Broadly, my research interests belong to *computational social science* and *social computing*—two tangled and fast growing fields typically situated at the intersection of computer and social sciences[1]—where much interest is placed on the analysis of large-scale datasets of online traces of human activity and social interactions such as social media messages, friendship links, usage or search logs. Such "social datasets" offer captivating insights into various dimensions of human phenomena, at both individual level and large scale. With their proliferation, the study of human phenomena has seen a remarkable shift in scale and level of detail, both in breadth and depth.

However, these datasets are more than just an observational tool, being used to make inferences about the physical or mental health of users, about their hireability, income, or political views, among many other applications. Thus, *as insights derived from these datasets are increasingly being used to drive policies, to shape products and services, and for automated decision making, it is increasingly important to understand and address the limitations around their use,* particularly when tackling significant societal challenges such as climate change, discrimination, or poverty.

This is where my core interests lie: I am interested in understanding *the boundaries to what we can learn from "social datasets", engendered by data, methodological, and ethical limitations*, as a way to gain insights about *when* a set of results are generalizable, and to better inform the design of dedicated tools that work around or harness these limitations. Yet, I am particularly concerned about the ramifications of these boundaries in the context of *addressing questions about issues that disproportionately affect the well-being of particular groups of individuals*, from systemic discrimination, inequality or public health to disaster relief and global-scale issues like climate change.

Alas, many studies rely on "social datasets", conjecturing that they are *adequate*, often *as-is*, for the problem at hand with little or no scrutiny. Yet, this is *rarely* the case. My work challenges such *"adequacy" assumptions* aiming to assess and address various types of limits that surface when using "social datasets", related to both the properties of the datasets, as well as of the methods for acquiring and leveraging them: Are we collecting the right data, or, in other words, are the working datasets representative of the targeted platform data? Can we generalize observations drawn from one dataset to other seemingly similar datasets? Are the datasets representative of the real-world phenomena we are interested in? Do our tools perform similarly regardless of data biases or variations? Under what conditions can we answer affirmatively to such questions? What mechanisms in the data processing pipeline are domain-agnostic and what elements would require domain-adaptation?

While these are all important challenges when working with social data, the sheer diversity of the components involved in the pipeline makes general solutions difficult and sometimes even unachievable. To tackle them, the approach behind my research is to focus on concrete, representative instances of each problem, and derive actionable insights with broader applicability. In addition, addressing them also requires a good grasp of the application domain and an inherently interdisciplinary and collaborative approach. In this regard, I have been fortunate to collaborate with researchers with diverse backgrounds (such as systems design, information retrieval, data mining, computational journalism, and social science), and I believe that continuing to foster such collaborations will help me in pushing the envelope in tackling such problems.

Retrospective: Probing the Limits of Social Data

This section focuses on my main efforts to quantify or improve the quality of the working datasets, and to appraise when a set of results or observations are generalizable across datasets, media, or applications.

Towards a Better Data Collection Pipeline: Monitoring Social Media in Crises

A first challenge when utilizing online social and activity traces to answer questions about human phenomena is to assemble data collections satisfying certain integrity conditions (e.g., completeness, representativeness, or precision). Alas, popular data sources add further complexity to this challenge, as their APIs set additional constraints via rate limits and rigid query languages. A standard practice is to query and specify what data is relevant (or not) through a bounded list of keywords (or hashtags) or geo-coordinates, setting additional boundaries to the working datasets. In addition, working

[1] In a nutshell, the main difference stems from the focus of *computational social science* on answering questions about individuals and society through the use of large-scale (typically, online) sources of social, behavioral or demographic data, in contrast to *social computing* which is rather concerned with the study of user interactions with online systems and the design of these systems to enhance such interactions or for performance gains [10, 11].

with incomplete or inaccurate such lists may further result in data loss or misrepresentation. But, how can we build more comprehensive collections without introducing too many false positives?

To alleviate this problem, in the context of social media data collection during time-critical events, my work takes advantage of *domain knowledge* and demonstrates that using a *domain-specific*, yet *generic*, lexicon consisting of terms that tend to frequently appear across different crisis events improves the datasets' quality [2]. Beyond the application domain, our results also indicate that *the amount of data that it is currently mined represents only a fraction of the relevant data* (as low as 18% in our study, and 33% on average). Our method increases the recall of social media samples (with up to over 30%), resulting in more complete data collections. It also helps better preserving the original distribution of message types and sources, reducing in this way the likelihood of introducing further biases in the resulting datasets. We also showed how this lexicon can be used to automatically learn new event-specific terms and adapt to the targeted event by applying simple pseudo-relevance feedback mechanisms.

