Natural Language Processing (2026), 1–6
doi:10.1017/nlp.2026.10027

CAMBRIDGE UNIVERSITY PRESS

NLP FOR SOCIAL GOOD

Natural language processing for social good: Where we are, what is missing, and where we should go

Tharindu Ranasinghe

Lancaster University, UK
Email: t.ranasinghe@lancaster.ac.uk

(Received 10 April 2026; revised 10 April 2026; accepted 12 April 2026)

Abstract

Natural language processing (NLP) technologies increasingly shape public life, yet their deployment for social good remains unevenly distributed across domains, languages, and geographies. This piece inaugurates the NLP for Social Good column in this journal. In this piece, I map the current state of NLP for Social Good (NLP4SG) across nine application domains. The picture that emerges is one of striking imbalance: AI harms, inclusion, and digital violence attract the bulk of research attention, while poverty, peacebuilding, and environmental protection remain critically underexplored. I argue that the field must address three structural gaps, domain coverage, linguistic diversity, and evaluation methodology, if NLP is to fulfil its potential as a force for equitable social progress. The piece concludes with five directions that I believe will define the next chapter of NLP4SG research.

Keywords: Machine learning; language resources; multilinguality; natural language processing; social good

1. Introduction

This piece inaugurates the Natural language processing *(NLP) for Social Good* column in this journal. The column aims to provide a regular forum for discussing how NLP technologies can, and sometimes fail to, contribute to positive social outcomes. My goal in this first instalment is to offer a landscape overview: where are we, what is missing, and where should we go?

The idea that NLP should serve the public good is not new. Hovy and Spruit (Hovy and Spruit 2016) raised the question of NLP's social impact a decade ago. Since then, the field has seen the establishment of recurring workshop series such as NLP for Positive Impact (NLP4PI), which has run five editions between 2021 and 2026, and domain-specific venues including ClimateNLP (since 2024), BioNLP (since 2002), and ClinicalNLP (now in its seventh edition). Shared tasks have also dominated across social good applications, from explainable detection of online sexism (SemEval 2023) (Kirk *et al.* 2023) to multilingual persuasion technique detection in memes (SemEval 2024) Dimitrov *et al.* (2024) and clinical text generation (BioNLP 2024) Xu *et al.* (2024). Furthermore, in 2025, EMNLP included "NLP for Social Good" explicitly in its call for papers.

Several recent surveys have attempted to chart this research area systematically. Fortuna *et al.* (2021) proposed an early map of NLP4SG, analysing around 50,000 ACL Anthology papers using keyword matching and finding that explicit social good papers accounted for under 10% of publications, with healthcare dominating and areas such as environmental sustainability and language disorders largely neglected. Jin *et al.* (2021) took a complementary approach, evaluating NLP tasks through the lens of social impact and finding significant gaps between what the community works on and what would have the greatest positive effect. Adauto *et al.* (2023) scaled this analysis to

over 76,000 papers using LLM-based annotation. Most recently, Karamolegkou *et al.* (2026) conducted the most comprehensive survey to date, analysing 47,000 ACL Anthology papers across nine domains aligned with UN Sustainable Development Goals and the World Economic Forum's Global Risks framework.

I draw on these surveys, supplemented by additional research, to construct the overview that follows. The remainder of this column is organised as follows. Section 2 maps the current research landscape and identifies the domains that receive the most and least attention. Section 3 examines how large language models have reshaped the possibilities and perils of NLP4SG. Section 4 addresses the persistent challenge of linguistic diversity. Section 5 discusses evaluation. Section 6 includes five directions I believe will define the field's next chapter.

2. The research landscape in NLP for social good

Karamolegkou *et al.* (2026) annotated 47,000 ACL Anthology papers (2019–2025) across nine NLP4SG domains using GPT-4.1 mini with zero-shot classification. The resulting map reveals that approximately 76.5% of papers received at least one social good domain label, but the distribution across domains is heavily skewed.

