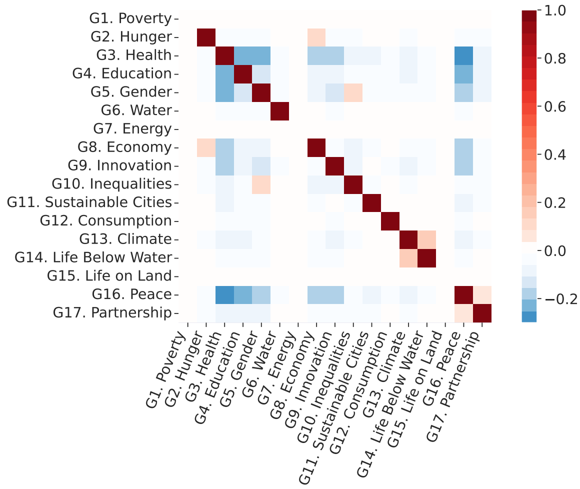


Figure 3: The correlation heatmap of the 17 UN SDGs.

[3] https://paperswithcode.com/

UN SDG	Proportion	# Samples (Before → After)
G1. Poverty	0.00%	0 → 0
G2. Hunger	0.00%	0 → 4
G3. Health	34.12%	202 → 202
G4. Education	16.22%	96 → 97
G5. Gender	3.04%	18 → 85
G6. Water	0.00%	0 → 1
G7. Energy	0.00%	0 → 0
G8. Economy	2.36%	14 → 58
G9. Innovation	13.85%	82 → 97
G10. Inequalities	5.24%	31 → 32
G11. Sustainable Cities	0.84%	5 → 16
G12. Consumption	0.17%	1 → 1
G13. Climate	0.17%	1 → 20
G14. Life Below Water	0.00%	0 → 2
G15. Life on Land	0.00%	0 → 0
G16. Peace	31.93%	189 → 201
G17. Partnership	2.70%	16 → 21

Table 2: Statistics of the data for Task 2. We first show the natural proportion of UN SDGs in the dataset, which motivates us to do upsampling. And we show the number of samples before and after upsampling. G1, G7, and G15 remain absent from the dataset even after upsampling.

Data Overview for Task 3. Since there has been significant previous work (Jain et al., 2020; Luan et al., 2018) on extracting salient information from scientific literature, we only annotate test data for this task to evaluate existing models in our domain of NLP4SG. On average, the spans of task terms are 3.6 words long, and the spans of method terms have 2.4 words long. Also, each paper has an average of 1.06 tasks and 1.5 methods, and the entire set has a total of 263 unique tasks and 335 unique methods.

4 NLP4SG *PaperAnalyzer*

Our *PaperAnalyzer* system works as shown in Figure 1, with three sequential steps: we first train a binary classifier (addressing Task 1), followed by a UN SDG goal classifier (addressing Task 2), and finally processing the paper with a pretrained task and method extractor (addressing Task 3).

4.1 NLP4SG Binary Classifier

Experimental Setup. We adopt several commonly used LLMs on our NLP4SGPAPERS dataset, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), SciBERT (Beltagy et al., 2019), and InstructGPT (Ouyang et al., 2022) text-davinci-002.

We use distant supervision (DS) for data augmentation when finetuning the models. To obtain distantly supervised data, we use two methods to identify additional NLP4SG papers from the unannotated set of almost 76K papers. We first apply keyword matching using a list of curated NLP4SG keywords such as "healthcare" and "education", and then we also include papers with high cosine similarity with the textual descriptions of the UN SDGs. See the full list of keywords and details of how we compile the distantly supervised data in Appendix D.1.1, and see our model implementation details in Appendix D.1.2.

	F1	Acc	P	R
Random (Proportional)	6.68	79.05	7.61	5.95
Random (Uniform)	15.65	46.65	9.77	39.29
InstructGPT (Zero-Shot)	28.83	80.25	27.49	30.30
InstructGPT (Few-Shot)	40.72	86.75	49.73	34.47
BERT	65.87	92.90	83.03	54.58
BERT + DS	73.14	94.05	84.38	64.54
RoBERTa	74.89	94.30	83.74	67.73
RoBERTa + DS	75.16	**94.35**	83.82	68.13
SciBERT	73.42	94.10	**84.90**	64.68
SciBERT + DS	**75.98**	94.15	78.39	**73.71**
Ablation Study				
SciBERT + DS by Similarity	62.96	89.35	55.86	72.11
SciBERT + DS by Keyword	74.68	94.10	80.93	69.32

Table 3: Model performance on Task 1, including the random baselines, whose distribution are proportional to the label distribution or uniform; as well as common LLMs, with and without our proposed DS technique. We report the F1 score, precision (P), and recall (R) of the positive class, as well as accuracy (Acc). For the best-performing model, SciBERT+DS, we also conducted ablation studies removing the different subsets of our DS data.

Results. We show our model performance for the binary classification task in Table 3. The best-performing model is SciBERT+DS, which is enhanced by the addition of our DS technique. DS contributes 2.56 percentage points (pp) improvement over the baseline SciBERT model on fully supervised data only. Moreover, with the ablation study, we show that by using only DS by keyword we get an improvement of 1.26 pp on the F1 score. Using only similarity DS does not improve the F1 score, but causes an increase in recall of 7.43 pp. Both DS techniques combined help the model to increase the F1 score to 75.98%. We choose the model with the best F1 score because we want to get a precise but also a complete report of the NLP4SG landscape.

Interpretability & Error Analysis. To understand the working of our model, we apply a commonly used interpretability tool, the local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) to our best performing model, SciBERT+DS. LIME outputs the words that our model relies on as positive and negative evidence for the classification. Some common positive keywords for our model include "biomedical", "students", and "social", while some negative keywords are "lexical", "parsing" and "grammar". The details of our LIME implementation and more visualized examples can be found in Appendix E.1.

Our model makes a few common error types, including: (1) For false positives (FP), many papers include the common word "social", which appears in 10.10% of the FP papers. We suspect that this is due to the frequent use of "social media", which is not necessarily related to social good, and the fact that "social" is a good indicator of an NLP4SG paper since it appears in 15.03% of the true positive papers. (2) Among false negative (FN) samples, a large subset is innovation and education-related papers, constituting 30.30% and 27.27% respectively among the FN, larger than their proportion (13.55% and 23.90%) in the entire test set. A reason might be that the frequent use of words like "learning" and scientific vocabulary makes our model struggle identifying a research paper that uses machine learning vs one addressing education topics.

4.2 UN SDG Goal Classifier

Experimental Setup. For Task 2, we evaluate zero-shot models using all the 759 annotated NLP4SG papers as our test set, as it is challenging to train classifiers for 17-class classification with large imbalance that results in several low-occurrence classes. Specifically, we first adopt the pretrained MNLI classifiers, including BART (Lewis et al., 2020), DistilBERT (Sanh et al., 2019), and DeBERTa (He et al., 2021). We also include the InstructGPT model (Ouyang et al., 2022), which we ask to generate a list of relevant SDGs for the paper. More experimental setup details are in Appendix D.2, including the prompts, implementation details for the models, and the evaluation metrics.

Results & Error Analysis. We evaluate all models and report their results in Table 4. Among all the models, InstructGPT has the highest performance, scoring 69.37% F1. It also has the highest performance on all other measures, such as partial list match with 74.57% of the samples, and exact list match with 66.66% of the samples.

Model	F1	PM	EM	P	R
BART	27.30	37.55	10.80	44.22	35.65
DistilBERT	26.09	58.23	3.95	17.78	56.66
DeBERTa	25.50	39.13	11.20	37.82	35.89
InstructGPT	**69.37**	**74.57**	**66.66**	**77.64**	**67.95**

Table 4: Model performance on Task 2. We report the weighted F1, percentage of samples whose ground-truth list of goals exactly matches (EM) with the model output, percentage of samples whose ground-truth list of goals partially matches (PM) with the model output, precision (P), and recall (R).

