ICWSM 2021 Workshop on Data for the Wellbeing of Most Vulnerable: Promises and Challenges

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 development of data sharing agreements and monitoring infrastructure before potential disasters strike, although the data collection should be performed such that the populations are not harmed in the process, and the resulting insights should contribute to a holistic understanding of the situation and subsequent decision-making. Finally, the researchers should always keep in mind that, just because there is a lack of data on a population, does not mean there is no problem, as there are still vulnerable populations (such as those not using social media, or having restricted access to technology) who may need assistance and study, and who are currently bypassed by the "big data" efforts.

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. 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 extending the benefits of big data revolution to these populations. As such, the topic is highly multidisciplinary, and the workshop attracted a diverse audience, including academia across computer science, medicine and social sciences, as well as representation from humanitarian organizations including UNICEF and UN World Food Programme.

The use of new data sources for the benefit of underserved and vulnerable populations presents several exciting research opportunities, as well as serious challenges in terms of reachability and privacy. Many new sources have been employed to study such populations, including social media, internet searches, and app usage. Most of these are owned by companies, which provide a limited access via their APIs. However, the APIs often put limits on data collection scope, some of which are reasonable in terms of privacy preservation. There have been recent attempts of companies providing special access (or producing special-purpose datasets) in the theme of "data for good", such as Facebook's Data For Good[1] and Twitter For Good[2] initiatives. Other data, such as phone traces that are owned by telephone companies, or search queries that are usually owned by a few large service providers, may require a connection to a particular research group inside the company. In this workshop, we saw that there is an increased collaboration between companies and humanitarian sector in terms of data sharing, as well as many resources that are becoming well-organized and documented, such as those hosted by the Humanitarian Data Exchange (HDX)[3].

On the other hand, the desire for more data must not overtake the need for privacy and an even more careful consideration of possible harms to the subjects the research ostensibly is trying to benefit. For instance, as one of the keynote speakers, Claudia Cappa, pointed out during open discussion, it is possible to directly harm the subjects when attempting to learn more, such as in cases of domestic violence or other abuse. Further considerations of privacy concern minors, as well as those who may be entangled in legal or bureaucratic systems, such as migrants and wage workers. Several researchers in the workshop have mentioned requests from governments and other agencies to pro-

[1]https://dataforgood.fb.com/
[2]https://about.twitter.com/en/who-we-are/twitter-for-good
[3]https://data.humdata.org/

vide surveillance services using the computational tools discussed, which had to be handled carefully and with the utmost consideration with the welfare of the target population.

Below, we summarize the presentations and discussions that took place at the workshop. Overall agreement among attendees was that it is an exciting area of research, which is becoming ever more important. The need for establishing methodologies and data sharing agreements is greatest *before* the disasters strike, and must be addressed by all stakeholders as soon as possible.

Keynotes

The first keynote speaker, Claudia Cappa, is a Senior Adviser for Statistics and Monitoring in the Data and Analytics Section, at the UNICEF headquarters. Coming from a qualitative research background, Claudia discussed her experience with leveraging social media data to assess issues addressed traditionally via surveys and interviews. She presented a study of Twitter and Reddit data which was reviewed for cases of abusive content and children's exposure to violence during the COVID-19 pandemic. Her findings show the potential of assessing important issues of vulnerable populations online, while she pointed out the need for creating collaborations with data owners prior to a disaster or a pandemic so that we create solid baselines when an emergency occurs.

Elisa Omodei, second keynote speaker of the workshop, is the Lead Data Scientist of the Hunger Monitoring Unit at the UN World Food Programme's Research, Assessment and Monitoring division. She presented the HungerMap, a near realtime tool to map and predict food insecurity with non-traditional data. She underlined that tools like this require important financial and organisational commitment initially, however, the potentials are enormous especially during emergencies due to pandemics, political conflict or simply places where survey data are not possible to obtain.

Bridging the two worlds, Vedran Sekara, the final keynote speaker of our workshop, Assistant Professor at the IT University of Copenhagen and a Principal Researcher & Machine Learning Lead at UNICEF, underlined the potentials but also limitations of AI to address issues of the most vulnerable. He pointed out that operationalizing AI models can be very challenging, while transferability of models is hard and should not be overseen. Importantly, he stressed that not everything that matters is or can be measured (marian woods paradox), while real breakthroughs in the humanitarian sector can only be accomplished by a close collaboration with local stakeholders.

Panel

Core part of our workshop was the panel discussion which was organised in an interactive way, giving the possibility to all participants to actively ask questions and participate in the discussion. Supported by an online Q&A management system, a variety of topics were considered, especially those concerning the integration of the insights achieved via data science methods into the decision-making process of those who may provide the direct benefit to the studied populations. To insure academic research has more chances of producing real-life benefits, the panelists suggested that it should be presented not only in the academic venues, but also those organized by the humanitarian organizations, such as Humanitarian Networks and Partnerships Weeks (HNPW)[4]. These, ideally, would then result in co-designed studies by several partners.

A question about under-studied vulnerable populations brought the discussion to the inherent biases of some datasets, such as social media, which reaches only certain geographic and demographic areas. Dr. Sekara recommended that data sources that are more "egalitarian" should be considered, such as satellite imagery and phone records – data which is more likely to capture all residents of a geography, for example.

Another question concerned the ways in which big data can be harnessed in order to improve the design of surveys. One way to do this was to merge big data insights with survey data, such as that by Blumenstock et al.(Blumenstock, Cadamuro, and On 2015) that models socioeconomic status using a combination of mobile usage and phone surveys. The survey questions themselves could be designed in the light of the knowledge extracted from data, such as satellite imagery, in order to hone the questions to the lived environment of the respondents.

