Probing the Limits of Social Data: Biases, Methods, and Domain Knowledge

THÈSE N^{O} 6892 (2016)

PRÉSENTÉE LE 22 JANVIER 2016

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE SYSTÈMES D'INFORMATION RÉPARTIS
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Alexandra OLTEANU

acceptée sur proposition du jury:

Dr P. Pu Faltings, présidente du jury
Prof. K. Aberer, directeur de thèse
Dr C. Castillo, rapporteur
Dr E. Kiciman, rapporteur
Dr A.-M. Kermarrec, rapporteuse

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2016

Abstract

Online *social data* has been hailed to provide unprecedented insights into human phenomena due to its ability to capture human behavior at a scale and level of detail, both in breadth and depth, that is hard to achieve through conventional data collection techniques. This has led to numerous studies that leverage online social data to model or gain insights about real world phenomena, as well as to inform system or methods design for performance gains, or for providing personalized services.

Alas, regardless of how large, detailed or varied the online social data is, there are limits to what can be discerned from it about real-world, or even media- or application-specific phenomena. This thesis investigates four instances of such limits that are related to both the properties of the working data sets and of the methods used to acquire and leverage them, including: (a) online social media biases, (b) assessing and (c) reducing data collection biases, and (d) methods sensitivity to data biases and variability. For each of them, we conduct a separate case study that enables us to systematically devise and apply consistent methodologies to collect, process, compare or assess different data sets and dedicated methods.

The main contributions of this thesis are:

(i) To gain insights into *media-specific biases*, we run a comparative study juxtaposing social and mainstream media coverage of domain-specific news events for a period of 17 months. To this end, we introduce a generic methodology for comparing news agendas online based on a comparison of spikes of coverage. We expose significant differences in the type of events that are covered by the two media.

(ii) To assess possible *biases across data collections*, we run a transversal study that systematically assembles and examines 26 distinct data sets of social media posts during a variety of crisis events spanning a 2 years period. While we find patterns and consistencies, we also uncover substantial variability across different event data sets, highlighting the pitfalls of generalizing findings from one data set to another.

(iii) To *improve data collections*, we introduce a method that increases the recall of social media samples, while preserving the original distribution of message types and sources. To

locate and monitor domain-specific events, this method constructs and applies a domain-specific, yet generic lexicon, automatically learning event-specific terms and adapting the lexicon to the targeted event. The resulted improvements also show that only a fraction of the relevant data is currently mined.

(iv) To test the *methods sensitivity*, to data biases and variability we run an empirical evaluation on 6 real-world data sets dissecting the impact of user and item attributes on the performance of recommendation approaches that leverage distinct social cues—explicit social links vs. implicit interest affinity. We show performance variations not only across data sets, but also within each data set, across different classes of users or items, suggesting that global metrics are often unsuited for assessing recommendation systems performance.

The overarching goal of this thesis is to contribute a practical perspective to the body of research that aims to quantify biases, to devise better methods to collect and model *social data*, and to evaluate such methods in context.

Keywords: Data biases, evaluation, social media, crisis computing, recommendation systems, data collection, domain knowledge

Abstract (Italian)

Ai *dati sociali* on-line é ormai riconosciuta la capacità di fornire approfondimenti senza precedenti riguardo i fenomeni umani, data la loro capacità di catturare comportamenti umani ad una scala e un livello di dettaglio, sia in ampiezza che in profondità, difficili da ottenere attraverso tecniche di collezione dei dati tradizionali. Questo ha portato numerosi studi che fanno leva su tali dati on-line per modellare o acquisire conoscenza sui fenomeni del mondo reale, allo stesso modo per dare informazioni sui sistemi o progettare metodi allo scopo di incrementare le prestazioni, o per fornire servizi personalizzati.

Purtroppo, indipendentemente da quanto grandi, dettagliati e variegati siano i dati sociali on-line, ci sono limiti riguardo cosa può essere dedotto da essi riguardo fenomeni del mondo reale, o anche riguardo fenomeni specifici dei media e delle applicazioni. Questa tesi investiga quattro istanze di tali limiti che sono relativi sia alle proprietà delle collezioni di dati prese in esame, sia ai metodi usati per acquisirli ed utilizzarli, includendo: (a) polarizzazioni degli on-line social media, (b) valutazione e (c) riduzione delle polarizzazione nella collezione dei dati, e (d) sensitività dei metodi alla polarizzazione e alla variabilità dei dati. Per ognuno di essi, condurremo un caso di studio separato che ci permetterà di concepire sistematicamente e applicare metodologie consistenti per collezionare, processare, confrontare o valutare sia collezioni di dati differenti che metodi dedicati.

I principali contributi di questa tesi sono:

(i) Per acquisire conoscenza nel merito delle *polarizzazioni specifiche dei media*, effettueremo uno studio comparativo giustapponendo la copertura dei social media con quella dei media tradizionali riguardo notizie ed eventi specifici dal dominio per un periodo di 17 mesi. A tale fine, introdurremo una metodologia generica per confrontare on-line le notizie basata sul confronto dei picchi di copertura. Esporremo differenze significative nel tipo di eventi che sono coperti dai due media.

(ii) Per valutare le possibili *polarizzazioni tra collezioni di dati*, eseguiremo uno studio trasversale che assembla ed esamina sistematicamente 26 collezioni di dati distinte, contenenti messaggi provenienti dai social media durante una varietà di eventi di crisi che coprono

un periodo di 2 anni. Mentre troveremo schemi ricorrenti e consistenze, scopriremo una sostanziale variabilità tra differenti collezioni di dati riguardo eventi, evidenziando le insidie nella generalizzazione dei risultati da una collezione di datiad un'altra.

(iii) Per *migliorare le collezioni di dati*, introdurremo un metodo che migliora il recupero dei campioni dei social media, preservando al contempo la distribuzione originale dei tipi di messaggi e delle sorgenti. Per localizzare e monitorare eventi specifici del dominio, questo metodo costruisce ed applica un lessico che sia al contempo specifico del dominio ma generico, imparando automaticamente termini che sono specifici dell'evento ed adattando il lessico all'evento mirato.

(iv) Per valutare la *sensitività dei metodi* alla polarizzazione e la variabilità dei dati, eseguiremo una valutazione empirica su 6 collezioni di dat provenienti dal mondo reale, sezionando l'impatto degli attributi dell'utente e degli elementi sulle prestazioni degli approcci raccomandati che fanno leva su indizi sociali distinti—collegamenti sociali espliciti vs. affinità implicita su interessi. Mostreremo variazioni di prestazioni non solo tra collezioni di dati differenti, ma anche all'interno di ogni data set tra differenti classi di utenti o elementi, suggerendo che le metriche globali sono spesse inadatte per valutare le prestazioni dei sistemi di raccomandazione.

L'obiettivo generale di questa tesi è quello di contribuire una prospettiva pragmatica al campo della ricerca che ha come scopo quello di quantificare le polarizzazioni, di concepire metodi migliori di raccolta e modellazione dei *dati sociali*, e di valutare tali metodi nel proprio contesto.

Parole chiavi: Polarizzazione dei dati, valutazione, social media, raccolta dei dati, conoscenza di dominio

Acknowledgments

"*With freedom comes responsibility. For the person who is unwilling to grow up, the person who does not want to carry his own weight, this is a frightening prospect.*"
—Eleanor Roosevelt, in You Learn by Living (1960)

First and foremost, I'd like to thank my advisor, Karl Aberer, for giving me the opportunity to do the PhD under his supervision. His patience, the trust, and the freedom he has granted me over these years have helped me immensely. I was able to pursue the topics I felt passionate about and I am grateful for it. This has often pushed me out of my comfort zone and has helped me grow.

I'd also like to thank my thesis committee members, Carlos Castillo, Anne-Marie Kermarrec, Emre Kıcıman, and Pearl Pu for agreeing to be part of my committee, for their time and effort to review this manuscript, as well as for their valuable and constructive feedback.

Throughout my PhD years I was extremely fortunate to work with and be mentored by many incredibly sharp and wonderful people. They have not only influenced my work and research taste, but have also influenced my world view. For my first year of PhD, I am grateful to Luke McDowell for research discussions, and to Guillaume Pierre (my master thesis advisor) for his advice and support. This has likely been the period most filled with moments of doubt, and their support helped me greatly. The doubts have not abandoned me in the following years, but the trust and continued support from Anne-Marie Kermarrec, along with her contagious optimism has kept me focused and motivated over my 2nd year.

In the mid of second year, an email from Ingmar Weber (thank you!) introducing me to Carlos Castillo was a truly fortunate stroke of serendipity. I owe Carlos some of the best times in my PhD and my professional growth. I am grateful to him for looking after me since then, and for everything he has thought me. My gratitude also goes to Fernando Diaz—I learned greatly from working with him, and I am thankful for his support. I was also fortunate to work with Emre Kıcıman, which has been one of the main highlights of my PhD from which I gained so much. I also want to thank my co-authors, Nick Diakopoulos, Daniel Gatica-Perez, Ingmar Weber,

and Sarah Vieweg, for a fruitful collaboration. I am grateful to Patrick Meier for his leadership. I was also fortunate to mentor, work with and learn from a group of highly talented students, including Stanislav Peshterliev, Anton Dimitrov, Zhicong Huang, and Büsser Alexander.

I spent most of my time as a PhD student in the lab, and I am grateful to my current and former LSIR colleagues for many moments of laughter and learning. I thank Berker Agir, Martin Benjamin, Burak Zeydan, Jean-Paul Calbimonte, Michele Catasta, Alex Constantin, Alevtina Dubovitskaya, Julien Eberle, Tian Guo, Amit Gupta, Hamza Harkous, Hoyoung Jeung, Nataliya Kryvych, Zoltan Miklos, Rammohan Narendula, Thanh Tam Nguyen, Julia Proskurnia, Rameez Rahman, Jean-Eudes Marie Ranvier, Sofiane Sarni, Saket Sathe, Matteo Vasirani, Zhixian Yan, Surender Reddy Yerva, Hao Zhuang. Many thanks go to Tri-Kurniawan Wijaya for discussions about work and life, for his endless energy, joy and friendship, but also for putting up with me as an office mate. I am also grateful to Mehdi Riahi and Quoc Viet Hung Nguyen for being great listeners and wonderful friends. A very special thank goes to Chantal François for her endless administrative and logistic support, but also for being always positive and a friend.

I'd also like to thank my EPFL friends for coffee breaks, long discussions, dry-runs and outings. Many thanks to Yu Chen, Sarvenaz Choobdar, Pamela Delgado, Florin Dinu, Mihai Dobrescu, Ana Petkovska, Amitabha Roy (and Mohita) and Valentina Sintsova. I am also really grateful to DSLAB for often treating me as a "family" member, including Silviu Andrica, Radu Banabic, George Candea, Amer Chamseddine, Vitaly Chipounov, Loïc Gardiol, Baris Kasikci, Johannes Kinder, Volodymyr Kuznetsov, Georg Schmid, Ana Sima, Cristian Zamfir and Jonas Wagner.

During my internships at Qatar Computing Research Institute and Microsoft Research Redmond I met many people that have made my experience so much wonderful. The home-cooked dinners, the dune bashing, the night safari, the happy hours, among many others, have made my time in Qatar unforgettable. I owe these to Pranay Anchuri, Kiana Calagari and Tarek Elganainy (for the laughter), Tarek M. El Gamal, Kiran Garimella, Muzammil Hussain, Suin Kim (for being the best "suffering" partner), Ji and Liliann Lucas (for the nights-over and being great friends), Fabiola Leyton, Yelena Mejova, Imran Muhammad, Kenton Murray, Nizi Nazar (for her care and kindness), Zhongyu Wei, Paco Guzman, and so many others. Likewise, my time at Microsoft Research was also filled with laughter and equally memorable moments because of many friends: Ozlem Aslan (for our long talks), David Graus (for endless debates), Victoria Lin, Bernhard Reinert and Yue Wang (for their friendship and kindness), Ke Tran (for understanding and laughing at my jokes), and Onur Varol, but also to my romanian friends Gentiana Coman (for always being there for me) and Andreea Sandu.

Many thanks also go to Matt Fitzgerald, Matthew Richardson, Jisun An, Haewoon Kwak, He-

mant Purohit, Aron Culotta and Patrick Meier for advice, feedback or data that has enabled parts of the work presented in this thesis. I am grateful to Francesco Fucci and Michele Catasta for their availability and help with the translation of this thesis abstract to Italian. My PhD work was financially supported by the grants *Sinergia (SNF 147609) Grant* and *Reconcile: Robust Online Credibility Evaluation of Web Content* from Switzerland through the Swiss Contribution to the enlarged European Union.

I am indebted to my parents for always believing in me, for their endless love, for being there when I needed them, and for forgiving my absence in so many difficult moments for them. This journey would have been so much difficult without Stefan's continued love and support. Over the years he has assumed all the roles I needed, being a partner, a friend, a mentor, a confidant, a cheerleader and so often the shoulder to cry over.

I deeply grateful to so many people, and I apologize in case I forgot to include anyone.
Thank you so much!

Contents

List of Figures

List of Tables

1. Introduction

> *"We live in a time when big data will transform society. Or so the hype goes."*
> —Boellstorff, 2013 [38]

And the hype goes a long way. Yet, the excitement around the potential of *big data*[1] has compelling arguments: It provides information at a scale and level of detail, both in breadth and depth, that would be hard to achieve through conventional data collection techniques, such as surveys and user studies [39]. This breadth, depth, and scale opened unprecedented opportunities to provide insights and to answer significant questions about society, policies, or health, by analyzing digital traces, social media interactions, query logs, health logs, and government records, among many other data sources [27, 39, 89, 130, 194, 268, 274, 311].

Social "Big" Data

Yet, while *big data* can come from a multitude of sources and can be used in a variety of applications, much interest is placed on the so called online usage, social or behavioral data [39, 130]—or, in other words, on the "*found data*", as Harford [130] calls it. This data typically includes digital traces produced by (or about) users, being often hailed to provide insights into how people communicate, connect, behave, what they like or whom they trust [119, 182, 194, 311]—and it is what interests us in this thesis. Thus, here, we make a distinction between this sort of data and other types of *big data*, such as the ones drawn from the Large Hadron Collider's experiments, from genetics, environmental sciences or astronomy [130, 141], and throughout the thesis we refer to it as *social data*, which we consider to be a broader umbrella concept.

The attention around online *social data* has particularly grown with the proliferation of "*a class of web sites and applications in which user participation is the primary driver of value*" [125], referred to as the *Social Web*. To highlight the collective and the user-driven nature of such data, researchers have coined a variety of terms to refer to it including "human traces", "usage data", or "wisdom of crowds" [25, 26, 89]—see Table 3.1 for a more comprehensive sample of references to social data. The core idea is that this sort of data can be used to understand both

[1] A quite poor, vague term [27, 130, 39].

Table 1.1.: Examples of references to *social data* in previous work.

user generated content [55]	behavioral logs [89]	social media data [22]
usage data [26] or logs [325]	personal data streams [299]	wisdom of crowds [25]
database of human activities [311]	crowd-sourced data [104]	activity tracking data [299]
social web [125]	query logs [268]	digital traces [39]

individual-level behavior and large human phenomena, as well as to offer users with personalized services tailored to their needs.

The diversity of social *platforms*—from recommendation [207] to social media sites [220], of *purposes*—from finding information [326] to keeping in touch with friends [191], as well as of *data points* meanings and semantics (e.g., clicks, likes, shares, social links) [311], has led researchers to explore the potential benefits of these data for a variety of domains and applications—from providing affected populations or response agencies with actionable information during crisis situations [262] to observing the variations in culinary preferences across geographical areas [6]. For instance, in the context of medical domain, the utility of *social data* has been probed by research investigating how to improve or replace traditional systems for detecting disease spread with either social media posts [190, 275] or search logs [114], for tracking suicide risk factors [157], for discovering drug side-effects [337], or for identifying recent mothers at risk of postpartum depression [73]. Social data can, in fact, address the issue of finding enough "participants" to conduct a sizable study that has often been an important impediment in fields like medicine or political science [268]. It can also allow an exhaustive comparison of users across groups or individuals being often used to personalize services, even if this means providing tailored health advice [299] or what movie to watch [33].

The Limits of Social Data

However, regardless of how large or varied the working data sets are, there are significant ethical and functional limitations to what can be discerned from *social data* about real-world phenomena (online or offline), or even about media or application dependent phenomena—which have yet to be rigorously addressed [274]. Additionally, while the properties of the data sets might vary, there are shared challenges independent of the peculiarities of the data source or application, the platform from which they are collected or the type of data that is collected—albeit some of them might be more prominent within specific contexts or for certain stakeholders. For instance, the data sets might not reflect the relevant offline or online populations in their entirety [311] since, often, different demographics tend to be drawn to different social platforms

leading to important population biases [274]. Additionally, the boundaries of the analyzed data sets are often set by a handful of keywords used to query them [311], or the data might not even be available since users are more likely to share information about their positive and extreme experiences, rather than about their average or negative ones [126, 170].

A Case-Study Driven Approach

The overarching goal of this thesis is to explore various types of limitations or biases that surface when working with *social data* across or within media, but also semantic or application domains, in order to quantify them and to provide insights into how they can be leveraged to build dedicated tools (as opposed to general purpose ones). To do so, we conducted four case studies, each focusing on a well defined working domain—social media in crises, climate change news, and recommendation systems—for which the required data is accessible, and that are representative of other domains, allowing us to explore various types of challenges that arise when working with *social data*.

Such a case-study driven approach is akin to the "model organisms" approach in biology that refers to the practice of selecting a few species that are widely studied due to various experimental advantages (e.g. accessibility, ease to obtain and maintain, short life-span), in order to understand fundamental biological phenomena [311].[2] While there are limits to such an approach—as the "model organism" (or, in our case, the working domain) might under- or over-represent their kind—by allowing researchers to focus on a common set of problems, tools and data, it facilitates a better understanding of basic, fundamental mechanisms and properties of their taxa.

1.1. Research Problems and Contributions

In this section, we formulate a set of broad research problems (RPs) concerning various challenges to leveraging online social traces in order to understand or predict human behavior, within the context of which we frame the contributions of this thesis. For each of them, we highlight the specific contexts (defined by domains and applications) in which we study them, taking what Tufekci [311] calls a "model organisms" approach.

There is a growing body of work [311, 64, 274, 39, 130, 121] that raises concerns about current practices of using online *social data* that have been used to answer a variety of complex

[2] We note that Tufekci [311] has made this comparison with respect to the dominance of Twitter as the main social platform of study, yet, we argue that the similarity also holds for popular application domains.

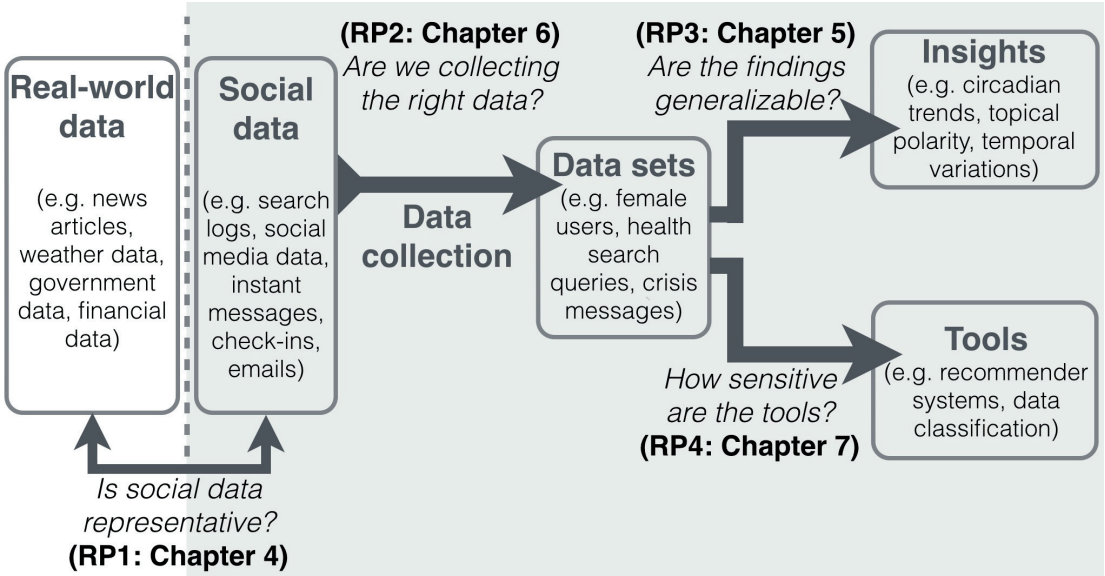


Figure 1.1.: Overview of the relationships between the research problems we formulate in this thesis. By real-world data we refer to data that can be obtained from both offline and/or online sources.

questions about human behavior, as well as to offer commercial services to their users (e.g. recommendations of places [238] or friends [233], prediction of election results [242], modelling of opinions [158]). The research problems we formulate here are grounded in this literature, which we categorize in three main classes based on previous work [311, 274, 39]: (1) *data collections*—stressing on issues of the working data sets, such as representativeness, biases or completeness; (2) *methods*—focusing on challenges related to the design, evaluation and accountability of methods that work with social data; (3) *privacy and ethics*—highlighting various ethical caveats such as avoiding discriminatory treatment or protecting users identity. We emphasize on the first two classes, while we discuss the third when we lay out the context of the research problems we address here, and at the end of the thesis.

RP1 (Data Collections): Social Media Biases. The prevalence of only a few media platforms for the study of human behavior without appropriately considering their structural biases[3] [311] has led to important concerns about what can be legitimately inferred from such online social traces. As Tufekci [311] emphasises, the focus on a certain platform it is not by itself inappropriate, yet, more effort needs to be put into understanding the behavioral norms that are specific to each online social medium [274]. Examples of using social traces to analyse or predict real-world phenomena include the spread of influenza [130], the life cycle of news events [51], the

[3]By structural biases we refer to biases as a result of platform-specific mechanisms that shape user behavior.

use of online platforms for advocacy [280], during crisis events [150], or chatter about weather events [170], just to name a few. Yet, *how reflective is the online behavior within a given social medium of other social media or of offline, real-world phenomena? Can the findings from a medium be generalized to other media?*

Contextualizing the problem. Empirical evidence suggests that social media communications are predominantly inspired by events in the news [309]; and, indeed, in this context, an important area of inquiry has been the way in which online social traces—with a focus on social media—echoes various types of events from sports [166] or economic events [273] to street movements [294] and other crisis situations [150], as it is believed that it provides cues about their impact [309].

Following these observations, our investigation focuses on events covered by mainstream media to understand how accurately social data mirrors them, and we are particularly interested in the following questions: *How much does social media reflect the news events covered by mainstream media? Does it focus more on a certain type of news events? What are the characteristics of those news events?*

To tackle these questions, we devise a *methodology for comparing news agendas online* that is based on the comparison of spikes of coverage. To operationalize what events of interest are across different media, we define them in relation to well-defined topics. To this end, for our investigation, we focus on *climate change*, and we use this methodology to compare the coverage of climate change related events in social and mainstream media over a period of 17 months. While our study covers only one social media source, Twitter—a large one and that is frequently associated with news [188]—our methodology helps to uncover a series of differences in the type of events covered in the two media. This work is discussed in Chapter 4 and has been published in:

> [244] *Comparing Events Coverage in Online News and Social Media: The Case of Climate Change.* Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, Karl Aberer. In Proceedings of 9th International AAAI Conference on Web and Social Media (ICWSM'15), Oxford, UK, May 2015.

RP2 (Methods): Improving Data Collections. However, even when running a study that is clearly confined within the context of a given platform, e.g. the representativeness, the completeness, or the precision of the working data sets with respect to the overall platform's data has been challenged [311, 121, 142]—these being often referred to as *sampling or self-selection biases* [234, 235]. Notably, for social media studies, a lot of emphasis is put on API limitations

(or data access limitations) [272, 121] and the problematic reliance on hashtag-based sampling [213, 311, 44]. A key issue is that the choice of keywords and hashtags "*is equivalent to specifying the boundaries of a data collection: working with the wrong list of keywords might cause relevant data to be missed*" [121]. Additionally, for ongoing efforts to advance theories on, e.g., how online platforms facilitate collective action, these practices and limitations have theoretical implications altering our understanding of how these technologies are used [64, 121].

Contextualizing the problem. Again, these issues have often been raised and investigated in the context of studying social platforms use during various events [41] such as natural disasters [64], protests [311], revolutions [235], elections [108], or political uprisings [121]. In the case of such studies, these issues are particularly important as, due to their sensitive nature, mistakes can be costly—e.g. inaccurate predictions, due to missing or faulty data, of disease spread or stock markets evolution can led to public frenzy, miss- or over-preparation, or to even losing assets [64, 130, 193].

We follow this direction and focus on crisis situations, being interested in the following questions: *Can we build more comprehensive collections without introducing too many false positives? Can we build collections that are representative of the overall platform data?*

To this end, our strategy is to take advantage of *domain knowledge* and show that using a domain specific, yet generic, lexicon containing terms that tend to frequently appear across various domain specific events we can improve the quality (with emphasis on recall) of the working data sets. We describe a systematic and general method to build the lexicon using existing data samples and crowdsourced labeling. We evaluate it using several data sets of social media communications during different crisis situations, and show that it leads to better trade-offs between precision and recall than when obtained with crisis-specific keywords manually chosen by experts. We also show that it helps to preserve the original distribution of message types. Chapter 5 details this study, which has appeared in:

> [245] *CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises.* Alexandra Olteanu, Carlos Castillo, Fernando Diaz and Sarah Vieweg. In Proceedings of 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14), Ann Arbor, US, June 2014.

RP3 (Data Collection): Data Collection Biases. Applying a consistent methodology to collect and sample data is good practice, yet there can still be confounding, external factors that impact the properties of the working data sets. Thus, to test the robustness of findings and to understand the generalizability of observations one should measure the social phenomena or methods on

multiple *distinct* data sets [274, 41, 323, 110, 103, 323].

Contextualizing the problem. Referring to research on social media use during disasters, the literature review by Fraustino et al. [103] indicates that it *"tends to examine one catastrophic event (...) and then imply that the findings are generalizable to other disasters."* This is particularly problematic as one important goal of this research is to reuse existing data assessment models for future disasters, yet research has shown that, e.g., prediction models do not generalize well from one data set to another, not even when the two data sets share common properties [152]. To this end, the questions that we are looking at here are: *Can we generalize observations from one data set to other data sets? What are the similarities and differences among the observed patterns across working data sets according to extrinsic properties of these data sets (e.g. type of event, duration, geographical spread)?*

To study these questions, we analyze social media use during 26 crisis events and unveil a set of challenges and opportunities related to the generalization of findings from one event to another. Our systematic examination of a diverse set of crisis events uncovers substantial variability across events, as well as patterns and consistencies. When automatically grouping events based on similarities in the distributions of different classes of tweets, we observed that despite the variability, similar events tend to be more similar to each other also in terms of the distribution of information sources and types. This work is covered in Chapter 5, and the results were published in:

> [249] *What to Expect When the Unexpected Happens: Social Media Communications Across Crises.* Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. In Proceedings of 18th ACM Computer Supported Cooperative Work and Social Computing (CSCW'15), Vancouver, BC, Canada, March 2015.