UN SDG	F1	P	R
G1. Poverty	–	–	–
G2. Hunger	75.00	75.00	75.00
G3. Health	88.78	87.50	90.10
G4. Education	73.45	64.34	85.57
G5. Gender	84.56	98.44	74.12
G6. Water	100.00	100.00	100.00
G7. Energy	–	–	–
G8. Economy	54.32	95.65	37.93
G9. Innovation	26.45	66.67	16.49
G10. Inequalities	27.40	24.39	31.25
G11. Sustainable Cities	44.78	29.41	93.75
G12. Consumption	50.00	33.33	100.00
G13. Climate	85.71	100.00	75.00
G14. Life Below Water	80.00	66.67	100.00
G15. Life on Land	–	–	–
G16. Peace	80.74	84.07	77.66
G17. Partnership	0.00	0.00	0.00
Weighted Average	69.37	77.64	67.95

Table 5: The class-specific performance of our best-performing InstructGPT method.

We also perform a more detailed breakdown of the performance of our best-performing model by different UN SDGs in Table 5. We can see that it is easy for the model to distinguish goals such as health, gender equality, and climate. However, other goals such as social inequalities and innovation are ambiguous for the model. One possible reason is that some NLP papers use domain-specific words such as "bias," which can also mean data bias, or spurious correlations in the data, so the model might confuse it with the general meaning of bias towards or against certain groups of people in society.

4.3 Task and Method Analyzer

Experimental Setup. Although our goal is to extract the main tasks and methods for our Sankey diagram, our setting is novel and distinct from traditional sentence-level named entity recognition (NER), which aims at span-level extraction of all named entities. Instead, our setting is document-level as we just need one mention of the same term in case there are co-references or paraphrases. We also accept generative answers if there are no appropriate existing spans. Thus, we evaluate Task 3 using common metrics for extractive and generative question answering (QA) (Rajpurkar et al., 2016; Chen et al., 2019; Sai et al., 2023), such as exact match, F1, and BERTScore (Zhang et al., 2020).

We first apply commonly used LLMs finetuned on QA datasets, such as BERT (Devlin et al., 2019), RoBERTa, and DeBERTa (He et al., 2021). We also include the scientific NER models, PURE (Zhong and Chen, 2021) and SciREX (Jain et al., 2020). Finally, we include the general-purpose model, InstructGPT (Ouyang et al., 2022). Model implementation details are in Appendix D.3.

Model	Task			Method		
	BS	F1	EM	BS	F1	EM
BERT	82.29	35.08	15.94	75.04	23.08	4.38
RoBERTa	81.39	33.43	14.74	76.47	27.30	9.56
DeBERTa	80.36	31.28	11.16	75.62	26.05	4.78
PURE	71.30	34.67	5.58	69.65	30.43	5.98
SciREX	84.30	**49.03**	14.74	72.14	**33.70**	6.37
InstructGPT	**84.65**	43.94	**19.12**	**77.95**	32.05	**11.55**

Table 6: Model performance on Task 3 by BERTScore F1 (BS), F1, and Exact Match (EM). Due to space limitations, we omit the "text-davinci" when describing the InstructGPT.

Results. We show the model performance for Task 3 in Table 6. The best model by BERTScore is InstructGPT-002, reaching 84.65% for task extraction, and 77.95% for method extraction. It also achieves a higher exact match score than other strong models such as SciREX.

5 Insights

Using our NLP4SG *PaperAnalyzer* system, we analyze all the 76K NLP papers from ACL Anthology to answer our driving questions Q1, Q2, and Q3. Additionally, we also build a demo website at `nlp4sg.vercel.app` to make our analysis results more accessible to all parties interested in NLP4SG. Our website starts with a homepage in Figure 4, whose visualization plots we will refer to for the related analysis in this section. For a full overview of various demos on the website, see Appendix A.

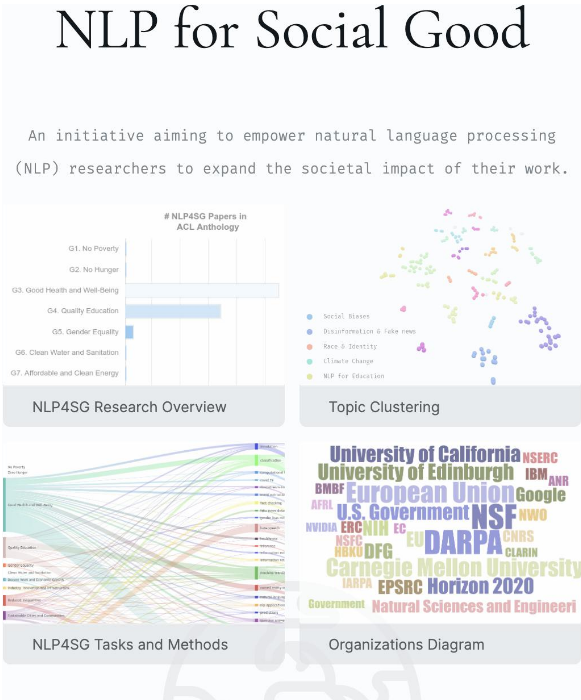


Figure 4: The homepage of our NLP4SG website.

Q1. How Many NLP4SG Papers Are There? We apply our NLP4SG binary classifier to the entire ACL Anthology with 76K papers. Overall, there are around 13.07% social good-related papers. For a more detailed overview, we plot the proportion of social good-related papers among all NLP papers every year from 1980 in Figure 5. We can see a clear rising trend of NLP4SG in recent years. Notably, the proportion of NLP4SG papers increased to 19.0% in 2020 from the initial percentage of 8.2% in 1980. It is important to note that while these percentages provide us with directional insight into the trend, their accuracy is contingent on the precision of the model used for analysis.

Q2. What Aspects of Social Good Do the Papers Address? Next, we are interested in the question of what social good aspects are addressed by NLP research.

In Figure 6, we show three bar charts side by side. The left one shows the importance scores of all the 17 SDGs according to Yang et al. (2020)'s worldwide survey of 360 sustainability researchers, where they incorporate all expert ratings into an overall priority score. The chart in

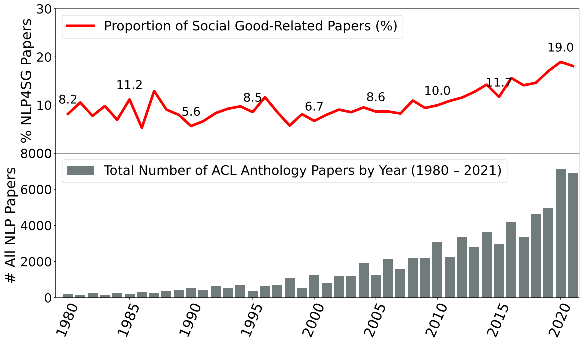


Figure 5: Percentage of social good-related papers (top) and the total number of papers (bottom) in ACL Anthology each year from 1980.

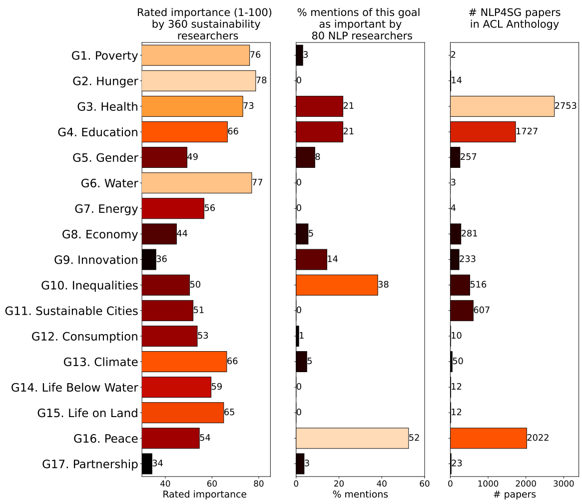


Figure 6: SDGs importance vs actual efforts. We use darker red for shorter bars, so the overlooked goals (especially on the right chart) can draw more attention.

the middle is compiled from our survey on "How current NLP researchers think about NLP4SG" at ACL 2021 of 80 NLP researchers, where we ask the subjects "What social good applications do you think the NLP community should consider?" and then map the answers (e.g., NLP for education, and combating misinformation) to corresponding SDGs. The survey details are in Appendix F.1. Finally, we apply our UN SDG goal classifier to all the ACL Anthology papers, and visualize the distribution of the 17 goals into the right panel of Figure 6. Our demo website[4] contains a version of this tri-view comparison in Appendix A.1, where the user can specify whether they want to look at the overall trend, or the trend in any specific year.

