Good Intentions Beyond ACL: Who Does NLP for Social Good, and Where?

Grace LeFevre[1], Qingcheng Zeng[1], Adam Leif[2], Jason Jewell[1], Denis Peskoff[1], Rob Voigt[3]
[1]Northwestern University, [2]University of California, Los Angeles,
[3]University of California, Davis
gracelefevre@u.northwestern.edu, robvoigt@ucdavis.edu

Abstract

The social impact of Natural Language Processing (NLP) is increasingly important, with a rising community focus on initiatives related to NLP for Social Good (NLP4SG). Indeed, in recent years, almost 20% of all papers in the ACL Anthology address topics related to social good as defined by the UN Sustainable Development Goals (Adauto et al., 2023). In this study, we take an author- and venue-level perspective to map the landscape of NLP4SG, quantifying the proportion of work addressing social good concerns both within and beyond the ACL community, by both core ACL contributors and non-ACL authors. With this approach we discover two surprising facts about the landscape of NLP4SG. First, ACL authors are dramatically more likely to do work addressing social good concerns when publishing in venues outside of ACL. Second, the vast majority of publications using NLP techniques to address concerns of social good are done by non-ACL authors in venues outside of ACL. We discuss the implications of these findings on agenda-setting considerations for the ACL community related to NLP4SG.

1 ACL and Social Good Research

As natural language processing rises in prominence throughout society, "NLP for Social Good" (Jin et al., 2021, NLP4SG) is an increasingly important topic of conversation in the NLP community: how can NLP methods be used to address questions of social concern and applied for positive social impact? Existing research on the ACL Anthology shows that while a substantial minority of papers have addressed social good since the 1980s, the proportion has increased over time. Under the definition of papers that address questions relevant to one of the 17 United Nations Sustainable Development Goals (SDGs),[1] Adauto et al. (2023) find that the proportion of papers in the ACL community addressing social good concerns has approached 20% in recent years.

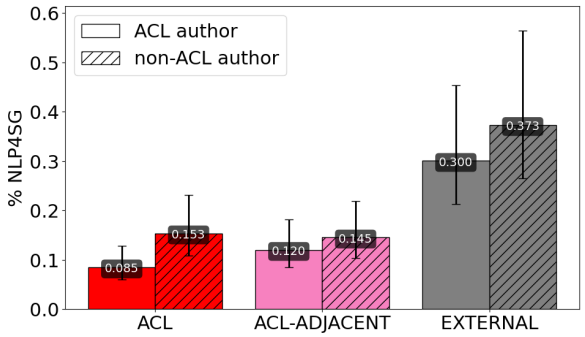

Figure 1: A higher ratio of NLP papers outside of the ACL Anthology (EXTERNAL) are characterized as social good than in ACL and ACL-ADJACENT venues. Moreover, NLP papers by non-ACL authors are more likely to focus on social good questions than those by ACL authors across all venue types.

However, many venues beyond ACL publish NLP4SG work. The proliferation of NLP techniques in other fields has led to substantial growth of papers that both use NLP methods and tackle social good questions, across dozens of conferences outside of the ACL Anthology (Movva et al., 2024). In this environment, critical questions of agenda-setting arise for the ACL community:

> (1) When ACL authors are doing work oriented towards social good, do they send that work to ACL venues?
>
> (2) Is ACL the place where most NLP4SG work is taking place?

Clear answers to these questions can help orient the community towards next steps moving forward. The first addresses whether ACL authors themselves consider our community to be a welcoming and appropriate venue for NLP4SG work; the second, the centrality of our community and the broader reach

[1]Gosselink et al. (2024) and Karamolegkou et al. (2025) also map NLP4SG work onto the UN SDGs.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5139–5151

Figure 2: Schematic of the metadata augmentation used in conducting our analyses. Papers are labeled for relevance to social good as defined by UN SDGs, author association with ACL, venue type, and neural vs. non-neural methods.

of our methods towards often inherently interdisciplinary social good goals.

In this work, we aim to gain a deeper understanding of the broader landscape of NLP4SG by adopting a scientometric approach that examines authors and publication venues (Ni et al., 2013).

We first augment an existing corpus of scientific papers with annotations for whether the work uses NLP techniques, whether the authors have published substantially within ACL, and the venue type (ACL, ACL-ADJACENT, or EXTERNAL; Section 2).

Analyzing these data, we find that papers by ACL authors in venues outside of the ACL Anthology are more than *three times as likely* to address social good topics as those inside it, and that the substantial majority of NLP4SG work takes place beyond ACL by non-ACL authors.

Furthermore, we examine and find substantial topical differences in content across venue types, differences in the probability that work using NLP techniques addresses social good when segmenting venues by discipline, and venue- and author-based differences in the proportion of NLP4SG work relying on neural methods (Section 3). We conclude by stressing the importance of defining social good and suggesting actions for closing the social good gap between ACL and other venues (Section 4).