RP4 (Methods): Assessing Methods Sensitivity. Another area of concern is how to reliably and systematically evaluate tools and algorithms to account for the biases across and within data sets, with some researchers advocating for testing the robustness of findings and showing results for more than one data collection or across different classes of data items in a collection [36, 274, 281, 311]; for adapting to platform changes (e.g. users might change how they interact with the platform due to functional changes such as a new strategy for ranking items or a new ability to share content) [75, 274, 323]; or for developing and using standard evaluation metrics when they do not exist [83].

Contextualizing the problem. One of the most popular (and long-standing) applications that leverage online social behavioral traces are the recommendation systems [305], which today

are inescapable in a wide range of web applications, e.g. Amazon or Netflix, to provide users with books or movies that match their interest. Particularly, the rise of the Social Web [125] has created new prediction opportunities, resulting in an ever-increased integration of a rich array of online social cues by the recommendation systems—cues that are acquired from users either explicitly (e.g., through friend lists, ratings or reviews) or implicitly (e.g., search logs, social interactions logs, visited websites) [36, 261]. Thus, given the diversity of social signals and their triggers (e.g. even explicit social relations might stem from friendship, shared interests, or trust), to understand if the conclusions about a certain recommendation strategy generalize beyond the context of a certain data set, it is important to run the experiments on distinct data sets, in order to understand their properties and how they impact performance [281]. As a result, the questions we explore here are: *Are the recommendation strategies performing similarly regardless of data biases or variations? Are global metrics (i.e. metrics aggregated over all data points) able to reflect the performance of a given recommendation strategy across various settings (e.g. different application domains, platforms or user demographics)? Are there specific data attributes that hint at the performance of one strategy with respect to another?*

To answer these questions, we conducted an extensive empirical analysis on 6 real-world publicly available data sets (including both the explicit social network among users and the collaborative annotated items), which dissects the impact of user and item attributes, such as the density of social ties or item rating patterns, on the performance of recommendation strategies relying on either the social ties or past rating similarity. Our results indicate that one cannot rely on global metrics to assess a given recommendation system performance not only across data sets, but also within each data set, across different classes of users or items. For instance, we see that when the basis of formulating connections among users stems from *plain* friendship, rather than from shared interests, the recommendation strategy relying on the social ties leads to less precise recommendations. In Chapter 7 we describe this study that has appeared in:

> [246] *Comparing the Predictive Capability of Social and Interest Affinity for Recommendations*. Alexandra Olteanu, Anne-Marie Kermarrec, and Karl Aberer. In Proceedings of 15th International Conference on Web Information Systems Engineering (WISE'14), Thessaloniki, Greece, October 2014 *(Best Paper Award)*.

We note that, while we can dissociate these research problems, they are often contingent on each other—e.g. the understanding of the biases of the working data sets can guide a tool evaluation, which, in turn, can help re-designing it to account for them. Equally important, as we discuss our case studies in detail, we also cover (although, with a lesser emphasis) a few other important (yet, orthogonal) challenges: (1) the use of *domain knowledge* to contextualize the problems

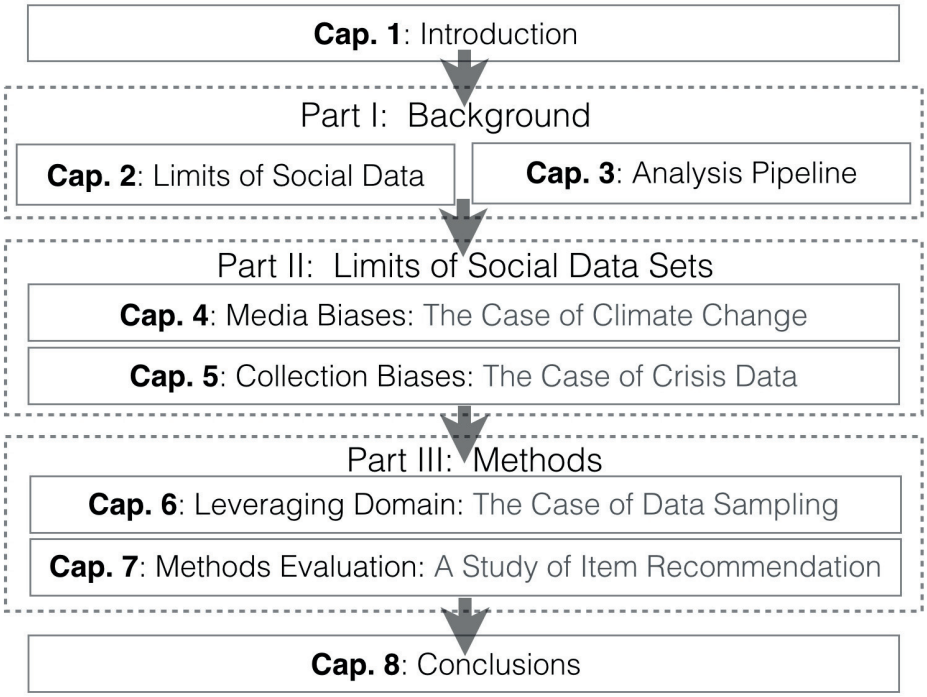


Figure 1.2.: Overview of the structure and the conceptual flow of this thesis.

and improve performance (a strong leitmotif of this thesis), and (2) *data sharing* which impacts the ability to run our analyses on multiple data sets, as well as to ensure and support their reproducibility and replicability.

Finally, we emphasise that, in this thesis, we seek to study the research problems described in this section in well defined contexts, highlighting the limitations of current approaches and outlining recommendations for studies that share similar challenges. We do not attempt to test or validate them across all existing platforms, algorithms or domains. We discuss in more depth the classes of challenges for the analysis of online social traces in the following chapters. This thesis contributes *a practical perspective to the body of research that aims to quantify biases, to devise better heuristics to collect and model social behavioral data, and to evaluate algorithms in context*.

1.2. Thesis Outline

We now give an overview of the thesis structure and highlight how the Chapters are grouped and how they relate to the research problems we have formulated in the previous section. The conceptual flow of the thesis is depicted in Figure 6.2, while the high-level relations between the research problems are highlighted in Figure 1.1. The thesis is organized in three parts, with **Part I** covering the broad context of the research problems we address in this thesis, **Part II** describing two case studies that characterize social data sets by contrasting them with real-world data or with each other (**RP1: Media Biases** and **RP3: Collection Biases**), and, finally, **Part III** focusing on methodological challenges related to data acquisition and methods evaluation (**RP2: Improving Data Collections** and **RP4: Methods Assessment**).

Part I: Background

Chapter 2 lays out the broad context in which we frame the research problems we tackle in this thesis, by surveying relevant prior work that raises concerns and scrutinizes the limits around the use of social data.

Chapter 3 briefly surveys social applications and platforms, along with examples of what kind of data is collected from users. It also broadly describes the prototypical pipeline for social data analysis.

Part II: Limits of Social Data Sets

Chapter 4 looks at media biases, exploring how accurately social data mirrors real-world data, by focusing on events covered by mainstream media. For this, it introduces a methodology for comparing news agendas online based on the comparison of spikes of coverage.

Chapter 5 studies to what extent observations made based on a single data set can be generalized to other similar data sets. To this end, it appraises the similarities and differences in social media communications that take place during different crisis events, according to specific characteristics of such events.

Part III: Methods

Chapter 6 investigates how we can improve social data sets at collection time. It introduces an approach for attaining more representative and comprehensive collections without introducing too many false positives.

Chapter 7 tests the methods generalization power and their sensitivity to social data sets biases and variability. By focusing on recommendation systems, it shows that the relative performance of different methods varies not only across data sets, but also within each data set, across different classes of users or items.

Chapter 8 concludes the thesis by summarizing the main contributions and outlining possible directions for future work.

Part I.

Background

2. On the Limits of Using Online Social Traces

> "*We must ask difficult questions of Big Data's models of intelligibility before they crystallize into new orthodoxies.*"—boyd and Crawford, 2012, [39]

This chapter lays out the broad context in which we frame the research problems we tackle in this thesis by surveying relevant prior work. Here we focus on limits and concerns around the use of social data as they are raised by previous work, while leaving to the Chapter 3 the discussion about the typical applications of *social data* and the prototypical analysis pipeline employed by such applications.

As the use of *social "big" data* flourishes as an area of inquiry about various dimensions of human behavior, the community has also started to ask important questions about good practices and the *limitations* of using social data sets originating from e.g. social media platforms, search engines, recommendation sites or location-based services, among others. Core issues include data biases as a result of collection or sampling strategies [274], the lack of coverage or representativeness with respect to the targeted populations [39, 121, 311], data access limits [41], biases due to platform and media specific norms [274], algorithmic stereotyping and profiling of users [57], or privacy risks [320], to name a few. We organize such issues highlighted by prior work in three main classes based on [39, 274, 311]:

(1) *Limits of data collections*: surveying existing challenges when working with social data sets such as representativeness, validity, population and sampling biases, completeness, or temporal variations (Section 2.1);
(2) *Methodological challenges*: surveying issues related to the design, evaluation, or generalizability of analyses or methods for collecting or utilizing social data set (Section 2.2);
(3) *Privacy and ethics*: discussing various ethical caveats when leveraging social data such as discriminatory treatment as a result of algorithmic reinforcement of human prejudice, or the risk of breaching users privacy (Section 2.3).

While some of the prior work we survey defines or addresses research problems and challenges of *social data* within a general framework [39, 125, 130], most studies discuss them in the context of specific platforms such as social media [311], search engines [326] or recommendation systems [281]. We note that challenges such as those related to publishing the research in this area (e.g. discussing space and format limitations, or the publication time frame) [41, 324] or the scalability of the analysis methodologies [27], are outside the scope of this Chapter. Additionally, we also recognize that the categories we identify and highlight are not mutually disjoint and often overlap, and that the suggested solutions for various challenges might pull in opposite directions (e.g. to solve privacy related issues one might need to compromise various performance metrics).

2.1. Limits of Social Data Sets

We start our prior work exploration with research that raises concerns about what can be legitimately inferred from existing social data sets [274, 311], as well as research that tries to quantify the limits of such data (e.g. the biases, the lack of coverage or representativeness) [234, 160]—with this Chapter focusing for the most part on the former. See Table 2.1 for a comprehensive overview of relevant prior work.

Media and Platform Specific Biases

Due to limitations in obtaining the needed data to conduct various observational studies, the research relying on online social signals to study and model human behavior is dominated by only a few social media and platforms. Alas, such research cannot be generalized, as each medium and platform exhibit its own structural biases [311] that can lead to very specific phenomena [274]; yet, this is often overlooked.

Functional biases & APIs. First, social data is typically proprietary and not directly accessible to the research community. As a consequences, the data access APIs restrictions and the terms of service set their own boundaries to what a working data set may or may not contain [39, 41, 212, 235]. For instance, a social media data set will typically include only content that is public and un-deleted, or to which the users have given explicit access (e.g. through agreements, by accepting a social connection) [39, 212], meaning that from the very beginning a fraction of relevant data is bound to be left out. Then, depending on the platform, the available APIs for accessing the data set further limits to what and how much of the public data can be accessed, often without clear guarantees about the properties of the provided data—e.g. much research

Table 2.1.: Challenges around working data collections and their mention in the related work. Note that the table either quotes, paraphrases or aggregates such mentions. Future references should be directed to the original papers.

Types	Sub-Types	Related categories from previous work
Media & platform biases	Functional biases & APIs	APIs restrictions (e.g. rate limits, filtering by content or geographical location) are potential sources of bias [121], violation of terms of service [41], missing the ecology for the platform, platform has its own suite of affordances [311], platform specific vs. platform independent phenomena due to how and what data is stored, or due to dynamic, proprietary, secret, undocumented, platform specific algorithms [274]
	Population biases	different demographics are drawn to different platforms [170, 274, 311]; not representative of the population at large, skewed to young and urban demographic groups [64], [Twitter] over-represents densely populated regions and men, and exhibits a a highly nonrandom sample of the overall race/ethnicity distribution [228], characteristics like gender, race and ethnicity, and parental educational background are associated with the use of online social platforms [131], representing the full breath of communicative activities taking place on the platform or the overall public debate, how well the platform data represents society [44]
	Behavioral biases	people behave differently on distinct platforms [274, 205, 251], humans alter their behavior when aware of being observed [274], status-related behaviors and norms may not translate to other platforms [311], the platform plays a role in shaping the representation of an event, representation and discussion of an event on social media are a constructed phenomenon, [Twitter] data is not a representative sample of people experiences [64], data is biased due to content ranking and user interface [27], most content is generated by a small fraction of active users [25]
Collection biases & representativeness	Data querying & sampling	limitations of hashtag-based research and sampling [41, 311, 64, 44], explicit and implicit bias inherent to a data collection approach [274], working with the wrong list of keywords might cause relevant data to be lost [121]
	Temporal considerations	change in the usage of a platform [274]; (for events) social media data sets typically depict a specific time period around the spike in the messages posted on a social platform, which overlooks e.g. the causes or the aftermath [64]; long-term logs of user behavior allows the observation of long-lasting effects of e.g. an experience [268], querying behavior depends on historical period, time of the day, time period [286], “Swiss cheese” decay of Twitter test collections due to content deletions [28]
	Context-specific biases	use of proxy populations to operationalize the definition of a group [274], different demographic and social groups may behave differently [311]; self-selection or self-reporting biases (e.g. choosing to use a convention or to talk about a topic) due to external, confounding factors (e.g. weather, social pressure) [170, 311]; opinions expressed on social media are not a random sample of those of all users as there exist self-report imbalances/reporting biases w.r.t negative sentiment and average feelings [126]; bias due to individual characteristics or privacy concerns [170]; self-cleaning and self-censorship [320, 69]; behavior changes as the task at hand becomes more difficult [23]; querying behavior varies across topics [286]; the temporal orientation of messages is dependent on factors like experience, number of friends, or mental health [279]
	Size & Representativeness	bigger data are not always better data [39, 112], issues with coverage and representativeness of messages and communication networks [121], social media data is always partial and incomplete [64], data redundancy, sparsity trade-offs, bigger data is not the same as having the right data [27], it is not just about the size of the data [193]
Data sharing, reproducibility & other challenges	Data sharing	independent data sets [274], proprietary platforms [311], about 5% of studies obtained their Twitter data from existing data sets originally collected by other researchers [342], replicability of results, full documentation of methods, sharing of data [41], collections are obtained through non-transparent sampling methods [203], proprietary nature of social media leads to two key problems: data replicability (when data sharing is prohibited) and data decay (when items are deleted as time passes) [161], data sharing can diminish the time and effort to collect data [324], only few [research] papers share their data at all [149]
	Digital divide	digital divide [63, 39], data “haves" and “have-nots" [41], the need for democratizing data science [58], reproduces the unequal power relations [64], privileged access [64], “embedded researchers” [274], sharing can alleviate inequalities in data access [324]
	Spam & Non-humans	non-humans, spammers, bots, organization accounts, “authentic" human users [274], non-human agents, bot activity, fake accounts [64], web spam [27]
	Other issues	online vs. offline behavior and binary [63, 315], there are no neat divisions between news and the personal, between public events and private effect [64], good, labeled data sets are hard to build [60], legal obligation to remove deleted content [212]

on Twitter relies on data endpoints that give access to at most 1% of the public tweets [120, 160, 234, 235]. Indeed, while some data access APIs seem to provide a random sample of the relevant content [234], others lead to biases with respect to e.g. the topical making of the relevant content [235] as well as the follower-followee network structure [120].

Second, each platform carries its own suite of affordances, i.e. the set of actions that can be performed and are encouraged, and the set of actions that are not supported or are hard to perform [311]. This set of possible actions tend to shape the behavioral norms on each platform, influencing this way the type of content that is shared or produced. Moreover, each platform uses proprietary, platform-specific algorithms to promote, or show content or users that affect what content or with which social connections users are likely to interact with on the platform, further biasing the "found data" [130, 274, 311].

Thus, in summary, such functional peculiarities of each online social environment directly impact what user demographics would be more likely to be drawn to them, as well as what kind of actions the users are likely to perform. To zoom into specific challenges, and following Ruths and Pfeffer [274], we further make a distinction between *population biases*—as a result of the misrepresentation of human *population* due to platform-specific characteristics, and *behavioral biases*—as a result of the misrepresentation of human *behavior* due to platform-specific characteristics.

Population Biases. The population biases typically refer to how well the working data sets collected from a given platform reflect either the corresponding offline or online populations, or those of distinct media or platforms. Indeed, yearly surveys from the Pew Research Center[1] of social media users demographics show that the demographic composition of the major social media platforms consistently differ both with respect to each other, as well as with respect to the offline or the Internet population [53, 54]. These observations have also been supported by academic observational studies through social media [228, 131, 274]. For instance, [228] finds that Twitter users significantly over-represent men and the population of regions that are densely populated. Another study, looking at social media use among a tech giant employees, shows that although growth in use and acceptance across social media platforms is not uniform, in time privacy and other user concerns regarding the use of these platforms tend to level off [18].

Behavioral Biases. To understand how differently people behave within different environments (e.g. different media or platforms) or how much the user behavior on social platforms reflects the real-world phenomena, a number of studies have contrasted them—yet, such studies are still scarce. For instance, [304] compares web search with microblogging search, finding that they

[1] Pew Research Center: http://www.pewinternet.org/

capture different use cases: queries on Twitter are shorter and more popular, focusing more on temporally relevant information and people, while search queries tend to change and develop as users learn more about a topic. Other research looks at the interplay between what people search and what people share on social media regarding health information [74], observing that information seeking and sharing practices are both dependent on the condition type (e.g. being a serious condition or not). Leskovec et al. compared news media with web-logs, showing that there is a few hours lag between the attention peak of a meme (short sentence or phrase) in mainstream media and web-logs [200]. A number of other studies look at the similarities and differences among different social media platforms along, e.g., adoption patterns [189], user personalities [147], news spreading [198], geographical and socioeconomic patterns [202], shared content [251] and behavioral patterns [205].

Collection Biases and Representativeness

Alas, the data collections are not shaped only by the platform-specific phenomena. Even when a study is clearly confined within the context of a given medium and platform, quality dimensions such as the representativeness, the completeness or the precision of the working data sets with respect to the overall platform's data has been challenged [311, 121, 142]. Depending on the problem at hand and the targeted context, other characteristics of the data collection such as the temporal parameters or the characteristics of the studied topic can also affect the reliability of the observations made on this data collection: e.g., as the general social context changes, users might also change how they use a given social platform, which might in turn render the observations from a past cross-sectional study as void.

Data Querying & Sampling. Often the data collections are obtained by retrieving the content matched by user-provided queries through public APIs. These queries typically contain to-be-matched parameters for content, time and/or geographical location. To this end, one recurrent discussion is the problematic reliance on keyword- or hashtag-based sampling [213, 311, 44]. A key issue is that the choice of keywords and hashtags "*is equivalent to specifying the boundaries of a data collection: working with the wrong list of keywords might cause relevant data to be missed*" [121]. Hashtags are often associated with different social, political and cultural frameworks, and, thus, the samples built on top of them can embed different dimensions [311]. Research has shown that data sampling affects the various communication networks that can be reconstructed based on the social media posts, and a poorly specified query to the sampling APIs exacerbate the biases in network properties (e.g. clustering, degree of correlation) more than the APIs limitations [121]. Ultimately, hashtags are just a form of social tagging

(or folksonomies[2]), and even if we assume that all relevant content is tagged, their usage is often inconsistent (e.g. different formats, spellings or word ordering) [259]. Thus, while some attempts to standardize the use of hashtags exists, e.g. see [241] for humanitarian emergencies, in order to better capture the main information types of interest for relevant stakeholders, the data collections built on top of them will still reflect only partially the relevant data, as it will overlook possible communications among actors that might not respect these standards.

Temporal Considerations. The temporal parameters of the data are also important, as often they are also part of a data set boundary specifications. For instance, data collections corresponding to real-world events are typically defined by the peak in the activity on a given social platform. However, different events may have different temporal fingerprints (e.g. disasters can have longer term consequences than sport events) that the corresponding social data sets defined around activity peaks might miss [64]. In addition, there may be also protracted situations such as wars or other long-lived events, whose temporal fingerprints may be characterized by multiple peaks.

However, events are not the only relevant case. For instance, even when tracking social signals at a more aggregate level, one will notice general variations regarding when and for how long the users focus on a certain topic—variations that can be triggered by current trends, seasonality or periodicity in activities, surprise factors, or even noise [265]. Looking at search query logs, researchers have observed that long-term query logs (as opposed to short-term, within session query information) provide better insights into the evolution of users interests, needs or experience over time (e.g. pregnancy or career evolution), or of the causes or impacts of certain personal events (e.g. having a medical condition) [268, 101].

Furthermore, it is important to note that neither the platform population or the subgroup engaged in a discussion topic, as well as the platform suite of affordances, are static. They often exhibit critical temporal dynamics. Regarding the former, in his ICWSM'11 keynote[3] shows how design decisions and platform chances influenced users behaviour—e.g. making the message composer much shorter lead to a significant decrease the length of the messages users write, but an increase in the number of messages they send. Regarding the latter, research has, for instance, shown that the demographic composition and participation of users posting about the 2012 Election cycle in the US is non-stationary and unpredictable over time [84].

An equality important problem is the "Swiss cheese" decay of social data as a result of content

[2] A form of ad-hoc categorization and labelling of the data within social systems [289].

[3] The keynote is available at: http://videolectures.net/icwsm2011_seligstein_trends/. ICWSM is a top-tier conference on computational social science

deletions and social platforms terms of service—rendering such content as unusable after a time window—which can leave important holes in the working data collections [28].

Context-specific Biases. Other challenges come from the innate properties of the applications, of the semantic domains or of the populations of interest. Common issues include so called *self-selection* or *self-reporting biases* [126, 170]. For instance, a study might be interested in the opinion of young college graduates about a new law. Yet, most users will not self-label themselves along such demographic criteria. Thus, to run the study researchers would typically rely on *proxy populations* that e.g. report on a social platform to be alumni of set of universities—and, this can end up being an important source of bias [274]. In fact, in the context of predicting users political orientation, researchers have shown that the choice of the proxy population drastically influence the performance of various methods [60]. Or, in other words, it highlights that the performance of such latent attributes inference varies across a given data set demographics, the study showing that existing classifiers tend to do much better on politicians than on "normal" users (with only a few political posts) [60]. This is important, as depending on various factors, such as social pressure, privacy concerns, topical interest, language, personality, culture, socioeconomic status or education, different users will adapt in a different way their online behavior (e.g. what they share, search or pay attention to) or even the choice of using a social platform [126, 320, 230, 69, 170, 311, 286].

Data Size & Representativeness. Notably, in the light of Google Flu Trends success in 2008 [114], Chris Anderson[4] remark in his provocative essay "The End of Theory" [16] have sparked intense academic debates:

> *"Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With* ***enough data****, the numbers speak for themselves."* (with emphasize added)

Yet, while the ability to capture large volumes of data brings along important scientific opportunities [175, 194, 113], size by itself is not enough. Indeed, such claims have soon been debunked by numerous critics [39, 112, 130, 193] which emphasise that they ignore, among others, that the size does not necessarily make the data better (e.g. more representative, more precise) [39], and that "*there are a lot of small data problems that occur in big data*" which "*don't disappear because you've got lots of the stuff. They get worse.*" [130]. Thus, more often than not, the problem is finding the *right data* [27] as, for instance, if adding more data also increases the level of noise, the quality and reliability of the results deteriorate.

[4]Editor in chief at Wired Magazine.

In fact, this makes the need to understand the data samples (particularly, their limitations) more important than ever [39]. For instance, social media users do not represent all people, nor are they representative of them [130]. It is even more essential to understand the missing data as it is to study the "*found data*" (as Harford calls it [130]).

Data Sharing, Reproducibility, and Other Challenges

Data Sharing & Reproducibility. The previous discussions about data collection also point to the importance of properly documenting the methods and parameters used to sample, collect, and process the data sets [41, 39]. This would not only better support the interpretation of the findings, but also assist other researchers in reproducing the same or similar data collections, when data sharing is not possible—typically due to the social platforms restrictive terms of service or due to proprietary data [324, 161]. Jurgens et al. [161] discern two key problems due to this proprietary nature of social data: *data replicability*—when data sharing is prohibited by terms of service or privacy constrains, and *data decay*—when the original data set cannot anymore be reconstructed due to content deletions.

However, when it is possible, data sharing makes it easier to both reproduce or replicate[5] results by preventing redundant, labor-intensive and time-consuming data collection [324]. Thus, while these various constrains hinder data sharing—which is rather an exception than a rule (with e.g. less than 10% of research papers on online social networks doing so [149]), there are efforts to aid reproducibility and future comparison by releasing and open-sourcing data and tools [172, 162, 218].

Digital Divide. Furthermore, one key concern related to data access is the potential to further deepen the divide between the data "*haves*" and the data "*have-nots*" [41], as well as between those that have or have not the computational skills needed to access and collect data [39, 324]. Additionally, the former is perceived as being further exacerbated by the raise of "embedded" researchers [274] that have privileged access [64] to certain social platforms resources. In this context, data sharing is seen as necessary means to make such research more inclusive and to reduce the existing data access gap [324].

Other challenges. Alas, there are many other challenges that we have not discussed, some of which are equally important and hard to address. For instance, another important challenge is how to separate non-humans (spammers, bots, organization accounts) from "authentic" human users and account for them [39, 64, 274]. Similarly, noise or content redundancy—lexical (e.g. duplicates, re-tweets, re-shared content) and semantic (e.g. near-duplicates, same meaning) of-

[5] See [88] for the difference among the two.

ten account for a significant-fraction of content [27], and might negatively impact the tools utility (e.g. showing redundant content to users) [266]. Other issues include the the difficulty to construct high-quality labeled data sets [60], to dissociate offline and online phenomena [315] or between public events and their private effect (e.g. what belongs to the public sphere and what to the private one) [64].

2.2. Methodological Challenges

Even when we understand the limitations of various working data sets of human traces, we still remain with methodological challenges to e.g. correct or at least account for biased, erroneous or missing data, to properly evaluate tools that leverage social data and validate observations, or interpret results in context (e.g. by considering the suitability of methods for the problem at hand or for the platform specific mechanisms) [39, 274, 60, 311]. See Table 2.2 for a comprehensive overview of relevant prior work.

Data management, processing and correction

Data management. One of the first challenge when working with big social data sets is how to adequately manage them. From a system perspective efficiently capturing, storing and maintaining ever growing volumes of social data sets are the main challenges (and very important ones) [27, 41] to enable the analysis of such data. In doing so, the systems rely on various data categorization, aggregation, and organization heuristics that optimize for certain (popular) types of data access (or queries), which might in turn favor e.g. the dominant classes of content (e.g. opinions produced by a certain demographic group) and popular data analyses, while neglecting or concealing others [25, 258, 32]. Alas, the way in which data is structured and categorized can have un-wanted consequences for the users associated with it, if it misrepresents them [258]. For instance, Amazon generated public outrage when it consistently mis-categorized LGBT material as "adult" material resulting in their demotion in search results [137]. The biases, assumptions or prejudice of those designing and building the systems can even make their way to how the data is organized at the lowest levels [91]. It is important to devise adequate data management methodologies to ease its use while also avoiding such pitfalls.