From the overall comparison in Figure 6, we can see a strong mismatch between the importance of the goals (left) and the actual size of research efforts (right). For example, NLP4SG papers mostly address health, peace, and education, which are admittedly important goals. However, they hardly address other important goals such as poverty and climate, which are pressing social issues. This might be because it may be more challenging to address these topics, e.g., lack of available datasets and expertise, trends dominating researchers' attention, funding climate (e.g., healthcare research being better funded), ease of publishing papers, the popularity of the field which may help gain citations more easily, the existence of good benchmark datasets, and so on. Note that we intentionally avoid getting into technological solutionism, and a more accurate framing should be trying to make the best use of the space where NLP can help (Appendix F.2), and the first two bar charts are all references to help gain better insights into this.

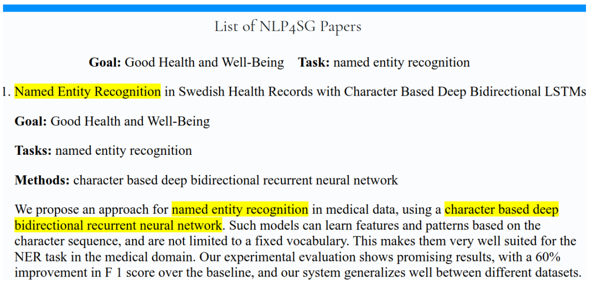
Q3. What Technical Solutions Are There? Diving deeper into our analysis, we study how existing research formulates tasks for each social good goal, and what methods they tend to adopt.

We visualize the distribution of tasks and methods in a Sankey diagram at the bottom of Figure 1, and add the connections to the SDGs. Briefly, the Sankey diagram takes as input the SDG-task-method information identified from each of the 9K NLP4SG papers from ACL Anthology. Hence, the diagram forms a comprehensive overview of what goals are addressed by what kind of task formulations, and then solved by what kind of models. Thus, researchers can quickly gain insights into NLP4SG and get an overview of the technical landscape in one place, and then spot opportunities where a better solution is yet to be proposed, e.g., new task proposals for an under-addressed goal such as poverty, or methodological upgrades for existing tasks so that the field can benefit from the latest advancement of NLP models.

The most common tasks in NLP4SG papers are: machine translation, which is present in 6.30% of the papers; classification in 4.06%; toxicity detection in 3.32%; named entity recognition in 2.19%; and sentiment analysis in 1.51%. The most common methods are: LLMs present in 12.18% of the NLP4SG papers, classifiers in 2.98%, deep neural networks in 2.51%, word embeddings in 1.80%, and SVMs in 1.62%.

[4]https://nlp4sg.vercel.app/sdg

On our Sankey diagram demo page,[5] we also enable clicks into each goal, task, or method to see the corresponding list of papers. We show the titles and abstracts of the papers and highlight text spans that contain the tasks and methods, as illustrated in Figure 7. This function is helpful to facilitate NLP4SG literature review, and saves time for future researchers.


List of NLP4SG Papers

Goal: Good Health and Well-Being **Task:** named entity recognition

1. Named Entity Recognition in Swedish Health Records with Character Based Deep Bidirectional LSTMs

Goal: Good Health and Well-Being

Tasks: named entity recognition

Methods: character based deep bidirectional recurrent neural network

We propose an approach for named entity recognition in medical data, using a character based deep bidirectional recurrent neural network. Such models can learn features and patterns based on the character sequence, and are not limited to a fixed vocabulary. This makes them very well suited for the NER task in the medical domain. Our experimental evaluation shows promising results, with a 60% improvement in F 1 score over the baseline, and our system generalizes well between different datasets.

Figure 7: An example paper in the paper list after clicking on “Health” and “Named Entity Recognition.”

6 Related Work

AI for Social Good. Recent years have seen a rising effort to apply AI technologies to various aspects of social good (e.g., Chen et al., 2020; Yu et al., 2018). To enhance the community which cares about AI for social good (Tomašev et al., 2020; Jin et al., 2021; Hager et al., 2019), it is crucial to gain insights about the progress overview. However, existing efforts mostly rely on time-consuming manual compilation, from surveying hundreds of papers (Cowls et al., 2021; Jin et al., 2021), conducting limited case studies (Floridi et al., 2020; Hager et al., 2019), to pairing existing papers on tasks and datasets with SDGs (Yeh et al., 2021). Some work uses keyword matching to extract social good-related papers (Shi et al., 2020; Fortuna et al., 2021), but our work is the first to build a comprehensive paper analysis system, and we are the only work designing a suit of various tasks, annotating a dataset, and using state-of-the-art NLP models to support such analysis.

NLP on Scientific Literature. is an active research area on building NLP models for analysing scientific text. Example research directions include dataset creation (Ammar et al., 2018; Lo et al., 2020; Wang et al., 2020); modeling for tasks such as information extraction (Jain et al., 2020; Luan et al., 2018; Zhong and Chen, 2021; Hope et al., 2021), summarization (Cachola et al., 2020), question answering (Dasigi et al., 2021), and general language modeling (Beltagy et al., 2019; Taylor et al., 2022). There are notable visualization platforms e.g. for COVID-19 research papers, such as search engines to browse papers (Bhatia et al., 2020), challenges and directions (Lahav et al., 2022), and associations between research groups and topics (Hope et al., 2020). Our work has a unique focus on identifying NLP4SG papers, aligning them with SDGs, and visualizing the research landscape, and it is the first visualization platform for NLP4SG based on NLP papers.

7 Conclusion

In this work, we make possible the reporting of the state of NLP4SG, with the help of several novel tasks specific to NLP4SG papers, the construction of a richly annotated NLP4SGPAPERS dataset, and the development of a *PaperAnalyzer* system based on state-of-the-art LLMs. We use this system to provide researchers insights into the landscape of NLP4SG research, paving the way for a more evidence-driven pursuit of NLP4SG that goes beyond good intentions. To make this information more accessible, we also create a website to visualize our analysis.

Limitations

One limitation of this work is that the models are not yet perfect, so the insights on the entire ACL Anthology are bottlenecked by the best model performance. To make up for this, we also provide in Appendix G.2 an additional analysis with only the gold, annotated data, which are accurate, but smaller and less representative of the entire landscape of NLP4SG. Since our system is modular, and we can plug in better models for each task, we welcome future work to further improve the current models. One example improvement direction is to model more inputs than the title and abstract, for instance by taking into consideration the full text and meta information such as the venue and publication year of the paper.

Our sampling technique is designed to ensure the representativeness of our sample within the larger population of academic papers in our dataset, in line with our primary objective to reflect the distribution of each SDG. We acknowledge that there may be variations in the distribution of papers over

[5] https://nlp4sg.vercel.app/sankey

the years, particularly in earlier years with limited number of papers. In future work, a more detailed analysis could be conducted to track the distribution of NLP4SG papers over time, providing a more nuanced understanding of how these papers have evolved and complementing our broader findings regarding SDG representation.

Another limitation is that the definition of social good might experience a gradual shift when society evolves its ethical standards. There is no fixed, definitive answer, so we welcome community discussions. This paper is a snapshot of existing thinking frameworks (Cowls et al., 2021; Jin et al., 2021) and our understanding by 2023. Updating the definition of social good is key to mitigating the risks of overlooking important societal problems. In the long run, we suggest community-wide discussions and iterations to update the definition of social good over time.

Finally, beyond the three tasks we propose for NLP4SG papers, there could be potentially other tasks, such as generating task formulations for under-addressed goals, suggesting new task-method combinations, and so on. For example, some NLP tasks that could address issues related to poverty could include sentiment analysis to understand public perceptions of poverty and related issues, information extraction from government documents to identify and track funding for poverty reduction programs, and machine translation to make information about poverty reduction programs and resources available to non-English speakers. Moreover, currently our work is limited to the knowledge within NLP, so it will be better if future work can collect data about the downstream real-world impact of each research trace, which can provide better feedback for the community.

Ethical Considerations

This work aims to provide insights about the state of NLP4SG research. It helps to increase the awareness of NLP researchers about applying NLP for positive social applications. In addition, the data we use is based on the open-sourced ACL Anthology dataset, so there are no user privacy concerns. To make this line of work more inclusive and comprehensive, we suggest a regular update of the social good identifier to keep up with the ethical standards of the community.

Author Contributions

The idea and mission of the paper were mainly devised by *Zhijing* and *Rada* as an important step in the NLP for Social Good initiative.

Fernando put tremendous efforts into running all the models, building the demo website, and various analyses. *Zhijing* guided the work, and supervised the entire implementation process. *Fernando* and *Zhijing* equally contributed to the dataset collection.