2 Augmenting a Corpus with Metadata

This type of analysis requires a collection of relevant and representative academic papers. Adauto et al. (2023), the work which we are most directly building upon, collect a dataset of 76,229 papers from ACL Anthology and analyze a sample of 5,000 papers for social good, as defined by the United Nations Sustainable Development Goals (SDGs). We aim to expand beyond ACL and use Semantic Scholar's Open Research Corpus (S2ORC) as a starting point (Lo et al., 2020).

The metadata of the papers is paramount for our analysis. While S2ORC maintains metadata in addition to full paper text, it is sourced from the paper text and due to nonspecific PDF text extraction, some is missing or incorrect. We expand our data with the Semantic Scholar Open Data Platform (Kinney et al., 2023) via the S2AG API, which stores additional metadata for each paper.[2] We further augment the corpus by linking each paper in S2ORC to a corresponding record in OpenAlex (Priem et al., 2022), which improves the distinction of authors with similar names and provides concepts for grouping our Semantic Scholar corpus.[3]

[2]https://api.semanticscholar.org/api-docs/datasets

[3]If a paper's MAG or DOI is available in the Semantic Scholar metadata, this is used to identify that same paper in OpenAlex; otherwise, papers are matched via their title and publication date/year (whichever is available).

For ACL papers, we also augment each paper's metadata with information from the ACL Anthology BibTeX. Finally, we use the resulting corpus to identify four key factors for each paper:

- NLP relevance
- venue type
- author classification
- social good relevance

The schematic in Figure 2 illustrates our metadata augmentation process for two representative papers. We release the resulting dataset, along with the code used to create it, to the community to encourage future work.[4]

2.1 NLP Relevance

OpenAlex provides a classification of relevant "concepts" for each paper in their database, which we leverage to categorize which papers involve NLP. We identify a set of core concepts associated with NLP by calculating the most frequently occurring OpenAlex concepts in ACL Anthology papers and removing ambiguous or multi-sense concepts (like "Computer Science"). Our full list of used concepts is in Table 1, including only clearly NLP-associated concepts like "Natural Language Processing," "Information Retrieval," and "Language Model."

Filtering our dataset to include only papers classified with at least one of these topics yielded our main dataset of 309,208 papers.

OpenAlex ID	Concept
C204321447	Natural Language Processing
C23123220	Information Retrieval
C203005215	Machine Translation
C119857082	Machine Learning
C186644900	Parsing
C28490314	Speech Recognition
C137293760	Language Model

Table 1: Concepts selected from OpenAlex.

2.2 Venue Type

We propose a distinction between three types of venues:

- ACL venues are those listed as "ACL Events" in the ACL Anthology.
- ACL-ADJACENT venues are those listed as "Non-ACL Events" in the Anthology.
- EXTERNAL venues belong to neither list and include a diverse range of journals and conference proceedings from other disciplines, such as general science, engineering, and interdisciplinary research. The top 10 most frequent external venues in our dataset are listed in Appendix A.

	VENUE		
AUTHOR	**ACL**	**ACL-ADJ.**	**EXTERNAL**
ACL	24,594	30,269	28,644
non-ACL	1,075	2,700	221,926

Table 2: Distribution of papers in our dataset. Unsurprisingly, ACL authors (those with three or more publications in an ACL venue) are responsible for the majority of ACL and ACL-ADJACENT papers. Far more NLP papers exist in EXTERNAL and are primarily written by non-ACL authors.

For EXTERNAL venues we further obtain coarse- and fine-grained disciplinary classifications, as well as venue-level h5-indices for a subset of EXTERNAL venues using the "top publications" metrics on Google Scholar.[5] Specifically, for each of their eight top-level disciplinary categories, we scrape the top 20 venues under every subcategory and align them to our existing venue names. To account for variation in formatting (abbreviations, punctuation, etc.), we normalize venue names and apply a token-based fuzzy matching algorithm[6] that compares similarity scores and retains only high-confidence matches. This process results in 3,281 venues in this subset. These correspond to 98,753 papers, or 32% of our original dataset, which are likely to be higher-impact and more reputable venues on average. We use this subset for detailed venue-level analysis in Section 3.3.

2.3 Author Classification

We propose an author-level distinction between "ACL Authors" and "non-ACL Authors." We define ACL authors as authors who, at any time, have published three or more papers in ACL venues, and define any given paper as being written by an ACL author if at least one author meets this criterion.

[4] https://github.com/asl7168/nlp4sg_beyond_acl

[5] https://scholar.google.com/citations?view_op=top_venues&hl=en

[6] We use the RapidFuzz library's token_sort_ratio scorer to align venue names based on string similarity, retaining only matches with a similarity score of 90 or higher.

This author-level distinction is important because it accounts for an author's experience within the ACL community in addition to considering whether a paper appears in an ACL venue.