Algorithmic variations and biases. This last point is also closely related to the need to gain understanding about how algorithms learn to discriminate from traces of human behavior, reflecting existing biases [225], and what are the possible trade-offs between performance, privacy or fairness [122, 225].

Table 2.2.: Methodological challenges when working with social data and their mention in the related work. Note that the table either quotes, paraphrases or aggregates such mentions. Future references should be directed to the original papers.

Types	Sub-Types	Related categories from previous work
Data management, processing & correction	Algorithmic variations & biases	algorithmic bias [176], algorithms can reinforce human prejudice [226], algorithmic invisibility [311], propagating the belief biases of users [326], data mining can inherit the prejudice, factors like the target variable and class labels to predict, the features, how the data was collected can bias the data mining outcomes [32], racial stereotyping in the behaviour of the algorithm and racial profiling in its applications [57], tag recommendation is an example of algorithm bias [27], it is feasible to observe how platform behaviour changes based on location or user behaviour (e.g. past searches) [193]
	Correct & account for biases & errors	appropriate statistical corrections informed by known biases [274], randomized studies [311], overlook of the way in which the data sets are assembled [60], selecting the most appropriate data access endpoint (e.g. from different APIs) [160, 234, 235], detect and remove duplicate content, non-humans, spammers, bots, fake accounts, organization accounts [274, 64, 27], running causal analyses [243, 87], selection of a representative and unbiased sample of the population [109]
	Data management	capturing, storing, and maintaining rapidly growing data collections would require very substantial infrastructure investment [41], transferring, storing and processing large amounts of data may not be feasible [27], data representation, naming and categorization [258]
	Exploiting context & biases	taken out of context, data loses meaning [39], clear meaning only in context [311], the similarity between users for a particular classification task requires knowledge of sensitive attribute [225, 90], self-reporting biases observed at a large scale may actually provide signals that ease the task of sentiment tracking in online environments [126]
Evaluation & validation	Adequate evaluation & generalizability	use of more distinct data sets, use of altered data sets that explicitly introduce or remove biasing factors, use a holdout data set, compare results to existing methods on the same data set, showing results on multiple platforms [274]; study of various phenomena on multiple platforms, multi-platform and multi-method analyses [311], applying same methodology across multiple hashtag events [41], replicate findings across time and data sources to ensure the patterns are robust and not evanescent trends [193], generalization of findings, under what conditions online behavior generalize to behavior in the general population [63], transferability of classifiers [60], findings based on social data are required to be systematically generalized and replicated [203]
	Negative results & disclaimers	unaware of failed studies, publication of negative results, pointing out the limitations in a study [274], methodological awareness, soliciting "limitations" sections [311], researchers need to assess and account for the gaps in the data sets [64], research with negative results refuting positive results are rarely published [107]
	Standards & documentation	develop baselines and guidelines [311], common ground in terms of methodological approach [63], standard set of Twitter metrics [41], documenting workflows and lack of standards [324], provenance of the data sets, description of home-grown tools [41], developing and using standard evaluation metrics when they do not exist [83]
Interpretation & inference	Interpretation	true meaning of various processes (e.g. what a retweet or a click mean?), qualitative pull-outs to help with interpretation [311], the cultural context can be difficult for geographically distant researchers to parse, need to account for what is in the data and what isn't [64], identify methodological choices which may affect and change the interpretation of the data [41]
	Opportunistic approaches & methods suitability	feature hunting, using the right method in the right place at the right time [274], opportunistic data gathering [41, 323]; [running studies] due to the availability of data, tools and ease of analysis [311], there is a lot of data mining for the sake of it [27], importing methods from other fields. flawed methodological analogies, selecting on the dependent variable [311], there should be a justification that appropriate state-of-the-art methods are used [274], social media should be analyzed as a communicative rather than representational system [271]
	Unintelligible actions & dedicated mechanisms	practices that are unintelligible to algorithms (e.g. subtweets, quoting via screen captures, hate-linking) [311], the platforms promote content based on platform specific metrics [64], each platform allows different mechanisms for information diffusion, sharing, etc. [311], platform-specific algorithms are constantly changing due to actions of engineers and customers [193]

Indeed, in recent years, an increasing number of researchers have steered their attention to measuring the impact and accounting for possible biases as a result of algorithms reinforcing the beliefs of users [32, 176, 226, 326]. For instance, search engines typically alter how they rank results based on what users click on, although a large fraction of the answers the users settled on are incorrect [326]. This is problematic, as users tend to accept the answers that confirm their biases, and the ranking algorithms tendency to mirror them can be counterproductive [326]. Other examples include delegating decision making such as "which candidate to employ?" to data mining techniques. Many of these techniques are designed to find correlations (that do not account for possible confounding factors) and are often dependent on what proxy attributes that are believed to be indicative of the desired qualities are predicted (e.g. predict the grades in annual reviews as a proxy for good employees), or on the choice of attributes to observe (e.g. the selected set of characteristics to represent the candidates). Yet, choosing different attribute to observe or predict impacts what data mining actually finds, and it can re-enforce existing prejudice about protected classes [32].

Correcting and accounting for biases and errors. Biases and errors can occur at any step in a data analysis pipeline (see Chapter 3 for an brief overview of such a pipeline). The first opportunity to account for them is at data collection time. As we mentioned in the previous section, much research relies on public APIs to retrieve the working data sets, and these APIs typically set limits to how much data one can retrieve (e.g. 1% for Twitter public API), and conceal the details about how the returned data points where selected from all relevant data points [120, 160, 234, 235] (see [267] for an overview of API limits). However, beside trying to pick the most advantageous API from several provided [235], many APIs support various types of predicates to query for data—typically referred to as the *query language*—that concede some flexibility (although, limited) to control the quality of the data collection [120]. Indeed, research has adapted information retrieval techniques to generate more optimal queries [272], to expand and adapt existing queries [213], or to split the queries and run multiple in parallel [276], with the goal of mitigating existing biases and improving the quality of the data collections (typically by improving their completeness).

Another opportunity to correct existing biases and errors is after the data is collected. Straightforward approaches to improve data quality, is to detect and remove spam or non-human accounts [274, 64], or duplicates (lexical or semantic) [27]—yet, such distortions can be hard to correct or remove as well [274]. Other (more preferable) alternative are to make appropriate statistical corrections that are based on known or identified (data set-specific) biases [274, 90], to conduct causal analyses whenever possible in order to adjust for e.g. selection biases [243, 87], or to apply re-weighting and post-stratification or multi-regression techniques [318, 335], which

effectiveness have been proved for adjusting survey and polling data. Biases can also be accounted for while designing algorithms, or when interpreting results, challenges that we will discuss next.

Exploiting context and biases. While biases in the data are typically perceived as sources of inaccuracy that needs correction [109, 274], if understood, they can also guide the design of more efficient tools [126, 199]. For instance, [126] shows that self-reporting biases provide useful signals for sentiment tracking by creating rich transient social contexts (e.g. for polarizing topics a happy event for a group of users might be bad news for another) that assist with labels acquisition and with dealing with sentiment prediction, even when sudden sentiment drifts occur. In the context of ranking peer recommendations, [199] shows that a quantitative understanding of position bias (the tendency to pay more attention to items at the top of a list) can be exploited to manipulate users attention.

In fact, most biases emerge as a result of various circumstances forming the specific settings (or context) of the analyzed phenomena—e.g. less data might be available about average feelings than extreme ones as users might find them more worthy to talk about [126]. Thus, if taken out of context the analyzed data sets lose from their meaning [39], observation also emphasised in [311].

Understanding the context can among others inform algorithmic design. For instance, context can aid in understanding what are the sensitive attributes for a certain user classification task which can, in turn, inform decisions such as who should be considered as similar to whom and help avoiding discriminatory treatment [90]. Further, culture and language act as barriers to social media communications and can explain some of the observed patterns (or the lack of them) with respect to the information flows [105].

A core dimension in understanding the context, is the *problem domain* (e.g. topicality, type of media, data semantics). For instance, a good example of leveraging the domain are the specialized search services (e.g. domain-specific search), known as search *verticals*, that focus on a specific information seeking task [19]. By focusing on a domain they can often leverage clear, un-ambiguous relationships between concepts specific to the domain, providing more precise and relevant results [331].

Evaluation and Validation

> "*There is (. . .) the need for increased awareness of what is actually analyzed.*"
> —Ruths and Pfeffer, 2014, [274]

Adequate Evaluation and Generalizability. A last opportunity to account for biases, as well as to test the robustness of findings and to understand the generalizability of observations, is at evaluation time. When the biases cannot be corrected as there is little knowledge about the level of biasing, or the biasing factors are too complex or hard to untangle, one should run e.g. longitudinal, comparative, multi-data sets, multi-platforms or cross-domains analyses [41, 103, 110, 274, 281, 323, 323]. When access to multiple, distinct data sets is limited, one can still re-run the analyses on data sets that are altered to introduce or remove noise or biases [274]. Further, when Liang and Fu [203] tried to generalize and replicate 10 known propositions made by prior work, they failed to do so for 6 of them due to variations of how the data sets were collected, but also due to inconsistencies in measurements, or differences in the analysis methodologies. Thus, when hypotheses about real-world phenomena are formulated and put forward to be tested, the results of different methods for collecting, measuring or processing the data should also be juxtaposed [274, 311].

Also, important data variations exist not only across platforms or data sets, but also within each data set across categories of users [60]. For instance, users from different geographical areas have different food consumption patterns [6, 325], thus analyses should account for such differences and evaluate tools across various user demographic criteria. Specifically, one can zoom in and also check if the findings hold across different relevant data sets demographics [36, 60]. Additionally, some researchers also advocate for considering the platform changes to ensure that the findings are not just temporal trends (e.g. users might change how they interact with the platform due to functional changes such as a new ranking strategy of items or a new ability to share content) [75, 323, 274], which can be done by replicating the findings also across time [203, 193]. While longitudinal, comparative, multi-data sets or multi-platforms research exists [18, 128, 174, 196, 6], they tend to be the exception, rather than the rule.

Negative Results and Disclaimers. Further, while failed studies or negative results are useful for learning about what hypotheses have been rejected, or what are the suitable data sets and methods, publications of negative results are scant [107, 274]. Additionally, disclaimers are also important. If errors or biases were not ruled out, researchers must discuss the gaps and limitations in the working data sets as well as of the employed methodologies and their studies [64, 274, 311].

Standards and Documentation. Emphasis has also been put on the need to develop baselines and guidelines [311, 324] and to find a common ground regarding the methodological approaches [63]. In addition, the provenance of working data sets, the workflows, the home-grown tools and methodologies often need to be better documented [41, 324]. While research in the area of text mining or information retrieval typically follows standard evaluation procedures and metrics, for many social media analysis tasks—specifically for domain-specific application such as crisis informatics—this is not the case [41, 83]. Thus, effort should be put in developing and using standardized experimental methodology when they do not exist [83].

Interpretation and Inference

Interpretation. Much research rests upon the assumption that online behavioral traces reflect in some quantifiable way real-world phenomena [22, 71, 173, 271]. On the one hand, such data can reveal fascinating insights about e.g. how people interact [317] or react to major life changes [71], and have helped confirming prominent social theories such as the small-world phenomenon [179, 312]. On the other hand, it has also been shown that there are significant differences even between what data explicitly gathered from users (e.g. the explicit social relationship among them) indicate and what the data implicitly acquired from them does (e.g. the interaction graph among users which exhibits a larger diameter and a smaller clustering coefficient than the graph drawn from explicit social links) [327].

Consequently, one first needs to critically question from what various social activity traces or processes stem from. For instance, social links between users can stem from friendship, trust or shared interests, and thus can embed very different social cues. The same holds for (re-)sharing content on social media: it can be a sign of endorsement, a result of finding the content interesting or amusing, yet, users also (re-)share content to ridicule or disapprove. Thus, the same mechanisms or processes might capture different signals depending on the context [271, 311]—yet these distinctions can be hard to make by automated methods or when looking at data in aggregate. As Tufekci advises [311], to understand the various signals behind the same mechanism or process, one might consider to pull-out and qualitatively analyze small data samples.

Further, when the researcher lacks the context, it might be difficult to account for what is or is not in the data—for instance, in the context of analyzing social media communication in crisis situations, a geographically distant researcher may have difficulties in understanding the cultural context and the event peculiarities [64]. Another aspect that should also be considered, is how different methodological alternatives may lead to different interpretations of what it is in the data [41].

Unintelligible actions and dedicated mechanisms. So, to reiterate, often the social signals can be unintelligible to algorithms as users may behave differently depending on the context (e.g. depending on to whom their content will be visible), yet such context is not always clear [311]. This unintelligibility is often dependent on the mechanisms available on each social platform (e.g. having a like button, but not a dislike one), but also on the variations in platform-specific algorithms and mechanisms in response to users actions [193]. This should be considered when interpreting the performance or the findings of tools and analysis methods.

Opportunistic Approaches and Methods Suitability. As discussed earlier in this Chapter, there are a number of challenges with respect to the working data sets, including their collection—yet, some types of data are are much easier to harvest than others [41, 323]. Alas, this has resulted in much research focusing on e.g. a few data sources [311] raising concerns about the research agenda being opportunistically driven by the access to data, tools or ease of analysis [41, 274, 323]—or how Baeza-Yates puts it: "*we see a lot of data mining for the sake of it*" [27].

A similar concern also regards the employed methods, like for content or user classification, when the performance is rather the result of e.g. "*feature hunting*" [274][6] instead of being based on a priori hypotheses—which encourages the practice of harking [169].[7] Hence, for each task at hand, even if the data or the methods are old, imported from other fields, adjusted or new, an argument should be made about their suitability [274, 311].

2.3. Ethical Challenges

"*Just because it is accessible does not make it ethical.*"—boyd & Crawford, 2012, [39]

But, boyd and Crawford are not unique in their call for caution [341]. Scientists [32, 57, 91, 167] and journalists [136, 176, 226] alike urge both scientists and practitioners to carefully scrutinize their use of social big data against a variety of possible ethical pitfalls such as breaching users privacy [122], or racial, socioeconomic status or gender-based profiling [32, 57]. Alas, only a small number of publications (between 2007 and 2012) relying on such data from Twitter was found to include any discussion or even an acknowledgement of any ethical challenges [342]. In the US, the Belmont Report puts forward a set of ethical guidelines for research that involves human subjects [100] that often serve as standards. The report is based on three ethical principles—respect to persons, beneficence (emphasizing on the "Do no harm" philosophy and

[6] Trying out multiple features until finding one that delivers important improvements.
[7] Harking refers to the practice of hypothesizing after the results are known.

Table 2.3.: Ethical challenges when working with social data and their mention in the related work. Note that the table either quotes, paraphrases or aggregates such mentions. Future references should be directed to the original papers.

Types	Related categories from previous work
Ethics (in general)	ethical perspective [63], ethical issues [64], ethical considerations [342, 212], how definitions of public and private information apply to Internet data, and whether an avatar or profile is a person [212, 214], indicators of legal and ethical risk [167]
User consent	consent [63], unaware their messages were made public, users consent [64], re-purposing such data for research may violate the expectations of content creators [users] [148], no specific consent was sought or received from the subjects, user profile information being considered freely accessible for collection and research [341], public content [. . .] consent is therefore implied [212]
Privacy & confidentiality	privacy concerns [65, 341], securing the privacy of the user, confidentiality [63], balancing privacy versus accuracy, misuse and protection of confidential and private data, anonymization [122], privacy self-management, concern that people identities would be exposed [64], privacy expectations of Twitter users [342], privacy risks, concerns and practices [320], privacy, anonymity, legal and ethical restrictions [27], machine learning presents new challenges for protecting individual privacy [140], data re-identification [124]
Data sensitivity & discrimination risks	multiple data feeds can be combined to generate intimate insights without a person [64], ethical appropriateness of archiving public tweets for research purposes [342], privacy management and ethical concerns for using deleted content [14, 212], user data is created in highly context-sensitive spaces [39], data sets can be combined, and thus sensitive knowledge can be inferred from benign data that are routinely shared [140], racial stereotyping and profiling [57], machine learning presents new challenges for ensuring fair use of data, need to increase technical expertise in consumer protection to address discrimination issues arising from big data [140], risks of bias or discrimination based on the inappropriate generation of personal data [65]
Others Concerns	[provide users with] access to the derivatives from their informational activities [64], the analysis process, along with the decision-making behind it are “black-boxed” [258], need to increase transparency into how companies use and trade data [140], building a digital dossier [124]

on minimizing risks to research participants while maximizing the benefits to the society), and justice—across three primary areas of application—informed consent, assessment of risks and benefits, and selection of subjects.

More recently, it was the Facebook contagion experiment [183]—where the researchers manipulated users social feeds to include more or less of a certain kind of content based on the expressed emotions—that sparked an intense debate on whether public sources of user data should be used only on the basis of being accessible [148].[8] This incident was followed this year by an unprecedented move from the SIGCOMM 2015 Program Committee[9] which decided to accept a paper on measuring censorship [46] on the condition of placing a prominent note at the top of the paper highlighting their ethical concerns [237]—which drew further attention to the issue. On the bright side, in our surveying we also noticed an increased inclusion of a discussion about (or at least a mention of) the ethical challenges in research papers (for good examples see [101, 184, 227]).

[8] Also note that a large fraction of the papers surveyed in Table 2.3 were published after this experiment.

[9] http://conferences.sigcomm.org/sigcomm/2015/, SIGCOMM is a top-tier conference on computer networking.

User consent. A main concern with these studies is that they leverage user data without any kind of consent from them [148, 341]. Although publicly available, user data is inherently sensitive as e.g. users might not anticipate a particular use of their data, especially when created in a context-sensitive space and time [39]. This becomes even more delicate when e.g. analyzing user demographic attributes [57]. While asking consent might be often seen as unpractical [39], we should note that there are a few efforts to design methodologies for acquiring consent while minimizing the burden on the participants [148].

Privacy and Confidentiality. A related concern regards the risk of breaching users privacy by e.g. exposing their identity or personal information [64, 341]. Often, social data can reveal more about an individual that what it appears on the surface: while people may choose not to reveal certain information about themselves (e.g. age, gender, sexual orientation, religious views), such information can be predicted using often easily accessible digital records of human behavior [182]. Such breaches can have harmful consequences [32] such as stalking, identity theft, discrimination or black-mailing [124].

Alas, even when flexible privacy settings exist, users rarely change the default privacy settings [320]. However, even as users consent is implied when they generate content in public online spaces, "*people privacy preferences depend on their circumstances*" [64]. Take the case of social media use during crisis situations by vulnerable populations that may publicly share personal information to either assist others, update their family and friends, or ask for help. Such disclose is closely coupled with the context, and, thus, data use and share should be extensively scrutinized and the privacy of these users should be protected [64]. Existing solutions that balance between privacy and accuracy should be considered [122].

Data Sensitivity and Discrimination Risks. Privacy breaches are often possible because publicly shared data sets can be combined to gain insights about private individuals without their knowledge [64, 122, 124, 140]. To address this, the sharing and archival of data embedding personal information [342], as well as the use of content that users have explicitly deleted should be cautiously handled and anonymization should be considered [14, 65, 212].

Additionally, as briefly discussed in the previous section, there is also the danger that the use of social data can result in some sort of discrimination against protected classes [32]. While some argue that the reliance on automated decision making processes that are trained on such data can lead to more objective and accurate decisions, many examples have shown that they can in fact inherit, propagate, or even amplify the biases and prejudice of past decision-makers with respect to various factors such as race, age, gender or socioeconomic groups [32, 65, 176, 226] (often referred to as *algorithmic discrimination* [176]). While such a result is often unintentional, it

can have a variety of consequences: companies could use such information to practice price steering and discrimination [128], or users employment, credit or housing prospectives may be affected due to being stereotyped and profiled based on their race [32]. This is concerning as the existing laws often cannot handle such issues [32, 65]. Further debate on regulations of personal data collection, use, or disclosure is thus required [65].

Finally, there is a need for more transparency [140]: users should be provide with information about how their data is used, or given access to the artifacts resulted from their personal data [64, 124]. Alas, more often than not, the way in which user data is processed and analyzed to support decision making remains "*black-boxed*" [258].

3. Social Data Applications and Analysis

In this Chapter, we highlight *when*—for what purposes and applications—and *how*—what is the data processing pipeline—the social data is used. To this end, after quickly reviewing the main goals when leveraging social data, we briefly survey a sample of online social applications and platforms along with examples of the type of data that is collected from users (Section 3.1). We then briefly describe the prototypical pipeline for social data processing and analysis that is employed by such applications and platforms (Section 3.2).

The use of online *social data* has particularly grown with the increased popularity of "*a class of web sites and applications in which user participation is the primary driver of value*" [125], often referred to as the *Social Web*. As a result, today, everything that users do online could be captured, recorded and mined for current or future potential uses, typically, by four main stakeholders, identified by Oboler et al. [239] as: (1) business clients, (2) government, (3) other users within the social media platform, and (4) the platform provider itself. The societal, commercial and academic value of this data stems largely from its personal nature, scale, variety, level of detail and accuracy or timeliness, and its use can be categorized depending on the main objective:

(1) *To study the human behavior*: when the focus is on predicting, modeling, or describing various, on-line or off-line, real-world phenomena with data sets of online digital traces of human behavior—with examples including the modeling of information diffusion or flow [115, 270], social influence [30], disease transmission [275], migration patterns [338], or language usage [230]. In this context, the next Chapters (belonging to Part II of this thesis) look at the extent to which social media data mirrors the news events covered by the online mainstream media, as well as how much the insights we gain about social media use in crisis situations from one crisis-event data set can be generalized to other similar data sets.

(2) *To aid design*: when the focus is rather on exploiting usage patterns in order to augment, optimize, design, build or evaluate systems, tools, methods or algorithms—with examples in-

cluding search results [304, 12], decision support systems [173], targeted advertising [322], urban planning [264], users modeling [85], content reliability assessment [143, 247], or peer-to-peer social networks [248]. In the last part of the thesis (Part III), we discuss two case studies highlighting how such usage patterns can guide the design and the evaluation of methods working with social data.

3.1. Social Applications and Platforms

To draw a more concrete picture of the applications of social data, in this section, we give a brief overview with specific examples.

The social applications and platforms can range from social media (e.g. Youtube, Pinterest), networking (e.g. LinkedIn, Facebook) or recommendation sites (e.g. Amazon, Booking.com) to Q&A (e.g. Stackoverflow, Quora) or collaborative sites (e.g. Wikipedia, Micromappers), among others. Given the diversity of value propositions that these sites put forward (e.g. "Meet your next favourite book" for Goodreads, or "The Free Encyclopedia" for Wikipedia, or "The best answer to any question" for Quora), when exploring the various types of such applications and platforms, which have at their core user participation, often, a first question is:

> *What are the main motivations of users for using these social applications and platforms?*

To answer this question, much research has explored it from both a quantitative and qualitative perspective, finding that users have a variety of reasons for using such applications or platforms, some of which tend to be more common across such social sites—e.g. to construct an online representation of self [206] or learn new things [159, 123]—while others tend to be rather specific to a class of applications—e.g. to keep up with friends on social networking sites [191, 159], to collaborate with others on collaborative sites [68], to seek domain-specific information on speciality sites [217], or to look for a romantic partner on dating sites [96].

This diversity is also indicative of the fact that various applications or platforms rather complement each other by supporting different functionalities, purposes or domains of focus. For instance, Teevan et. al [304] show that when it comes to search, users tend to use web search and search on a microblogging site for different purposes: microblogging search being used to monitor content about specific transient topics or from given users, while web search is used to develop and learn about a topic. Similarly, other researchers have shown that, on different social media or networking platforms, users may have different attitudes [287] and tend to use

Table 3.1.: Examples of social data, applications types and existing platforms.

Types of applications		
• bookmarking	• crowdsourcing	• recommendation systems
• citizen journalism	• data analytics	• search
• collaborative knowledge building	• expert finding	• social networking
• collaborative problem solving	• match making	• social coding
• crowdfunding	• question answering	• user-generated maps
Types of data		
• check-ins	• instant messages	• user comments & reviews
• crowdsourced annotations	• location traces	• user generated content
• emails	• product rating & review	• social media content
• explicit social interactions	• mobile phone tracking	• web search logs
• implicit social interactions	• online commercial transactions	• web page visit & usage logs
Platforms		
• Amazon.com	• Github.com	• Pinterest.com
• Alibaba.com	• Goodreads.com	• Quora.com
• Bing.com	• Google.com	• Stackoverflow.com
• Booking.com	• Kickstarter.com	• Tencent.com
• Digg.com	• Linkedin.com	• Twitter.com
• Facebook.com	• Medium.com	• Weibo.com
• Flickr.com	• Micromappers.com	• Wikipedia.org
• Foursquare.com	• Netflix.com	• Youtube.com
• Ebay.com		

them for different purposes as well [313]: e.g. Facebook is more often used for personal self-presentation, while Linkedin tends to be used for professional self-promotion. The variety of purposes and functions can also be observed in Table 3.1 which highlights a variety of applications types, as well as of types of data and existing social platforms.

Furthermore, the many use cases and motivations for using these sites, as well as the multi-purpose nature of some of them, has resulted in users searching, creating or sharing information on a diversity of topics including work [93], food [6], health [74], relations [106], weather events [170], and many others. This topical diversity promises to enable researchers to observe and learn about both personal, everyday experiences [171, 173], as well as large-scale, collective events [166, 114, 294, 250, 280].