Mrinmaya supervised Fernando's Master thesis, from which this work extended. *Mrinmaya* and *Bernhard* gave high-level feedbacks of the project. *Tom* gave insightful suggestions about how to design the paper and the dataset, as well as the common NLP models to use for scientific text. *Zhijing*, *Fernando*, *Mrinmaya* and *Rada* led the paper writing, and all authors contributed to improving the writing.

Acknowledgments

We thank Niklas Stoehr for his feedbacks on our earlier work (Jin et al., 2021) and his suggestion that we could further automate that framework. We appreciate Dan Lahav for his insightful discussions to improve the orientations of this paper and its role in NLP4SG initative. We acknowledge Jad Beydoun for helping with the dataset collection process, and Flavio Schneider for deploying the basic structure of the website based on which we built the visualizations. We thank Ariel Holtzman for the brilliant brainstorming of the paper title, Laura Biester for insightful suggestions when proofreading the paper, and LIT labmates Ashkan Kazemi and Artem Abzaliev for suggestions to improve the messages that this paper convey.

This material is based in part upon works supported by the Swiss National Science Foundation (Project No. 197155); a Responsible AI grant by the Haslerstiftung; an ETH Grant (ETH-19 21-1); the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; and by the John Templeton Foundation (grant #61156). Zhijing Jin is supported by PhD fellowships from the Future of Life Institute and Open Philanthropy.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics. 9

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, 9, 20

Parminder Bhatia, Kristjan Arumae, Nima Pourdamghani, Suyog Deshpande, Ben Snively, Mona Mona, Colby Wise, George Price, Shyam Ramaswamy, and Taha A. Kass-Hout. 2020. AWS cord19-search: A scientific literature search engine for COVID-19. *CoRR*, abs/2007.09186. 9

Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, and Jieyu Zhao, editors. 2022. *Proceedings of the 2nd Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online. 1

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, Mateusz 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. 1

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics. 9

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg. 15

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics. 7

Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278. 9

Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. A definition, benchmark and database of AI for social good initiatives. *Nat. Mach. Intell.*, 3(2):111–115. 9, 10, 15, 16

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics. 9

Rafael Ferreira Leite de Mello, Maverick Andre Dionisio Ferreira, Anderson Pinheiro Cavalcanti, Evandro Costa, and Cristóbal Romero. 2019. Text mining in education. *WIREs Data Mining Knowl. Discov.*, 9(6). 1

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. 1, 2, 5, 7

Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett, editors. 2021. *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online. 1

Luciano Floridi, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. 2020. How to design AI for social good: Seven essential factors. *Sci. Eng. Ethics*, 26(3):1771–1796. 9

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. 9

Gregory D. Hager, Ann W. Drobnis, Fei Fang, Rayid Ghani, Amy Greenwald, Terah Lyons, David C. Parkes, Jason Schultz, Suchi Saria, Stephen F. Smith, and Milind Tambe. 2019. Artificial intelligence for social good. *CoRR*, abs/1901.05406. 9

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced

Bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. 6, 7

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2021. Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4489–4503, Online. Association for Computational Linguistics. 9

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. 9

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. 5, 7, 9, 21

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. 1, 9, 10, 15, 16

Zhijing Jin and Rada Mihalcea. 2023. Natural language processing for policymaking. In Eleonora Bertoni, Matteo Fontana, Lorenzo Gabrielli, Serena Signorelli, and Michele Vespe, editors, *Handbook of Computational Social Science for Policy*, chapter 7, pages 141–162. Springer Cham. 1

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. 2022. A search engine for discovery of scientific challenges and directions. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11982–11990. AAAI Press. 9

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. 6

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. 2, 5

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics. 9

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. 5, 9, 21

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 2013, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc. 24

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155. 1, 5, 6, 7

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. 24

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv*. 14

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI. 1, 2

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*. 21

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and

Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*. 7

Paulo E. Rauber, Alexandre X. Falc
textasciitilde ao, and Alexandru C. Telea. 2016. Visualizing time-dependent data using dynamic t-sne. In *18th Eurographics Conference on Visualization, EuroVis 2016 - Short Papers, Groningen, The Netherlands, June 6-10, 2016*, pages 73–77. Eurographics Association. 15

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. 15

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics. 6

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):26:1–26:39. 7

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108. 6, 20

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *CoRR*, abs/2001.01818. 9, 19

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085. 9

Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. 2020. AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, 11(1):1–6. 9

United Nations. 2021. *The Sustainable Development Goals Report 2021*. United Nations. 2

United Nations. 2022. *The Sustainable Development Goals Report 2022*. United Nations. 2

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706. 9

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. 2020. Deep learning in clinical natural language processing: A methodical review. *J. Am. Medical Informatics Assoc.*, 27(3):457–470. 1

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. 7, 14

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Janel Lee, Marshall Burke, David B. Lobell, and Stefano Ermon. 2021. SustainBench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*. 9

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. 21

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731. 9

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. 4, 7

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, 9, 21

A Overview of the Demo Website

We create a demo website at https://nlp4sg.vercel.app/. Our website starts with a homepage in Figure 4, which provides access to all the analysis visualization plots made in Javascript. We introduce each visualization function in detail below.

A.1 Visualization 1: SDG Alignment

We first provide a visualization in Figure 8 of the SDGs (Section 5) from three aspects: the importance scores of SDGs collected in a survey on sustainability researchers (Yang et al., 2020), NLP researchers' opinions, and the number of ACL Anthology papers addressing the goals. In addition to the static plot in Figure 6, we enable a scroll bar on the top to select the view of papers only in a certain year at https://nlp4sg.vercel.app/sdg.

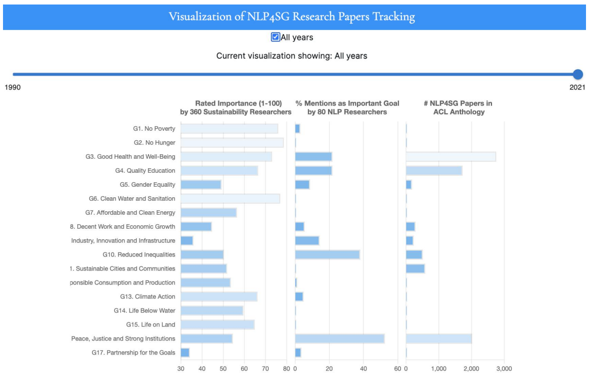


Figure 8: Visualization of the importance of/efforts on the SDGs by the sustainability researchers, NLP researchers, and actual ACL Anthology papers.

A.2 Visualization 2: Sankey Diagram of the Research Landscape

The Sankey diagram of the research landscape is shown in the bottom of Figure 1 or it can be directed accessed at https://nlp4sg.vercel.app/sankey. We include in Appendix G.1 the details of the scientific term normalization to clean the data for the Sankey diagram.

A.3 Visualization 3: Paper Browser

Corresponding to the Sankey diagram, we also enable clicks into each goal, task, or method to see the corresponding list of papers, an example of which is shown in Figure 7 (https://nlp4sg.vercel.app/sankey). We show the titles and abstracts of the papers and highlight text spans that contain the tasks and methods. This function is helpful to facilitate NLP4SG literature review, and saves time for future researchers.

A.4 Visualization 4: Supporting Organizations

To understand the support structure behind the pursuit of the SDGs, we collect the organizations supporting each research paper. We ran named entity recognition models by the Stanza Python package (Qi et al., 2020) over the Acknowledgment section. In case no organizations can be identified, then we use hand-crafted rules such as taking the main organization name of the authors if it is not a research institute, as these organizations could be self-supported. We visualize the distribution of organizations in the forms of a word cloud in Figure 9 and a Sankey diagram corresponding to each SDG in Figure 10.

Figure 9: Word cloud of the organizations mentioned in the NLP4SG papers.

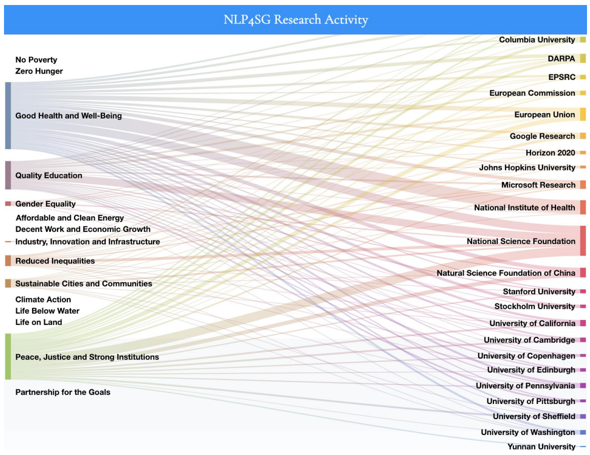


Figure 10: Sankey diagram of the organizations supporting the NLP4SG research work on each SDG.

