

International Workshop on Data for the Wellbeing of Most Vulnerable at ICWSM 2025

Yelena Mejova¹, Kyriaki Kalimeri^{1,2}, Daniela Paolotti¹

¹ ISI Foundation, Via Chisola 5, Torino, Italy
{first.last}@isi.it

² UNICEF, New York, NY, USA
kkalimeri@unicef.org

Abstract

The sixth installment of this workshop continued to provide a platform for the sharing of ongoing efforts in applying data science to issues involving vulnerable populations, including children, elderly, racial or ethnic minorities, socioeconomically disadvantaged, under-insured or those with certain medical conditions, who are often absent in commonly used data sources. Ten contributed talks presented a variety of case studies using social media, satellite, and other data sources, whereas two keynotes outlined algorithmic and organizational challenges of using data and AI for decision-making. Below we outline the proceedings of the sixth International Workshop on Data for the Wellbeing of Most Vulnerable that took place in Copenhagen, Denmark in 2025.

Introduction

Digital technologies—especially the Internet and social media—have dramatically expanded our ability to observe and understand the vulnerabilities that shape modern societies. Real-time, large-scale data from the web offers unprecedented visibility into issues like health disparities, social inequality, and collective stress. For instance, online behavior has been used to track the spread of seasonal illnesses such as influenza, and to explore the lived experiences of individuals coping with mental health conditions like anorexia and bulimia. Yet, despite this promise, significant blind spots remain. Many vulnerable populations—including children, older adults, racial and ethnic minorities, those with limited financial resources, and people with inadequate insurance or chronic medical conditions—are often missing from traditional datasets. Their digital traces are sparse or excluded altogether, leading to skewed understandings. Adding to this challenge, the rise of data-driven systems and generative AI introduces new risks: biases embedded in algorithms and training data can perpetuate exclusion rather than resolve it. As these technologies play a growing role in shaping public knowledge and crisis response, ensuring their fairness and inclusivity becomes not just a technical priority but a moral one.

The sixth edition of the ICWSM Workshop on Data for the Wellbeing of the Most Vulnerable aims to foster

deeper engagement with emerging data sources and innovative methodologies to better understand and support at-risk populations. Central to this mission is the careful selection of relevant datasets, thoughtful identification of vulnerable communities, and the ethical handling of data throughout the research process. Extending the benefits of the big data revolution to those most often overlooked requires not only technical precision but also a strong commitment to equity and reproducibility. Reflecting the complexity of this challenge, the workshop invites a truly multidisciplinary conversation—drawing on expertise from computer science and AI to epidemiology, demography, linguistics, and beyond.

The workshop took place on June 23, 2025 in Copenhagen, Denmark. The program included two invited keynotes, and 10 contributed talks. Below, we outline the themes and outcomes of the workshop.

Keynotes

The first keynote speaker was Patrick Michael Brock, who is a Senior Data Scientist at the World Bank-UNHCR Joint Data Center on Forced Displacement (JDC). His talk, “Innovating for better socioeconomic data on forced displacement” gave an overview of the World Bank UNHCR Joint Data Center’s work on collecting and analysing socioeconomic data on vulnerable people affected by forced displacement. To date, the Center has supported UNHCR to safely make 979 datasets publicly available. Because the data often concerns vulnerable populations, Dr. Brock emphasized the efforts on privacy protection and anonymization of the data (you can read more here¹). He also described a case study where satellite and building data were used to fill data gaps to construct a representative sample household survey sample frame to compare refugee and national population wellbeing in South Sudan. During discussion, the workshop audience was excited to learn more about the metadata provided about the datasets, especially how they were collected and the potential future updates. A question arose regarding the mismatch between the data available (or the “solutions”) and the research problems. Dr. Brock agreed that there is a language problem in communication between problem owners and data scientists, and

flagged JDC work-in-progress on an online platform, or a “marketplace”, where data and problems can be matched up and diverse teams created for fast results. For more information, see <https://www.jointdatacenter.org/>.