3.2. Analysis Pipeline Overview

Even with the diversity of social applications, types of data, and uses we highlighted in the previous section, there are a number of data processing and analysis steps that are typically shared across applications. To this end, in this section, we broadly describe the prototypical analysis pipeline when working with social data, which we break down into 6 generic steps high-

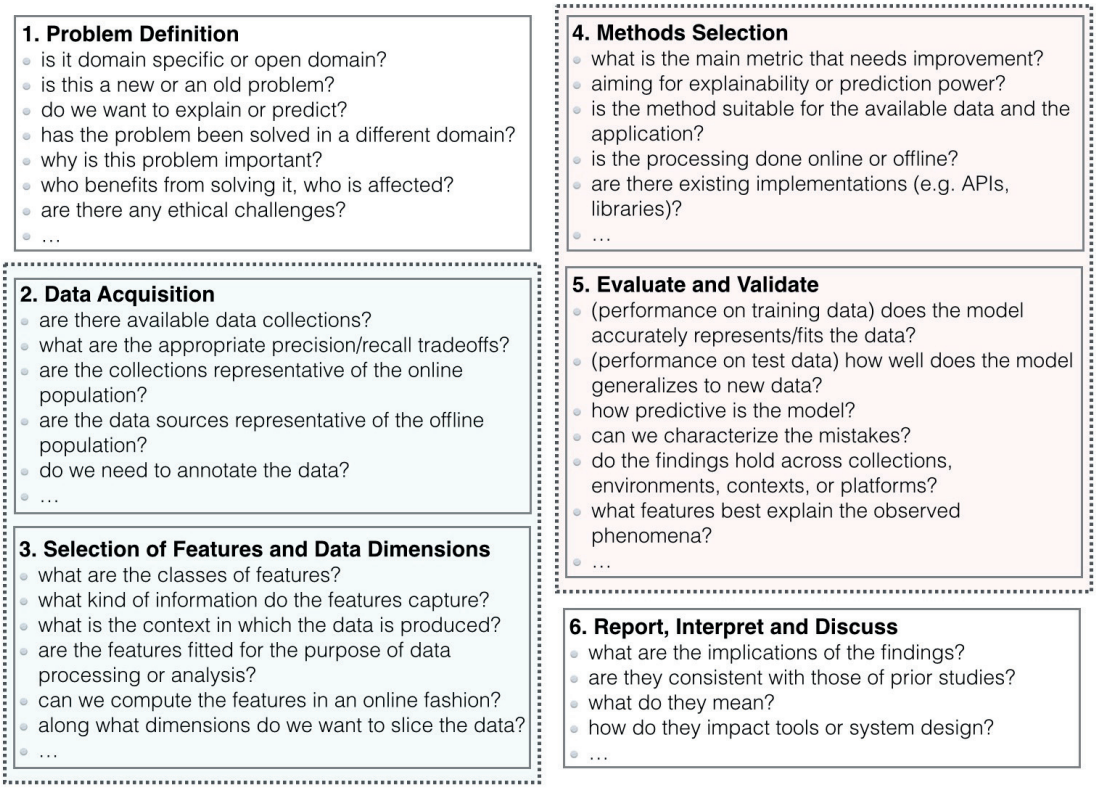


Figure 3.1.: Common steps in (social) data analysis along with example questions that can influence the decisions at each step. The steps enclosed by the dashed line are often coalesced as they are closely coupled to each other.

lighted in Figure 4.1. For each of these steps, depending on the main objective of the analysis, we discuss a few possible differences regarding alternative e.g. data acquisition and selection strategies, methods or evaluation strategies:

Step 1: Problem Definition. For each study, one would typically articulate the goal as either (1) designing, augmenting, optimizing or evaluating a method or tool, or (2) understanding, explaining or describing a real-world phenomenon; and the context as either domain-specific or open-domain, general-purpose. Such aspects are important as, for instance, when building a dedicated tool for a given domain one may actually leverage the usage patterns or knowledge that are specific to this domain since improving the performance for this domain is the main goal. However, when aiming to develop a general-purpose method or to describe a broad real-world phenomena with online social data one may instead need to validate the performance or observations across different domains, and to correct or account for various domain specific biases.

Step 2: Data Acquisition. Once equipped with a problem, the next step in addressing it is to locate the necessary data, which typically needs to satisfy certain characteristics and quality constraints. In the case of sources of online social data, the data is typically acquired through public APIs [267]—with some studies re-using existing data collections, yet this is rare. An important objective here is to build or use data collections that reflect the true incidence of e.g. the attributes or activities of the targeted data items (e.g. users, topics, messages, events) [32].

Depending on the problem at hand—e.g. depicting a certain class of users, measuring general usage trends on a platform or evaluating a new method—different data acquisition strategies that offer different trade-offs between various quality dimensions (e.g. precision, recall, representativeness) may be considered. For instance, if the goal is to study eyewitness accounts, using geo-located content, rather than content matched by a set of keywords, might help identify more eyewitness accounts although the proportion of relevant content might be lower—as the geo-based collections tend to have a lower precision (we discuss this particular aspect in Chapter 6). If the goal is to devise a new method for a well known problem, ideally, one would either test it on existing data sets that were used to evaluate previously proposed solutions, or would test it on several different data sets [274]—thus, multiple data collections of different properties might be needed.

Step 3: Selection of Features and Data Dimensions. Then, the constructs (or data items) of interest (e.g. topics, users, messages, events) are represented by a set or vector of features (e.g. n-grams for content, various demographic criteria for users)—this is typically referred to as feature definition, extraction and selection; and are typically classified along various dimensions (e.g. positive or negative messages, high- or low-impact events). In other words, the features are usually used to describe each data item (e.g. all words in a message) and represent attributes chosen to be observed, whereas the classes are used to group and organize the data items (e.g. all messages on a given topic).

To avoid pitfalls like selecting a feature representation that fails to capture all relevant details about the data items (e.g. ignoring users age when grouping users by education level), one might consider exploring various ways of capturing the same characteristic of a data item (e.g. both a user diet and her daily activities may be indicative of her health). Additionally, the future analyses and the selection of a set of classes should consider aspects such as how well separated these classes are (e.g. are there data items that match the definitions of different classes?), or if the dimensions used to determine them are independent (e.g. having a high income and having a college degree).

Step 4: Methods Selection. After the data is collected, represented along a set of features, and sliced along a set of classes of interest, various computational and statistical methods can be applied. Selecting the methods to be used depends first and foremost on the specific settings of each problem. For instance, the problem of assessing web content credibility can be cast as both a binary classification problem as well as a regression problem, and, thus, various classification and regression algorithms can be explored [247, 52]. Yet, if the goal is to test a certain hypothesis, say, men are more active on social media than women, statistical tests and models are used to either measure the strength of association or correlations among the variables of interest (e.g. being a men and being active on social media), or to test casual relationships [285]. There are other factors that guide or influence the selection of methods as well, including data quality and quantity, available data annotations or ability to annotate, available APIs, or if the processing can be done offline or needs to be done in an online fashion, among others.

Step 5: Evaluate and Validate. After the appropriate methods are selected, the models that they generate are tested to see if they are accurately representing or fitting the data—this can be seen as evaluating the models performance on the training data. Next, in many cases it is also desired to build and use methods or models that generalize well to other distinct data collections—this is particularly relevant when the goal is to solve an open-domain problem—yet, it might be required under other settings as well. Among the alternatives to test the generalizability methods or models are the use of different data sets (e.g., collected from different social platforms, acquired in different contexts or via different collection strategies), or performing a detailed assessment of the variation in performance across different data set demographics.

At this step, it is also important to understand when and why the methods succeed or fail. When the problem can be cast as a classification problem, this can mean exploring which are the attributes that best predict a certain class. When one aims to model a given phenomenon like the variations in the reactions of a group of people, this would be framed as what are the attributes that best explain or account for these variations.

Step 6: Report and Interpret. Then, the last step is to report, interpret, and discuss the implications of the findings. Besides reporting raw numbers, to aid the interpretation the findings and provide further insights, one would typically consider qualitative pull-outs of data about particular instances in which the models succeed or fail. Often, it is also important to consider the different social cues that the same (online) social mechanisms might embed in different contexts, and how this might have affected the methods performance. Finally, mainly for observational studies, the implications of the observations are typically discussed: e.g. how does an observed power-law distribution of items popularity might influence a recommendation system

design? how the polarization of users on a topic or the emergence of polarized groups affects the information diffusion in a social network?

We note that the examples and the brief descriptions of the steps are meant to be broadly indicative, and they do not necessary capture all the relevant details for all types of studies. For an in-depth discussion on the different design decisions that are taken when aiming to describe vs. predict a phenomenon see [285]. For a detailed discussion on how various decisions that are taken at each of these steps can bias the final results see [32].

Part II.

Limits of Social Data Sets

4. Social Media Biases: The Case of Climate Change

Social media is becoming more and more integrated in the distribution and consumption of news. Yet, are social media news similar to mainstream news? To understand to what extent social media mirrors mainstream media, this Chapter presents a comparative analysis covering a span of 17 months and hundreds of news events that have generated spikes of coverage in the mainstream media, social media, or both, by using a method that combines automated and manual annotations.

We focus on *climate change*, a contemporary topic that is frequently present in the news through a number of aspects, from current practices and causes (e.g. fracking, CO_2 emissions) to consequences and solutions (e.g., extreme weather, electric cars). The coverage that these different aspects receive is often dependent on how they are framed—typically by mainstream media. Yet, evidence suggests an existing gap between what the news media publishes online and what the general public shares in social media. Through the analysis of a series of events, including awareness campaigns, natural disasters, governmental meetings and publications, among others, we uncover differences in terms of the triggers, actions, and news values that are prevalent in the two types of media. For instance, we find that actions by individuals, legal actions involving governments, and original investigative journalism feature more frequently as viral events in social media, while meetings and publications by governmental and inter-governmental agencies tend to receive less attention in social media than they do in mainstream media.

The methodology we developed for this study can be extended to other important topics present in the news such as immigration, pandemics or human rights issues in order to uncover coverage differences among the two media.

4.1. Background

The study of anthropogenic (human induced) climate change goes back more than 100 years,[1] with a scientific consensus on the topic beginning to emerge in the 1980s. By 2014 our planet had registered the warmest year since 1880, when records began to be kept, and 14 of the 15 warmest years on record have all fallen in the first 15 years of this century.[2] Climate change is an issue with myriad impacts being felt and discussed across the globe. The increased salience of the topic has lead to many publications in scientific journals and in the general press, campaigns for legal reforms, and high-profile meetings and talks including the establishment of the IPCC, the Intergovernmental Panel on Climate Change [321]. These various events and publications vie for attention around the issue of climate change—each seeking to define and frame the problems, causes, or potential solutions that are worthy of consideration.

The steady presence of climate change as a topic discussed in media, due to its huge potential consequences,[3] creates a valuable research opportunity for *an in-depth comparative study on how news are communicated through different types of online media*, in particular mainstream news media (MSM) and social media. Understanding these differences offers insights into how such a complex and multi-faceted topic is comparatively covered and framed in these different media, hinting at existing biases between them. Why might some events or actors in the climate change discourse receive more attention in the mainstream media versus on social media, or vice versa? What are the types of news events that receive more attention in both? Ultimately, agenda setting serves to define the problems that are worthy of public attention [97], and we seek to understand and compare the agenda that emerges from traditional MSM attention as compared to the agenda that organically emerges on a social media platform.

4.1.1. Contributions

The main contribution of this Chapter is *a comparison between social media and mainstream news on climate change.* While, typically, it is hard to draw absolute boundaries between topics, the definition of climate change includes very specific elements of interest that allows us to operationalize what are the relevant events (§ 4.2.1). This comparison uncovers significant differences between triggers, actions, and news values of events covered in both types of media. For instance, mainstream news sources frequently feature extreme weather events framed

[1] http://en.wikipedia.org/wiki/History_of_climate_change_science, accessed 01.2015.
[2] http://www.huffingtonpost.com/2015/01/16/2014-hottest-year-on-record_n_6479896.html, accessed 01.2015.
[3] http://en.wikipedia.org/wiki/Media_coverage_of_climate_change, accessed 01.2015.

as being a consequence of climate change, as well as high-profile government publications and meetings. In contrast, actions by individuals, legal actions involving governments, and original investigative journalism, feature frequently as viral events in social media.

We also introduce *a methodology for comparing news agendas online.* This methodology is based on a comparison of spikes of coverage. We analyze two large-scale data sets, both covering a period of 17 months, on news (a global database of about 30 million news articles) and social media postings (a sample of about 2 billion tweets, corresponding to 1% of all Twitter posts). We perform automatic processing to discover terms and topics related to climate change using an iterative procedure. Next, we automatically detect a set of candidate events which are curated through a crowdsourced step of manual annotation. Along this process, we attempt to keep a uniform treatment of both media under analysis. This process offers a starting point for future comparative studies extending to other issues of global attention such as pandemics, global terrorism, or human rights issues.

Next, we outline previous work related to this study. Sections 4.2 and 4.3 present our data processing and annotation methodology. Section 4.4 presents the analysis of results we obtained from this study. The last section summarizes our findings and describes future work to extend this methodology to other domains and social media platforms; it also highlights several challenges and limitations of this study.

4.1.2. Related Work

In this study, we compare media coverage of a broad and long lived social issue: *climate change*. We outline relevant work on climate change discourse (§4.1.2), and describe other comparative studies of social media and news media (§4.1.2).

The Discourse on Climate Change

Climate change has been singled out as one of the most urgent global challenges [139], generating a great deal of interest from communication scholars in recent years. Schmidt et al. [278] perform a transversal study regarding news media coverage of topics related to climate change across 27 countries over a 15-year time frame. They look at the mainstream newspapers of each country and define the relevant articles as matching a specific search query. They found that events such as governmental meetings and report releases trigger increased conversations on climate change, and that such debates are more intensive in carbon-dependent countries.

Across all countries, they observe that media attention about climate change fluctuates and peaks around specific events, which are usually of global interest. This pattern is typical of media reporting in general, which is often characterized by topic peaks [269, 40]. In contrast with these studies, that typically focus on a handful of news outlets, in this Chapter we analyze the news coverage of events across global news media. Furthermore, looking at the coverage volume alone does not reveal nuances about the actors involved in the debate or how climate change is framed.

Molodtsova et al. [231] show that the number of tweets on climate change correlates with extreme weather events, a correlation that also holds for opinion polls on climate change [86]. Along with weather events, [178] and [280] found that other major events of global or local interest ignited discussions about climate change on Twitter as well, including political elections, governmental meetings, and climate-related demonstrations. The study by Kirilenko et al. [177] is closest to ours, as they look at both mainstream media coverage (14 news outlets) and attention patterns in Twitter. They analyze the influence of local weather anomalies on the volume of climate change publications in mainstream media and Twitter. In contrast, we juxtapose these media across a wide range of issues (not only weather) to understand the selection gap between them. Given our goal of *comparing climate change agendas*, we look at certain types of events that are often related to climate change, to seek an answer to which types of events are more prominent in one media or another.

Discourse Comparisons

There is a well-documented difference between what news journalists select to publish, and what their readers consume and share [37]. Journalists have to adhere to deontological ethics and balance between "*public interest and what the public is interested in*" [300], which, in turn, might lead to different attention patterns between social and mainstream news media. Users tend to rely more on their social entourage to filter the news rather than on journalists [134]. Such research motivates our current study that focuses on the comparison of climate change events that emerge in mainstream news media and on Twitter.

Comparative research [98] of Twitter communications includes studies on hashtag life-cycles [197], usage across users of different languages [138], or food consumption [6]. In this work we study the prominence of different types of news events as found on Twitter and in online news media by focusing on climate-related news.

Newspapers vs blogs.

In contrast to Twitter, comparative research has been applied much more frequently to the analysis of the coverage and framing of various issues across newspapers and/or blogs, including religion [29], surveillance [80], and immigration [82]. Some studies have also examined and compared the climate change discourse between clusters of blogs corresponding to climate change acceptors and skeptics [81, 95]. Instead, the focus of this study is not the debate between acceptors and skeptics, but the ways in which different news events feature in different media.

Other studies compare news media with blogs, showing that there is a few hours lag between the attention peak of a meme (short sentence or phrase) in mainstream media and blogs [200]. The media frames—the different ways of communicating about an issue—have also been studied to gain understanding into their impact on the perceptions about news [253], as they are one important tool to shape public opinion [76]. In this Chapter we depict news events to reveal nuances about the factors related to spikes in coverage of an event in mainstream news media and social media.

4.2. Data Collection and Candidate Events

In this section, we define the class of events we are interested in (§4.2.1), explain how we collected news articles (§4.2.2) and social media postings (§4.2.3) and describe the event detection framework used to generate candidate events (§4.2.4).

4.2.1. Defining "Climate Change" News

Our analysis is grounded in the current understanding of the discourse on climate change. For the purposes of this study, by the *discourse on climate change* we mean the discussion around its anthropogenic causes, adopting the definition used in the United Nations Framework for Climate Change (emphasis added): "a change of climate which is attributed directly or indirectly to *human activity* that alters the composition of the *global atmosphere* and which is in addition to natural *climate variability* observed over comparable time periods."[4] The three elements we have emphasized in this definition delimit the scope of the news we consider:

[4] http://unfccc.int/key_documents/the_convention/items/2853.php, accessed 03.2015.

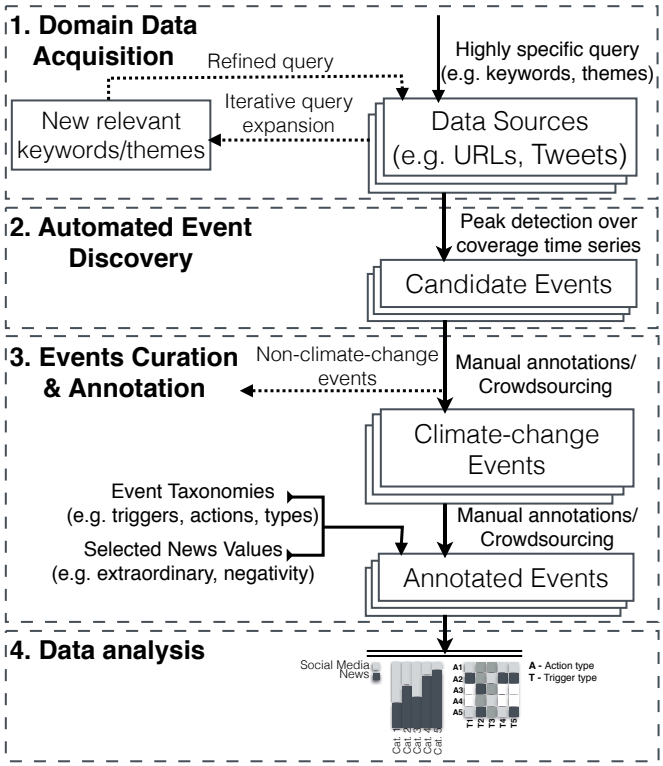


Figure 4.1.: The main steps of the analysis framework employed for this study: (a) domain data acquisition (§4.2.2 and §4.2.3), (b) automated event discovery (§4.2.3), (c) events curation and annotation (§4.3), and (d) data analysis (§4.4).

(i) the presence of human activity as causes,
(ii) effects in the global atmosphere, and
(iii) variations of climate.

Even with this definition, the boundary delimiting which news are related to climate change and which are not, is by no means absolute. Articles about climate change cover a large number of topics that vary from causes (e.g., CO_2 emissions, deforestation) to consequences (e.g., melting Arctic ice, extreme weather), current practices (e.g., fracking, coal use) and actions to stop it (e.g., electrical cars, recycling), just to name a few. Topics such as climate-induced migration and risks to food security, among others, are also frequently included in a long list of consequences of climate change: "we will continue to see rising oceans, longer, hotter heat waves, dangerous droughts and floods, and massive disruptions that can trigger greater migration, conflict, and hunger around the globe."[5]

[5]US President Obama on climate change in the State of the Union Address: http://whitehouse.gov/sotu, accessed 01.2015.

We consider that a news article is about climate change if *it operates within the climate change frame*, in which framing is defined as a set of actions described by [186]:

(i) defining the problem,
(ii) diagnosing its causes,
(iii) making a moral judgment, or
(iv) suggesting a remedy.

We do not look for causation links between a news event and climate change, e.g. whether a severe storm is caused by climate change or not. Instead, we examine the way in which the storm is framed in the news, in this case, if it is described as being part of the climate change problem.

Our data sampling attempts to have a broad coverage of events related to climate change, which results in a set of *candidate events*, including several false positives. In Section 4.3 we describe a manual annotation process by which we remove spurious events.

4.2.2. News Data Acquisition

We use news data collected by GDELT (Global Data on Events, Location, and Tone)[6] and social media data from Twitter covering the same time interval.

Mainstream Media Collection. We use GDELT, currently the largest global event catalog, to automatically discover relevant events with high mainstream media coverage. GDELT releases data about daily media coverage in two formats: the Event Database and the Global Knowledge Graph (GKG).[7] GDELT covers a "cross-section of all major international, national, regional, local, and hyper-local news sources, both print and broadcast, from nearly every corner of the globe, in both English and vernacular"[8] including major international news sources.

We use GDELT's GKG, as it provides the number and the list of news articles covering each event from their database, to discover the list of climate-change related events that received moderate to high media coverage between 1st April 2013 to 31st September 2014, barring January 2014 for which Internet Archive missed Twitter data [9]; this covers 17 months. However, given that we are interested in the peak in the coverage, rather than in the number of events, for

[6] http://www.gdeltproject.org accessed 03.2015
[7] http://gdeltproject.org/about.html accessed 03.2015
[8] http://tm.durusau.net/?p=47505 accessed 03.2015
[9] To analyse Twitter data we rely on the historical archive available at Internet Archive https://archive.org .

this study we directly use the news articles, not the events automatically mapped by GDELT; applying a consistent methodology for detecting events.

To locate the URLs corresponding to news articles relevant to climate change, we rely on GDELT themes and taxonomies, which are topical tags that automatically annotate events. To systematically identify all the GDELT themes and taxonomies that are related to *climate change* we first built the co-occurrence graph among them. We start with a set of relevant themes/taxonomies containing only the `ENV_CLIMATECHANGE` theme (used to annotate news articles discussing climate change and global warming in GDELT), and iteratively add themes, respectively taxonomies, that co-occur for at least 25% of their corresponding URLs with the ones already in the set (the relevance test). We do so until no new theme/taxonomy is added. This results in a set of 39 themes (full list in our data release, details in Appendix A.3). Then, we extract all the unique URLs corresponding to events annotated in GDELT with one of these themes for each day. The resulting collection of 561,644 URLs contains an average of about 30,000 URLs per month, with over 80% of the tags being tagged with the theme `ENV_CLIMATECHANGE`.

4.2.3. Social Media Data Acquisition

Next, we use data from Twitter, a common place for news consumption and conversation, which is also monitored by United Nations as "*measuring these conversations can help reveal what climate issues are discussed most, and where such topics are prioritized.*"[10]

We rely on publicly available data covering about one and a half years of Twitter's Sample API[11], which we then retrospectively sub-sample. The quality of such sub-samples is discussed in the Chapter 5.

To locate relevant tweets we start with a set of highly-specific terms about climate change, e.g. *climatechange, global_warming* [254, 178]—see Appendix A.3—which we then expand in a snowball fashion as we did for themes/taxonomies in GDELT.

Candidate Term Selection. Given a Twitter collection obtained by sampling with a set of keywords $K_{climate}$—deemed relevant for climate change—we detect new relevant keywords by (1) extracting uni-grams and bi-grams that co-occur with terms in $K_{climate}$, and (2) rejecting

[10] http://unglobalpulse.net/climate/about/, accessed on 01.2015

[11] These tweets are collected via Twitter's Sample API and can be found in the Internet Archive: https://archive.org/details/twitterstream, accessed on 01.2015.

those infrequent (occurring in less than 25 unique tweets[12]) or that contain only verbs, adjectives or adverbs (e.g. verb: *run*, adj.: *beautiful*, adv.: *often*)—typically not specific to any domain. When both a bi-gram and the uni-grams contained on it appear in this set, we keep only the bi-gram (the more specific term) if it accounts for more than one third of the uni-grams' frequency, otherwise we keep the uni-grams. Such automated approaches tend to mis-detect less precise terms e.g. *year, park, hell, light* [245], which we manually filter out. We refer to the remaining set as the *candidate terms*, K_{cand}.

Then, using the remaining terms, we build the co-occurrence graph with the terms in $K_{climate}$ and K_{cand}, and select from K_{cand} the terms that co-occur in at least 25% of unique tweets[13] with terms from $K_{climate}$ (the relevance test). The creation of a co-occurrence sub-graph in Twitter is done in a stream processing fashion, avoiding loading the entire data in memory. Thus, we extract the tweets matched by terms in K_{cand}, mimicking the way in which Twitter does keyword tracking on both tweet text and the URLs contained on it. Then, we test each term from K_{cand} for relevance to climate change as described above. We keep repeat this process 5 times, discovering a total of 230 terms (full list available in our data release). Qualitatively, terms discovered in the last passes are less obviously about climate change than the terms discovered in the initial passes. This results in a collection of 482,615 tweets, an average of about 28,000 tweets per month. Given that this is a 1% sample, our estimate is that the tweets in our sample are representative of a larger set of around 2.8M or more tweets per month related to climate change.[14]

4.2.4. Events Discovery

We analyze attention patterns in the scale of days and roughly follow the heuristic for activity peak detection used by [197]. To identify coverage peaks we compute the time series of the aggregated daily coverage in GDELT (respectively Twitter)—where the coverage is the number of URLs (respectively, tweets), c_i for each day d_i—and use a sliding window of $2m + 1$ centered around day d_i, with $m = 15$—resulting in a month-long time window. Then, within each window we juxtapose the volume on d_i, v_i, with a baseline represented by the median volume within the window. We declare a peak if v_i deviates more than 1.5 median average deviations

[12] We correct term frequency to account for cases when their prominence is caused by frequent bi-grams in $K_{climate}$.
[13] For this study we compute statistics over the set of unique tweets to avoid biases due to viral tweets.
[14] While sampling from the entire set of tweets might yield a slightly different set of terms, given that the terms co-occurance threshold is set as a fraction of the unique tweets, we expect the variations to be small.

(MAD) from the mean;[15] and $v_i > t_r$, where $t_r = 50$ is an arbitrary value used to filter out low-frequency peaks which tend to be vague.

This resulted in 218 peaks represented as a ⟨date, theme⟩ pair for GDELT, and 428 ⟨date, keyword⟩ pairs in Twitter.

Detectable events. The attention patterns of Twitter keywords have been described as belonging to three classes [197]: (i) continuous, i.e. having a relatively constant volume, (ii) periodic, i.e. having spikes at regular periods, and (iii) isolated, i.e. having singular peaks. Similar observations hold for news consumption [200, 51]. As detailed by Lehmann et al. [197], the method we discussed above will miss events that do not peak when observed at a granularity of one day, e.g. events that build slowly over weeks or months, or smaller phenomena occurring at a finer granularity (i.e., at the level of hours, minutes or seconds).

Events identification. We annotated each detected peak with the most likely event that triggered it. This annotation often takes the form of a news headline. To assist the event identification, we computed the frequency of uni-grams, bi-grams and tri-grams based on the text of the corresponding URLs (respectively, tweets). Then, we manually checked the items containing the most frequent n-grams, based on which we annotate the event. When two different sets of frequent n-grams referred to different events (e.g., the peak was due to two concomitant events) we add both of them; otherwise, if there were not clear sets of frequent n-grams referring to a single event, we mark the peak as *ambiguous*. When two different pairs ⟨date, theme/keyword⟩ referring to the same event co-occured within a half of month time window we map them to a single entry in our event list (e.g. typically a meeting or a natural hazard that lasted for several days).

This resulted in 195 candidate events in GDELT, out of which we marked 14 as *ambiguous* (possibly related to more than one news event); and 202 candidate events in Twitter, with 22 marked as *ambiguous*. Further, we note that many of the candidate events in Twitter were duplicates. For instance, a cartoon of a polar bear mending an iceberg with duct tape[16] peaked on 4 non-consecutive days in June and July 2014. We mark 12 such cases as duplicates. Thus, after removing duplicates and ambiguous events, we remain with 181 events in GDELT, and 168 events in Twitter.

[15] We chose MAD for its' robustness [201], but also experimented with standard deviation, and the deviation function used by [197], obtaining similar results.

[16] https://twitter.com/thereaibanksy/status/526438158742081537 among many others.

4.3. Events Filtering and Annotation

As noted in the previous section, some of the automatically-identified events are not related to climate change. Two annotators, the author of the thesis and another co-author of this study, reviewed each event to remove false positives (§4.3.1) and to classify each event according to a taxonomy we present in this section (§4.3.2). Finally, we annotate events according to how they are perceived in terms of news values (§4.3.3). This section describes the annotation process, with the analysis deferred to the next section.