A.5 Visualization 5: Paper Cluster Visualization

We also make a bottom-up plot of all the papers at a glance at https://nlp4sg.vercel.

app/papers. Specifically, we use the text embeddings of the papers by sentence transformer (Reimers and Gurevych, 2019),[6] project them using t-SNE (Rauber et al., 2016), and cluster by the HDBSCAN clustering algorithm (Campello et al., 2013). A screenshot of the paper clustering visualization page is in Figure 11.

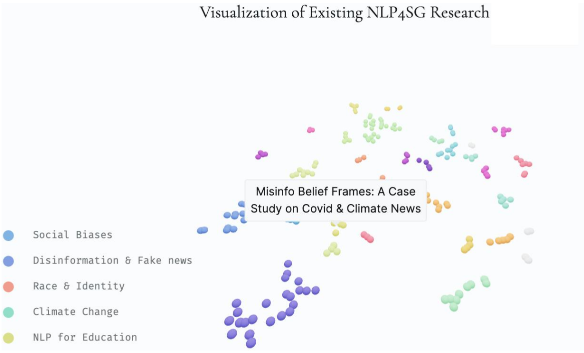


Figure 11: Clustering of the NLP4SG papers.

B Data Collection Details

B.1 Data Source

We identify ACL Anthology as a quality source of data, since it contains a large set of peer-reviewed NLP papers. There are other potential sources of peer-reviewed NLP papers, such as papers submitted to the NLP track at other venues such as NeurIPS and ICLR; as well as non-peer-reviewed papers on arXiv under the category cs.CL. However, for the former, it is hard to algorithmically identify NLP papers with high precision and recall, not to mention the different copyright regulations; for the latter, it is hard to control the paper quality if we open to non-peer-reviewed data. After balancing all the concerns and also thinking of our target audience, the NLP community, we think ACL Anthology, which is peer-reviewed and maintained by the NLP community, is a good data source for our work.

B.2 Preprocessing

We conduct the following preprocessing on the ACL Anthology data. To identify the paper entries in the entire anthology, we remove those that are workshop proceedings, invited talks, and book reviews, by a set of hand-crafted rules, such as identifying the title starting with "Book Reviews: ..." or ending with "... Workshop." To clean the titles, we remove the special symbols "{" and "}" in titles such as "{N}o{P}ropaganda at {S}em{E}val-2020 Task 11: A Borrowed Approach to Sequence Tagging and Text Classification."

Moreover, not all entries on the anthology provide a non-empty abstract, so for papers with missing abstracts, we parse their PDFs using the PDF-to-JSON converter doc2json Python package[7] to extract the abstracts.

B.3 Annotation Scheme

We introduce the annotation scheme for the three tasks below.

B.3.1 Annotating Task 1: NLP4SG Classification

Inspired by previous frameworks defining NLP4SG (Jin et al., 2021) and AI for social good (Cowls et al., 2021), we design a decision flowchart to identify whether an NLP paper addresses a social good-related problem in Figure 12.

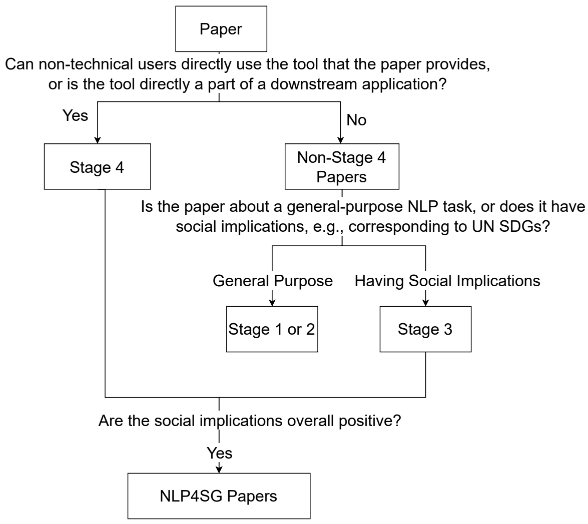


Figure 12: A decision flowchart used for the annotation of Task 1, NLP4SG binary classification.

Specifically, there are two important decision criteria. One is about the relevance to social good topics, e.g., the 17 UN SDGs. We also include social problems in the digital era, such as online toxicity, fake news, and privacy, into their corresponding SDG, i.e., (Goal 16) Peace, justice and strong institutions. The second criterion is about how directly the paper addresses the social problem. For

[6]We use the all-mpnet-base-v2 model of huggingface at https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[7]https://github.com/allenai/s2orc-doc2json

example, general linguistic tools such as syntax parsing could be an important technology, but not carrying direct social implications. We adopt the categorization of the stage of a technology in Jin et al. (2021), and only count those with a clear implication for downstream applications (i.e., Stage 3 and 4).

To get the idea through to the annotators, we provide both the formal guideline above, and also concrete examples as follows.

What is included: (Example set 1) Directly related to the high-level definition of SDGs, mentioning, e.g., healthcare; mental health (psychocounseling, hope speech); education; facilitating efficient scientific research (which belongs to Goal 9 Industry, innovation and infrastructure); helping employment (job matching, training job skills); helping collaboration among decision-makers. (Example set 2) Related to the fine-grained subcategories of SDGs, e.g., encouraging civic engagement, and enabling social problem tracking for the goal of (Goal 16) Peace, justice and strong institutions. (Example set 3) Social problems in the digital era: e.g., online toxicity, misinformation, privacy protection, and deception detection.

What is excluded: (Example set 1) General-purpose, coarse-grained NLP tasks: machine translation, language modeling, summarization, sentiment analysis, etc. (Example set 2) General-purpose, fine-grained NLP tasks: news classification; humor detection; technologies to increasing productivity, e.g., email classification, report generation, meeting note compilation (because they are application-agnostic which could be used for both good and bad purposes, and also a bit too general); textbook-related QA but using it as a benchmark to improve the general modeling capabilities; tasks whose data is socially relevant, but the task is neutral (e.g., POS tagging for parliament speech); NLP to help other neutral disciplines, e.g., chemistry; tasks a bit too indirectly related to SDGs, e.g., parsing historical language document, or cultural heritage-related tasks, low-resource MT, which bridges resources from one community to another, but is a bit too indirect, and also depends case by case on the actual language community, plus there is a tradeoff between efficiency and equality. (Example set 3) Tasks with controversial nature or unknown effect (varying a lot by how people use them in the future): e.g., news comment generation; financialNLP, which could be used in either way to help the economy, or perturb the market for private profits; simulated NLP tools for the battlefield; user-level demographic prediction.

As mentioned in the Limitations and Ethical Implications sections, the definition of social good might experience a gradual shift when society evolves its ethical standards. There is no fixed, definitive answer, so we start with relatively inclusive criteria here, and welcome community discussions and iterations to improve the definitions. This paper is a snapshot of existing thinking frameworks (Cowls et al., 2021; Jin et al., 2021) and our understanding by 2023. In the long run, we suggest community-wide discussions and iterations to update the definition of social good from time to time.

B.3.2 Annotating Task 2: SDG Classification

We let the annotators read the official *descriptions* and *indicators* of the 17 SDGs on the UN website such as `https://sdgs.un.org/goals/goal1`. For each paper identified as NLP4SG in Task 1, we ask the annotators to label all relevant SDGs.

We present examples of NLP4SG paper titles corresponding to each SDG below.

