ICWSM 2022 Workshop on Data for the Wellbeing of Most Vulnerable: Global Problems, Local Solutions

Yelena Mejova[1] **Kyraki Kalimeri**[1] **Daniela Paolotti**[1] **Rumi Chunara**[2]
[1]ISI Foundation, Turin, Italy
[2]New York University, NY, USA
yelenamejova@acm.org, kyriaki.kalimeri@isi.it, daniela.paolotti@isi.it, rumi.chunara@nyu.edu

Abstract

This workshop focused on applying new data analytics to address the needs of the most vulnerable populations, introduce resilience in vulnerable situations, and help battle new sources of vulnerabilities. The aim was to highlight latest developments in the use of new sources of data, including web and social media, in the efforts to address the health and other needs of most vulnerable, including children, families, marginalized groups, and those at the threat of poverty, conflict, natural disaster, or epidemic risk. The workshop brought together practitioners from the humanitarian sector and researchers from around the world. Main themes from the keynotes, paper and abstract presentations, and an interactive panel concerned the implications of data in socio-technical systems. This broad topic is explored in the Keynote talks through questions regarding online systems, data gathered to measure health outcomes as well as analyses methods for policy implications. The presented papers also considered similar topics, in relation to online speech on social media, coverage of topics in online media, and the use of data and measures of different forms to measure poverty and social vulnerability. Overall, the continued relevance and infusion of topical areas along with the lively discussion during many aspects of the workshop, showed a great level of interest and research on this significant topic.

Keywords— social media, humanitarian, disaster, epidemiology, misinformation, vulnerable

Introduction

The scale, reach, and real-time nature of the Internet is opening new frontiers for understanding the vulnerabilities in our societies, including inequalities and fragility in the face of a changing world. From tracking seasonal illnesses like the flu across countries and populations, to understanding the context of mental conditions such as anorexia and bulimia, web data has the potential to capture the struggles and wellbeing of diverse groups of people. Vulnerable populations including children, elderly, racial or ethnic minorities, socioeconomically disadvantaged, underinsured or those with certain medical conditions, are often absent in commonly used data sources. The very absence of these populations in data can reveal areas of concern, indicating potential lack of access to vital technologies, and potentially being overlooked by algorithms trained on such data. The recent developments around COVID-19 epidemic makes these issues even more urgent, with an unequal share of both disease and economic burden among various populations.

Thus, the aim of this workshop was to encourage the community to use new sources of data to study the wellbeing of vulnerable populations including children, elderly, racial or ethnic minorities, socioeconomically disadvantaged, underinsured or those with certain medical conditions. The selection of appropriate data sources, identification of vulnerable groups, and ethical considerations in the subsequent analysis are of great importance in the extension of the benefits of big data revolution to these populations. As such, the topic is highly multidisciplinary, bringing together researchers and practitioners in computer science, epidemiology, demography, linguistics, and many others. Building on the success of two previous workshops on the same topic at ICWSM, this year's workshop brings a specific focus to timely areas of data and algorithm fairness. Around 35 people attended the workshop, with about 20 online, and 15 in person (it was held in a hybrid format).

Keynotes

The first keynote was given by Dr. Karrie Karahalios, who is a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign where she runs the Social Spaces Group. In her talk, Dr. Karahalios described her work in auditing technosocial systems. An example of such system is one around the housing access, where discrimination continues to be revealed. Similar discrimination can be found in the targeting of job advertising on social media. Dr. Karahalios talked further about the law suit that was filed in order to allow access to web data for the purpose of research. On the other hand, her studies found that there is a high desire of web users for control over their data and algorithm settings, but knowledge about such settings is poor, showing the need for a clearer communication to online users. Another facet of control is contestability, where users are able to express their opinions about platform be-

havior, possibly raising alarms around unfair or biased treatment, and call for improvements.