We use a mixture of annotation done by the same two annotators and by crowdsource workers through the Crowdflower platform,[17] selecting workers in countries having a majority of native English speakers, collecting 5 independent annotations for every element (3 for the easier task of false positives removal), resolving disagreements by majority voting, and using a set of unambiguous test questions provided by the author of the thesis to catch inattentive workers, following standard recommendations from this platform.

4.3.1. False Positives Removal

The automatic data collection described above was designed to be inclusive, which has the disadvantage that some non-climate-change events get included in both the mainstream news and the social media collection.

Again, the two annotators review each one of these candidate events to remove false positives, i.e. events that do not match the definition given in Section 4.2.1. Two URLs were sampled from each event, including a Wikipedia entry or official activity/publication page when available, or the URL of a tweet, when no URL was available. Some cases are trivial to label, for instance when "climate change" or "global warming" are mentioned in the headline of a news article linked from the event. In many cases, however, the reference to climate change is indirect, e.g. a protest by Greenpeace against Procter and Gamble which is presented as an action against deforestation, a cause of climate change according to the manifesto inviting to this demonstration.

Out of the 181 non-ambiguous news candidates, 122 (67%) were accepted, 43 (24%) rejected and 16 (9%) marked as borderline.[18] From the 168 non-ambiguous and non-duplicate[19] Twitter

[17] http://www.crowdflower.com/

[18] When the event is only marginally associated with climate chance; e.g. while Greenpeace is often involved in climate change campaigns, the "Court Hearing: Greenpeace Activists to stay in jail" story in our event list rather focuses on the trial outcomes.

[19] Duplicate processing is described in §4.2.4.

candidates, 119 (71%) were accepted, 46 (27%) rejected and 3 (2%) marked as borderline.

Next, we contrasted our labels with annotations provided by crowdsourced annotators on the same events. The options given to them were: coding (A) related to climate change, (B) weakly related to climate change, (C) not related to climate change, (D) cannot judge (e.g. broken links, not in English, or other issues).

Mapping them to our assessment of the same events (A and B correspond to accept and borderline, C and D correspond to reject), overall, we observe a 77% agreement with the annotations from crowdsource workers. Specifically, for news (respectively Twitter) 38.1% (resp. 48.2%) of events were labeled as related to climate change, 19.9% (resp. 20.8%) as weakly related, 41.4% (resp. 27.4%) as not related, and 0.5% (resp. 3.6%) as not in English, etc. In general, crowdsource annotators applied a more narrow definition of climate change events, which often overlooked some elements of the news being analyzed. For instance, news about the development of a "Stem cell hamburger" were accompanied by statements from the scientists, in which they indicated that the development of this synthetic meat is motivated by reducing the number of farm animals and hence the methane released to atmosphere that causes climate change. This was missed by annotators who instead indicated this news was not related to climate change.

Disagreements in which crowdsource annotators labeled an event that we accepted as "not related to climate change" were further reviewed by a third annotators[20]. This annotator rejected a further 30 events from that set (23 from news and 7 Twitter). The final list contains 211 events, out of which only 25 events (about 25% from News, and 22% from Twitter) appear in both lists.

4.3.2. Event Annotations

We annotate each event according to a series of types and sub-types from previous work, as summarized in Table 4.1. According to the literature we cite in the table, climate change coverage in the news is often triggered by either a disaster, or by statements or actions of a group of people, or in some cases an individual. In the case of disasters, we further classify them as natural or human-induced [99]. In the case of statements and actions of people, we divide them into the following categories of actors, following observations from previous work cited in Table 4.1:

– *Governmental organization:* Any institution belonging to any government branch (executive, legislative, judicial), or any inter-governmental agency, or any government employee acting

[20] Also involved in the study.

Table 4.1.: Typology of events covered in media, in relation with categories described in previous work.

Type	Sub-Type	Examples	Related categories in previous work
Disaster	Natural Hazards	Typhoon, Drought	(extreme) weather events [278, 231, 86, 178]
	Human-Induced Hazards	Deforestation, Oil Spill	deforestation [40]
Government (all branches) and inter-governmental agencies	Legal actions	New legislation	legislation, policies, international agreements [278, 135], executive actions, police arrests [135]
	Publication/studies/research	Government-sponsored study	(inter-)government reports [278, 254], public surveys [178]
	Meetings/Conferences	IPCC meeting	(inter-)government meetings [278, 178, 280]
	Other (e.g. campaigns, statements)	City installs recycling bins	political elections, climate change adaptation database launch [178, 254], Keystone project [135]
Groups, NGOs, and universities	Legal actions by NGOs	Lawsuit initiated by NGO	petitions [178, 135]
	Publication/studies/research	Academic research	scientific studies [178]
	Other (e.g. campaigns, statements)	Direct action, e.g., cleaning a beach	labor unions, environmental groups statements [278], awareness campaign [178, 135], march/protest organization [280]
For-profit (excl. media, universities)	Legal actions by for-profit entity	Lawsuit by for-profit group	—
	Publication/studies/research	Reports by for-profit group	scientists funded by carbon-based industries, memos [40]
	Other (e.g. campaigns, statements)	Google invests in solar energy	energy industry activities [278]; paid media campaigns [40]
Media	Publication/studies/research	Newspaper investigation	reports/investigation by news media [278, 178, 254]
	Other (e.g. campaigns, statements)	Campaign by newspaper	media activism [280], campaigns [254],
Individuals	Legal actions by individuals	Lawsuit by individuals	—
	Publication/studies/research	New book	opinion/editorial [178]
	Other (e.g. campaigns, statements)	Bill Gates funds climate research	elite person campaign [278], statements/changes in opinion by individuals [178]

in official capacity.
- *Non-governmental organization:* Any non-profit, non-governmental group, formally established or not. We include in this category educational and research institutions, which are all universities in our data set.
- *For-profit organization:* Any for-profit organization, including business and corporations but excluding media and universities, which appear in the other categories.
- *Media organization:* Any media organization.
- *Individual:* Any individual that is not acting as a representative of any of the organization types listed above.

We further categorize the actions of organizations or individuals as follows:

- *Legal actions:* Any action that is legally binding, including new executive orders and new laws, plus any action brought to a court of law, such as lawsuits.
- *Publications:* Any release of a document to the public, including reports, studies, memoranda, infographics and cartoons.
- *Meetings:* Any meeting, conference, convention, etc.
- *Other:* Other types of actions not belonging to the categories above, in our data this corresponded mostly to campaigns and brief public statements.

The annotation was done by the two annotators. We noted that an event can have more than one trigger, and we took this into account in our annotation, associating a second trigger to some events when deemed necessary.

Then, we again contrasted our labels with annotations provided by crowdworkers on this set of events. Workers were provided the same categories detailed above and were asked to choose the most likely one for each event (i.e., only one type and sub-type). Mapping our assessment to theirs—we consider agreement if they choose one of our labels (either the first or the second type/sub-type)—we observe a 80.1% agreement for sub-types.

4.3.3. News values

Finally, to understand *why* a certain event is covered prominently, we annotate the events according to *news values*. News values are factors that determine the prominence with which an event is covered in the news. There are many news values, see e.g. the lists by Harcup and O'Neill [129] and Stovall [297]. For the purposes of this analysis, after inspecting the list of events labeled as related to climate change, we decided to study the following six:

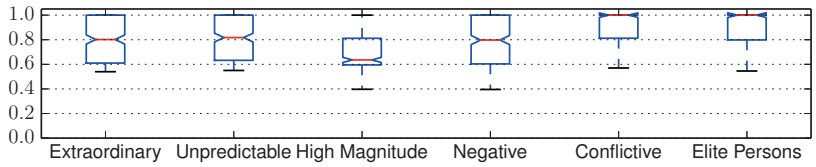


Figure 4.2.: Distribution of confidence (a weighted measure of agreement among workers) in annotations, with 1.0 indicating complete agreement.

– *Extraordinary:* the event is out of the ordinary or rare.
– *Unpredictable:* the event could not have been anticipated.
– *High magnitude:* the event has large global consequences.
– *Negative:* the event represents bad news.
– *Conflictive:* the event involves two persons/groups in antagonism.
– *Related to elite persons:* the event involves someone rich, powerful or famous.

We do not claim these are all the news values that matter in this case, but given limited resources for annotation, bounding the number of them is necessary. This annotation is done through crowdsourcing by using instructions that echo the list above (full text of instructions, plus examples given to annotators for each class, are available in Appendix A.4).

We note that some of these tasks are more subjective than others, and hence elicit a lower level of agreement, as measured by the distribution of the agreement of annotators on each task (a value reported as the *confidence* on each annotation by the crowdsourcing provider). For instance, In Figure 4.2 we see that while references to elite persons and conflictive news are labeled with higher confidence (*median*=1.0), whether a news item is of high magnitude is a judgment in which there is less agreement among annotators (*median*=0.6). Other news values have in general a high level of agreement (*median*=0.8).

4.3.4. Examples

The full annotated data set is available for research purposes (see Section 4.5.4). Some examples are the following:

- *"Climate refugee fighting stay in New Zealand,"* covered by news media and discussing the legal actions taken by a man from Kiribati Islands and the New Zealand government regarding an asylum request, was annotated as *neutral news* of *low magnitude*, yet *extraordinary* and *unpredictable*, and depicting a *conflict* between two entities.

Table 4.2.: Types and sub-types of events found in our data set. Numbers add up to more than 100% because one event may have more than one type. Distributions are significantly different at $p < 0.01$.

	Disast.	Gov.	Non gov.	For-profit	Media	Indiv.
News	20.2%	62.6%	32.3%	6.1%	1.0%	4.0%
Twitter	7.1%	52.7%	29.5%	5.4%	8.9%	14.3%

	Disaster		Actions			
	Hum.	Nat.	Legal	Publ.	Meet	Other
News	3.0%	17.2%	19.2%	46.5%	13.1%	27.3%
Twitter	0.9%	6.3%	22.3%	37.5%	3.6%	47.3%

- *"Climate change expert pleads guilty for fraud,"* debated on Twitter and discussing the fraud committed by a climate expert and former employee of the US Environmental Protection Agency, was annotated as *bad news* of *low magnitude*, yet *extraordinary* and *unpredictable*.

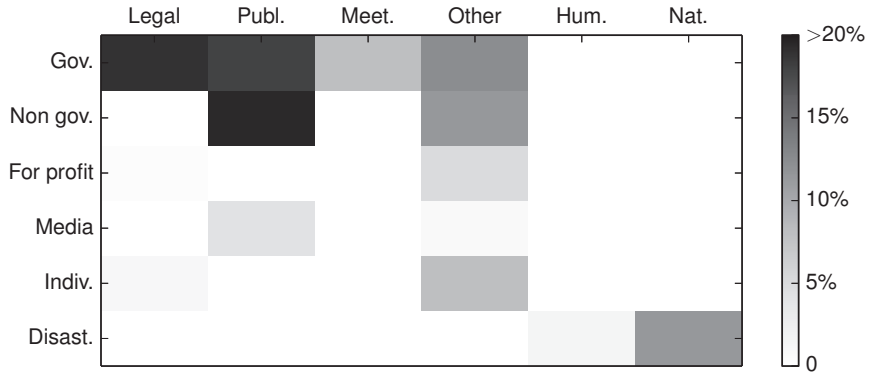
- *"Typhoon Haiyan,"* covered by news media and debated on Twitter as an event related to climate change leading to significant human and material loss, was annotated as *bad news* of *high magnitude*, *extraordinary* and *unpredictable*.

4.4. Data Analysis

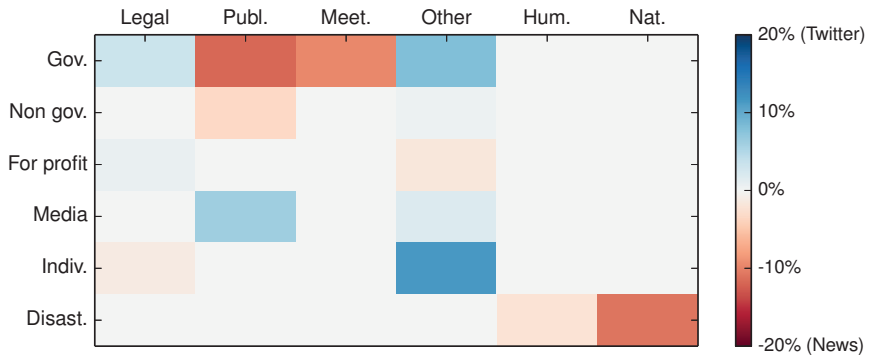
This section presents our observations regarding event types (§4.4.1), news values (§4.4.2) and their interaction (§4.4.3).

4.4.1. Event types

Table 4.2 presents differences in coverage between mainstream media (MSM) and Twitter as present in our data set. We observe significant differences in terms of coverage of disasters, which MSM favors much more than Twitter (20% vs. 7%); in the presence of media-triggered events—such as the publication of an investigation by a newspaper, which is an infrequent event in terms of global news coverage but does trigger significant reactions in Twitter (1% vs. 9%); and in the coverage of individual actions, which are given less prominence in news compared to Twitter (4% vs. 14%).



(a) Distribution of events into types and sub-types. Darker cells contain more events.



(b) Comparison of the distributions of events in main stream news and Twitter. Red indicates more events in mainstream news, blue indicates more events in Twitter (best seen in color).

Figure 4.3.: Distribution of types and sub-types.

There are interesting similarities and differences between the types of actors and actions covered in both types of media, as depicted in Figure 4.3:

– Government/inter-governmental agencies, which receive the largest amount of coverage in both (top row of Figure 4.3(a)), are discussed in relation to a broad range of action types. The main difference seems to be a larger coverage of publications and meetings in MSM, contrary to coverage of legal and other types of actions which are covered more often in social media (top row of Figure 4.3(b)).
– Non-governmental groups (and universities), are covered in both cases mostly due to publications, and also through other actions (second row of Figure 4.3(a)) such as campaigns and public statements.
– For-profit organizations are covered mostly due to other actions (third row of Figure 4.3(b)), which are usually advertising and announcements of projects.

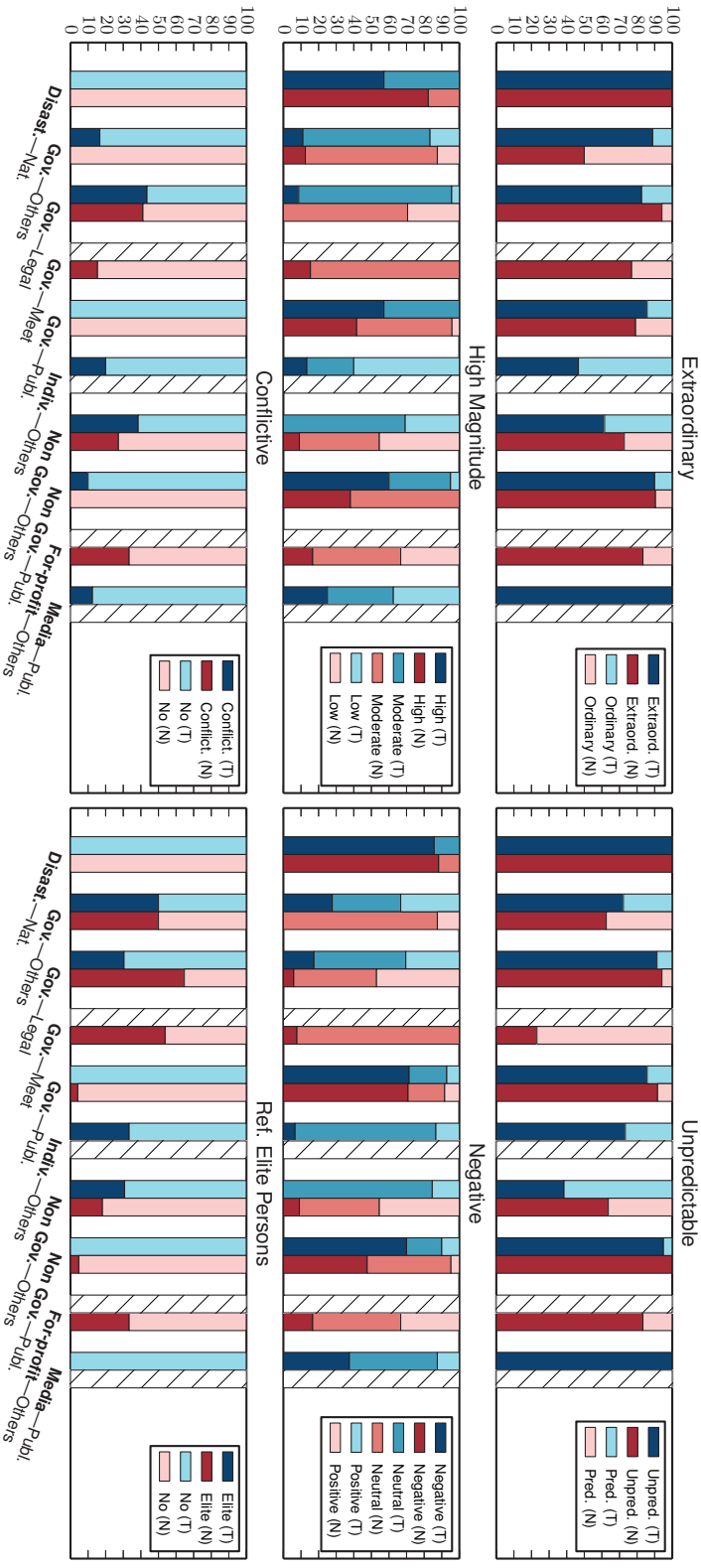


Figure 4.4.: Distribution of news values for types/sub-types of events in Twitter (T, in blue) and mainstream news (N, in red). Hatched bars indicate insufficient data (less than 5 events). (Best seen in color.)

Table 4.3.: Analysis in terms of news values of events covered in our mainstream news and Twitter data sets. Asterisks in the last row highlight statistically significant differences at $p < 0.01$ (***), $p < 0.05$ (**), $p < 0.10$ (*).

	Extraordinary		Unpredictable		High Magnitude		
	Extraord.	Ordinary	Unpred.	Pred.	High	Moderate	Low
News	83.8%	16.2%	82.8%	17.2%	34.3%	54.5%	11.1%
Twitter	75.9%	24.1%	80.4%	19.6%	25.0%	55.4%	19.6%
Both	79.6%	20.4%	81.5%	18.5%	29.4%	55.0%	15.6%
	**					**	

	Negative			Conflictive		Ref. Elite Persons	
	Negative	Neutral	Positive	Conflict.	No	Elite	No
News	43.4%	40.4%	16.2%	10.1%	89.9%	22.2%	77.8%
Twitter	34.8%	46.4%	18.9%	18.8%	81.3%	21.4%	78.6%
Both	38.9%	43.6%	17.5%	14.7%	85.3%	21.8%	78.2%
				**			

– Media organizations become protagonists with respect to climate change through their original investigative reporting (fourth row of Figure 4.3(a)), yet, the number of events they create in news is lower than in Twitter (fourth row of Figure 4.3(b)). This is because an original investigation by one news source will rarely be quoted by many other news sources, but it can have a significant impact in Twitter.
– Individuals are covered in both media occasionally with respect to actions (fifth row of Figure 4.3(a)), which are usually public statements. Individuals receive much more attention in Twitter than in traditional news media (fifth row of Figure 4.3(b)).
– Disasters in general are covered more by mainstream news than by Twitter, mostly due to their coverage of natural disasters (last row of Figure 4.3(b)). Disasters have been observed to be a prominent subject in international news articles collected by GDELT [187].

As regards the 25 events that were prominently covered by both media, 60% were primarily triggered by government/inter-governmental agencies (e.g. *"UN Climate Summit 2014"*), 24% were campaigns or publications by non-governmental groups (e.g. *"2013 Earth Day/Week"*) and 16% natural disasters (e.g. *"Typhoon Haiyan"*).

Overall, our results serve to quantify the thematic gap between the type of news events the mainstream media focuses on, and the types of events that gain interest and attention on social media [37].

4.4.2. News values

The analysis of news values in our data, shown in Table 4.3, indicates that both media tends to cover events that are

(i) extraordinary,
(ii) unpredictable,
(iii) of moderate and high magnitude, and
(iv) negative or neutral.

Contrary to what one could assume given the literature on news values, events involving conflict or referencing elite persons are not the majority of news about climate change.

There are significant differences between MSM and Twitter, with relatively more coverage given in Twitter to ordinary events in comparison to MSM. Twitter also has relatively more coverage of events that are considered of relatively low magnitude and that involve two groups or individuals in antagonism.

4.4.3. Event types and news values

The differences in terms of news values that we observe between MSM and social media are largely correlated with the selection of events they cover, as can be observed by comparing both media in the same category. This is depicted in Figure 4.4, which only includes cases when there are at least 5 events. News values for the same type of event are often similar between these two media, save for small differences.

While most of the reported events on climate change are extraordinary (rare), there is one exception in which more ordinary news events are the majority, which are individual actions, featured significantly (5 events or more) on Twitter, but not in MSM. While most events are also unpredictable, there are events announced well in advance, such as governmental and inter-governmental meetings. These news events only feature significantly on MSM, but not on Twitter.

As regards magnitude, disasters and publications are often the ones linked to the largest effects. In the case of disasters, it is by their consequences: most of the disasters that are associated by the press or Twitter to climate change are severe weather phenomena affecting large areas. In the case of publications, this has more to do with the content of the publications, sometimes describing existential threats to humans as a whole. The overall lower magnitude of events

covered in comparison to news may be explained by the confluence of two observations:

(1) Twitter focuses more on events with individual triggers which tend to have lower magnitude ratings, and
(2) MSM focuses more on disaster events which tend to have higher magnitude ratings.

In terms of negativity, most news are neutral or bad. Proportionally, the most negative news are those related to disasters and to publications.

While most news events do not involve people/groups in conflict, the cases in which they have a more conflictive content are legal actions by governments (which usually are targeted at a specific group, such as a mining corporation), and statements and actions by non-governmental organizations (e.g. statements by an NGO against a certain industry).

Finally, references to elite persons (famous, rich, or powerful) are almost never included in publications (governmental or non-governmental sources), but are present in some minority amount in the remaining categories.

4.5. Conclusions

This section outlines our main conclusions regarding how similar social media news are to online mainstream media news. We include a summary of the observations about the differences in the discourse on climate change in the two media by emphasising on different types of news they tend to focus on (§4.5.1). We, thus, show that at least for online mainstream news media, social media, in particular Twitter, is by no means a perfect proxy. We also discuss how to extend the methodology for comparing news coverage across media that we introduce in this Chapter to other domains, as well as other media (§4.5.2).

4.5.1. Climate Change in Mainstream News and Social Media

From the domain and application side, there are interesting similarities between mainstream news and social media, both in terms of the types of events they cover and the news values of those events. However, there are also striking differences, as the activity of the two media tends to peak around different news events, with an overlap in their peaks of attentions of about 22%–25% of the events.

Disasters. A key trigger of news coverage on climate change are disasters, both natural and human-induced. Disasters covered with respect to climate change tend to be severe atmospheric events affecting large parts of the globe. There is an important difference between online mainstream news media and Twitter, with mainstream news media covering these events much more than Twitter.

Publications, meetings, and legal actions. News events on climate change are usually triggered by publications describing negative, global-scale consequences of climate change. News coverage of climate change is also triggered by legal actions initiated by governments, like passing new laws and bringing lawsuits against corporations. The coverage of these events differs in online mainstream news media and Twitter. In mainstream news media, government/intergovernmental meetings and publications receive comparatively more attention than in Twitter, where legal actions and official statements have a greater impact.

Individual actions. Actions by individuals appear prominently on Twitter. In about half of the cases, these individuals do not belong to the elite: they are neither rich, nor powerful, nor famous. Twitter indeed allows those individuals, in many cases, to generate peaks of attention as large as the ones that are obtained by large organizations or governments.

Recommendations. For *activists and advocates*, publications highlighting high-impact negative effects of climate change feature prominently across both types of media. Additionally, they also seem to be picked up by social media even when they do not include endorsements by elite persons or references to them. For *public relations or for-profit corporations*, discussions about lawsuits involving corporations, while not appearing so prominently in mainstream media, circulate in social media. For *media organizations*, the alignment between mainstream news and social media news on this topic is significant, but there are many gaps. It would not be unreasonable to look at what are the news events in which there is the larger gap in favor of social media, particularly actions and public statements by individuals, as opportunities to disseminate information that may appeal to social media users.

4.5.2. Towards a General Method for Comparing Online Media

The method we have presented here can be extended to a variety of topics in the news. GDELT associates news articles to hundreds of themes, enabling analysts to perform the same procedure we have described for other themes (e.g. `HEALTH_PANDEMIC`, `IMMIGRATION`). In the case of Twitter, any topic from which a subset of initial hashtags can be identified is amenable to the

same event discovery process. To extend this method to other social media platforms (e.g. Reddit), these platforms need to minimally support data access allowing to query content through a list of keywords—unless one has full access to a social media platform data.

Two important elements require adaptation. First, the triggers and actions should be specific to the topic, although some overlap with the ones we have used here is expected (i.e. government, non-government, for profit, etc.). Second, the selection of the relevant news values also requires some familiarity with the topic, and as in this Chapter, it is hard to claim it is in any sense an optimal selection.

Applying this event-driven methodology to the discovery of differences between mainstream media and social media in other domains may lead to findings as interesting as the ones we uncover here. These findings can be contrasted with those from qualitative analysis, particularly of events that generate peaks of attentions in both media simultaneously.

4.5.3. Limitations

Finally, our approach too is not devoid of *challenges and limitations*, which include:

- We seek a deeper understanding of the climate change discourse, and we do not attempt to test and validate our methodology across multiple domains—we have outlined how this can be done in the previous section (§4.5.2).
- We do not describe patterns of consumption attention, but rather patterns of coverage, or output attention. In other words we are not claiming that a certain issue is more read, but that it is more written about.
- We use a period of time of 17 months, while climate change has been discussed in the news for decades.
- We cover only one language (English), but we note that it is the language in which most reports triggering this debate are written, including the ones by IPCC.
- We cover only one social media source (Twitter), but it is a large one and it is frequently associated with news [188].
- While we study the coverage trends in aggregate, the users on Twitter do not represent a monolithic community and social norms might vary across sub-populations. However, we argue that general trends across all users still provide important insights about significant differences among the two media.
- Both data collections have their own biases. For instance, while the global database of news we use (GDELT) is considered a reliable source of news media coverage across the world [20],

it may also be biased towards US news media, which are comparatively more active media organizations [187].
– Finally, as the absolute number of viral events in both media is small in our study, it is hard to confidently make claims about the interplay between mainstream media and social media—e.g. when they reference or trigger each other.

4.5.4. Reproducibility & Data Release

To ensure and support the reproducibility and replicability of this case study, the data we obtained from conducting it, including themes, keywords, news events, and labels, is available for research purposes at `http://crisislex.org/`.