G2. Hunger

- A Gold Standard for CLIR evaluation in the Organic Agriculture Domain
- CRITTER: a translation system for agricultural market reports

G3. Health

- A Treebank for the Healthcare Domain
- Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters

G4. Education

- An MT learning environment for computational linguistics students
- Salinlahi III: An Intelligent Tutoring System for Filipino Heritage Language Learners

G5. Gender

- An Annotated Corpus for Sexism Detection in French Tweets
- Mitigating Gender Bias in Machine Translation with Target Gender Annotations

G6. Water

- A conceptual ontology in the water domain of knowledge to bridge the lexical semantics of stratified discursive strata

G8. Economy

- Multilingual Generation and Summarization of Job Adverts: the TREE Project
- Situational Language Training for Hotel Receptionists

G9. Innovation

- An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols
- Retrieval of Research-level Mathematical Information Needs: A Test Collection and Technical Terminology Experiment

G10. Inequalities

- Analyzing Stereotypes in Generative Text Inference Tasks
- Recognition of Static Features in Sign Language Using Key-Points

G11. Sustainable Cities

- FloDusTA: Saudi Tweets Dataset for Flood, Dust Storm, and Traffic Accident Events
- Trouble on the Road: Finding Reasons for Commuter Stress from Tweets

G12. Consumption

- Multiple Teacher Distillation for Robust and Greener Models

G13. Climate

- CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims
- Tackling Climate Change with Machine Learning

G14. Life Below Water

- Marine Variable Linker: Exploring Relations between Changing Variables in Marine Science Literature
- Literature-based discovery for Oceanographic climate science

G16. Peace

- On Unifying Misinformation Detection
- Fully Connected Neural Network with Advance Preprocessor to Identify Aggression over Facebook and Twitter

Since Goal 16 is a big category including tasks such as toxicity detection, and fighting misinformation, we provide an additional label when a paper within Peace addresses more fine-grained tasks. We collect statistics for the main subcategories: 44 samples (23.28%) for toxicity detection, 44 samples (23.28%) for fighting misinformation, 11 samples (5.82%) for privacy protection, and 6 samples (3.17%) for deception detection.

G17. Partnership

- MEDAR: Collaboration between European and Mediterranean Arabic Partners to Support the Development of Language Technology for Arabic
- The Telling Tail: Signals of Success in Electronic Negotiation Texts

B.3.3 Annotating Additional Data for Task 2: Upsampling

To enrich the data for Task 2, we upsample the low-occurrence classes. We denote all the classes with less than 5% proportion in our dataset as the low-occurrence classes. We first automatically extract candidate papers that might correspond to the low-occurrence classes, and then manually annotate their SDG classes. To automatically identify a candidate pool of papers regarding the low-occurrence classes, we first run the best model of Task 1 on the entire ACL Anthology to extract 9K NLP4SG papers, and among them, we use our curated keywords corresponding to the SDGs to extract candidate papers.

In Table 7, we can see that by the automatic matching algorithm, we get 261 candidate papers in total, including 1 candidate paper for poverty, 25 for hunger, 67 for gender, ..., and 40 for partnership. And after manually annotating these 261 papers, we obtain 4 confirmed papers for hunger, 67 for gender, 1 for water, 44 for economy, ..., and 5 for partnership. We use this upsampled data as additional data for Task 2, resulting in 759 total NLP4SG papers for Task 2.

UN SDG	# Matched Papers	# Papers after Annotation
G1. Poverty	1	0
G2. Hunger	25	4
G3. Health	High Occurrence	0
G4. Education	High Occurrence	0
G5. Gender	67	67
G6. Water	3	1
G7. Energy	2	0
G8. Economy	67	44
G9. Innovation	High Occurrence	15
G10. Inequalities	High Occurrence	1
G11. Sustainable Cities	33	11
G12. Consumption	0	0
G13. Climate	21	19
G14. Life Below Water	2	2
G15. Life on Land	0	0
G16. Peace	High Occurrence	12
G17. Partnership	40	5
Total	261	167

Table 7: Number of upsampled papers by SDGs. We first identify the candidate papers (# Matched Papers), and then manually annotate them (# Papers after Annotation). In total, there are 261 candidate papers, from which 167 are identified as relevant for NLP4SG.

B.3.4 Annotating Task 3: Salient Scientific Term Analysis

We present the title and abstract of social good papers to the annotator and ask them to write down a list of the primary tasks and methods addressed by each paper. In case there are hyponymy and hypernymy relationships among the scientific terms, we take the more specific term that is still common enough to show up in the Sankey diagram. For example, we would use the term LSTMs, but not neural networks (too broad), or stacked BiLSTMs with co-attention (too specific to be frequent enough to show up on the Sankey diagram). We utilized this information to enable the annotators to choose the relevant task and method spans from the title and abstract, along with an extra text box as an option in case an exact span was not available.

Below are the list of some common tasks and methods.

List of common tasks: automatic speech recognition, COVID-19-related analysis, text classification, data collection, event extraction, fact checking, fake news detection, gender bias mitigation, information extraction, information retrieval, linguistic analysis, machine translation, NLP applications, named entity recognition, natural language generation, question answering, relation extraction, rumor detection, sentiment analysis, stance detection, text summarization, toxicity detection, etc.

List of common methods: BERT, classifiers, computational models, conditional random fields, convolutional neural networks, ensemble methods, LSTM, multi-task learning, RoBERTa, supervised learning, support vector machines, topic models, transfer learning, word embeddings, etc. In the absence of more specific methods, it is also acceptable to annotate more coarse-grained methods if they are the only ones mentioned in the paper: language models, language technology, machine learning methods, machine translation system, NLP models, neural networks, recurrent neural networks, transformers, etc.

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 **autoencoders** 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: predict healthcare-associated infections
Method: restricted Boltzmann machines, autoencoders

Table 8: Example annotation of Task 3.

B.4 Annotator Background

The data associated with each task was annotated by two NLP researchers who are proficient in English. There are both female and male researchers, as well as white and Asian demographic backgrounds.

B.5 Inter-Annotator Agreement

For Task 2, we computed the Cohen's Kappa score of the binary classification of each SDG as shown in Table 9, and calculate the weighted average as 88.67%. Moreover, the partial match score is 96.46% and the exact match is 85.35%.

C More Dataset Information

C.1 License

License or terms for use and/or distribution: The dataset is open-sourced with the Creative Commons Attribution Share Alike 4.0 license. We provide the ACL Anthology paper IDs, title, abstract, publication year, and acknowledgements section from ACL Anthology, with our annotations. Ma-

UN SDG	Cohen's kappa
G1. Poverty	–
G2. Hunger	–
G3. Health	92.48%
G4. Education	95.19%
G5. Gender	86.32%
G6. Water	–
G7. Energy	–
G8. Economy	73.49%
G9. Innovation	84.92%
G10. Inequalities	68.96%
G11. Sustainable Cities	33.15%
G12. Consumption	100.00%
G13. Climate	100.00%
G14. Life Below Water	–
G15. Life on Land	–
G16. Peace	89.63%
G17. Partnership	79.31%

Table 9: Cohen's kappa coefficients for each SDG.

terials prior to 2016 in the ACL Anthology are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License, and materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

C.2 NLP4SG through the years

While papers addressing social problems in NLP have been predominantly published in recent years, it's important to include older papers in our analysis to get a complete picture of the growth of NLP4SG. Some examples of NLPSG papers for every decade are:

- 1960s: Automated Processing of Medical English. https://aclanthology.org/C69-1101
- 1970s: Powerful ideas in computational linquistics - Implications for problem solving, and education. https://aclanthology.org/P79-1028.pdf
- 1980s: Temporal Inferences in Medical Texts. https://aclanthology.org/P85-1002
- 1990s: ANTHEM: advanced natural language interface for multilingual text generation in healthcare (LRE 62-007). https://aclanthology.org/1996.amta-1.27
- 2000s: Collaborative Annotation of Sign Language Data with Peer-to-Peer Technology. https://aclanthology.org/L04-1280/
- 2010s: Detecting Offensive Tweets in Hindi-English Code-Switched Language. https://aclanthology.org/W18-3504/
- 2020s: Gender Bias in Machine Translation. https://aclanthology.org/2021.tacl-1.51/

C.3 Task 2 Co-Occurrence Matrix

In Figure 13, we show the co-occurrence matrix with the absolute number of co-occurrences.

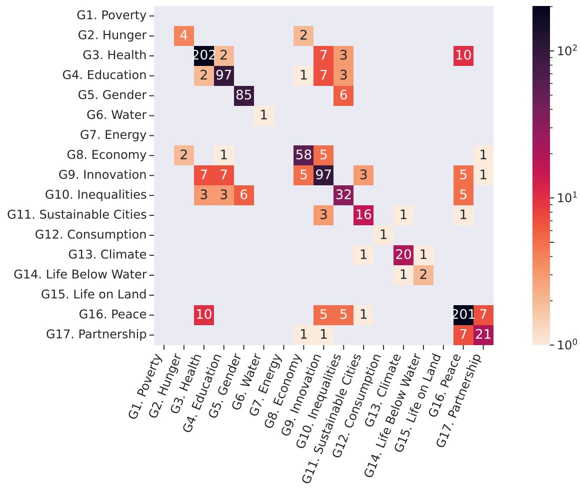


Figure 13: Heatmap of co-occurrences of UN SDGs on NLP4SG papers.