The second keynote was Roberta Sinatra, a Professor in Computational Social Science at the Center for Social Data Science, University of Copenhagen, with a talk titled “Failing our Youngest: Fairness, Bias, and Accountability in a Decision Support System”. In it, she presented a case study from Danish child welfare services, which used an algorithm to identify and assess children at risk from abuse. She showed that this algorithm used protected attributes in its calculation, displayed age bias, and suffered of a potential information leak because of standardization before separating train and test data and use of proxies of abuse also as features. Dr. Sinatra emphasized the need for rigorous protocols when deploying algorithms having to do with human well-being. Further, the training data may have been biased, since older children were more likely to self-report abuse than the younger ones. Several questions were raised by the workshop attendees about the design and training of the model, to which Dr. Sinatra responded that a greater transparency into the exact code used in the system is needed to answer these questions.

Contributions

All contributions were reviewed by 2–3 members of the program committee (10 in total), most of whom specialize in computer or data science, and especially in the studies relevant to a vulnerable population.

Effect of AI on the Labor Market

Choi et al. (2025) propose a Weighted AI Index that captures the a job openings’ exposure to AI-related innovation. To do this, they utilized LinkedIn and Wipo platforms and used GPT-4o to extract job tasks that may be mentioned in recent patents. Industries found to be most susceptible to automation were Wholesale, Transportation and Warehousing, Manufacturing, Information, and Retail. A thematic analysis also revealed how higher exposure to current AI technologies may be associated with lower wages in some sectors.

Public health

Bickham et al. (2025) studied a TikTok dataset gathered using eating disorder-related keywords and hashtags. Within this dataset, the authors captured spikes in activity around the Eating Disorder Awareness Week sponsored by the (U.S.) National Eating Disorders Association (NEDA). However, after mid-2023 there is a gradual decline in the volume, possibly due to the platform’s increased moderation of eating disorders-related content. Perhaps as a consequence, the most prominent hashtags in the dataset were that around recovery and awareness, and those addressing specific difficulties such as #fearfood, wherein those in recovery confront foods that were associated with anxiety and weight gain fears. The dataset is available at <https://github.com/cbickham3232/EDTok-A-Dataset-for-Eating-Disorder-Content-on-TikTok>.

Gender, Inclusion and Bias

Janczy et al. (2025) presented an abstract of a study examining the way people of different genders are referred to in the press—either by first or last name. They find that last name usage is most prevalent in politics, where women are disproportionately referred to by their first name. When controlling for the gender of the speaker, they find that both male and female speakers predominantly use last names when referring to men. Further, they find that naming conventions also vary by (the U.S.) party affiliation: Republicans refer to Democrats by their first names more frequently than vice versa. Using a matched experimental design, the authors found a confirmation of gender bias in professional communication.

Brown (2025) examined the public reception of the introduction of pronoun-sharing features that allow platform users to explicitly share their preferred pronouns. The framing found in articles in North America about such features was found to be overall positive. Further, the author presented the results of 12 interviews of trans people working in the Technology sector, which point to an uneven and sporadic decision-making and implementation of such features. However it is not clear whether the public’s positive attitude will persist as the political environment in the region changes.

Political Action

The abstract presented by Maggioni et al. (2025) compared Twitter discourse expressing belief in climate change with the adoption of pro-environmental behaviors around Europe. The authors found a positive relationship between these two variables, even when taking into account education and the average level of difficulty in paying bills of the local population. The direction of causality, as is often the case, was brought up by the audience members, and the authors agreed that this relationship may be best used for the monitoring, but not necessarily the prediction, of offline pro-environmental actions.

Pinto and Ferrara (2025) used a multimodal TikTok dataset EcuadorCrisisTikTok capturing the 2024 political crisis and organized crime escalation in Ecuador. The videos were transcribed using OpenAI’s Whisper, which also identifies the most prominent language. The majority of the dataset was in Spanish, while also having a prominent English and Portuguese component, indicating an international reach. A topic discovery process identified four main themes: public safety and security, violent crimes and assassinations, corruption and narcotrafficking, and political activism and debate. The dataset is available at <https://github.com/gabbspinto/EcuadorCrisisTikTok>.

Verdi Do Amarante and Pierri (2025) presented an abstract concerning political ads in Brazil’s 2024 municipal elections and their content, reach and engagement. Using the Meta Ad Library, the authors collected 50,000 ads from 1,000 candidates advertising on Facebook and Instagram. They find that ads from center-aligned candidates, candidates seeking reelection, and female candidates tend to achieve higher engagement per Brazilian Real (BRL) spent.

The results suggest that the candidates do A/B testing in the beginning of the campaign, and spend most of their budget closer to the election. Better disclosures of funding is necessary to improve the transparency of such influence campaigns.