5. Data Collection Biases: The Case of Crisis Data

To assess possible biases across data collections, in this Chapter, we use a systematic methodology to collect, sample and analyze social media data, and focus on crisis events, as the use of social media to communicate timely information during crisis situations has become a common practice in recent years. Particularly, the one-to-many nature of Twitter has created an opportunity for stakeholders to disseminate crisis-relevant messages, and to access vast amounts of information they may not otherwise have. Additionally, on the application side, our goal is to understand what affected populations, response agencies, and other stakeholders can expect—and not expect—from these data in various types of disaster situations. Anecdotal evidence suggests that different types of crises elicit different reactions from Twitter users, but we have yet to see whether this is in fact the case. In this Chapter, we investigate 26 crisis situations—including natural hazards and human-induced disasters—in a systematic manner and with a consistent methodology. This leads to insights about the prevalence of different information types and sources across a variety of crisis situations—e.g., the intrinsic characteristics of the crisis situations lead to biases with respect to both information type and sources across data collections.

5.1. Background

When a disaster occurs, time is limited and safety is in question, so people need to act quickly with as much knowledge of the situation as possible. It is becoming more common for affected populations and other stakeholders to turn to Twitter to gather information about a crisis when decisions need to be made, and action taken. However, the millions of Twitter messages ("tweets") broadcast at any given time can be overwhelming and confusing, and knowing what information to look for is often difficult.

One way to help those affected by a disaster to benefit from information on Twitter, is to provide

an indication of *what* information they can expect to find. The capacity for affected populations to know what types of information they are likely to see on Twitter when particular kinds of mass emergencies occur, can potentially help them be more efficient in their information-seeking and decision-making processes.

To explore this idea, we collected tweets that were broadcast during 26 different crisis situations that took place in 2012 and 2013. For each crisis, we examine the types of information that were posted, and look at the sources of the information in each tweet. Our specific aim is to measure the prevalence of different types of messages under different types of crisis situations.

Our results suggest that some *intrinsic characteristics* of the crisis situations (e.g. being instantaneous or progressive) produce *consistent effects* on the types of information broadcast on Twitter. The results are of interest to members of the public, emergency managers, and formal response agencies, who are increasingly trying to understand how to effectively use social media as part of their information gathering processes.

5.1.1. Related Work

We know that tweets sent during crisis situations may contain information that contributes to situational awareness [314], and though disaster situations exhibit common features across various events [307], previous research has found that *information shared on Twitter varies substantially from one crisis to another* [165, 236, 245]. Indeed, some variability across disasters is expected. For instance, data from the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) shows that disasters in high-income countries cause significantly more economic damage, but affect fewer people and have fewer fatal casualties, compared to disasters in countries with low or middle incomes [240].

Comparative research is an established discipline in communication studies [98], but to date, this method has not been extensively applied to the study of social media communications during crises. There is little overlap in the crises examined across research groups, and no attempt to date to apply the same methodology consistently to a large and diverse set of crises. The literature review by Fraustino et al. [103] indicates that research on social media during disasters "*tends to examine one catastrophic event (...) and then imply that the findings are generalizable to other disasters.*"

In our attempt to fill this gap, we examine tweets that were broadcast during a broad range of different crisis situations, and systematically apply the same methodology to the analysis of

each event. To understand the information posted in Twitter during disasters, this methodology is based on previous work that categorizes tweets by type (including [7, 48, 153, 245, 316]) or by source (including [72, 79, 185, 221, 232, 293]). The taxonomy we employ in this Chapter to categorize tweets builds upon both of these categorizations.

5.1.2. Contributions

For decision-makers and other stakeholders to be as prepared as possible, knowing what information they are likely to gain from social media can save time and help them decide where to direct their often limited resources. When stakeholders know what types of content to expect (e.g., advice, support, damage reports), and which information sources will be prevalent (e.g. news organizations, eyewitnesses, NGOs), they do not have to sift through masses of social media posts; instead, they have a reasonable expectation of what they will find, and can then make more informed decisions regarding their situational assessment process.

Based on our goal to ease the information overload wrought by social media during crisis situations and to gain insights about possible biases across data collections, the question we address in this Chapter is: *what are the similarities and differences in Twitter communications that take place during different crisis events, according to specific characteristics of such events?* To answer this question, we study the prevalence of different information types and sources found on Twitter during different types of crises, and correlate this with some of their intrinsic characteristics.

5.1.3. Methodology Overview

To perform this study, we employ the following methodology:

Step 1: We determine a set of dimensions that allow us to characterize different crises: hazard type, temporal development, and geographic spread. This choice is grounded on the emergency-response literature.

Step 2: We determine a set of dimensions to characterize social media messages during a crisis: informativeness, information type, and source. This choice is grounded on the literature on social media use for emergency management.

Step 3: We collect Twitter data corresponding to 26 crises that took place in 2012 and 2013, using retrospective sampling on the 1% public data stream which is publicly available in

the Internet Archive.[1]

Step 4: We create, run, and evaluate a series of crowdsourcing tasks to perform content annotation on approximately 1,000 messages from each of the crises. This is informed on the crowdsourcing and emergency management literature.

Step 5: We perform a statistical analysis of the dependencies between types of crises and types of messages to uncover both similarities and differences across data collections.

5.2. Step 1: Determining Crisis Dimensions

Given that our research question connects two domains: disaster studies, and social media content analysis, the framework we use is composed of two parts. We categorize the crises according to a series of dimensions that characterize them. Next, we annotate tweets from each crisis according to dimensions that characterize different types of content.

When considering how to organize our data and approach our annotation process, we turned to dimensions used in the sociology of disaster research (p. 50 in [255]). For each crisis, we consider hazard type (natural vs. human-induced), sub-type (e.g. meteorological, hydrological, etc.), temporal development (instantaneous vs. progressive), and geographic spread (focalized vs. diffused).

5.2.1. Hazard type

Hazard type is the first dimension we examine that may impact the types of contents disseminated through social media. The specific hazard types we consider are based on two taxonomies used in Europe[2] and the US,[3] as well as the traditional hazard categories listed by Fischer [99].

The first distinction is between those that are *natural* and those that are *human-induced*. Subcategories and examples of each one are listed in Table 5.1. All sub-categories are covered by crises analyzed in this study, with the exception of the "biological" category, which we were unable to sufficiently account for regarding Twitter communications.

[1] https://archive.org/details/twitterstream
[2] http://www.emdat.be/classification
[3] http://www.ready.gov/be-informed

Table 5.1.: Hazard categories and sub-categories.

Category	Sub-category	Examples
Natural	• Meteorological • Hydrological • Geophysical • Climatological • *Biological (N/A)*	• tornado, hurricane • flood, landslide • earthquake, volcano • wildfire, heat/cold wave • *epidemic, infestation*
Human-Induced	• Intentional • Accidental	• shooting, bombing • derailment, building collapse

5.2.2. Temporal Development

When considering the temporal development of crises, we classify them as *instantaneous* (e.g. an earthquake or a shooting), or *progressive* (e.g. a hurricane or a heat wave) [8, 49, 252]. As we qualitatively coded the temporal aspects of the crises, we labeled a disaster *instantaneous* if it "*does not allow pre-disaster mobilization of workers or pre-impact evacuation of those in danger*," and *progressive* if it is "*preceded by a warning period*" [8].

5.2.3. Geographic Spread

We look at the geographic spread of a crisis, and specify if it is *focalized* (such as a train accident) or *diffused* (such as a large earthquake) [8, 260]. A focalized crisis affects and mobilizes response in a small area, while a diffused disaster impacts a large geographic area and/or mobilizes national or international response.

Other Crisis Dimensions. We recognize that this list of crisis dimensions is not exhaustive. In particular, linguistic and cultural differences have been shown to influence message content, and the adoption of certain conventions in Twitter, e.g. [138, 257]. We also recognize that these dimensions are not independent from one another. For instance, with the exception of war and large-scale nuclear disasters, most human-induced crises tend to be focalized, while meteorological hazards are often diffused. Additionally, the interplay between these dimensions may yield complex results in terms of the types of information included in Twitter messages, and the source of that information. For example, hazard type combined with geographic spread can affect both the public access to firsthand information about a crisis, as well as their ability to post information about their whereabouts.

Table 5.2.: Typologies of content used in this chapter, and their relationship to some aspects mentioned in previous work

This work	Related categories from previous work
Informativeness:	
Informative	informative (direct or indirect) [153]; curating or producing content [221]; contribute to situational awareness [316]; situational information [282]; contextual information to better understand the situation [290]
Not informative	trolling [221]; humor [195]; off-topic [245, 263, 295]; rumor [145]; humor or irrelevant/spam [290]
Information type:	
Affected individuals	medical emergency, people trapped, person news [48]; casualties (and damage), people missing, found or seen [153]; reports about self [7]; fatality, injury, missing [314]; looking for missing people [263];
Infrastructure & utilities	(casualties and) damage [153]; reports about environment [7]; built environment [314]; damaged, closures and services [145]; collapsed structure, water shortage/sanitation, hospital/clinic services [48]; road closures and traffic conditions [308];
Donations & volunteering	donations of money, goods or services [153]; donations or volunteering [245]; requesting help, proposing relief, relief coordination [263]; donations, relief, resources [145]; help and fundraising [42, 282]; shelter needed, food shortage/distribution [48]; volunteer information [316]; help requests [7]
Caution & advice	caution, advice [153]; warnings [7]; advice, warnings, preparation [245]; warning, advice, caution, preparation [316]; tips [195]; safety, preparation, status, protocol [145]; preparedness [329]; advice [42]; advice and instructions [282]; predicting or forecasting, instructions to handle certain situations [290];
Sympathy & emotional support	concerns and condolences [7]; gratitude, prayers [245]; emotion-related [263]; support [145]; thanks and gratitude, support [42, 282];
Other useful information	fire line/emergency location, flood level, weather, wind, visibility [316]; smoke, ash [308]; adjunctive and meta-discussions [282]; other informative messages [245]; information verification, explanation of particular problems [290];
Source:	
Eyewitness	citizen reporters, members of the community [221]; eyewitnesses [42, 79, 185, 245]; local, peripheral, personally connected [295]; local individuals [291, 316]; local perspective, on the ground reports [306]; direct experience (personal narrative and eyewitness reports) [282]; direct observation, direct impact, relayed observation [308];
Government	(news organizations and) authorities [221]; government/administration [245]; police and fire services [145]; police [77]; government [42]; public institutions [306]; public service agencies, flood specific agencies [295];
NGOs	non-profit organizations [72, 306]; non-governmental organization [245]; faith-based organizations [295];
Business	commercial organizations [72]; enterprises [306]; for-profit corporation [245];
Media	news organizations (and authorities), blogs [221]; journalists, media, and bloggers [72, 79]; news organization [245]; professional news reports [195]; media [42]; traditional media (print, television, radio), alternative media, freelance journalist [306]; blogs, news-crawler bots, local, national and alternative media [295]; media sharing (news media updates, multimedia) [282];
Outsiders	sympathizers [185]; distant witness [50]; remote crowd [291]; non-locals [295, 306].

5.3. Step 2: Determining Content Dimensions

When assessing the tweets that were broadcast during each disaster event, we turned to previous research on information broadcast via social media in disaster. We constructed a coarse-grained categorization that covers the categories of information that are highly represented in previous work (including [48, 153, 221, 314, 316] among others). Due to the large number of events and messages we consider, and the limitations of using crowdsourcing workers to perform the annotation (as opposed to experts, who would be prohibitively expensive at this scale), we formulated basic information categories broad enough to be applicable to different crisis situations. The resulting categories and the previous research represented by them, are shown in Table 5.2: *informativeness*, *information type*, and *source*.

5.3.1. Informativeness

We recognize that informativeness is a subjective concept, as it depends on the person who is asking for or receiving information. In addition, as with any communication, the context in which the information exchange is taking place is critical to understanding its implications. We capture this dimension following Vieweg et al. [316], by checking whether the tweet contributes to better understanding the situation on the ground. Accordingly, we use the following annotation options:

A. Related to the crisis and informative: if it contains useful information that helps understand the crisis situation.
B. Related to the crisis, but not informative: if it refers to the crisis, but does not contain useful information that helps understand the situation.
C. Not related to the crisis.

5.3.2. Information Type

As we closely analyzed a set of samples of messages communicated via Twitter during disasters, we found that the type of content often varies substantially across hazards; a finding corroborated by many other studies [7, 48, 153, 245, 316].

To identify a set of broad categories whose incidence (though with different degrees of occurrence) is to a large extent independent of event specificities, and to obtain a manageable coding

scheme, we first identified the list of information categories used in related work studying various types of events (e.g., wildfires [316], drug wars [221], floods [42, 295], earthquake [263], nuclear power plant [306], to name a few). Then, we proceeded with merging in a bottom-up fashion those categories that overlap and/or are related. Finally, we gathered the remaining categories, typically accounting for information specific to each crisis or type of crisis (e.g., flood level, weather, wind, visibility [316]), into a "catchall" category—*other useful information*. The exact matching of information types present in the related work to each of the categories used in this chapter is depicted in Table 5.2. The information types that we use are:

A. Affected individuals: deaths, injuries, missing, found, or displaced people, and/or personal updates.
B. Infrastructure and utilities: buildings, roads, utilities/services that are damaged, interrupted, restored or operational.
C. Donations and volunteering: needs, requests, or offers of money, blood, shelter, supplies, and/or services by volunteers or professionals.
D. Caution and advice: warnings issued or lifted, guidance and tips.
E. Sympathy and emotional support: thoughts, prayers, gratitude, sadness, etc.
F. Other useful information not covered by any of the above categories.

5.3.3. Source

When people turn to Twitter to learn about a disaster, they are often concerned with the source of information. Hence, we focused on *content source*, which may be different from tweet author; e.g. if the Twitter account of a large media organization quotes a government official, the "source" is the government official. Sources are categorized as: *primary sources* (eyewitness accounts) or *secondary or tertiary sources* (typically mainstream media or others engaged in journalistic acts) [72, 79, 185, 221, 232, 293].

For the former, we chose to broaden the definition of an eyewitness account as originating from "*a person who has seen something happen and can give a first-hand description of it*"[4] to also accommodate those cases when the account does not include a direct observation, yet the user is personally impacted by the event, or it "*is about a direct observation or impact of a person who is not the micro-blogger*" [308]—typically relaying the observations of friends or family.

In the latter case, we can find several organizations who often aggregate information about a crisis, including business, governmental, and non-governmental sources:

[4] http://www.oxforddictionaries.com/definition/english/eyewitness

Table 5.3.: List of crises studied, sorted by date, including the duration of the collection period for each dataset, the number of tweets sampled from the 1% Twitter stream, and several dimensions of the crises

Year	Country	Crisis Name	Days	Tweets	Hazard category	Hazard subcategory	Hazard type	Development	Spread
2012	Italy	Italy earthquakes	32	7.4K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2012	US	Colorado wildfires	31	4.2K	Natural	Climatological	Wildfire	Progressive	Diffused
2012	Philipinnes	Philipinnes floods	13	3.0K	Natural	Hydrological	Floods	Progressive	Diffused
2012	Venezuela	Venezuela refinery explosion	12	2.7K	Human-induced	Accidental	Explosion	Instantaneous	Focalized
2012	Costa Rica	Costa Rica earthquake	13	2.2K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2012	Guatemala	Guatemala earthquake	20	3.3K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2012	Phillipines	Typhoon Pablo	21	1.9K	Natural	Meteorological	Typhoon	Progressive	Diffused
2013	Brazil	Brazil nightclub fire	16	4.8K	Human-induced	Accidental	Fire	Instantaneous	Focalized
2013	Australia	Queensland floods	19	1.2K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Russia	Russian meteor	19	8.4K	Natural	Others	Meteorite	Instantaneous	Focalized
2013	US	Boston bombings	60	157.5K	Human-induced	Intentional	Bombings	Instantaneous	Focalized
2013	Bangladesh	Savar building collapse	36	4.1K	Human-induced	Accidental	Collapse	Instantaneous	Focalized
2013	US	West Texas explosion	29	14.5K	Human-induced	Accidental	Explosion	Instantaneous	Focalized
2013	Canada	Alberta floods	25	5.9K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Singapore	Singapore haze	19	3.6K	Mixed	Others	Haze	Progressive	Diffused
2013	Canada	Lac-Megantic train crash	14	2.3K	Human-induced	Accidental	Derailment	Instantaneous	Focalized
2013	Spain	Spain train crash	15	3.7K	Human-induced	Accidental	Derailment	Instantaneous	Focalized
2013	Phillipines	Manila floods	11	2.0K	Natural	Hydrological	Floods	Progressive	Diffused
2013	US	Colorado floods	21	1.8K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Australia	Australia wildfires	21	2.0K	Natural	Climatological	Wildfire	Progressive	Diffused
2013	Phillipines	Bohol earthquake	12	2.2K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2013	UK	Glasgow helicopter crash	30	2.6K	Human-induced	Accidental	Crash	Instantaneous	Focalized
2013	US	LA Airport shootings	12	2.7K	Human-induced	Intentional	Shootings	Instantaneous	Focalized
2013	US	NYC train crash	8	1.1K	Human-induced	Accidental	Derailment	Instantaneous	Focalized
2013	Italy	Sardinia floods	13	1.1K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Phillipines	Typhoon Yolanda	58	39.0K	Natural	Meteorological	Typhoon	Progressive	Diffused

A. Eyewitness: information originating from eyewitnesses of the event or of response/recovery operations, or from their family, friends, neighbors, etc.
B. Government: information originating from the local or national administration.
C. Non-governmental organization: information originating from NGOs.
D. Business: information originating from for-profit business (except news organizations).
E. Traditional and/or Internet media: information coming from sources such as TV, radio, news organizations, web blogs, or journalists.
F. Outsiders: information originating from individuals that are not personally involved/affected by the event.

5.4. Step 3: Data Collection

5.4.1. List of Events

Table 5.3 shows our data sets, which were made available for research purposes (see Section 5.8.2). They correspond to a set of 26 events during 2012 and 2013, and which spawned significant activity on Twitter. Table 5.3 also includes crisis dimensions of hazard type, development, and spread (we consider the Singapore haze to be partially human-induced due to intentional fires to clear land). We note that in our data set, all human-induced crises are focalized and instantaneous, while all natural hazards are diffused, but may be instantaneous or progressive.

To obtain our list of events, we started with a set of disasters compiled mainly from Wikipedia.[5] We then filtered it by choosing events that had at least 100,000 tweets associated with them—which is reflected by at least 1,000 tweets in the 1% public data stream we used.

Floods are the most frequent type of natural hazard in our data, and also the natural hazard that affects the most people in the world. According to data from the United Nations for the 2002–2011 period, an average of 116 million people were affected by a flood every year, followed by 72 million people a year affected by drought, 40 million by storms, 9 million by extreme temperatures, and 8 million by earthquakes [240].

[5] From the list of significant events per month, e.g. for January 2013 we consulted http://en.wikipedia.org/wiki/January_2013

5.4.2. Data Sampling

Our data collection method is shaped by limitations to data access through Twitter, and is based on first collecting a base data sample and then retrospectively sub-sampling it. The base data sample was obtained by constantly monitoring Twitter's public stream via Twitter's Sample API, which consists of a sample of approximately 1% of all tweets[6] and it is accessible via Internet Archive[7], allowing full reproducibility of this work. In the 2012-2013 period, this collection contains on average about 132 million tweets (amounting to 38 GB of compressed data) per month. The quality of Twitter data samples acquired via the publicly available APIs that offer limited access to the full Twitter stream has been studied extensively, to understand the nature of the biases of such data samples [111, 120, 160, 234, 235]. Yet, while [235] have shown biases with respect to hashtag and topic prevalence in the Streaming API (which we do not use in this study), [234] shows that the data obtained via the Sample API closely resemble the random samples over the full Twitter stream, which corroborates the specifications of this API. Additionally, given the daily volume of tweets "*the 1% endpoint would provide a representative and high resolution sample with a maximum margin of error of 0.06 at a confidence level of 99%, making the study of even relatively small subpopulations within that sample a realistic option*" [111].

The sub-samples are obtained by running keyword searches over the base data—keyword searches that mimic the way in which Twitter does keyword tracking to obtain a sample of the data that one can obtain in real time.[8] An advantage of this retrospective sampling method is that one can capture the entire period of the event, which is not the case for other collections built during the disasters, which generally lack the first minutes or hours of the event.

Keywords were selected following standard practices commonly used for this type of data collection by practitioners [42, 151, 245, 306], and typically include hashtags or terms that pair the canonical name of the disaster with proper names of the affected locations (e.g., Manila floods, #newyork derailment), the proper names of the meteorological phenomena (e.g., Hurricane Sandy), or, at times, hashtags promoted by governments, response agencies, or news media. These terms can be either in English or in the language of the population affected by the disaster. In Chapter 6 we show that this method produces a sample of messages whose distribution of information categories closely resembles the sampling by other methods e.g. *geofencing*, which samples all tweets from users in the affected area.

[6] https://dev.twitter.com/docs/api/1.1/get/statuses/sample
[7] https://archive.org/details/twitterstream
[8] https://dev.twitter.com/docs/streaming-apis/parameters#track

To identify the keywords/hashtags used during each event, the author of the thesis used a search engine to lookup for "Hashtags ⟨Event Name⟩." The search results often included news articles discussing the social media use during the searched event,[9] resources from NGOs using social media for crisis management,[10] Internet media platforms,[11] governmental resources on social media use[12] or research papers [42, 245]. Using these resources, we built an initial list of hashtags/keywords, which we further validated and iteratively improved by manually searching for them on Twitter. In those cases in which the hashtag/keyword had been used for other purposes, we also looked for the combination of the hashtag/keyword, and the event name. When other keywords frequently appear with those already on our list, we also searched for them in Twitter. If there were at least a few instances in which they appeared in relevant tweets without the other keywords, we added them to the list—essentially simulating what a practitioner can do at the time of the event. The size of the resulting keywords lists vary, yet our results in Chapter 6 suggest that across events keywords lists of various sizes retrieve collections which exhibit comparable representativeness with respect to a reference sample. Specifically, in Chapter 6, we see that for all the analyzed events, although the keywords list varies from 4 to 36 terms and that the keyword-based collections tend to bias the collections towards e.g. media reports, the representativeness with respect to a reference sample is similar across the keyword-based collections corresponding to different events.

For the *instantaneous* hazards we start the collection from the moment when the event happen, while for the *progressive* hazards we start from the moment the hazard was detected (e.g., when a storm formed for a hurricane, or when the first fires were detected for wildfires). The volume of tweets in each collection decreases after onset, but we continue collecting data until that volume stabilizes to a low value (specifically, when the standard deviation of the daily number of tweets becomes less than 5).

As a post-processing step, we remove very short tweets (i.e. those made up of 3 tokens or less), as they are in general hard to classify and rarely contain any useful information. We do not remove near-duplicates or re-tweets (RTs) because we are interested in the extent to which people repeat and pass along existing messages.

[9] http://mashable.com/2012/06/29/colorado-wildfire-social-media/, http://www.techinasia.com/singapore-haze-infographic/ and others.

[10] http://wiki.crisiscommons.eu/wiki/Crises, http://crisiswiki.org/

[11] http://twitchy.com/2014/07/07/earthquake-hits-southern-mexico-and-guatemala-fatalities-damage-reported-pics/, https://storify.com/ABC13Houston/plant-explosion-in-west-texas

[12] http://www.gov.ph/2013/11/09/online-efforts-for-typhoon-yolanda/

5.5. Step 4: Crowdsourced Data Annotation

We employed crowdsource workers to perform manual annotation of our data sets in April and May 2014.[13] The workers were provided with detailed instructions and examples of correctly labeled tweets, so they could successfully complete the annotation task.

5.5.1. Task Description

Below are the instructions given during the annotation phase to crowdsource workers. "You," in the task description refers to the crowdsourcing worker. The underlined parts, and the examples, changed for each crisis.

M1. Informativeness. The instructions used for this annotation task are shown below, and include examples for each class.

> Categorize tweets posted during the 2013 Colorado floods. Please read them carefully, following links as necessary, and categorize them as:
> A. Related to the floods and informative: if it contains useful information that helps you understand the situation:
> – "RT @NWSBoulder Significant flooding at the Justice Center in #boulderflood"
> – "Flash floods wash away homes, kill at least one near Boulder via @NBCnews"
> B. Related to the floods, but not informative: if it refers to the crisis, but does not contain useful information that helps you understand the situation:
> – "Pray for Boulder, Colorado #boulderflood"
> C. Not related to the floods:
> – "#COstorm you are a funny guy lol"
> D. Not applicable; too short; not readable; or other issues.

M2. Information Type. Instructions and examples:

> Categorize tweets posted during the 2012 Colorado wildfires. Please read them carefully, following links as necessary, and categorize as:
> A. Affected individuals: information about deaths, injuries, missing, trapped, found or displaced people, including personal updates about oneself, family, or others.
> – "Up to 100,000 people face evacuation in Colorado"
> B. Infrastructure and utilities: information about buildings, roads, utilities/services that are damaged, interrupted, restored or operational.

[13] We employed workers through the crowdsourcing platform *CrowdFlower: http://crowdflower.com/*

– “Officials working the #HighParkFire confirmed that several roads are closed”
C. Donations and volunteering: information about needs, requests, queries or offers of money, blood, shelter, supplies (e.g., food, water, clothing, medical supplies) and/or services by volunteers or professionals.
– “#Offer Storage Space http://t.co/... #COwildfire”
D. Caution and advice: information about warnings issued or lifted, guidance and tips.
– Wildfire warnings issued for six counties Sunday - http://t.co/...”
E. Sympathy and emotional support: thoughts, prayers, gratitude, sadness, etc.
– “Pray for Boulder #COwildfire”
F. Other useful information NOT covered by any of the above categories.
– “To track fire activity in CO, check this site @inciweb Colorado Incidents http://t.co/...”
G. Not applicable; not readable;not related to the crisis.

M3. Source. Instructions and examples:

Categorize tweets posted during the 2013 Queensland floods (Australia). Please read them carefully, following links as necessary, and indicate the most likely source of information for them as:
A. Eyewitness: if the information originates from eyewitnesses to the event or to response/recovery operations, or from their family, friends, neighbors, etc. :
– “Just found out my mum is trapped at home, no water, no power, tree’s down across roads out of her property near glasshouse mtns”
– “Outside sounds like it is going to shatter my bedroom windows any sec now #bigwet #qld”
B. Government: if the information originates from national, regional or local government agencies, police, hospitals, and/or military.
– “PRT @theqldpremier: UPDATE SCHOOL CLOSURES: An updated school closures list is available now at http://t.co/...”
C. Non-government: if the information originates from non-governmental and not for profit organizations such as RedCross, UN, UNICEF, etc.
– “RT @RedCrossAU: Everyone affected by #qldfloods, let people know you’re safe: http://t.co/...”
D. Businesses: if the information originates from for-profit business or corporations such as Starbucks, Walmart, etc.
– “RT @starbucks: With many partners impacted by OLD floods, consider making (or increasing) donations”
E. Traditional and/or Internet news or blogs: if the information originates from television channels, radio channels, newspapers, websites or blogs such as CNN, KODA, New York Times, etc.
– “RT @ABCNews24: #QLDfloods watch: Authorities are preparing for tornadoes in southeast Queensland.”

F. Outsiders: if the information originates from individuals that have NO acquaintances affected by the event, nor are they associate with any organization.
– "RT @TheBushVerandah: Just heard a farmer had to shoot approx 100 sows at mundubbera ... In preference to them drowning"
G. Not applicable; not readable; not related to the crisis.