Generalizability and Data Biases: Social Media Use During Crises

Applying a consistent methodology to collect social media data is good practice, yet there can still be extraneous factors that impact the properties of even seemingly similar datasets. Alas, such factors are often ignored. For example, a governmental report reviewing research on social media use during crises indicates that there is a tendency "to examine one catastrophic event (...) and then imply that the findings are generalizable to other disasters" [2]. This is problematic, as one important goal of this research is to reuse existing data assessment models for future crises. To assess the extent to which observations drawn from a single event can be generalized to other events, I ran a transversal study that systematically assembled and examined 26 datasets of social media posts (on Twitter) from a variety of natural or human-induced crisis events in 2012 and 2013 [7]. This has been singled out as the *most comprehensive study of social media use during crises*,[3] and has also served as the basis for the "Social Media and Natural Disasters" chapter in the UN OCHA yearly report on World Humanitarian Data and Trends 2014.[4]

My work uncovered substantial variability across events in terms of information type (e.g., donations and volunteering, infrastructure damage) and sources (e.g., eyewitness accounts, media)—*highlighting the difficulty of generalizing observations from one dataset to another*. In some cases, the most common type of messages during one crisis was absent in another. Even two events of the same type and in the same country may look quite different vis-à-vis the information on which people tend to focus: In the Philippines, during Typhoon Yolanda in 2013, the Twitter stream was dominated by donation messages from outsiders, while during Typhoon Pablo in 2012, messages about caution and advice relayed via news media sources were most prevalent. Yet, when we looked at the data at a meta-level, our analysis also revealed *patterns regarding the type of information people are concerned with, given particular properties of an event* (e.g. being instantaneous or progressive, or impacting a small or a large geographic area). For instance, messages from government sources about caution and advice are more common in progressive than in instantaneous crises.

Online Media Biases: Events Coverage in Mainstream vs. Social Media

A key assumption of many social media-based studies is that online social or activity traces reflect real-world phenomena. For instance, empirical evidence suggests that social media communications are often driven by events in the news, with social media being increasingly integrated in the distribution and consumption of news. As a result, social media is seen both as the place to capture the public opinion to news events, but also as an important outlet for news consumption. My work has investigated *the extent to which one source of news can substitute the other:* How similar are the news that circulate on social media to those published by mainstream media? Does social media focus more or less on a certain type of news events?

To this end, my work introduces a *methodology for comparing news agendas online that is based on the comparison of spikes of coverage* [1]. To operationalize what events of interest are across different media, we defined them in relation to well-defined topics, and demonstrated how the methodology can be applied to compare the coverage of *climate change*

[2] http://reliefweb.int/report/world/social-media-use-during-disasters-review-knowledge-base-and-gaps

[3] http://irevolutions.org/2015/02/16/comprehensive-crisis-tweets-study/

[4] http://www.unocha.org/data-and-trends-2014/

related events in mainstream and social media (Twitter) using two large-scale datasets that cover a period of 17 months: a global database of about 30 million news articles (GDELT[5]) and a sample of about 2 billion tweets, corresponding to about 1% of all Twitter posts. We showed that social media significantly deviates from mainstream media, focusing more on actions by individuals, original investigative journalism, and legal actions involving governments. More broadly, it tends to pay more attention than mainstream media to news events considered ordinary, predictable, and of low-impact. Thus, while there is also some overlap between the two types of media, *social media is not always a good proxy for online mainstream news media*.

Methods Assessment: A Look at Item Recommendation

Another important area of concern is *how robust our tools or methods are to data biases and variations across and within datasets*. What are the data attributes that determine when our methods succeed or fail? Does their performance vary across different classes of users? Do current evaluation metrics appropriately reflect the tools performance across various application scenarios?