AI harms (bias, fairness, toxicity, interpretability) and inclusion/inequalities together account for the largest share of NLP4SG research, with over 15,000 and 19,000 papers, respectively. These are followed by education (around 13,700 papers), healthcare (around 5,400), digital violence (around 4,800), and misinformation (around 4,200). At the other end of the spectrum, peacebuilding (around 1,100 papers), poverty (around 1,000), and environmental protection (around 380) remain what the survey's authors call areas that are "only starting to gain traction in response to real-world crises."

Furthermore, the co-occurrence heatmap in Karamolegkou *et al.* (2026) shows that AI harms and inclusion research frequently overlap with multiple other domains, making them central hubs in the NLP4SG network, while poverty and peacebuilding appear relatively isolated. Research on environmental protection, which might seem naturally connected to sustainability discourse more broadly, has the weakest inter-domain connections of all.

Between 2019 and 2025, NLP4SG publications have grown across nearly all domains, with particularly rapid growth in AI harms, inclusion, and education. The methodological landscape has also evolved significantly: dataset creation, model analysis, and interpretability have become the dominant tasks, while transfer learning, prompting, and in-context learning have risen to prominence as methods, gradually displacing older paradigms based on supervised learning and classical machine learning.

3. LLMs in NLP for social good

The rise of large language models since 2022 has fundamentally reshaped the NLP4SG landscape. On the opportunity side, zero-shot and few-shot transfer enables NLP applications without the large labelled datasets that were previously required, which was precisely the bottleneck that has historically limited NLP deployment in low-resource social good settings. Open-source models have made sophisticated NLP accessible to organisations without massive compute budgets.

Concrete LLM-powered social good applications have emerged across domains. In healthcare, LLMs now support clinical note summarisation, patient education, and diagnostic decision-making, with domain-specific models like MedPALM and MEDITRON showing strong performance. In mental health, human-AI collaboration has been shown to enable more empathic conversations in peer support settings (Sharma *et al.* 2023). In crisis response, lightweight LLM frameworks using parameter-efficient fine-tuning maintain over 99% of full fine-tuning performance at half the memory cost for disaster tweet classification. In education, LLM-powered tutors are being adapted for specialised populations, including learners with hearing impairments.

However, the risks are substantial and domain-specific. Hallucination in clinical settings is the most acute concern: a 2025 study in Communications Medicine found that six LLMs tested with physician-validated clinical vignettes repeated or elaborated on planted errors in up to 83% of cases. Even with explicit mitigation prompts, the error rate was only halved. A 2026 benchmark across 37 models found general hallucination rates of 15–52%, reaching 64% in healthcare contexts without safeguards. For NLP4SG applications in high-stakes domains such as healthcare, legal aid, and crisis response, these error rates are not merely inconvenient, as they can cause direct harm to vulnerable populations.

The environmental cost of LLMs presents a further tension. Training GPT-3 consumed approximately 1,287,000 kWh and produced around 552 tonnes of CO2 equivalent. Inference costs are also increasingly dominant. For NLP4SG, this argument has particular force: many communities served by social good applications, such as those in crisis zones, rural healthcare settings, or developing economies, have poor or no internet connectivity, making cloud-dependent systems impractical regardless of their technical capabilities.

4. The language gap in NLP4SG

The intersection of language resources and social good creates a great inequality: communities most in need of NLP4SG applications, such as those facing poverty, health crises, and conflict, disproportionately speak languages with the fewest NLP resources.

Several grassroots communities have emerged to address this gap, and their models deserve attention. Masakhane ("We build together" in isiZulu) has grown to over 2,000 researchers from more than 30 African countries, producing datasets for named entity recognition, sentiment analysis, and LLM evaluation across dozens of African languages. AI4Bharat, led from IIT Madras, has created the world's largest Indic parallel corpus (BPCC, with 230 million sentence pairs across 22 Indian languages) alongside neural translation systems and speech datasets covering 13 languages. IndoNLP addresses Indonesia's 700+ languages through resources like NusaX and IndoBERT. The LoResLM workshop series supports indigenous building language models for low-resource languages (Hettiarachchi *et al.* 2026).