By defining ACL authors based on their publication history, we aim to better differentiate between researchers with sustained engagement in the field and those with more limited exposure. Notably, our threshold of three ACL papers aligns with ACL's own requirement for becoming a reviewer, reflecting a level of experience that the community considers sufficient for evaluating research.

The distribution of papers in our dataset across each author and venue type is shown in Table 2.

2.4 Social Good Relevance

We apply the NLP4SG model from Adauto et al. (2023) to classify each paper as relevant or not relevant to social good. However, applying this model to papers from outside the ACL Anthology may have limited accuracy due to domain transfer issues. Therefore, we conduct a manual evaluation of the accuracy of this model on papers in non-ACL venues. Three authors with backgrounds in NLP annotated 200 randomly selected papers published by an ACL author at a non-ACL venue for their association with social good, as defined by association with UN SDGs. Annotation instructions are provided in Appendix B and results in Table 3.

Overall, the NLP4SG classifier aligns with the human judgment 77.5% of the time, achieving a precision of 71.0% and a recall of 66.2% for an overall F1 score of 68.5%. As expected for out-of-domain classification, this represents some loss of performance over Adauto et al. (2023), who reported an F1 score of 75.9% for papers in the ACL Anthology. However, we find the model is not dramatically unbalanced in its predictions. Due to lower recall, it is conservative at assigning NLP4SG labels to papers in EXTERNAL venues. Therefore, it is sufficient for the large-scale trends we track in this work and we use its labels at face value. To reflect the uncertainty introduced by the classifier's performance, we include error bars in Figure 1, estimating the lower bound as $value \times precision$ and the upper bound as $value \div recall$. These account for potential over- and under-estimation due to false positives and false negatives, respectively.

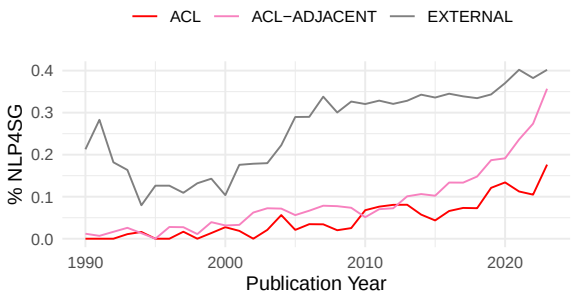

Figure 3: Proportion of NLP papers identified as NLP4SG by venue type and year. Applications of NLP techniques to social good questions are increasing as a share of all NLP papers across all venues.

	HUMAN	
CLASSIFIER	NLP4SG	Other
NLP4SG	49	20
Other	25	106

Table 3: Out-of-domain NLP4SG classification confusion matrix. Most non-NLP4SG papers are properly detected. Comparable amounts of social-good papers are spuriously detected and missed by the classifier compared to the human annotators.

3 Findings

Using this metadata-augmented corpus, we identify large-scale trends in the landscape of NLP4SG.

3.1 Distributional Differences by Author and Venue Type

As the NLP community seeks to interrogate its role in advancing social good concerns, a natural question is where different types of authors engaged with our community choose to publish social good work. Figure 1 visualizes broad distributional trends in NLP4SG publication patterns, showing the proportion of NLP papers classified as NLP4SG among work using NLP techniques by both ACL authors and non-ACL authors in ACL, ACL-ADJACENT, and EXTERNAL venues.

Our first set of analyses regards author behavior, with a particular focus on ACL authors, where we make two key observations. First, papers by ACL authors are more likely to address concerns of social good when they appear in ACL-ADJACENT venues rather than core ACL venues (one-sided two-proportion z-test, $z = 13.5035$, $p < 0.001$). Second, we find that papers by ACL authors are also significantly more likely to address social good when they appear in EXTERNAL venues compared to ACL