Satellite Imagery for Mobility

Aidoo et al. (2025) propose to use low-cost, 3-meter resolution satellite imagery to train a model for remote sensing applications. Specifically, they propose to estimate parking lot occupancy around stores in Germany which are usually open on Saturdays but not on Sundays. The authors extend their system to monitor mobility patterns at a major bus terminal in Sudan, successfully capturing a drop in activity during the onset of an armed conflict. As typical of such data, cloud cover is one of the biggest limitations of using this data, especially for events which are temporally fine-grained. However, if the weather allows, the authors see promise in using such models to monitor human mobility especially in cities bordering countries undergoing conflicts or other disasters.

Although satellite imagery has been extensively used in the research around vulnerable populations, Musienko et al. (2025) have found that the coverage of such data may not be ubiquitous. Considering specifically major satellite constellations that provide optical satellite imagery with a ground sampling distance below 10 meters, the authors found that the locations farther away from the equator are generally revisited more frequently, resulting in the fact that less developed, less populated places had fewer satellite images available. The coverage may change, however, around major conflicts, pointing to possible business decisions in the selection of surveillance targets. Whether such increased focus can be predictive of conflicts is an intriguing idea. Further, findings for technology that is less affected by cloud cover may be different.

Opinion Dynamics with Minority Groups

Janik, Michalski, and Arora (2025) has presented an extension of network influence mechanisms to consider the extent a minority group is represented. Using the CoD-ING opinion diffusion model, the authors examine different versions of networks by removing minority nodes using degree-preserving network rewiring. They show that minorities are an important component of opinion dissemination processes, and that they often act as hubs and bridges in the networks, and that their removal significantly impairs F1 scores. These findings, authors argue, resolve the Granovetter-Centola paradox: “Although weak ties allow widespread dissemination of information, minorities form strong local bridges that allow the social reinforcement required for behavior adoption.”

Workshop Organization

The workshop was organized by researchers from ISI Foundation, a nonprofit research institute in Turin, Italy, and UNICEF HQ, NYC, USA:

- Yelena Mejova, a Senior Research Scientist at ISI Foundation specializing in tracking health-related signals in

text and social media.

- Kyriaki Kalimeri, who is a Senior Research Consultant at UNICEF, and researcher at ISI Foundation working in the areas of Computational Social Science and Humanitarian AI.
- Daniela Paolotti, a Senior Research Scientist ISI Foundation, an expert in computational epidemiology and participatory disease surveillance.

For more information on the workshop, please see its website at <https://sites.google.com/view/dataforvulnerable25>.

References

- Aidoo, T.; Koebe, T.; Maurya, A.; Shrestha, H.; and Weber, I. 2025. A Weak Supervision Learning Approach Towards an Equitable Mobility Estimation. *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Bickham, C.; Ramirez-Gonzalez, B.; Chu, M. D.; Lerman, K.; and Ferrara, E. 2025. EDTok: A Dataset for Eating Disorder Content on TikTok. *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Brown, C. 2025. “This actually goes beyond just pronouns”: Shifting discourses on trans-affirming features in tech. *Abstract at the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Choi, E. C.; Cao, Q.; Guan, Q.; Peng, S.; Chen, P.-Y.; and Luceri, L. 2025. Mapping Labor Market Vulnerability in the Age of AI: Evidence from Job Postings and Patent Data. *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Janczy, J.; Čuljak, M.; Spitz, A.; and Arora, A. 2025. Gender Bias in How Public Figures are Referenced in News Quotes at Scale. *Abstract at the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Janik, M.; Michalski, R.; and Arora, A. 2025. Computationally Establishing Causality: The Impact of Demographic Minorities on Bias in Opinion Diffusion Models. *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Maggioni, E.; Maria Aiello, L.; Garlaschelli, D.; and Mastandrea, R. 2025. Twitter anticipates adoption and change in pro-environmental behavior. *Abstract at the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Musienko, V.; Jacquet, A.; Weber, I.; and Koebe, T. 2025. Coverage Biases in High-Resolution Satellite Imagery. *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Pinto, G.; and Ferrara, E. 2025. A Multimodal TikTok Dataset of Ecuador’s 2024 Political Crisis and Organized Crime Discourse. *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Verdi Do Amarante, D.; and Pierri, F. 2025. Targeting and Messaging in Political Ads on Meta platforms During Brazil’s 2024 Municipal Elections. *Abstract at the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.