Multilingual models have improved, but they still remain insufficient. Aya from Cohere for AI represents perhaps the most inclusive effort to date: Aya 101 was followed by Aya Expanse and notably Tiny Aya (February 2026), which is a 3.35-billion-parameter model designed to run locally on consumer devices, covering 70+ languages with regional specialisation.

Funding for low-resource NLP4SG also remains modest. The Lacuna Fund, the world's first collaborative fund for creating ML datasets in low- and middle-income contexts which allocated approximately $1 million for NLP datasets in Africa and Latin America in 2024, a fraction of the billions invested in English-centric AI. Community events like the Deep Learning Indaba (held in Dakar in 2024 with over 600 attendees) provide vital infrastructure, but the structural funding asymmetry remains.

5. Evaluation in NLP4SG

The inadequacy of standard NLP metrics for evaluating NLP4SG systems is one of the field's most persistent problems. NLP benchmarks lack clear alignment with user needs and encourage what the authors call "pointless SOTA-chasing." For NLP4SG, this means evaluation must demonstrate connection to actual social outcomes, not merely performance on proxy tasks.

Several promising evaluation approaches are emerging. Disaggregated evaluation, as proposed by Pfohl *et al.* (2024), surfaces health equity harms through stratified assessments across patient subpopulations. They argued that equal performance across subgroups is an unreliable measure of fairness when data reflects real-world disparities. Participatory evaluation is also gaining traction:

Wilson, Atabey, and Revans (2025) reported on participatory design engagements at the NLP4PI workshop, arguing that what constitutes a "good" NLP output depends entirely on the context of use, which is something standard metrics cannot capture.

On the institutional side, the ACL community has strengthened its ethical infrastructure. The Responsible NLP Research Checklist is mandatory for all ACL Rolling Review submissions and was updated in October 2024. Since December 2024, inappropriately completed checklists can result in desk rejection. The ACL Publication Ethics Policy requires disclosure of all generative AI use. The EU AI Act, the world's first comprehensive AI law, classifies AI in healthcare and education as "high-risk," requiring conformity assessments and human oversight—directly affecting many NLP4SG applications.

6. Future directions

Based on the analysis above, I identify five directions that I believe will define NLP4SG's next chapter.

- **Small language models on edge devices** This may be the most transformative direction. Many communities served by NLP4SG, such as crisis response teams, rural healthcare workers, and refugee settings, have poor or no internet connectivity. On-device processing keeps sensitive data local, eliminates prohibitive cloud costs, and enables offline operation. Small language models consuming 60–70% less energy than their larger counterparts align the social good mission with environmental responsibility. Tiny Aya's 3.35-billion-parameter model running locally in 70+ languages signals this future.
- **Addressing the neglected domains** Poverty, peacebuilding, and environmental protection lack the community infrastructure that healthcare NLP has built over two decades. The ClimateNLP workshop series demonstrates how quickly a dedicated venue can catalyse research. Poverty and peacebuilding need equivalent institutional support.
- **Multilingual and multicultural grounding** NLP4SG systems must work in the languages spoken by the communities they aim to serve. This requires not only multilingual model development but also culturally grounded evaluation. The participatory models pioneered by Masakhane and AI4Bharat, in which communities are partners in research, offer templates for the broader NLP4SG community to adopt.
- **Human-centred evaluation** frameworks We need evaluation that connects benchmark performance to real-world outcomes. This means field trials, participatory evaluation design, disaggregated fairness assessments, and explicit measurement of deployment impact.
- **Cross-disciplinary partnerships** Successful NLP4SG collaborations such as CLEAR Global's humanitarian translation, the DEEP platform for crisis analysis, GhanaNLP's Khaya translation app share a common feature: sustained engagement between NLP researchers and domain practitioners. However, NLP4SG papers tend to appear in lower-impact venues, suggesting a prestige penalty that discourages such work. The NLP community must actively counteract this through dedicated tracks, awards, and career incentives that value deployment and impact alongside technical novelty.