To explore these questions, I focused on the item recommendation problem—one of the longest-standing applications of social data—and conducted an extensive empirical analysis on 6 real-world datasets (including both the explicit social network among users and the collaborative annotated items), dissecting the impact of user and item attributes on the performance of recommendation strategies relying on either the social ties or past rating similarity [3]. My work demonstrates that the reliance on popular global metrics[6] to evaluate and compare the performance of various approaches is inconclusive as they provide little insights into when the approaches succeed or fail. Our results show significant performance variations across different classes of users and items, and that the different social signals captured by the same online mechanism may as well significantly impact the recommendation performance. For instance, we found that when the basis of formulating social connections among users stems from *plain* friendship, rather than from shared interests, leveraging the social ties leads to less precise recommendations. Broadly, *these results make a case for more extensive and fine-grained evaluations, not only across different datasets, but also within each dataset, across different classes of users and items.*

Current Work and Prospective: Social Issues and Data Pitfalls

Moving forward, I believe that the need to identify, quantify, and tackle current limitations around the use of social datasets will remain a persistent and important issue for years to come. However, it is also worth noting that eliminating all data biases, noise or other limitations is *unlikely*, perhaps even *undesirable*. In some cases, data biases that bound the applicability of general solutions might help boost the performance of dedicated solutions [2] or may inform system design [4]. In other cases, the solutions to various limitations might pull in opposite directions—e.g., in a user classification problem, one may be faced with the decision to err against a minority group to ensure accurate results for the majority or vice versa. It may often be unclear what are the *fair trade-offs* to make even when having a in-depth understanding of the problem domain.

In this light, I believe that the implications of various limitations are more perilous in certain instances or application domains than in others (e.g., systemic discrimination against a protected class vs. sub-optimally ranking items in a shopping list). Thus, I am interested in studying them in the context of tackling questions related to social issues, with a particular focus on the treatment and well-being of vulnerable individuals. These are the general research directions that I am passionate about, and that I hope to be able to pursue for the longer term. In the following, I elaborate in more detail a few more specific research directions in this area that I am interested in.

The Interplay of Personal and Collective Experiences, Perceptions and Outcomes

From Personal Experiences to Societal Impact. A key assumption is that by aggregating the social-behavioral cues we capture online from numerous users, we can draw broader insights about the collective makeup, predispositions

[5] http://gdeltproject.org

[6] Metrics computed or averaged over all predictions.

or actions—and, a large part of my work relies on this assumption. Moving forward, I believe there is still untapped value in applying it to understand *what is the impact on the society at large of individual experiences that are a result of e.g. systemic discrimination, inequality or poverty*. A particular direction where such insights may be beneficial is to understand when and why opinion gaps arise between the majority and a particular minority within a society. For instance, in my recent work on a movement on racial equality [8], I found preliminary evidence in favor of the observation that while there is an increasing number of discussions about minority group[7] issues, these discussions are largely held among the minority group members. Such studies can have implications for community sensing, policy-making, public relations or activists.

Cross-cultural experiences. I am also interested in the broader impact of individuals whose experience makeup is inherently cross-cultural (e.g. expats, migrants, bilinguals). The rise of such groups has important economic, social and cultural implications, and I believe there is a lot of value in studying their emergent roles in the society—e.g., whether they play a role in bridge-building across communities or cultures, or how their roles change with their position within the society (w.r.t. income, profession, etc.).

From Collective Experiences to Personal Impact. Conversely, as online traces increasingly capture rich and longitudinal data about each users' actions or opinions, we can attempt to monitor and measure the impact of collective experiences (e.g. social movements, law implementations, or disasters) at a personal level: Can we distinguish between their short- and long-term effects? Do they disproportionally impact different user demographics? When? Why?

Quantifying and Controlling for Biased Data and Other Pitfalls

As I emphasized throughout, as social media, mobile apps, and other services increasingly capture data about all aspects of users lives, as we use this data to understand human phenomena or for automated-decision making, there are important questions that arise about the accuracy, the reliability, or the fairness of the conclusions we draw from it. To this end, there are three main directions that interest me:

First, I am interested in understanding what are the main triggers that lead to biases in social datasets (e.g. functional or normative elements such as platform-specific mechanisms or conventions), and quantifying their effects (e.g. population, behavioral or temporal biases): What are the factors that make users more or less likely to share information about their experiences or opinions? Can we explain the gaps in platform feature-use across distinct classes of users? Can we better assess what kind of signals social data actually captures, what they represent or whom they represent?

Second, we need to develop frameworks and best practices for observational studies on social data: How can we efficiently distill true associations in the data from the spurious ones due to confounding and lurking factors? Can we develop off-the-shelf tools to quantify and control for various types of biases, or other data pitfalls? Can we separate data analysis elements that need domain-adaptation from those that are domain-agnostic? The project I was involved during my internship at Microsoft Research, introducing an *open-domain* framework that distills outcomes likely to be mentioned after *any* common and critical situations based on users social media posts, has undertook important steps in this direction [5, 6].