D Experimental Details

D.1 Models for Task 1

D.1.1 Distant Supervision

For the keyword matching, we curate a list of social good-related keywords provided in our GitHub repository which extends the AI for social good keywords by Shi et al. (2020). We compare the performance of keyword matching based on the title vs. title and abstract in Table 10, and find that keyword matching in titles gives higher performance, so we adopt the title keyword matching as the first DS method.

For the second DS method, cosine similarity matching, we calculate the cosine similarity scores 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 take the most similar UN SDG. We add a paper to the positive set if its text similarity is among the top 1% most similar papers in the unlabeled data from the ACL Anthology. We add a paper to the negative set if its similarity is in the bottom

50% percentile. For the text embedding similarity, we use a general-purpose sentence similarity model, all-distilroberta-v1[8] which is a pre-trained distilroberta-base (Sanh et al., 2019) model fine-tuned on a 1B sentence pairs dataset.

Merging both DS sets, the entire DS set consists of positive samples from papers whose titles match the keywords, and papers which are very similar to the UN SDG descriptions; and negative samples from papers that are very different from the UN SDG descriptions.

We report the quality of data extracted by these DS methods evaluated on the training set in Table 10, which correlates with the ablation study results in Table 11. We can see that the performance of the entire DS set is higher than that of the keyword-only set, which is higher than the similarity-only set.

	F1	P	R
Keyword Match (Title)	66.67	**79.90**	57.19
Keyword Match (Title+Abstract)	55.17	42.50	**78.60**
Cosine Similarity	8.33	47.06	4.57
Keyword Match (Title) + Cosine Similarity	**72.02**	76.85	67.76

Table 10: Quality evaluation of the distant supervision heuristics reported on the training set.

	F1	Acc	P	R
SciBERT with No DS	73.42	94.10	84.90	64.68
SciBERT + DS by Similarity	62.96	89.35	55.86	72.11
SciBERT + DS by Keyword	74.68	94.10	80.93	69.32

Table 11: For the best-performing model, SciBERT+DS, we conduct ablation studies by removing the different subsets of our DS data. The resulting performance correlates with the DS data quality reported in Table 10, where keyword + similarity gets the best quality, and similarity only gets a very low quality.

D.1.2 Models Implementation Details

BERT, RoBERTa, and SciBERT We finetune SciBERT (Beltagy et al., 2019) on the NLP4SG text classification task. We freeze the embedding layer and the first 11 encoder layers, and we fine-tune the parameters of the 12th encoder layer, the pooler, and the classifier layer.

We use one GPU model NVIDIA Quadro RTX 6000 with 24 GiB. We train the classifier for 15 epochs using a batch size of 32. The training time for our best-performing model is approximately 1.5 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. All the runs together make a total of around 18 GPU hours. We tried with a similar setting using BERT-large and RoBERTa-large, obtaining lower performances.

InstructGPT For the second approach, we design a prompt and ask InstructGPT 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
- Final question: We use the sentence "Is this paper contributing to the UN Sustainable Development Goals? Answer yes or no."

We use the OpenAI API[9] to access GPT. We use "davinci-text-002." We keep the default values of the API, and set the temperature to zero and maximum number of tokens of the response as 50.

Few-shot InstructGPT 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.

> 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:
> `[InstructGPT response]`

[8] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[9] https://beta.openai.com/overview

D.2 Models for Task 2

Model 1: InstructGPT We design a prompt to get which UN goal a paper contributes to. Since InstructGPT 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 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. We also tested distilbert (typeform/distilbert-base-uncased-mnli) and DeBERTa (MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli).

D.3 Models for Task 3

PURE and SciREX PURE (Zhong and Chen, 2021) is a model for end-to-end relation extraction. We use PURE trained on the SciERC (Luan et al., 2018) dataset of scientific abstracts.

SciREX (Jain et al., 2020) is a document level IE model that not only identifies mentions, but also models their saliency, and their coreference links.

For SciREX and PURE, we run them on a GPU model NVIDIA Quadro RTX 6000 with 24 GiB. It takes around one hour to extract the information from 9K social good papers.

BERT-based Models We used LLM fine-tunned on the SQUAD dataset (Rajpurkar et al., 2018). We provide as context the title and abstract of the paper and asked "Which NLP task does this paper address?" and "Which NLP method(s) does this paper use?". In particular we used the following models "bert-large-uncased-whole-word-masking-finetuned-squad" , "deepset/roberta-large-squad2", and "deepset/deberta-v3-large-squad2".

InstructGPT We use InstructGPT 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 InstructGPT 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.

Prompt to obtain the tasks

> Identify the NLP task(s) that this paper is dealing with. Select a text span that is an appropriate answer, or if no span serves as a good answer, just come up with a phrase. Examples of tasks are: fake news detection, name entity recognition, question answering, etc.
> `[Title + Abstract]`
> The primary NLP task(s) addressed in this paper are:
> `[InstructGPT response]`

Prompt to obtain the methods

> Identify the NLP method(s) used in this paper. Select a text span that is an appropriate answer, or if no span serves as a good answer, just come up with a phrase. Separate the methods with commas and don't include NLP tasks. Examples of methods are: BERT, SVM, CNN, etc.
> `[Title + Abstract]`
> The primary NLP method(s) used in this paper are:

```
[InstructGPT response]
```

E Details of the Interpretability Analysis

E.1 Task 1: Example LIME Outputs

Using the best model, we show examples of LIME output for true positive (TP), true negative (TN), false positive (FP), and false negative (FN) data samples in Figures 14 to 17, respectively.

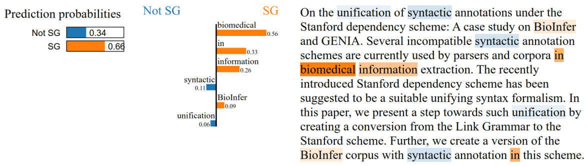


Figure 14: LIME output for a TP example.

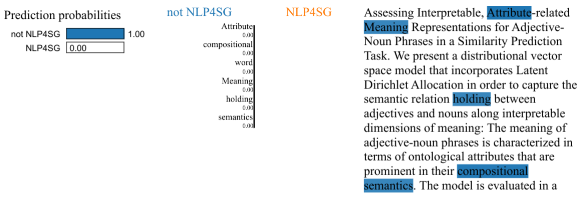


Figure 15: LIME output for a TN example.

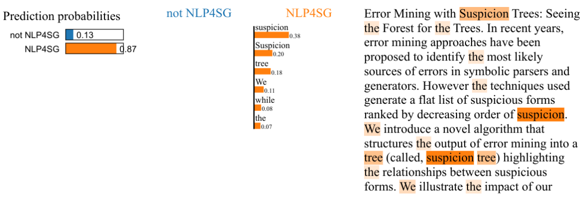


Figure 16: LIME output for an FP example.

E.2 Task 2: Analyzing the Reasoning Chain of InstructGPT

For the correctly classified samples by the best InstructGPT method, we perform further analysis to check the *explanation quality* of the model.

We show two motivating examples in Table 12, where the first example shows sufficient intermediate reasoning steps starting from what the paper contributes, to intermediate effects of the paper, and finally to the end social impact, while the second example lacks intermediate reasoning steps, and mentions only what the paper does, followed the UN SDG right after, without intermediate explanations.

Specifically, we denote a good "impact chain" as a complete and cohesive reasoning chain from the direct description of the paper's research output

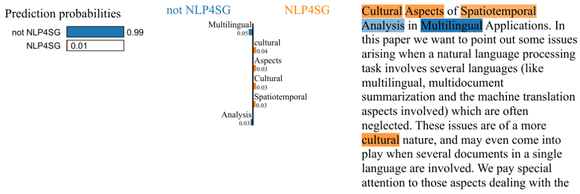


Figure 17: LIME output for an FN example.

to intermediate impact, and end impact. Then we randomly select 200 test samples, and use the following two criteria to check the explanation quality of InstrutGPT: (1) whether the mentioned "impact chain(s)" sound cohesive and reasonable, e.g., not missing necessary intermediate reasoning steps, which is analogous to the precision score, and (2) whether there are other obvious impact chains that are missing, which is analogous to the recall score. For each criterion, we assign 1 for good performance, 0 for bad performance, and 0.5 for partially good performance.