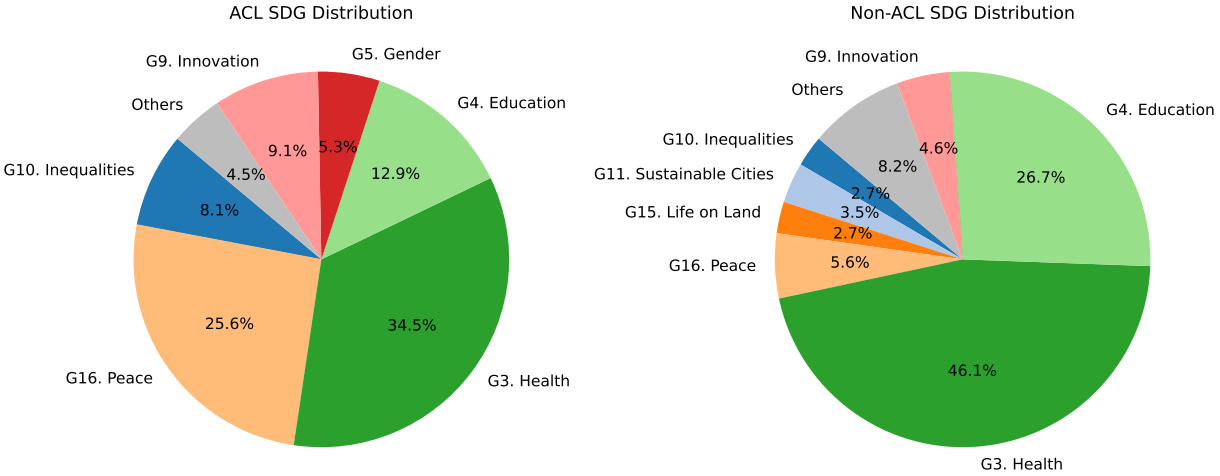

Figure 4: Proportions of SDG topics across venue types. The targets of social good research have different distributions between ACL-associated venues and other conferences. ACL venues have a greater focus on topics related to peace, innovation, and inequality while EXTERNAL venues have a greater focus on those related to health and education.

venues of any kind ($z = 71.4087$, $p < 0.001$); distributionally, we find that for ACL authors, the proportion of NLP4SG papers is roughly three times higher outside of ACL than within it.

Looking to the behavior of non-ACL authors, we find that NLP papers by non-ACL authors in EXTERNAL venues represent the category most likely to address social good. This is true both in terms of the proportion of these papers that tackle social good concerns, and dramatically so in the raw number of such papers that appear: the number of NLP4SG papers by non-ACL authors in EXTERNAL venues approaches an order of magnitude more than those by ACL authors.

Considering change over time (Figure 3), while we observe the same increase in NLP4SG within ACL identified by previous research (Adauto et al., 2023), but find that this trend is yet stronger in ACL-ADJACENT venues. In EXTERNAL venues the share of NLP4SG papers among NLP papers has always been relatively high but showed a dramatic growth in the mid-2000s and continues to increase in recent years.

3.2 Topical Differences Across Venue Types

A natural follow-up question is whether NLP4SG papers that appear in ACL venues differ in the types of social good concerns they address compared to those outside it. We use an LLM-based approach to estimate the category of social good for each paper classified as NLP4SG in our corpus.

As discussed in Adauto et al. (2023), InstructGPT (Ouyang et al., 2022) demonstrated superior performance in classifying papers according to the UN SDGs. These include categories such as "Health," "Peace," and "Education." Building on this, we employ GPT-4o (OpenAI and other authors, 2024) to classify two sets of NLP4SG papers—those within the ACL Anthology (including both ACL and ACL-ADJACENT venues) and those outside of it (EXTERNAL)—into one of the 17 SDGs. The prompt template used for classification and the full list of SDGs is provided in Appendix C. We conduct the classification using a greedy decoding strategy.

We note topical differences between social good detected inside and outside of the ACL Anthology, illustrated in Figure 4. The ACL dataset has a comparatively stronger focus on peace (25.6%), innovation (9.1%), and inequalities (8.1%) compared to the EXTERNAL dataset. The higher proportion of peace-related research in ACL may reflect the community's focus on hate speech detection and related areas that aim to foster peaceful interactions. By contrast, the EXTERNAL dataset has a predominant focus on health (46.5%) and education (25.1%), likely reflecting applications of NLP work to targeted practical domains. Overall, despite some similarities, these results suggest that NLP4SG work within ACL and NLP4SG work outside of ACL tend to prioritize different aspects of social good.

3.3 Differences by Venue Discipline

We leverage Google Scholar metadata and venue-level disciplinary classifications (described in Sec-