A large part of the debate concerned privacy consideration when such research is performed. The panelists agreed that it is best to avoid individual-level data, and to work on the aggregated statistics as much as possible. All agreed that the project aims must also align with everybody involved, such that there is no doubt that no harm will come to the subjects. Here, Dr. Cappa brought up an example of data collections that may actually trigger harm, if they are not performed extremely carefully. Finally, a participant pointed out that privacy can be not only an individual, but also a communal concern, and vulnerable communities must be considered.

Finally, the panelists were asked about summer school or internship programs that young researchers interested in the area could take advantage of. UNICEF has an internship program, and Data Analytics section may look for internships in the future, but they may be open to work jointly to have an informal collaboration to write papers. This summer there will be a summer school on Behavioral Digital Trace Data in Response to the COVID-19 Pandemic[5] also visiting PhD students. Also, Center for Humanitarian Data has summer program that welcomes young researchers[6]. Finally, the panelists encouraged people to reach out to people who work in the field directly.

In summary, the following insights emerged from the discussion:

1. Decision-making for vulnerable populations should not be automatic
2. It is fundamental to establish a collaboration with governments, statistics departments, and academics from the rel-

[4]https://www.unocha.org/humanitarian-networks-and-partnerships-weeks-hnpw

[5]https://socialcomquant.ku.edu.tr/summer-school-2021/

[6]https://centre.humdata.org/data-fellows/

evant geographic locales, a priori and not during an emergency.

3. Data holders should listen to the needs of the community and provide actionable data. Ideally, the data should be fine-grained, recent, community-specific.
4. At the same time, attention should be paid not to harm the vulnerable community by collecting data.
5. Importantly, especially in the context of vulnerable populations, just because there is no data, doesn't mean there is no problem.

Contributions

All submissions were reviewed by at least three multidisciplinary program committee (PC) members in the fields of computer science, digital epidemiology, and computational social sciences. Two full papers and three abstracts were accepted based on the quality of the rigor of analysis, results and presentation, and we provide a brief description for each contribution below.

The role of vulnerability in mediating the relationship between threat perception and the use of face masks in the context of COVID-19, presented by Emanuele Del Fava. In this study, the authors conducted a multi-country survey employing Facebook as recruitment tool. The performed stratified advertisement by demographic group, stratified by sex (2), age (4), and region. The survey assessed several aspects concerning the pandemic including threat perception. Their findings showed increased threat perception and wearing masks for people with vulnerabilities: those highly threatened were 2.25 times more likely to wear a face mask than those little to moderately threatened. The association between high perceived threat and wearing face masks is weaker among women and to a lesser extent older adults. Threat perception and wearing face masks was higher among women and other vulnerable groups. The authors find it to be a timely and cost-effective way to collect data, even performing several different studies at the same time to compare results, which would then complement findings from surveys.

Tactical Reframing of Online Disinformation Campaigns Against The Istanbul Convention, presented by Tugrulcan Elmas. The study deals with the Instanbul convention signed by 34 countries in 2011 to protect women from domestic violence. Turkey announced withdrawal from convention in 2021, triggered a campaign. The authors perform a case study of tactical reframing using Facebook Data by Crowdtangle, capturing most public pages and groups around the convention. They show a bottom up campaign against the convention which develops over time, with increased mentions of the rights of divorced men, citing homosexuality, and the joining of religious and political groups. In future work, the authors would like to understand whether the users change their narrative around the issue using Twitter.

The Role of Data-Driven Discovery in Detecting Vulnerable Sub-populations, presented by Girmaw Abebe Tadesse and Skyler Speakman. The author began by stressing the need to be careful about how we define vulnerability, arguing that it's a discovery question, not a modeling one. Thus, they propose using tools from anomalous pattern domain to detect vulnerable sub-populations in data, as a pattern detection task. Their aim was to overcome limitations of human-driven confirmatory analysis, using data-driven techniques. In several case studies, the authors illustrate how well-defined sub-populations can be found in data that have abnormal target statistic. For instance, in sub-Saharan Africa they find abnormal under-5 child mortality for a particular set of women. In another study in Ghana examining neonatal mortality, they find a particular scenario when the mortality rate is abnormally high. The work is in collaboration with Bill & Melinda Gates foundation, and code will be made publicly available.

Getting "Clean" from Nonsuicidal Self-injury: Addiction Language and Experiences on the Subreddit r/selfharm, presented by Himelein-Wachowiak. The study deals with two vulnerable populations: People with addiction and people who self-injure. The authors looked for words concerning addiction in the Reddit platform at the dedicated page r/selfharm. They used diagnostic criteria for substance use and addiction: 11 criteria which can have levels. They coded posts for symptoms of addiction, finding the top ones to be urges/cravings, escalating severity/tolerance, physically hazardous non-suicidal self injury (NSSI), consistent efforts to quit or cut back, and causing interpersonal problems. The authors conclude that clinicians who treat NSSI may want to adapt techniques from addiction treatment. The authors also point out that, in their experience, Reddit may contain more "honest" self-expression than Twitter and Facebook, but unfortunately there is a lack of informed consent for such research.

For a complete listing of papers and recordings of the presentations, check out the workshop website at https://sites.google.com/view/dataforvulnerable21

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.

Acknowledgement

We thank our workshop program committee members[7] for their helpful reviews and support.

References

Blumenstock, J.; Cadamuro, G.; and On, R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350(6264):1073–1076.

[7] https://sites.google.com/view/dataforvulnerable21