5.5.2. Task Characteristics

For all annotation tasks, we provide examples both in English and the language most commonly used to communicate about the event (if there was a common language used other than English.) Regarding worker selection, the platform we used for crowdsourcing allows us to select workers by country (but not at a sub-country level), so we specified that workers must be from the country where the event took place. In a few cases when there were not enough workers to perform the task, we also included workers from neighboring countries having the same official language. We selected workers in this way to ensure that they understand the tweets posted by individuals local to the event, and that they would be more likely to be able to understand dialects, references to regional and/or local places, and overall be versed in the culture of the area in which the event took place. Additionally, following standard guidelines from this crowdsourcing platform, 20 to 30 tweets per crisis and task were classified by the author of this thesis. We consider as *untrusted* all workers whose assessments differ significantly from ours on these tweets (less than 70% of agreement), otherwise we consider them as *trusted*.

Workers were presented with the tweet text, including any links (which they were invited to follow), and then asked to choose a single category that best matched the content of the tweet. To avoid potential ethical concerns on behalf of Twitter users who are likely unaware that their tweets are being collected and analyzed, workers did not have access to the author username, nor the time at which the tweet was sent. In addition, we avoid possible privacy violations by not displaying the username nor the profile picture of persons affected by a given disaster. This practice follows customary procedures used for using crowdsourced annotation of text messages for both information type [17, 52, 153, 245] and information source [79, 245].

Trusted workers took from 10 to 12 seconds to label each tweet (in terms of interquartile mean, which is the figure reported by the crowdsourcing platform). We collect labels from at least 3 different trusted workers per tweet and task, and determine the final label of the tweet by simple majority.

About 15-20 trusted workers participated in each classification step (i.e. a set of 1,000 tweets from a single event and with a single question M1, M2, or M3), with the bulk of the work being done by about 10 of them in each case—with no worker labeling more than 300 items in a classification task, a limit set by us following recommendations from the crowdsourcing provider. The total amount paid to the crowdsourcing platform for the 3 classification tasks was approximately $35 (USD) per event. Payments to specific individual workers depend on how many tasks they performed and on their agreement with the test questions, following an internal procedure of the crowdsourcing provider.

The first classification task is to identify tweets which are related to a crisis. A tweet may contain a crisis' keywords but be unrelated to it, as some keywords may be quite general, and refer to any number of topics other than the disaster situation. In addition, unscrupulous spammers sometimes exploit the popularity of a crisis hashtag to post promotional content [35]. As a result, the first labeling phase (M1) also has a data cleaning role. For each event we label a set of 1,000 tweets selected uniformly at random. We imposed a minimum threshold of 900 crisis-related tweets per crisis, and in the cases where it was necessary (9 out of 26 crises), we continued labeling random samples of tweets until passing the threshold. Next, we kept only the tweets that were related to the crisis (independently of whether they were deemed informative or not), and classified them with respect to *information types* (M2) and *sources* (M3).

5.5.3. Task Evaluation

Tweet classification is a subjective process, especially when performed at a large scale, and with a focus on tweet content. To evaluate to what extent subjectivity affects our results, we performed the following experiment: Two annotators[14] *independently* labeled 200 tweets sampled uniformly at random from all the crises. They classified tweets according to information types and sources, by looking at the content of the tweets as displayed in the Twitter platform, including conversations (if any), and looking at links in the tweets, and user profile information from its authors. We also note that the annotators had background information about each of the events.

We measure inter-assessor agreement with Cohen's Kappa, resulting in $\kappa = 0.80$ for information type (95% confidence interval CI: $[0.73, 0.87]$) and $\kappa = 0.73$ for source (95% CI: $[0.64, 0.81]$). Customarily, values in this range indicate substantial agreement.

[14]The author of the thesis and another co-author of this study.

Next, we take all tweets in which both annotators agree and compare their joint label with those provided by crowdsource workers. The results are $\kappa = 0.81$ (95% CI: $[0.73, 0.88]$) for information type and $\kappa = 0.72$ for source (95% CI: $[0.62, 0.83]$). Again, these values reflect substantial agreement. The *individual* agreement of annotators with workers (which includes cases in which the labels given by annotators do not agree) is lower but still substantial ($\kappa = 0.69$ and $\kappa = 0.74$ for information type, $\kappa = 0.57$ and $\kappa = 0.63$ for source).

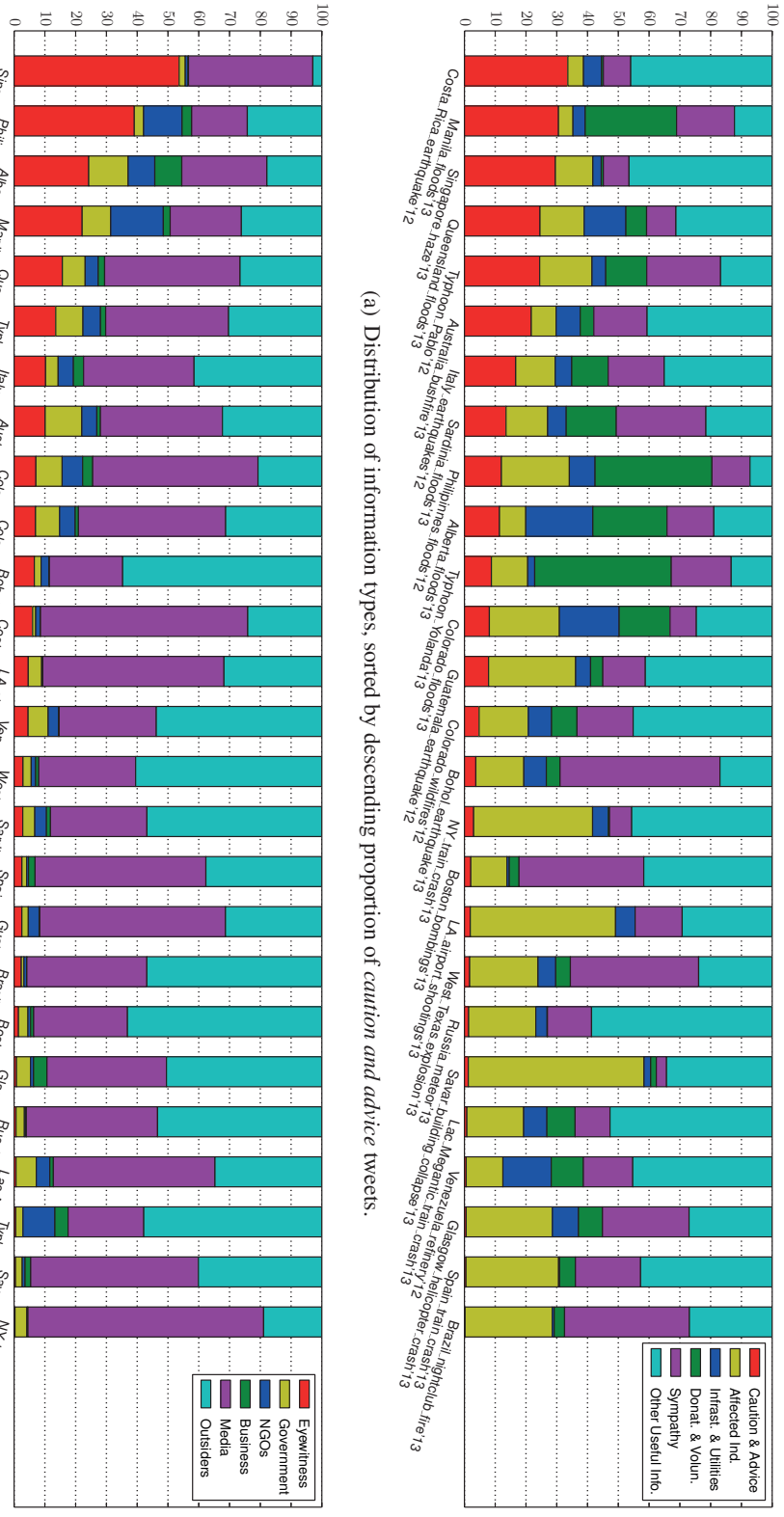
The conclusion is similar to that of previous work using crowdsourcing labeling (e.g. [79, 288]), crowdsource workers collectively provide reliable labels for social media annotation tasks, at a volume that would be very costly to achieve by other means (in our case, $26 \times 1,000 \times 3 = 78,000$ labels).

This experiment also allows us to evaluate the biases of crowdsourcing labeling. For information type, in 15% of the cases the crowdsourced label does not correspond to the one given by the two annotators (among the annotators this discrepancy is 16%). The most common error of crowdsourcing workers is labeling "Caution and Advice" messages as either "Donations and Volunteering" or "Other Useful Information"—e.g. messages advicing where the affected population can send information or submit rescue requests to govermental agencies were confused with requests for help from volunteers and NGOs. For information source, in 17% of the cases the crowdsourced label did not agree with the one of the annotators (among annotators this discrepancy is 18%). The most common error was labeling "Eyewitness" as "Outsiders" or "Media." This means that in the analysis, we have to consider that "Caution and Advice" and "Eyewitness" may be underrepresented categories, while the other categories we mentioned may be overrepresented. The extent of the total under-representation/over-representation across all categories, however, is about 15%-17%, and more importantly, is not larger than the discrepancy among the two annotators who performed this evaluation.

5.6. Step 5: Data Analysis

The final step is to perform an analysis of the data annotated by the crowdsourcing workers. We begin by presenting results about the overall distribution of content types across crises, which we connect to the crisis dimensions by mining association rules. Then we consider temporal aspects, as well as the interplay between content dimensions.

Finally, we show that while substantial variability exists, similar crises tend to have a similar distribution of message types (§5.6.6). Though we make no claims that these 26 crises are



(a) Distribution of information types, sorted by descending proportion of *caution and advice* tweets.

(b) Distribution of information sources, sorted by descending proportion of *eyewitness* tweets.

Figure 5.1.: Distributions of information types and sources (best seen in color)

representative of every event of every type we consider, we do note patterns and consistencies in the proportion of different messages, and present potential explanations about them, to serve as foundations for future explorations.

5.6.1. Content Types vs. Crisis Dimensions

We first present our results regarding relationships between crisis dimensions and the prevalence of categories of information found in tweets.

Informativeness

The proportion of messages in each collection that were found to be about the crisis at hand (i.e. classified in the first two categories of M1) was on average 89% (min. 64%, max. 100%). In this case, one of the most significant factors is whether the keywords/hashtags adopted by people tweeting about the crisis are specific to the event, or were used for other purposes. For instance, `#yolandaph` was specifically used for Typhoon Yolanda, while `#dhaka` (the name of the capital of Bangladesh) was used after the Savar building collapse, but also for other purposes.

Among these messages, the proportion of informative messages (i.e. those in the first category of M1) was on average 69% (min. 44%, max. 92%). Most of the messages considered "not informative" contained expressions of sympathy and emotional support (e.g. "thoughts and prayers").

Information types

Figure 5.1(a) shows the distribution of information types found in the tweets related to each crisis. Below, we sort the categories in decreasing order of average prevalence, noting the (wide) range on the proportion of each type.

- *Other useful information*: 32% on average (min. 7%, max. 59%). This "catchall" category is the largest among the information types. An analyst interested exclusively in the remaining categories can skip these messages on the initial pass of analysis. We note that the events in which this category was the least prevalent (i.e., the other categories accounted for more than 80% of the messages) were all diffused. While we do not claim that all, or even most, diffused events will have fewer-than-average amounts of "other

useful information" tweets, it is potentially useful to know that this type of tweets is not prevalent in the diffused events we studied.

The information captured by the "other useful information" category varies significantly across events. For instance, in the Boston bombings and LA Airport shootings in 2013, there are updates about the investigation and suspects (e.g. "*Boston bomber, Dzhokhar Tsarnaev has finally been arrested and is in custody.*"); in the West Texas explosion and the Spain train crash, we find details about the accidents and the follow-up inquiry (e.g. "*'black box' on Spain train that crashed, killing 79, shows conductor was on phone, traveling nearly twice speed limit*"); in earthquakes, we find seismological details (e.g. "*5.1 earthquake, 54km SW of Champerico, Guatemala. Nov 7 16:42 at epicenter depth 35km*").

- *Sympathy and emotional support*: 20% on average (min. 3%, max. 52%). Tweets that express sympathy are present in all the events we examined. The 4 crises in which the messages in this category were more prevalent (above 40%) were all instantaneous disasters. Again, we make no hard-and-fast claims about all instantaneous disasters, but this finding leads us to conjecture that people are more likely to offer sympathy when events are not predicted, take people by surprise, and may cause additional distress due to their unforeseen occurrence (e.g. "*Wow can't believe what just happened at the Boston Marathon. Praying for everyone down there*").

- *Affected individuals*: 20% on average (min. 5%, max. 57%). The 5 crises with the largest proportion of this type of information (28%–57%) were human-induced, focalized, and instantaneous. These 5 events can also be viewed as particularly emotionally shocking. They resulted in casualties, but a small enough number of casualties to generate many reports regarding individuals who lost their lives, suffered injuries, or were missing or trapped (e.g. "*at least 2 have died in train derailment in the Bronx. #MetroNorth*").

- *Donations and volunteering*: 10% on average (min. 0%, max. 44%). The number of tweets describing needs or offers of goods and services in each event varies greatly; some events have little or no mention of them (e.g. Singapore Haze), while for others, this is one of the largest information categories. In our data, tweets about donations and volunteering were more prevalent in Typhoon Yolanda in 2013 (44%) and in the floods in Sardinia, Colorado, Alberta, and Manila in 2013, and in the Philippines in 2012 (16%–38%)—tweet examples include "*We've opened residences to those evacuated due to #yycflood. If you need lodging visit an #ABemerg Response centre*" or "*#RescuePH please help the residents of 16 purity st. Remmanville bicutan. They need food and water.* " In contrast,

they were 10% or less for all analyzed crises that were human-induced, focalized, and instantaneous.

- *Caution and advice*: 10% on average (min. 0%, max. 34%). In instantaneous crises, there is unsurprisingly little information of this type (0%–8%), as these events are often not predicted and only post-impact advice can be present. The only exceptions in our data are the Italy earthquakes in 2012 (17%)—in which the collection covers two consecutive earthquakes plus a number of significant aftershocks which happened over an interval of less than 10 days, and Costa Rica earthquake in 2012 (34%)—when tsunami alerts were issued across Central America and parts of South America including even distant countries like Chile. Apart from these two events, the events with the most tweets that include information about caution and advice are caused by diffused natural hazards, and the 5 with the highest fraction from this set are all progressive (22%–31%). Tweet examples include "*yellow advisory for Metro Manila. Moderate-heavy rains in next 3 hours, possible floods in low-lying areas*" or "*The Burnett Highway, six kilometres south of Mt Morgan has water over it. Proceed with caution.*"

  Further, barring the meteor that fell in Russia in 2013, we can see a clear separation between human-induced hazards and natural hazards: all human induced events have less caution and advice tweets (0%–3%) than all the events due to natural hazards (4%–31%). The meteor was a rare event that felt like a bomb whose shock wave shattered windows and damaged thousands of buildings, remaining undetected before its atmospheric entry.[15]

- *Infrastructure and utilities*: 7% on average (min. 0%, max. 22%). The crises where this type of information represents more than 10% of tweets were the Queensland, Alberta, and Colorado floods of 2013, and the Venezuela refinery explosion in 2012. In flood situations, it is common for public institutions and roads to be closed and for electricity and water supplies to be cut (e.g. "*#cuboulder campus will also be closed tomorrow (9/13) due to #boulderflood*" or "*@XcelEnergyCO reports most electric customers back on line, but 2000 gas customers still out, half in Boulder County*"). In the case of the refinery explosion, which was an important industrial site, many living in the area were suddenly without electricity due to the massive impact of the discharge. We note that even when ability of affected populations to tweet might be affected by e.g. power outages, other actors (e.g. media, governmental agencies or NGOs) also tend to tweet more frequently information about infrastructure.

[15] http://en.wikipedia.org/wiki/Chelyabinsk_meteor

Sources

In Figure 5.1(b), we see the distribution of tweet sources, and we observe the following:

- *Traditional and/or Internet media*: 42% on average (min. 18%, max. 77%). Regardless of the event, traditional and Internet media have a large presence on Twitter, in many cases more than 30% of the tweets. The 6 crises with the highest fraction of tweets coming from a media source (54%–76%) are instantaneous, which typically make the "breaking news" in the media (e.g. "*@Reuters: BREAKING NEWS: 6.3 magnitude earthquake strikes northwest of Bologna, Italy: USGS*" from Reuters during the earthquakes in Italy, or *"BREAKING: Reports of shots fired at LAX Airport, says senior government official. Stay with @msnbc for the latest."* during LA airport shooting).

- *Outsiders*: 38% on average (min. 3%, max. 65%). Depending on the event, the number of "outsiders" can vary. This was in general about 18% or more, with the exception of the Singapore haze in 2013 that had only 3% of tweets from outsiders. The Singapore haze was an event that strongly disrupted the city, but did not result in life-threatening injuries or deaths. Examples of tweets from outsiders include reactions to the news e.g. "*Crazy news to hear a helicopter has crashed in Glasgow city centre, can only hope everyone involved is ok!*" or actions taken to help those affected by the disaster e.g. "*I just donated to the @britishredcross #Typhoon Haiyan Appeal. Please donate at http://link_donations*"

- *Eyewitness accounts*: 9% on average (min. 0%, max. 54%). In general, we find a larger proportion of eyewitness accounts during diffused disasters caused by natural hazards. The 12 events with the highest percentage of eyewitness accounts are all diffused (6%–54%) and the top 6 are also progressive (13%–54%). Tweet examples include observations "*View from my desk earlier #sydneyfires #smoke #scary #city #orangeApocalypse http://link_instagram*" or "*never ending rains hounds us in 3 days strait now. Flood waters r inside our house.*"

- *Government*: 5% on average (min. 1%, max. 13%). A relatively small fraction of tweets include information sourced by government officials and agencies—only for two of the crises we analyze this exceeds 10%. We surmise that this is because governments must verify information before they broadcast it, which takes considerable time [144]. Therefore, government accounts may not have the most up-to-date information in crisis situations. The 7 events with the highest percentage of tweets from governmental agencies are due to natural-hazards, progressive and diffused (7%-13%), which are the cases when the governments typically intervene to issue or lift warnings or alerts: e.g. "*Total Fire Bans*

will be in place in a number of areas tomorrow including Greater Sydney, the Greater Hunter, and the Illawarra" from the New South Wales rural fire service in Australia, or "*Maps of the Sunnyside and Bowness evacuations can be found here: http://link_maps* " from City of Calgary.

- *NGOs*: 4% on average (min. 0%, max. 17%). Like governments, NGOs are also careful to broadcast only verified information. In the human-induced crises we studied there is little NGO activity on Twitter ($\approx$ 4% or less). The highest levels of NGO tweets are seen in natural disasters and all those in which the fraction of such tweets was 6% or more are typhoons and floods (e.g. "*Filipinos abroad may donate to the #YolandaPH fund via Paypal through PDRF's #BrickByBrick Program*" during Typhoon Yolanda), which are diffused events typically affecting large areas and populations.

- *Business*: 2% on average (min. 0%, max. 9%). For the most part, we do not see a large amount of tweet activity from businesses in the disaster situations we studied. The proportion is below 5% for all crises except the Alberta floods in 2013 with 9% of tweets coming from businesses. Furthermore, with only one exception—the Glasgow helicopter crash when e.g. a taxi company offered free rides to hospital to the victims families: "*anyone with loved ones in hospital from #Clutha and struggling for transport to visit, we'll provide free taxis*"—the crises with 3% or more tweets from business were diffused.

5.6.2. Association Rules

To systematically search for relationships between the characteristics of crises and the messages on Twitter, we applied an association-rules mining method [61]. To err in the side of caution, we report only the automatically-discovered association rules that are valid for more than 20 out of the 26 crises. To apply this method to numerical data, each category in the information types and sources was divided into two classes: above the median, and below the median.

For information types, we found one rule that is valid for 24 out of 26 of the crises: when the geographical spread is diffused, the proportion of caution and advice tweets is above the median, and when it is focalized, the proportion of caution and advice tweets is below the median. For sources, we found one rule that is valid for 21 out of 26 of the crises: human-induced accidental events tend to have a number of eyewitness tweets below the median, in comparison with intentional and natural hazards.

Both rules are possibly related to different levels of access to the area affected by the event and to its surroundings.

5.6.3. Content Redundancy

We next look at content redundancy. Heuristically, we consider two tweets to be near-duplicates if their longest common subsequence was 75% or more of the length of the shortest tweet. Among the sources of information, messages originating from non-governmental organizations and government sources tended to show more redundancy, with the top 3 messages (and their near-duplicates) accounting for ≈20%-22% of the tweets. Among information types, messages of caution and advice, and those containing information about infrastructure and utilities, were the most repeated ones, with the top 3 messages (and their near-duplicates) comprising ≈12-14% of the tweets.

5.6.4. Types and Sources

Some information types are more frequently associated with particular sources, as shown in Figure 5.2, in which each ⟨type, source⟩ cell depicts the probability that a tweet has that specific combination of information type and source. NGOs and business are more frequently the source of tweets related to donations and volunteering, mostly to ask for resources and request volunteer work (NGOs), or to announce free or discounted goods or services for those affected by a disaster (businesses).

Tweets from governments are often messages of caution and advice, such as tornado alerts or evacuation orders (e.g. "*#HighParkFire evacuation orders issued for Pingree Park area*" from the Larimer County Sheriff's Office)—which agrees with observations in [329] where "preparedness" is the larger category used by government communications. Instead, eyewitness tweets focus on affected individuals (e.g. "*Feels strange being evacuated from the 33rd floor. Trust me, if the water gets there we're all in big trouble. #yyc*"). Both government and eyewitness tweets also frequently include a variety of messages that belong to the "other useful information" category (e.g. "*Clarifying a rumour for #yyc. There are NO zoo animals being sheltered at the Courts. #yycflood*" from police, or "*Sunset in my hometown Fort Collins. We can see the smoke from High Park Fire.*" from an eyewitness). Outsider messages are predominantly about sympathy and support (e.g. "*prayers go out to the people in Russia! #RussianMeteor*").

Finally, tweets from traditional and Internet media offer a variety of information types including

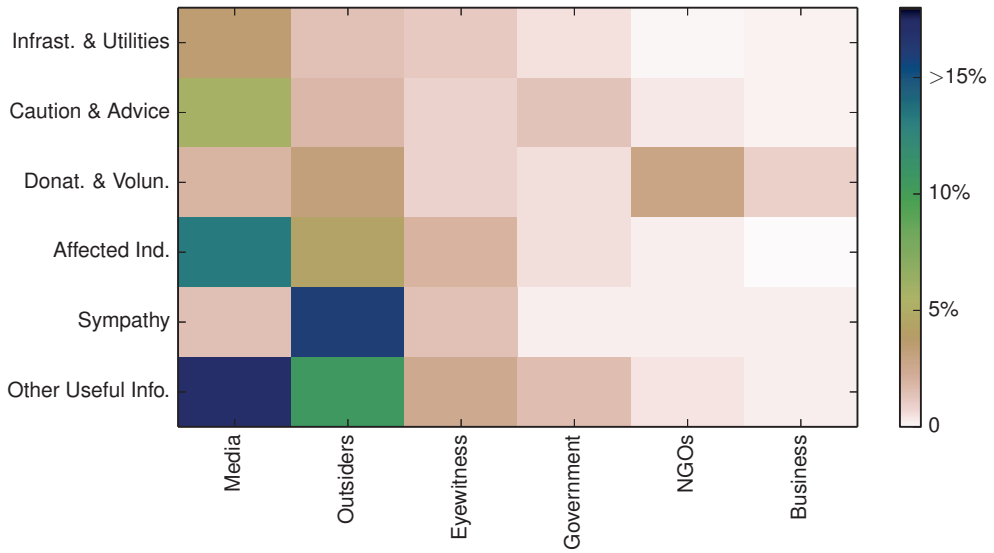


Figure 5.2.: Average distribution of tweets across crises into combinations of information types (rows) and sources (columns). Rows and columns are sorted by total frequency, starting on the bottom-left corner. The cells in this figure add up to 100%.

information about affected individuals, and messages of caution and advice. Media are also the most prominent source of information regarding infrastructure and utilities.

5.6.5. Temporal Aspects

We study how the volume of different categories of messages evolves over time, as shown in Tables A.4 and A.5 in Appendix A.5. We separated crises according to their temporal development (instantaneous vs. progressive), depicting using "spark lines" the total volume of messages over time, and the total volume of messages in each information type and source.[16] This analysis focuses on the differences between the average timestamps of messages in different information categories.

In terms of information types, the messages that are likely to arrive first are those of caution and advice, and sympathy and support, roughly in the first 12–24 hours after the peak of the crisis. This is particularly evident in instantaneous crises. Then, messages about affected individuals and infrastructure are most frequent. Typically, the last messages to appear are those related to donations and volunteering. Interestingly, this follows the progression in the stages of a crisis

[16]Each point in the spark line corresponds to a *calendar* day, which explains why in some instantaneous crises the overall curve goes up at the beginning (when the crisis occurs at night).

from emergency response to early recovery actions [307].

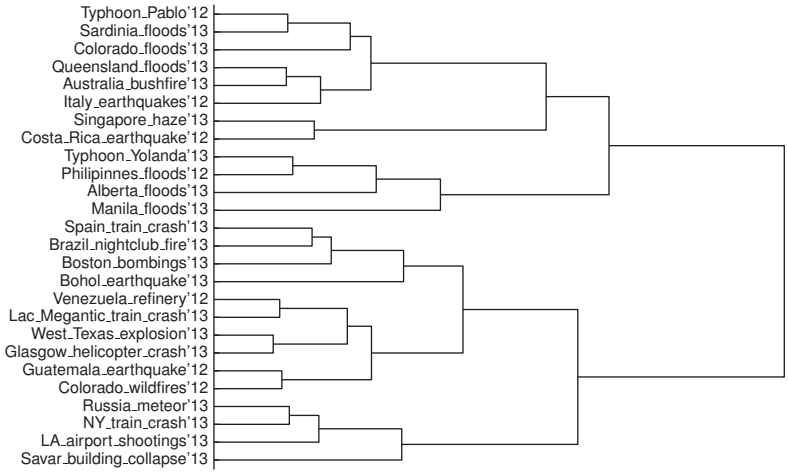
In terms of sources, there are differences depending on the type of temporal development. In *instantaneous* crises, outsiders, media and NGO messages tend to appear early, with other sources following (the temporal position of eyewitness messages varies substantially depending on crisis type). On the other hand, during *progressive* crises, eyewitness and government messages are the ones more likely to appear early, mostly to warn and advice those in the affected areas, while NGO messages appear relatively late. In addition, there is an interesting temporal complementarity between messages from governments and NGOs that merits to be studied in depth in future work.