	Total NLP Papers	NLP4SG Papers	NLP4SG %
Social Sciences	6,288	3,470	55.2%
Top 10 NLP4SG venues: International Journal of Environmental Research and Public Health, English Language Teaching, SAGE Open, International Journal of Emerging Technologies in Learning (iJET), Cogent Education, Reading and Writing, Scientometrics, Information, Frontiers in Education, Journal of Environmental and Public Health			
Health & Medical Sciences	26,734	14,021	52.4%
Top 10 NLP4SG venues: PLOS ONE, Frontiers in Psychology, BMC Medical Informatics and Decision Making, International Journal of Environmental Research and Public Health, Frontiers in Genetics, Nucleic Acids Research, Diagnostics, Frontiers in Oncology, JMIR Medical Informatics, Frontiers in Neuroscience			
Life Sciences & Earth Sciences	26,325	11,680	44.4%
Top 10 NLP4SG venues: PLOS ONE, BMC Bioinformatics, Scientific Reports, Bioinformatics, Nucleic Acids Research, Database, PLOS Computational Biology, Sustainability, Heliyon, Frontiers in Plant Science			
Humanities, Literature & Arts	3,104	1,281	41.3%
Top 10 NLP4SG venues: English Language Teaching, Journal of Language and Linguistic Studies, Cogent Arts & Humanities, Languages, English Education, Studies in Second Language Learning and Teaching, Journal of Child Language, Language Policy, Applied Psycholinguistics, Journal of Psycholinguistic Research			
Business, Economics & Management	649	239	36.8%
Top 10 NLP4SG venues: Natural Hazards and Earth System Sciences, Natural Hazards, Ekonomska Istrazivanja-Economic Research, Cogent Business & Management, International Research Journal of Tamil, Cogent Economics & Finance, Journal of Artificial Societies and Social Simulation, International Journal of Genomics, Disaster Medicine and Public Health Preparedness, International Journal of Forecasting			
Chemical & Material Sciences	3,173	1,020	32.1%
Top 10 NLP4SG venues: Nucleic Acids Research, Journal of Computer Science, Molecules, Materials, Biomolecules, Molecular & Cellular Proteomics, Science Advances, Biosensors, RSC Advances, Chemical Science			
Engineering & Computer Science	41,619	13,168	31.6%
Top 10 NLP4SG venues: BMC Bioinformatics, Sensors, IEEE Access, BMC Medical Informatics and Decision Making, Bioinformatics, Applied Sciences, Remote Sensing, Database, JMIR Medical Informatics, PLOS Computational Biology			
Physics & Mathematics	15,618	2,825	18.1%
Top 10 NLP4SG venues: Journal of Physics, Entropy, Mathematics, Symmetry, Physical Review, Chaos Solitons & Fractals, Discrete Dynamics in Nature and Society, Trends in Hearing, The Annals of Applied Statistics, Statistics in Medicine			
Overall	**98,753**	**35,398**	**35.8%**

Table 4: Disciplines differ in their propensity to use NLP methods to address social good questions. Over half of NLP work in Social Sciences and Health & Medical Sciences is focused on social good compared to less than 20% of work in Physics & Mathematics.

tion 2.2) to examine fine-grained venue-level differences in NLP4SG papers.

One question of interest is whether other computer science venues outside of ACL show similar trends to ACL regarding the use of NLP methods for social good. To quantify this, we identify CS venues outside of ACL as those appearing in one of the following subcategories of 'Engineering & Computer Science': 'Artificial Intelligence', 'Computational Linguistics', 'Computer Graphics', 'Computer Vision & Pattern Recognition', 'Computing Systems', 'Data Mining & Analysis', 'Databases & Information Systems', 'Human Computer Interaction', 'Library & Information Science', 'Robotics', 'Signal Processing', 'Software Systems', 'Theoretical Computer Science'.

Searching our dataset for these subcategories, we identified a set of 161 CS venues associated with 10,688 papers. Of these, 10% of the 2,902 ACL-authored papers were classified as NLP4SG, compared to 16.9% of the 7,766 by non-ACL authors. These numbers are quite similar to those we observed for our ACL and ACL-ADJACENT venue categories (see Figure 1), suggesting that ACL is comparable to other large computational communities in terms of the proportion of NLP research addressing social good.

Moving beyond only computer science venues to the full spectrum of disciplines, we aim to ask whether the diverse disciplines that compose our EXTERNAL venues differ in their propensity to use NLP methods to address social good questions. Table 4 shows the overall distribution of papers identified as corresponding to a known venue in the Google Scholar metadata, classified according to the eight major top-level categories.[7] We

[7]Note that some venues are associated with more than one of the eight categories, so the sum of the individual categories "total NLP Papers" exceeds the overall.

Category	NLP4SG Coefficient	p-value	significance
Business, Economics & Management	-2.357	0.4336	
Chemical & Material Sciences	26.199	0.0000	***
Engineering & Computer Science	-24.170	0.0000	***
Health & Medical Sciences	-22.944	0.0000	***
Humanities, Literature & Arts	2.883	0.0000	***
Life Sciences & Earth Sciences	-19.423	0.0000	***
Physics & Mathematics	-4.152	0.0000	***
Social Sciences	19.493	0.0000	***

Table 5: Regression results predicting venue-level h5 index from paper NLP4SG classification across eight coarse disciplinary categories. Broadly speaking, NLP work focused on social good tends to be published in lower-impact venues than other NLP work. However, for certain disciplines like Social Sciences, this trend is reversed.

find that disciplinary areas differ in the regularity with which they use NLP methods to address social good concerns. More than half of papers using NLP techniques in venues classified as Social Sciences and Health & Medical Sciences are identified as NLP4SG, while NLP papers in Physics & Mathematics address social good less than 20% of the time.

Lastly, we leverage h5-index metadata from Google Scholar to examine the relationship between NLP4SG work and venue-level impact as measured by citation patterns. The h5-index that we use is defined by Google Scholar as "the largest number h such that h articles published in 2019-2023 have at least h citations each."[8] For each of the eight coarse disciplinary categories, we fit a regression with papers as observations predicting the venue's h5-index against whether a given NLP paper is NLP4SG, and report coefficients of the NLP4SG variable from each regression in Table 5.