The second keynote, Dr. Elaine O. Nsoesie, is an Assistant professor and Data Science Faculty Fellow at Boston University. Her keynote concerned "Data and Health Equity" and emphasised the need to close data gaps to reveal racial, gender, income and other inequalities, which can then be used to motivate concrete policy changes. Her work focuses on how socioeconomic indicators explained the way different populations were responding to COVID interventions, using many data sources, such as search, social media, consumer reviews, remote sensing, news, and crowdsourcing. Examples of inequity in digital health include the lack of representation in datasets (like social media), poor performance of devices on non-white body types, and race adjustments in clinical algorithms. Since COVID impacts minorities differently, Dr. Nsoesie described efforts, such as the COVID Racial Data Tracker (https://covidtracking.com/race/) that advocate for standardised, fine-grained data collection that includes markers of interest. In general, she called for policies around data collection that includes race, ethnicity, and gender dimensions, develop an understanding of the types of biases in health data, disaggregate by race, age, gender, income, and other characteristics. Finally, we should consider the root causes of biases in data and algorithms in order to address them.

The third keynote, Dr. Enrique Delamonica, is a Senior Statistics Specialist on Child Poverty and Gender Equality in the Division of Data, Research and Policy at UNICEF. An economist and political scientist, he previously served as Chief of Social Policy and Gender Equality, UNICEF, Nigeria. He outlined three challenges for the data for social good, mainly concerning social policy and wellbeing, eliminating child poverty, and socio-economic change. In particular, Dr. Delamonica introduced the POZE paradigm for studying social change, which captures individual human behavior and social change, and the feedback loops in either direction. As an example, he showed the interrelationships between wellbeing indicators, including nutrition, education, water and sanitation, and health, and called for a complex model that would capture the synergistic structure of these variables.

Panel

Following the keynotes, the workshop included a panel discussion with the three invited speakers, and questions asked by the audience. The discussion began with vulnerable populations that may have been helped by increased attention, and those that may still benefit from additional data resources. Dr. Delamonica mentioned intrafamily violence and that it is a challenge for those involved to reach out during lockdowns. Dr. Karahalios mentioned strides made in hiring or admitting to universities more women, but that populations at the intersection of some demographics may not have been addressed (such as women of color). She also pointed to better definitions of geographic locations (such as "rural" and "urban") that affect the EPA protection laws around water and food safety. Further, she referred to the work by Anna Lauren Hoffmann "Where fairness fails" (Hoffmann 2019) which points to the difficulties to engaging with the target populations. This engagement is important in defining the labels or classes one would study, which is an important decision each researcher must make. Dr. Nsoesie also pointed out that she has noticed the lack of resources for diverse students that may be increasingly admitted to the university. So mentoring programs are important in filling in the gaps that may exist in knowledge or social networking.

Further, a question was asked on collecting data on populations for whom data collection has historically been adversarial or dangerous, such as the LGBTQ+ community. Dr. Karahalios confirmed that in her study, some respondents from this community were concerned about whether the researchers would be able to protect their privacy. Beyond this concern, there was also concern about using machine learning to infer people's orientations, especially in countries where such information may pose a threat to life and liberty. Thus they try to use participatory practices in their methodology, in order to involve the target community in the research process. Dr. Delamonica pointed out that, in some countries, some data collection is illegal, and some questions just cannot be asked. Once the data is collected (if it can be), then the researchers need to deal with the transparency / privacy trade-off. Dr. Karahalios mentioned the important role of community organisers, who are a valuable resource in connecting to the communities. She also cautioned against using some privacy preserving techniques, as they may result in erroneous research results. For instance, a dataset such as census data, with differential privacy applied to it, may have some of the communities disappear after anonymization. Dr. Nsoesie seconded the importance of community organisers, who were important during the COVID pandemic. The conclusion of this discussion was that "to solve global problems, one needs to go local".