7. Concluding remarks

NLP for Social Good stands at a turning point. The field has achieved institutional recognition, built community infrastructure, and demonstrated real-world deployments. But three structural tensions will determine whether this momentum translates to genuine social impact. First, the domain imbalance: as long as poverty and peacebuilding lack the community infrastructure that

AI harms and hate speech detection enjoy, NLP4SG will underserve the communities facing the greatest need. Second, the geographic and linguistic asymmetry: NLP4SG research remains overwhelmingly concentrated in institutions with the least proximity to the problems being addressed. Third, the evaluation gap: without frameworks that connect performance to outcomes, the field risks producing sophisticated tools that fail on deployment.

I hope this column will serve as a regular venue for advancing these conversations. I particularly invite contributions that address the underexplored domains identified here, that report on field deployments rather than benchmark improvements, and that centre the perspectives of communities that NLP4SG aims to serve. The technical capabilities of modern NLP are remarkable. The question is whether we can direct them where they are most needed.

References

Adauto F., **Jin Z.**, **Schölkopf B.**, **Hope T.**, **Sachan M.** and **Mihalcea R.** (2023). Beyond good intentions: Reporting the research landscape of NLP for social good. In Houda B. Juan P. and Kalika B. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 415–438. https://doi.org/10.18653/v1/2023.findings-emnlp.31. https://aclanthology.org/2023.findings-emnlp.31/

Dimitrov D., **Alam F.**, **Hasanain M.**, **Hasnat A.**, **Silvestri F.**, **Nakov P.** and **Da San Martino G.** (2024). SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In Atul K.O., A. Seza D., Harish T. M., Giovanni D.S.M., Sara R. and Aiala R. (eds), *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 2009–2026. https://doi.org/10.18653/v1/2024.semeval-1.275. https://aclanthology.org/2024.semeval-1.275/.

Fortuna P., **Pérez-Mayos L.**, **AbuRa'ed A.**, **Soler-Company J.** and **Wanner L.** (2021). Cartography of natural language processing for social good (NLP4SG): Searching for definitions, statistics and white spots. In Anjalie F., Shrimai P., Maarten S., Zhijing J., Jieyu Z. and Chris B. (eds), *Proceedings of the 1st Workshop on NLP for Positive Impact. Association for Computational Linguistics*, pp. 19–26. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.3. https://aclanthology.org/2021.nlp4posimpact-1.3/. Online: August.

Hettiarachchi H., **Ranasinghe T.**, **Plum A.**, **Rayson P.**, **Mitkov R.**, **Gaber M.M.**, **Premasiri D.**, **Tan F.A.** and **Uyangodage L.** (2026). Overview of the second workshop on language models for low-resource languages (LoResLM 2026). In Hansi H., Tharindu R., Alistair P., Paul R., Ruslan M., Mohamed G., Damith P., Fiona A.T. and Lasitha U. (eds), *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*. Rabat, Morocco: Association for Computational Linguistics, pp. 651–661. https://doi.org/10.18653/v1/2026.loreslm-1.56. https://aclanthology.org/2026.loreslm-1.56/.

Hovy D. and **Spruit S.L.** (2016). The social impact of natural language processing. In Katrin E. and Noah A.S. (eds), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 591–598. https://doi.org/10.18653/v1/P16-2096. https://aclanthology.org/P16-2096/.