We find interesting discipline-specific patterns. Broadly, NLP4SG work incurs a cost in venue-level impact, such that relative to other papers in a given discipline using NLP methods, NLP4SG papers tend to be published in lower-impact venues. This large-scale pattern mirrors the trend for more NLP4SG in ACL-ADJACENT venues, which tend to be more specialized or regional with higher acceptance rates than ACL venues. However, for a subset of disciplines, particularly in the Chemical & Material Sciences and Social Sciences, the opposite is true and NLP4SG work is published in higher-impact venues on average relative to other work using NLP methods in that field.

[8] https://scholar.google.com/intl/en/scholar/metrics.html

3.4 Methodological Differences

Another difference worth exploring involves the NLP methods used in a paper, which may vary significantly by author type or venue type. For example, we find that of all NLP4SG papers since 2020 that explicitly mention LLMs in the abstract, 68.7% have an ACL author.

We explore this possibility of methodological differences by using an LLM to annotate the methodologies used in 20k of the abstracts in our dataset: 10k randomly sampled from ACL venues (either core or adjacent), and 10k randomly sampled from EXTERNAL venues. Both random samples were balanced, with 5k NLP4SG and 5k non-NLP4SG papers each. We classify the methods with zero-shot LLM annotation as either neural or traditional in nature. This is of course a highly simplified view into the complex methodological landscape of NLP, which we employ with the goal of understanding high-level trends in how NLP4SG work may or may not make use of the most contemporary methods in the field. Our prompt is provided in Appendix C. We validated the results with expert human annotations of 50 abstracts, yielding a classification accuracy of 86%.

Figure 5 shows the results for papers published in 2017 or later, grouped by venue type and author type. We see clear differences in methodological focus emerge across these three venue types. Unsurprisingly, core ACL venues predominantly rely on neural methods, with approximately 80% of papers classified as neural. This makes sense given the venue's emphasis on cutting-edge, model-driven methods. Although a majority of papers published at ACL-ADJACENT venues also use neural methods, the overall proportion is somewhat lower. Compared to core ACL venues, ACL-ADJACENT venues

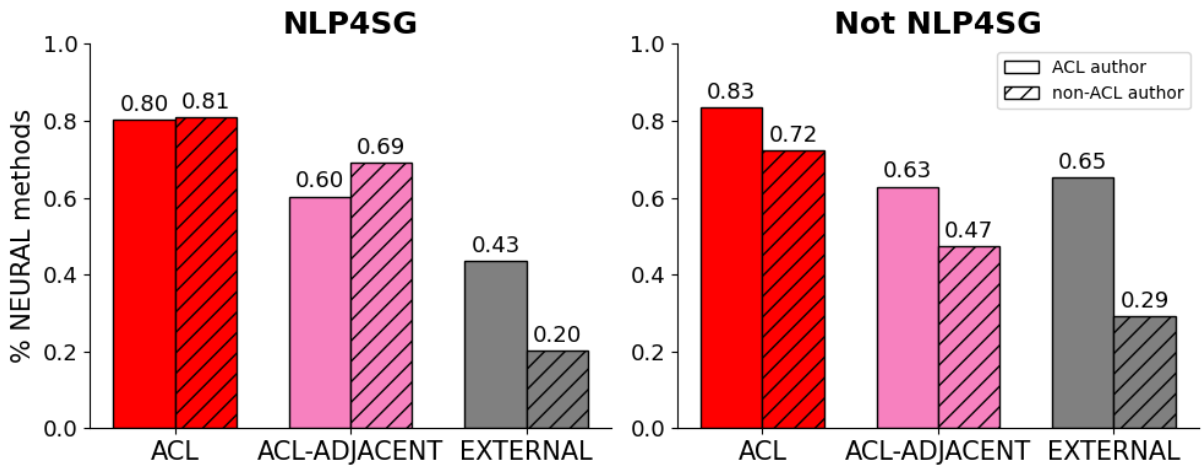

Figure 5: Distribution of neural methods within papers across venue and author types since 2017. Within the ACL Anthology, NLP4SG papers are just as likely to rely on neural methods as those not focused on social good. In external venues, NLP4SG papers are relatively less likely to use neural methods.

are often regarded as supporting a broader range of methodologies or placing less emphasis on state-of-the-art results.

While EXTERNAL venues do often use neural methods, non-neural methods are more prevalent. This trend holds across all eight disciplinary categories in the EXTERNAL data except "Engineering & Computer Science" which, much like ACL, is predominated by neural methods in the modern era. However, a clear distinction by author type emerges here: when publishing in an EXTERNAL venue, ACL authors are using neural methods much more than non-ACL authors.

In both core ACL and ACL-ADJACENT venues, NLP4SG papers have the same proportion of neural work as all other papers; however, this is not the case for EXTERNAL venues, where papers not focused on social good are more likely to use neural methods than NLP4SG papers across both ACL and non-ACL author types.