Finally, I also believe that the popular metrics we rely on to evaluate systems and frameworks that work with social data need further scrutiny. For instance, should we interpret precision and recall in the same way, irrespective of trying to classify people or cat pictures? How does the cost of an error change with the application domain? Can we classify applications based on the trade-offs they require among various performance metrics?

Overarching Goals: Data Sharing, Ethics, and Standards

I surely share the excitement about the opportunities opened by the current technological environment to answer critical questions about society or to make advancements in domains like health or education. Nonetheless, I also believe that we need to persistently and critically scrutinize existing data sources and tools, as well as our assumptions about them. Beyond specific problem instances and application domains, there are shared challenges related to data, methods, and ethics that need further attention—which I began to thoroughly document [9].

[7] By minority group, I refer to a group that is subordinate to a more dominant group in society.

Ethics and Automated Decision Making. Alas, the ethical issues when working with social datasets are often overlooked. Consistent standards across the community about how to handle such data are needed: e.g. when can we disclose user identifiers in papers? what about disclosing social media posts as-is? Related concerns regard the increased reliance on automated decision making and the risk of algorithmic discrimination which require further attention.

Standards and Data Sharing. The need for standards extends beyond data manipulation and assessment of findings. There is a need for sustained efforts to maintain repositories of datasets and tools, or develop mechanisms that ease the task of sharing tools and data. I believe data sharing is cornerstone for the replicability and reproducibility of results, but also for helping others build upon existing work. To this end, I am maintaining `CrisisLex.org`, a repository of crisis-related social media data and tools, where we release data and scripts from our studies[8], including crisis lexicons, and annotated social media datasets and other data derivatives. Our ongoing efforts go beyond our own data release, as we actively document and promote other publicly available datasets and resources, `CrisisLex.org` currently hosting and/or documenting datasets corresponding to about 40 crisis-events.

References

[1] **Alexandra Olteanu**, Carlos Castillo, Nicholas Diakopoulos, and Karl Aberer. Comparing events coverage in online news and social media: The case of climate change. In *Proc. of 9th International AAAI Conference on Web and Social Media (ICWSM'15)*, 2015.

[2] **Alexandra Olteanu**, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In *Proc. of 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, 2014.

[3] **Alexandra Olteanu**, Anne-Marie Kermarrec, and Karl Aberer. Comparing the predictive capability of social and interest affinity for recommendations. In *Proc. of 15th International Conference on Web Information Systems Engineering (WISE'14)*, 2014 **(Best Paper Award)**.

[4] **Alexandra Olteanu** and Guillaume Pierre. Towards robust and scalable peer-to-peer social networks. In *Proc. of the Fifth EuroSys Workshop on Social Network Systems*, 2012 **(Best Paper Award)**.

[5] **Alexandra Olteanu**, Onur Varol, and Emre Kıcıman. Towards an open-domain framework for distilling the outcomes of personal experiences from social media timelines. In *Proc. of 10th International AAAI Conference on Web and Social Media (ICWSM'16)*, 2016.

[6] **Alexandra Olteanu**, Onur Varol, and Emre Kıcıman. What does social media say about the outcomes of personal experiences? In *Proc. of the 20th ACM Computer Supported Cooperative Work and Social Computing (CSCW17)*, 2017.

[7] **Alexandra Olteanu**, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proc. of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'15)*, 2015.

[8] **Alexandra Olteanu**, Ingmar Weber, and Daniel Gatica-Perez. Characterizing the demographics behind the #BlackLivesMatter movement. In *AAAI Spring Symposia on Observational Studies through Social Media and Other Human-Generated Content (AAAI SSS'16)*, 2016.

[9] Carlos Castillo, Fernando Diaz, Emre Kıcıman, and **Alexandra Olteanu** *(alphabetical order)*. A critical review of online social data: Limitations, ethical challenges, and current solutions. In *Tutorials at 10th International AAAI Conference on Weblogs and Social Media (ICWSM'16)*, 2016.

[10] Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach. Computational social science and social computing. *Machine Learning*, 95(3):257, 2014.

[11] Andre Oboler, Kristopher Welsh, and Lito Cruz. The danger of big data: Social media as computational social science. *First Monday*, 17(7), 2012.

[8] All data releases are done according to the ToS of the social media platforms from which the data was collected.