Jin Z., **Chauhan G.**, **Tse B.**, **Sachan M.** and **Mihalcea R.** (2021). How good is NLP? a sober look at NLP tasks through the lens of social impact. In Chengqing Z., Fei X., Wenjie L. and Roberto N. (eds), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, pp. 3099–3113. https://doi.org/10.18653/v1/2021.findings-acl.273. https://aclanthology.org/2021.findings-acl.273/. Online: August.

Karamolegkou A., **Borah A.**, **Cho E.**, **Choudhury S.R.**, **Galletti M.**, **Gupta P.**, **Ignat O.**, **Kargupta P.**, **Kotonya N.**, **Lamba H.**, **Lee S.-J.**, **Mangla A.**, **Mondal I.**, **Moudakir F.Z.**, **Nazar D.**, **Nemkova P.**, **Pisarevskaya D.**, **Rizwan N.**, **Sabri N.**, **Samway K.**, **Stammbach D.**, **Schulten A.S.**, **Tomás D.**, **Wilson S.R.**, **Yi B.**, **Zhu J.H.**, **Zubiaga A.**, **Søgaard A.**, **Fraser A.**, **Jin Z.**, **Mihalcea R.**, **Tetreault J.R.** and **Dementieva D.** (2026). NLP for social good: A survey and outlook of challenges, opportunities and responsible deployment. In Vera D., Kentaro I. and Lluís M. (eds), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Rabat, Morocco: Association for Computational Linguistics, pp. 5110–5170. https://doi.org/10.18653/v1/2026.eacl-long.238. https://aclanthology.org/2026.eacl-long.238/.

Kirk H., **Yin W.**, **Vidgen B.** and **Röttger P.** (2023). SemEval-2023 task 10: Explainable detection of online sexism. In Atul K.O., A. Seza D., Giovanni D. S. M., Harish T.M., Ritesh K. and Elisa S. (eds), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 2193–2210. https://doi.org/10.18653/v1/2023.semeval-1.305. https://aclanthology.org/2023.semeval-1.305/.

Pfohl S.R., **Cole-Lewis H.**, **Sayres R.**, **Neal D.**, **Asiedu M.**, **Dieng A.**, **Tomasev N.**, **Rashid Q.M.**, **Azizi S.**, **Rostamzadeh N.**, **McCoy L.G.**, **Celi L.A.**, **Liu Y.**, **Schaekermann M.**, **Walton A.**, **Parrish A.**, **Nagpal C.**, **Singh P.**, **Dewitt A.**, **Mansfield P.**, **Prakash S.**, **Heller K.**, **Karthikesalingam A.**, **Semturs C.**, **Barral J.**, **Corrado G.**, **Matias Y.**, **Smith-Loud J.**, **Horn I.** and **Singhal K.** (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine* **30**(12), 3590–3600.

Sharma A., **Lin I.W.**, **Miner A.S.**, **Atkins D.C.**, **Althoff T.** (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* **5**(1), 46–57.

Wilson C., **Atabey A.** and **Revans J.** (2025). Towards child-centred AI in children's learning futures: Participatory design futuring with SmartSchool and the co-design stories toolkit. *International Journal of Human-Computer Studies* **199**, 103431.

Xu J., **Chen Z.**, **Johnston A.**, **Blankemeier L.**, **Varma M.**, **Hom J.**, **Collins W.J.**, **Modi A.**, **Lloyd R.**, **Hopkins B.**, **Langlotz C.** and **Delbrouck J.-B.** (2024). Overview of the first shared task on clinical text generation: RRG24 and "discharge me!. In Dina D.-F., Sophia A., Makoto M. Kirk R. and Junichi T. (eds), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Bangkok, Thailand: Association for Computational Linguistics, pp. 85–98. https://doi.org/10.18653/v1/2024.bionlp-1.7. https://aclanthology.org/2024.bionlp-1.7/.

Cite this article: Ranasinghe T. Natural language processing for social good: Where we are, what is missing, and where we should go. *Natural Language Processing* https://doi.org/10.1017/nlp.2026.10027