4 NLP4SG Within and Beyond ACL

In the broader context of AI for Social Good (Tomašev et al., 2020, AI4SG), there has been a concerted effort to apply artificial intelligence methodologies to address societal challenges. Shi et al. (2020) provide an extensive overview of AI4SG applications, discussing various domains where AI has been effectively utilized, such as healthcare (Sarella and Mangam, 2024), environmental sustainability (Thulke et al., 2024), and education (Ferreira-Mello et al., 2019). They also identify common challenges in implementing AI solutions for social good, including ethical considerations, data accessibility, and the need for interdisciplinary collaboration.

The work that we present here directly builds on Adauto et al. (2023), who introduce the NLP4SGPapers dataset and categorize papers based on their relevance to social issues and the UN SDGs. Our results replicate their work: we find a similar level of papers in the ACL Anthology classified as NLP4SG (e.g., 17.4% since 2020).

Yet as we look beyond the confines of ACL-associated venues, we see that ACL is not where most NLP4SG work is taking place, regardless of who is doing it. Papers by ACL authors are dramatically more likely to be classified as NLP4SG when published outside ACL, even in the modern era: less than 5% of all NLP4SG papers since 2020 appear in ACL or ACL-ADJACENT venues. The rest are in EXTERNAL venues, the majority of which are written by non-ACL authors. Our findings on the distribution of NLP4SG papers further evidence the breadth and depth of NLP as a tool for social good. Indeed, in some disciplines like Social Sciences and Health & Medical Sciences, when NLP is used (and it is used frequently), it is most often in the service of social good questions.

There are likely substantial differences between NLP4SG work by author and venue that remain to be understood. To better understand the landscape of NLP4SG across various academic venues, future work can conduct comparative analyses that examine how different venues contribute and relate to one another—for example, via citation net-

works (Mosbach et al., 2024; Wahle et al., 2023), a dynamic we do not address here. Though most NLP4SG work happens outside of ACL, it may be that much external work in this area looks to ACL as a core source of novel methods.

To conclude, we return to the agenda-setting questions for the ACL community posed in the introduction:

> (1) When ACL authors are doing work oriented towards social good, do they send that work to ACL venues?
>
> (2) Is ACL the place where most NLP4SG work is taking place?

Surprisingly, we find that the answer to both of these questions is a clear "no". The evidence strongly suggests that when ACL-associated authors seek to publish work targeting questions of social good, they tend to look to venues beyond ACL. Moreover, the majority of work using NLP methods for social good happens outside of ACL. The reasons for these trends are less clear. One place to look is the NLP Community Metasurvey (Michael et al., 2023). While this survey did not explicitly ask about social good, the findings suggest that the ACL community greatly underestimates its own belief in the value of interdisciplinary science, which includes work addressing social good questions almost by definition. Perhaps researchers feel that NLP4SG work is undervalued in the community, when this may not even be the case.

As a first step, we propose that these findings can serve not as an indictment but rather an inspiration. As ACL researchers, we should recognize that many members of our community do publish substantial amounts of NLP4SG work in other venues, and we should continue to advocate that such work be encouraged at *CL conferences by concrete mechanisms such as theme tracks, workshops, and the selection of keynote speakers. A small but important indicator of these ongoing developments in our community is that EMNLP 2025 is the first major NLP conference in which "NLP for Social Good" appears explicitly in the call for papers as a part of what was previously the "Computational Social Science and Cultural Analytics" track.

Ultimately, we hope that these findings help encourage a discussion within the ACL community about our role and continued growth in research advancing social good. To this end, we release our augmented metadata for replication and extended work in this area.

Limitations

Most importantly, the evidence presented here is associational rather than causal: we do not know, for example, whether ACL authors publish relatively more social good work outside of ACL because they preferred to do so due to publication incentives, or because that work was first rejected at ACL before finding another venue, or some other reason.

We considered the Semantic Scholar corpus appropriate as a source for broad research publications because it is the largest available corpus of open scientific papers, but beyond that we cannot make any strong claims about the representativeness of our dataset. For example, when linking venues to Google Scholar we find poor representation for the "Business, Economics & Management" category, where several top NLP4SG venues are relatively low-impact. However, we note that perfect representativeness of the full scholarly record would be very challenging to achieve under any definition. We also believe differences in representativeness of the corpus would be very unlikely to change any of the core findings in our work.

We note that conditioning on ACL authorship may overlook newer contributors to the field. However, the "ACL author" constraint requires only that one author meet the three-paper criteria – and early-career researchers are typically co-authoring papers with more experienced PIs. In addition, we consider an author an "ACL author" if they have ever written 3 ACL papers. Taken together, we argue that these requirements are largely reasonable, but may mean that we may be failing to consider papers as "ACL authored" when they ultimately should be if any author later goes on to publish additional papers in the community.