We find that 77.6% of the explanations contain cohesive and convincing reasoning chains, and 18% of the explanations miss at least one obvious impact chain of how the paper contributes to UN SDGs.

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.
> `[InstructGPT response]`

F More Insights

F.1 Survey

We conduct a survey from July 31, 2021 to August 31, 2021 about "How current NLP researchers think about NLP4SG." We receive 80 responses from NLP researchers, among which more than 30% are NLP professors. We reach the subjects at the ACL conference and also through emails. The demographics of the survey subjects are shown in Figure 18.

An Example with Sufficient Reasoning Steps for the Impact Chain

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. [...] We propose two novel deep learning approaches to infer the symptom names and their status: [...] This task stems from a realistic application of assisting medical providers in capturing symptoms mentioned by patients from their clinical conversations. [...]

InstructGPT 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*.

An Example Lacking Intermediate Reasoning Steps for the Impact Chain

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. [...]

InstructGPT 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 12: Examples of sufficient and insufficient reasoning about the impact chain of NLP4SG papers.

The question in the survey that contributes to the SDG visualization is "What social good applications do you think the NLP community should consider?", for which we manually parse the open-ended answers to corresponding SDGs. Example answers are NLP for Education (22.9%), physical and mental health (18.6%), and combating misinformation (17.1%).

F.2 Ideal SDG alignment we want to push for

We show an ideal framework in Figure 19, where for each goal, NLP can help with a subset of the things. And the meaningful action space is the gap between the current NLP research landscape vs. what NLP can potentially do.

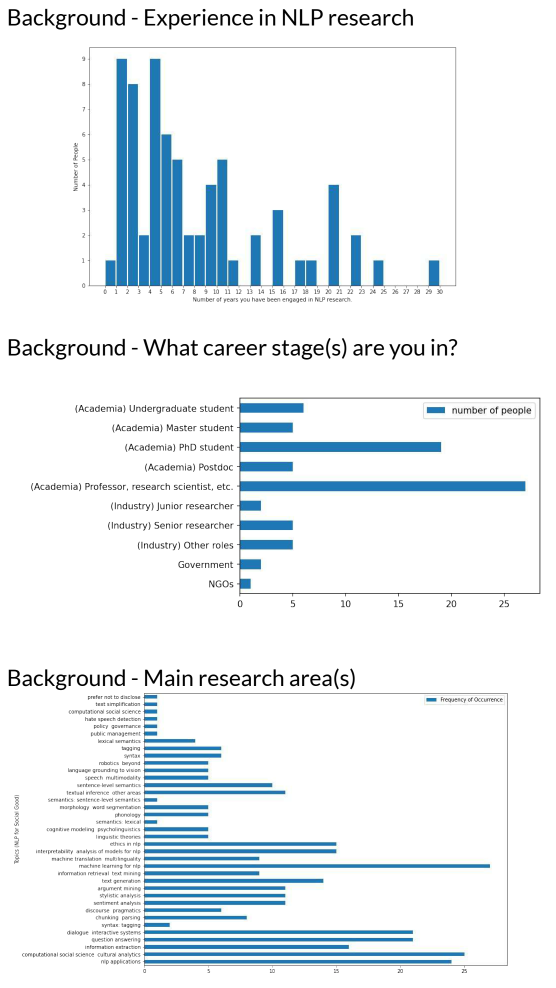


Figure 18: Backgrounds of the survey subjects, including the number of years in NLP research, career stage, and research areas.

G More Website Details

G.1 Implementation Details

Although we run the *PaperAnalyzer* on the entire ACL Anthology. For visualization, we mostly visualize 1980 to 2021, because there were not many papers before 1980, and also the progress tracking of NLP4SG in the recent decades might be more relevant for the audience. Also, since our data was crawled in the middle of 2022, the trend of 2022 is not comparable to the previous year 2021, so we omit 2022 for the moment, but it could be added in the future.

For the visualization of the Sankey diagram, we conduct some preprocessing over the salient scientific terms produced by the model. Since there are variances in the expression of the same task or method, such as "long short-term memory networks," and "LSTMs", we group similar terms to-

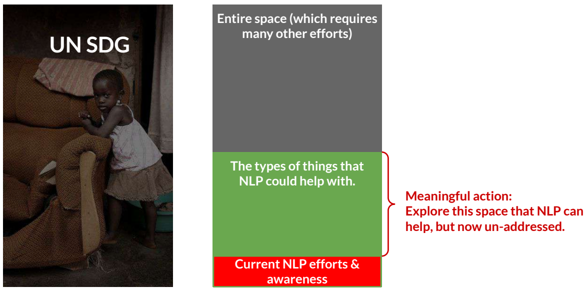


Figure 19: Illustration of the action space NLP researchers can aim for.

gether with the following procedure:

- We get a vector representation for each word with Word2Vec (Mikolov et al., 2013). We train Word2Vec with the abstracts 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.
- We measure the ratio score similarity between each word of a group to its centroid and just keep words in the cluster that have a similarity score greater than 75. We used the library "TheFuzz"[10] which uses Levenshtein distance to compute its similarity measure.
- 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.

G.2 Demo on the Gold Data

As mentioned in the Limitations section, one constraint is that the visualizations are bottle-necked by the model performance. As a supplementary, we also provide the visualization based on our annotated NLP4SGPapers dataset, which is smaller but accurate.

We produce all the visualization figures mentioned in Section 5, but based on the ground-truth annotation in our NLP4SGPapers dataset. Figure 20 visualizes the proportion of NLP4SG papers over the years, where the proportion of NLP4SG papers steadily increased from 2000.

We also show the Sankey diagram, using the gold data only, in Figure 21 and the SDG visualization in Figure 22.

[10]https://github.com/seatgeek/thefuzz

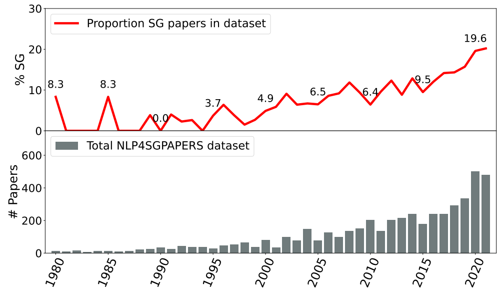


Figure 20: Percentage of social good-related papers (top) and the total number of papers (bottom) in NLP4SGPapers each year from 1980.

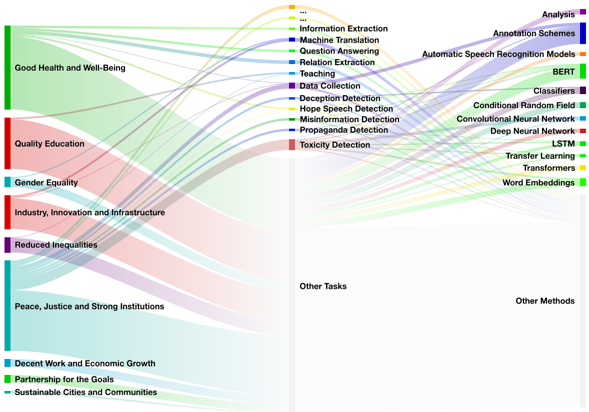


Figure 21: Sankey diagram based on gold annotations at https://nlp4sg.vercel.app/sankey_gold.

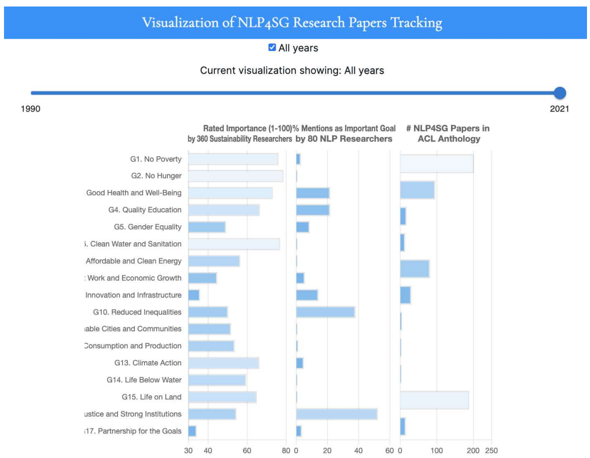


Figure 22: SDG visualization based on gold annotations at https://nlp4sg.vercel.app/sdg_gold.