Contributions

All submissions were reviewed by a multidisciplinary program committee (PC), with members in the fields of computer science, digital epidemiology, and computational social sciences. Three papers were accepted to be presented at the conference, as well as three abstracts, presented as shorter talks.

The first contributed work "The Coverage of Sexual Violence in Spanish News Media" was by Marilena Budan and Carlos Castillo, analysing sexual violence published by online news media in Spain. They collected a corpus of 120,000 messages on Twitter posted during 2020 by 13 of the most popular online news outlets in Spain, and employed a supervised classifier to detect tweets pointing to articles related to sexual violence. The authors then clustered and extracted useful information from the violent comments using regular expressions. They showed that the Spanish media covers sexual assault cases much more often than sexual harassment cases, despite the latter being more frequent. More worryingly, crimes happening at home are underrepresented in the media, and crimes happening in leisure spaces are over-represented. In general, rather than presenting a balanced view of different types of sexual violence, media outlets perpetuate and reinforce harmful preconceptions and myths. These findings were in line with the biases

described in the literature.

The second contribution, entitled, "Characterizing Anti-Asian Rhetoric During The COVID-19 Pandemic: A Sentiment Analysis Case Study on Twitter" by Juan Banda, Ramya Tekumalla, Luis Alberto Robles Hernandez, Zia Baig, Michelle Pan and Michael Wang, showed that the COVID-19 pandemic introduced a measurable increase in the usage of sinophobic comments or terms on online social media platforms. In the United States, Asian Americans have been primarily targeted by violence and hate speech stemming from negative sentiments about the origins of the novel SARS-CoV-2 virus. The authors combined publicly available resources to train a machine learning classification model that predicted sinophobic behavior. Then they applied the model to a longitudinal dataset spanning two years of pandemic related tweets predicting sinophobic behavior, overlaying their findings with news events.

The third contribution of our workshop, "Age dataset: A structured general-purpose dataset on life, work, and death of 1.22 million distinguished people" by Issa Annamoradnejad and Rahimberdi Annamoradnejad proposed a dataset providing new insights into the demographics of mortality in relation to gender and profession in history. The dataset was obtained via a framework based on machine learning and text mining models to collect, predict, and cleanse online data with a focus on age and gender. Their five-step method inferred birth and death years, binary gender, and occupation from community-submitted data to all language versions of the Wikipedia project. The obtained dataset is the largest to date on notable deceased people and includes individuals from a variety of social groups, including but not limited to 107k females, 124 non-binary people, and 90k researchers, who spread across more than 300 contemporary or historical regions.

[1]http://workshop-proceedings.icwsm.org/index.php?year=2022

The lightning presentations also brought varied yet significant contributions to different vulnerable populations and measures. For example, the first paper "Strengths and limitations of big-data derived poverty indices in Indonesia" Daniele Sartirano illustrated the process of creating maps of poverty in Indonesia. Another paper, on "Language Modeling and NLP Challenges in the Humanitarian Sector" Nicolo Tamagnone presented a new dataset for training humanitarian BERT model for classification. In the final paper presentation, Sophia Lou presented work on "Comparing United States Small Area Composite Vulnerability Indices with Mortality, Health, and Well-Being" on behalf of authors, which examined different social vulnerability indices in the United States. The main conclusion of this work illustrated how different indices correlate with different outcomes, such as mortality and mental health.

The papers presented at the workshop can be accessed at the workshop proceedings webpage.[1]

Workshop Organization

This workshop was organized by:

Yelena Mejova is a Senior Researcher at ISI Foundation, in Turin, Italy,

Kyriaki Kalimeri is a Researcher at ISI Foundation, in Turin, Italy,

Daniela Paolotti is a Senior Researcher at ISI Foundation, in Turin, Italy, and

Rumi Chunara is an Assistant Professor in the departments of Computer Science and Engineering and Epidemiology/Biostats at New York University, USA.

References

Hoffmann, A. L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22(7):900–915.