In this work we only looked at NLP4SG as a topic area. The ACL community is small relative to all of academic publishing, so it is logically possible that NLP+X for other X could follow similar trends where the total amount of published work on NLP+X is larger outside of ACL than within it. Therefore, our finding about the total quantity of publications cannot tell the whole story; however, we feel it is an important piece of the large-scale context of NLP4SG amidst the other findings presented here. An approach similar to that employed here could be used in future work to examine author- and venue-level differences for other areas of NLP.

As a parting thought, social good in NLP is not

limited to just published work, but includes other forms of public engagement that may ultimately be more impactful, such as public applications and tools, community and participatory engagement, educational outreach, and publications intended for broad readership. In this case, we argue that looking at published work provides a feasible and meaningful subset of NLP4SG activity to examine.

Ethical Considerations

Social good is a hard-to-define concept. We rely on the existing SDG-based definition, but it is possible alternative definitions—and concordant alternative decisions about values related to what constitutes a social good—could substantially shift the findings reported here.

The NLP4SG classification model from Adauto et al. (2023) is licensed under Apache 2. A code assistant was used for visualization iterations.

References

Fernando Adauto, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. Beyond good intentions: Reporting the research landscape of NLP for social good. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 415–438, Singapore. Association for Computational Linguistics.

Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Brigitte Hoyer Gosselink, Kate Brandt, Marian Croak, Karen DeSalvo, Ben Gomes, Lila Ibrahim, Maggie Johnson, Yossi Matias, Ruth Porat, Kent Walker, et al. 2024. Ai in action: Accelerating progress towards the sustainable development goals. *arXiv preprint arXiv:2407.02711*.

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.

Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Rajarshi Ghosh, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, et al. 2025. Nlp for social good: A survey of challenges, opportunities, and responsible deployment. *arXiv preprint arXiv:2505.22327*.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

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.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2023. What do NLP researchers believe? results of the NLP community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada. Association for Computational Linguistics.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105.

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2024. Topics, authors, and institutions in large language model research: Trends from 17k arxiv papers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1223–1243.

Chaoqun Ni, Cassidy R. Sugimoto, and Jiepu Jiang. 2013. Venue-author-coupling: A measure for identifying disciplines through author communities. *Journal of the American Society for Information Science and Technology*, 64(2):265–279.

OpenAI and 78 other authors. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *Preprint*, arXiv:2205.01833.

Prakash Nathaniel Kumar Sarella and Vinny Therissa Mangam. 2024. Ai-driven natural language processing in healthcare: transforming patient-provider communication. *Indian Journal of Pharmacy Practice*, 17(1).

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

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):2468.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif Mohammad. 2023. We are who we cite: Bridges of influence between natural language processing and other academic fields. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12896–12913.

A Top 10 EXTERNAL Venues

- Journal of physics
- PLOS ONE
- Scientific Reports
- Frontiers in Psychology
- Sensors
- IEEE Access
- IOP conference series
- Lecture Notes in Computer Science
- BMC Bioinformatics
- Proceedings of the AAAI Conference on Artificial Intelligence

B Annotation Instructions

Annotators were provided with NLP4SG binary classification instructions taken directly from Adauto et al. (2023), including their Task 1 decision flowchart (Figure 12 in their paper) and the following text.

Included: 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. 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. Social problems in the digital era: e.g., online toxicity, misinformation, privacy protection, and deception detection.

Excluded: General purpose, coarse-grained NLP tasks: machine translation, language modeling, summarization, sentiment analysis, etc. General purpose, fine-grained NLP tasks: news classification; humor detection; technologies for 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 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; tasks with controversial nature or unknown effect (varying a lot by how people use them in the future): e.g., news comment generation; financial NLP, 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.

C Prompt Templates

Task: Classify the paper into ONE of the SDG categories.

Instructions:

- You will be given a paper title and abstract.
- Classify the paper into ONE of the SDG categories.
- You should **ONLY** return ONE SINGLE SDG label, no other text.

Categories with Example Papers:

G1. Poverty

- Role of AI in poverty alleviation: A bibliometric analysis

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

G7. Energy

- Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities

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

G15. Life on Land

- Harnessing Artificial Intelligence for Wildlife Conservation

G16. Peace

- On Unifying Misinformation Detection
- Fully Connected Neural Network with Advanced Preprocessor to Identify Aggression on Social Media

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

Paper Title: {paper-title}
Paper Abstract: {paper-abstract}

Instructions Given the below academic paper's abstract, extract the primary methodological paradigm applied by the researchers.

text: {abstract}

The possible two responses for [methodology] are: NEURAL TRADITIONAL

- Examples of methods in the NEURAL category are large language models (LLMs), generative AI, deep learning, neural networks, LSTMs, vector representations, and embeddings.
- Examples of methods in the TRADITIONAL category are statistical NLP, logistic regression, support vector machines, random forests, structure prediction, lexicons, regular expressions, and rule-based heuristics.

Return only: [methodology]
Do NOT return anything other than NEURAL or TRADITIONAL.