5.6.6. Crisis Similarity

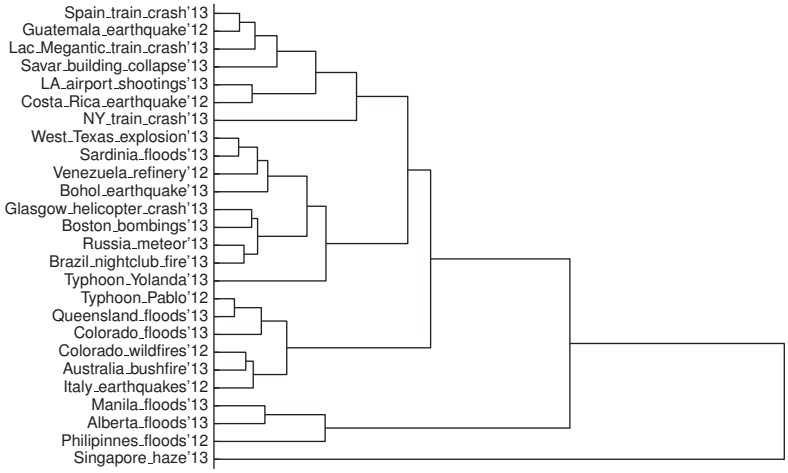
In further seeking links between disaster characteristics and tweet content and source, we apply an unsupervised method—specifically, we use hierarchical agglomerative clustering. Performing this clustering uncovered groups of crises that have similar content distribution. Given that we compare probability distributions, to measure the similarity between two crisis events we use Bhattacharyya distance (for two discrete distributions p and q this is $-\ln(\sum_{c\in C}\sqrt{p(c)q(c)})$ where C is the set of all classes) which quantifies the overlap between two statistical samples. To combine clusters of crises, we used complete-linkage clustering, which merges those clusters for which the distance between their furthest elements is the smallest.

Figure 5.3(a) shows the resulting dendrogram when the clustering is done according to the distribution of information types. We see two large clusters: first, the cluster on the bottom is dominated by human-induced crises, while in the one on the top there are only natural hazards. This indicates that, despite the significant variations we have shown, human-induced crises are more similar to each other in terms of the types of information disseminated through Twitter than to natural hazards.

Second, events also cluster depending on how they developed. The cluster at the bottom includes instantaneous events, with one exception: the Colorado wildfires in 2012. This exception may be due to the nature of this particular fire. The combination of heat, drought conditions, and high winds caused the fire to quickly develop, and it claimed 350 houses in just over 12 hours. The cluster on the top includes progressive disasters, with two outliers: Italy earthquakes in 2012—a sequence of earthquakes and aftershocks—and the Costa Rica earthquake in 2012—during which a Caribbean-wide tsunami watch was issued, resulting in a large volume of caution and advice messages that are typically more prominent in progressive crises.



(a) Clusters by information type.



(b) Clusters by source.

Figure 5.3.: Dendrograms obtained by hierarchical agglomerative clustering of crises. The length of the branch points reflect the similarity among crises. We remark that the clusters do not reflect similar messages, but instead similarities in terms of the proportion of different information types and sources in each crisis.

A similar picture emerges in the case of clusters by distribution of sources, shown in Figure 5.3(b). In this case, there is a large cluster dominated by human-induced crises (on the top), followed by two small clusters encompassing only natural hazards, and the Singapore haze 2013 as an outlier (this haze was caused by a mix of natural and human causes). Further, the large cluster on the top is dominated by instantaneous events (with two exceptions, Typhoon Yolanda and Sardinia Floods in 2013), while in the other clusters the events are progressive, excepting Italy earthquakes in 2012.

Furthermore, while the events development and type arise as the main factors impacting the clusters composition, in both Figures 5.3(a) and 5.3(b) we also notice that the clusters are being dominated by either diffused (top cluster by information type and bottom clusters by information source) or focalized events (the remaining clusters). The clusters tendency to encompass events that are similar along all these dimensions is likely explained by the dependency among the crisis dimensions (e.g., typically, the progressive events are also diffused and human-induced crises tend to be focalized).

5.7. Discussion: Social Media

Disasters are common events that occur regularly; the United Nations Office for Coordination of Humanitarian Affairs recorded 394 disasters caused by natural hazards in the 2002–2011 period [240]. While disasters take place often, and may be caused by similar hazards and/or human actions, each event is unique [255] (pag. 5). Regardless of their distinct nature, and of variations in individual reactions and responses, commonalities across crises exist. Sociologists of disaster point out that despite the differences among disaster agents (e.g. flood, earthquake, bomb, fire), there are actions that planning and emergency response teams must take that are independent of these differences [307].

This brings us to an interesting juxtaposition; the types and amounts of information broadcast on Twitter differ across each of the 26 specific crises we studied. This can be viewed as a display of the uniqueness of each event. In some cases the most common tweet in one crisis (e.g. eyewitness accounts in the Singapore haze crisis in 2013) was absent in another (e.g. eyewitness accounts in the Savar building collapse in 2013). Furthermore, even two events of the same type in the same country (e.g. Typhoon Yolanda in 2013 and Typhoon Pablo in 2012, both in the Philippines), may look quite different vis-à-vis the information on which people tend to focus.

Yet, when we look at the Twitter data at a meta-level, our analysis reveals commonalities among the types of information people tend to be concerned with, given the particular dimensions of the situations such as hazard category (e.g. natural, human-induced, geophysical, accidental), hazard type (e.g. earthquake, explosion), whether it is instantaneous or progressive, and whether it is focalized or diffused. For instance, caution and advice tweets from government sources are more common in progressive disasters than in instantaneous ones. The similarities do not end there. When grouping crises automatically based on similarities in the distributions of different classes of tweets, we also realize that despite the variability, human-induced crises tend to be more similar to each other than to natural hazards.

This leads us to believe that we can view Twitter as a medium through which the nuance of disaster events is highlighted or amplified; it is a tool that becomes incorporated into the social construction of the disaster event, and through which we can understand the detailed differences on a large scale when we look closely at Twitter data. At the same time, when we look at those same data at a higher level, we see commonalities and patterns.

Practitioners, including emergency managers, public information officers, and those who develop the tools used by them, should consider that the proportion of tweets that are relevant for a specific purpose will almost invariably be smaller than the proportion of the tweets that are not. For instance, if an analyst or an application focuses on content that is not present in mainstream or other Internet media sources, and wants to exclude content provided by outsiders who are not affected by the crisis, then it will have to skip through 80% of the tweets on average. The same holds for information types. If we group together the four main types we used (affected individuals, donations and volunteering, caution and advice, and infrastructure and utilities), they cover on average 47% of the tweets related to a crisis. This implies that if an application wants to focus on these information types, at least 53% of the messages will have to be discarded. These are lower bounds, as often not all of the tweets of a given type will be relevant for a particular application. *Noise* is a natural consequence of the diversity of information in this medium.

Developers should consider that emergency response includes a set of actions that have to be taken in preparation of any crisis event, plus a broad space for adaptability in response to specific events [307]. Hence, tools to process social media in disaster should consider that there are broad classes of information that are likely to be prevalent, and can be anticipated to occur. At the same time, a substantial volume of messages will depend on specificities of every event, and tools must incorporate methods to adaptively detect and process them.

5.8. Conclusions

The overarching goal of this case study is to test how much the observations made on one data set can be generalized to other data sets by identifying systematic similarities and differences among them in the context of social media use during crisis situations.

Our systematic examination of a diverse set of crisis situations finds patterns and consistencies across crises, but it also uncover substantial variability across different event data sets, highlighting the pitfalls of generalizing findings from one data set to another. Specifically, we see that intrinsic characteristics of the crisis situations (e.g. being instantaneous or progressive) produce *consistent effects* on the types of information broadcast on Twitter. Additionally, on the application side, to the best of our knowledge, this is the largest transversal study on tweets broadcast in response to various international disaster and crisis situations.

5.8.1. Limitations and Future Work

Social media communications in crisis. On the application side, the high-level patterns we have found lay the foundations for future studies that go into the detail of each specific crisis or each specific information category analyzed.

However, we note that we did not cover all possible crisis situations. For instance, we did not include human-induced progressive or diffused situations, which are less common than the classes we did study. The former (human-induced progressive) mostly refers to politically-driven crises, such as instability leading to demonstrations, riots, and/or civil wars. The latter (human-induced diffused) in recent years have been mostly wars affecting an entire country or region, or less-common, large-scale industrial accidents such as the oil spill in the Gulf of Mexico in 2010. Additionally, the management of a crisis is typically divided into phases: mitigation, preparedness, response and recovery [252, 307]. This case study is concerned mostly with the response phase and partially with the recovery phase, as these attract the bulk of social media activities [151, 64]. Language and cultural differences could also be included as explicit crisis dimensions [138, 257], together with temporal factors. Microblogging practices are likely to evolve over the years, and our collections cover a period of just about 20 months. The study of other crisis dimensions, other types of crises and other phases, will certainly deepen our findings. Additionally, extending this research to other social media platforms will help understanding what are the similarities and differences in how various platforms are used during crisis situations.

Methodologically, we asked crowdsource workers to match each tweet to one specific class. This simplifies the labeling process and makes the presentation of the results clearer. When workers associate a tweet to multiple classes, it may be possible that the distributions change. Employing professional emergency managers as annotators instead of crowdsource workers may lead to further results. Finally, assessing the quality, credibility, or veracity of the information in each tweet is relevant for most of the potential consumers of this data. However, we note that in these cases the cost of the annotation would certainly increase—or the amount of labeled data would decrease.

Social Media Collection Biases. We found that the biases and consistencies in the data collections we uncovered tend to depend on the intrinsic characteristics of the events being analyzed (e.g. involving a small or a wide area). Future work should test if similar patterns hold across other domains as well. To this end, the analysis methodology that we have described in this Chapter, can be extended to a variety of other topics such as sport events or political elections. To conduct the study on a different domain, the main elements that require adaptation are the event taxonomy and the message taxonomy (although this might require some familiarity with the topic).

5.8.2. Reproducibility & Data Release

To ensure and support the reproducibility and replicability of this case study, the tweets used in this research, and the labels collected through the crowdsourced annotation, are available for research purposes at `http://crisislex.org/`.

Part III.

Methods

6. Leveraging Domain: The Case of Data Sampling

To explore how we can improve social data sets at collection time, we focus back on social media use during crisis situations. For this application the quality of the data collections is particularly important as locating timely, useful information during crises and mass emergencies is critical for those forced to make potentially life-altering decisions. Yet, as the use of Twitter to broadcast useful information during such situations becomes more widespread, the problem of finding it becomes more difficult. We describe an approach toward improving the recall in the sampling of Twitter communications that can lead to greater situational awareness during crisis situations. First, we create a lexicon of crisis-related terms that frequently appear in relevant messages posted during different types of crisis situations. Next, we demonstrate how we use the lexicon to automatically identify new terms that describe a given crisis. Finally, we explain how to efficiently query Twitter to extract crisis-related messages during emergency events. In our experiments, using a *crisis lexicon* leads to substantial improvements in terms of recall when added to a set of crisis-specific keywords manually chosen by experts; it also helps to preserve the original distribution of message types.

6.1. Background

As discussed in Chapter 5, the popular microblogging platform Twitter is a frequent destination for affected populations during mass emergencies. Twitter is a place to exchange information, ask questions, offer advice, and otherwise stay informed about the event. Those affected require timely, relevant information. In Chapter 5 we showed that information broadcast on Twitter can lead to enhanced situational awareness, and help those faced with an emergency to gain valuable information, this observation is also consistent with [314].

[0]The study described in this Chapter was done while the thesis author was an intern at Qatar Computing Research Institute, and was published in [245].

The velocity and volume of messages (*tweets*) in Twitter during mass emergencies makes it difficult to locate situational awareness information, such as road closure locations, or where people need water. Users often employ conventional markers known as *hashtags* to bring attention to specific tweets. As detailed in Section 5.4.2, the idea is that those looking for emergency information will search for specific hashtags, and tweets that contain the hashtag will be located. In crisis, hashtags are often adopted by an information propagation process [293], but in some cases, they are suggested by emergency response agencies or other authorities. Alas, even with several dozen such hashtags, only a fraction of the information broadcast on Twitter during mass emergencies is covered [42, 311]. Therefore, automatic methods are necessary to help humans cull through the masses of Twitter data to find useful information.

6.1.1. Contributions

In this Chapter, we explore the problem of improving the quality of the working data collections with respect to the overall platform data. To do so, here, we tackle the specific problem of how to locate tweets that contain crisis-relevant information during mass emergency situations: our goal is to improve query methods, and return more relevant results than is possible using conventional manually-edited keywords or location-based searches.

Problem definition. Given a crisis situation that occurs within a geographical boundary, automatically determine a query of up to K terms that can be used to sample a large set of crisis-related messages from Twitter.

Our approach. Create a *crisis lexicon* consisting of crisis-related terms that tend to frequently appear across various crisis situations. This lexicon has two main applications:

1. Increase the recall in the sampling of crisis-related messages (particularly at the start of the event), without incurring a significant loss in terms of precision.
2. Automatically identify the terms used to describe a crisis by employing pseudo-relevance feedback mechanisms.

Generalizability. Our approach is presented with respect to crises, but it can be applied to *any domain*. We describe a systematic method to build the lexicon using existing data samples and crowdsourced labeling; the method is general and can be applied to other tasks (e.g. to build a sports-related or a health-related lexicon). The lexicon, along with the data and the code we used to build it are available at `http://crisislex.org/`.

6.1.2. Related Work

Mining social media in crises. During crises, numerous disaster-related messages are posted to microblogging sites, which has led to research on understanding social media use in disasters [292, 263], and extracting useful information [152].

The first challenge in using microblog data is to retrieve comprehensive sets of disaster-related tweets [43]. This is due to Twitter's public API limitations (described in §6.2.1) that make this type of data collection difficult. To the best of our knowledge, data collection during crises usually falls in two categories: keyword-based collections and location-based collections , with the former being more common. In a keyword-based collection, a handful of terms and/or hashtags are used to retrieve tweets containing those terms [146] ignoring other posts [43]. While the resulting samples might have little noise [316], they are typically constructed around visible topical hashtags and might omit a significant number of disaster-related tweets [42]. Furthermore, *keywords are only as responsive as the humans curating them and this method may lose relevant tweets due to latency*. Location-based sampling, on the other hand, is limited to tweets that are either geo-tagged or mention the places affected by the disaster; both of these conditions occur in a small portion of tweets.

Once collected, it is necessary to process the data in a meaningful way. Imran et al. [152] automatically identify tweets contributing to situational awareness and classify them according to several types of information. Yin et al. [336] designed a system for leveraging microblog data during disasters; their data capture module is close in scope with our work, yet it makes no distinction between disasters and other events. In turn, our lexicon could enhance their burst detection mechanisms to better identify disasters.

Query generation and expansion. Our problem resembles *deep-web crawling*, the process by which web crawlers access public data (belonging to large online retailers, libraries, etc.) on the web that is not accessible by following links, but only by filling in search forms. To this end, it performs *query generation*: identify a set of keywords that are entered in search forms to return such data [328, 339].

The goal of exhaustively retrieving documents hidden behind web interfaces has been approached as a *minimum weighted dominating set* and *set-covering* graph problems [339, 328]. We reuse the idea of representing document or term co-occurrences as a graph, but we formalize our problem as finding the *maximum weighted independent set* as we look for *discriminative* queries that maximize only the volume of retrieved documents *relevant* to given topics (Section 6.3.1). In web search, reformulating the initial query such that it returns documents from

the domain of interest is known as *vertical selection and aggregation*. Arguello et al. [19] reuse past knowledge to predict models for new domains by focusing on *portability* and *adaptability*. We use their idea of supervision and use knowledge on past crises to generate queries for future ones.

The query generation step can be followed by *query expansion* that after searching with an initial query adds to it new terms [67]. For this, *pseudo-relevance feedback (PRF)* is typically used. It scores and selects new terms according to their distribution in the feedback documents (i.e., those retrieved with the initial query), or according to the comparison of their distribution in these documents and the entire collection [330]. Re-sampling PRF terms by combining PRF results from several query sub-samples downturns the chance of adding noisy terms to the query [62]. However, Twitter API terms of use do not allow us to run similar queries simultaneously, and running them sequentially might lead to data loss at the beginning of the crisis. Hence, we cluster tweets based on which terms matched them, treating each term as a different query [330].

Adaptive information filtering. Unlike classic query generation and expansion on static collections, the data stream relevant to crisis events evolves over time. Our query is maintained over long periods, performs a binary selection rather than compiling a ranked list of documents, and is limited in size—akin to *information filtering over streams of documents* [13, 192].

In contrast to current approaches that exploit the time dimension of a static microblog collection [222, 229], we collect data as it is produced, rather than searching in a historical repository. Wang et al. [319] expands a user-provided query with new hashtags to retrieve more microblog data related to given events. We automate the entire retrieval process by exploiting knowledge on past crises to generate a query, which is then expanded with terms specific to new crises.

Lexicon building. We exploit the fact of having a single domain by creating a lexicon that captures crisis-relevant terms frequently used in crises tweets, which is then adapted to a specific event (Section 6.3). Typically there are two design decisions regarding lexicons: categorize terms in a number of predefined categories (e.g., WordNet, VerbNet), and/or weight terms across one or more dimensions (e.g., SentiWordNet). The former is adopted for building broad linguistic resources with numerous dimensions. If the application domain is more focused (e.g., sentiment extraction) the latter is used [163], which we also adopt here.

6.2. Data sets and Evaluation Framework

In this section we describe the input data sets we use, and the evaluation method and metrics by which we compare different alternative strategies.

6.2.1. API limits

Twitter's API for accessing tweets in real-time (the *streaming* API) has several limitations. The two that are most relevant for the work we describe in this Chapter are the following.

First, tweets can be queried by content *or* by geographical location. Specifically, if both content and geographical criteria are specified, the query is interpreted as a disjunction (logical *OR*) of both. The content criterion is specified as the disjunction of up to 400 terms, in which each term is a case-insensitive conjunction of words without preserving order. The location criterion is specified as the disjunction of a set of up to 25 rectangles in coordinate space.

Second, independently of the method used to query, the resulting set is limited to 1% of the stream data. If the query matches more than 1% of the data, then the data is sub-sampled uniformly at random. As a result, even if we use a "blank" query (collect everything), we never obtain more than a sample of 1% of tweets. As a query becomes broader (i.e, by including more terms or a larger geographical region) at some point we start losing tweets because of this limitation. This means that "collecting everything and then post-filtering" is an ineffective sampling method: at least part of the selection must be done at query time.

6.2.2. Data sets

We use data from 6 disasters between October 2012 and July 2013, occurring in English-speaking countries (USA, Canada, and Australia) which affected up to several million people. Crisis keywords were defined by two research groups: Aron Culotta's "Data Science for Social Good" team [21], and the NSF SoCS project group at Kno.e.sis using the Twitris tool [284], who shared partial lists of tweet-ids with us. Location-based data was partially collected using Topsy analytics. As detailed in Table 6.1, for each disaster we use two sets of data collected from Twitter: (1) a keyword-based sample[1] and (2) a location-based sample. We note that filtering by

[1]The West Texas explosion keyword-collection was obtained from GNIP, which allows more expressive query formulation than the Twitter API. We used an estimated query that approximates this collection with a precision and recall higher than 98%.

Name / Type	Start / Duration	Keyword-based sampling (# of terms): Examples of terms	# of tweets	Location-based sampling Region(s)	# of tweets
Sandy Hurricane	2012-10-28 3 days	4: hurricane, hurricane sandy, frankenstorm, #sandy	2,775,812	NY City; Bergen, Ocean, Union, Atlantic, Essex, Cape May, Hudson, Middlesex & Monmouth County, NJ, US	279,454
Boston Bombings	2013-04-15 5 days	17: boston explosion, BostonMarathon, boston blast, boston terrorist, boston bomb, boston tragedy, PrayForBoston, boston attack, boston tragic	3,375,076	Suffolk and Norfolk Counties, Massachusetts, US	88,931
Oklahoma Tornado	2013-05-20 11 days	36: oklahoma tornado, oklahoma storm, oklahoma relief, oklahoma volunteer, oklahoma disaster, #moore, moore relief, moore storm, #ok, #okc	2,742,588	long. $\in [-98.25, -96.75]$ $\wedge$ lat. $\in [34.5, 35.75]$	62,237
West Texas Explosion	2013-04-17 11 days	9: #westexplosion, #westtx, west explosion, waco explosion, texas explosion, tx explosion, texas fertilizer, #prayfortexas, #prayforwest	508,333	long. $\in [-97.5, -96.5] \wedge$ lat. $\in [31.5, 32]$	16,033
Alberta Floods	2013-06-21 11 days	13: alberta flood, #abflood, canada flood, alberta flooding, alberta floods, canada flooding, canada floods, #yycflood, #yycfloods, #yycflooding	370,762	Alberta, Canada	166,012
Queensland Floods	2013-01-27 6 days	4: #qldflood, #bigwet, queensland flood, australia flood	5,393	Queensland, Australia	27,000

Table 6.1.: Summary statistics of the six disasters and the two data samples (keyword-based and location-based). The set of crisis-specific keywords were manually chosen by the data providers.

a conjunction of keywords and locations is *not* possible using Twitter's current streaming APIs. In addition, both of these conditions occur in only a fraction of the relevant tweets (Section 6.2.3).

The *keywords-based samples* use keywords chosen by the data providers following standard practices for this type of data collection. This typically includes hashtags suggested by news media and response agencies,[2] terms that combine proper names with the canonical name of the disaster (e.g., *oklahoma tornado*), or the proper names given to meteorological phenomena (e.g., *typhoon pablo*).

The *location-based samples* are obtained by collecting all the postings containing geographical coordinates inside the affected areas. Geographical coordinates are typically added automatically by mobile devices that have a GPS sensor, in which their users have allowed this information to be attached to tweets. Location-based samples were obtained through two data providers: GNIP,[3] which allows to specify a region through a rectangle defined by geographical coordinates, or Topsy,[4] which additionally allows to indicate the names of the places of interest (counties, states, etc.)

6.2.3. Evaluation Framework

Our filtering task can be seen as a binary classification task. The positive class corresponds to messages that are related to a crisis situation, while the negative class corresponds to the remaining messages. This is a broader, more inclusive definition than being informative [152], or enhancing situational awareness [314].

Labeling crisis messages. The labeling of messages was done through the crowdsourcing platform Crowdflower[5]. For efficiency and to improve the quality of the data we use to train our models, we perform a pre-filtering step. We first eliminate messages that contain less than 5 words as we deem them too short for training our lexicon. Next, we eliminate messages that are unlikely to be in English by checking that at least 66% of the words were in an English dictionary.[6]

The task is designed to encourage crowd-workers to be *inclusive*, which is aligned with the goal

[2] http://irevolution.net/2012/12/04/catch-22/
[3] http://www.gnip.com/
[4] http://www.topsy.com/
[5] http://www.crowdflower.com/
[6] NLTK's English dictionary and the English database WordNet.

Categorize tweets posted during the 2013 Oklahoma Tornado:
Read carefully the tweets and categorize them as:
A. In English and directly related to the tornado.
– "The tornado in Oklahoma was at least a mile wide"
B. In English and indirectly related to the tornado.
– "The nature power is unimaginable. Praying for all those affected."
C. In English and not related to the tornado.
– "Oklahoma played well soccer this night"
D. Not in English, too short, not readable, or other issues.
– "El tornado en Oklahoma ..."

"*Seeing everyone support #Oklahoma makes my heart smile!#oklahomatornado*"
This tweet is:
A. In English and directly related to the tornado.
B. In English and indirectly related to the tornado.
C. In English and not related to the tornado.
D. Not in English, too short, not readable, or other issues.

Figure 6.1.: Example instructions (top) and example crowdsourcing task (bottom) used for labeling crisis messages.

of having high recall. We present workers a tweet and ask if it is in English and:

(A) directly related to a disaster,
(B) indirectly related,
(C) not related, or
(D) not in English or not understandable.

For purposes of our evaluation, the positive class is the union of tweets found to be directly and indirectly related, and the negative class is the set of tweets found to be not related.

For clarity, we include the type of disaster in the question. Example instructions appear in Figure 6.1. We showed crowd-workers 15 tweets at a time, out of which one tweet was labeled by the author of this thesis, and used to control the annotation quality of crowdworkers. Given the subjectivity of the task, tweets used to control quality were selected to be obvious cases for each category.

From each crisis we labeled 10,050 tweets selected uniformly at random from the keyword-based sample (50% of labels) and location-based sample (50% of the labels). On average, about 100 workers participated in each crowd-task. We asked for at least 3 labels per tweet and kept the majority label. On average, 31.5% tweets were labeled as directly related, 22.2% as indirectly related, 45.8% as not-related, and 0.5% as not in English, etc.

Measuring precision and recall. Evaluating *precision* is straightforward, as it corresponds to the probability that a message included in a sample belongs to the positive class. Evaluating *recall* is more difficult as it requires a complete collection containing all the crisis-related mes-

Disaster	Keyword-based Prec.	Keyword-based Recall	Location-based Prec.	Location-based Recall
West Texas Explosion	98.0%	29.0%	6.7%	(100.0%)
Alberta Floods	96.0%	41.9%	8.0%	(100.0%)
Boston Bombings	86.3%	25.3%	15.9%	(100.0%)
Sandy Hurricane	92.1%	39.3%	26.1%	(100.0%)
Queensland Floods	71.2%	17.9%	8.8%	(100.0%)
Oklahoma Tornado	66.2%	45.4%	9.0%	(100.0%)
Average	85.0%	33.1%	12.4%	(100.0%)

Table 6.2.: Precision and recall of keyword-based and location-based sampling. The task is finding crisis-related messages.

sages for each disaster. Yet, such a collection may require to label up to 300K messages to cover a single minute of Twitter activity.[7]

Since our methods rely on selecting tweets based on keywords, we evaluate them on the *location-based sample*. According to this definition, the recall of a keyword-based sampling method is the probability that a positive element in the location-based sample matches its keywords.

Table 6.2 evaluates the keyword-based and location-based samples using the crowdworker labels. Both precision and recall vary significantly across crises. In general, the precision of keyword-based sampling (66% to 98%) is higher than that of location-based sampling (7% to 26%). We note that the average recall of about 33% that we observe in the keyword-based samples means that about two thirds of the crisis-related messages in the location-based samples do not contain the specified keywords – that is the main motivation for the methods we describe in Section 6.3.

Further metrics. We regard the problem of collecting crisis messages as a *recall-oriented task*. Our solution should accept messages when in doubt, without accepting all messages which yields a trivial 100% recall.

There is a significant imbalance between the positive and negative classes, as seen in Table 6.2. Due to this, we use the metric *G-mean* – the geometric mean of the recall of the positive class and the recall of the negative class—often used to assess the classification performance on imbalanced data [298]. Furthermore, we measure the F_2 and F_1 scores, where F_k is $\frac{(1+k^2)PR}{k^2P+R}$ with P and R being precision and recall, with emphasis on the F_2 score which weights the recall more heavily for reasons we have explained.

We also evaluate the proportion of different classes of messages (e.g. related to donations, warn-

[7] https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

ings) in each sample to understand the representativeness of the data samples yield by each of the approaches we test in this Chapter. We defer the explanation of that evaluation to Section 6.4.2.

6.3. Proposed Method

Our method is based on creating a generic *crisis lexicon*: a list of terms to be used instead of a manual query to sample crisis-related messages. This crisis lexicon can be expanded with terms specific to a given crisis, either manually, or by using a mechanism similar to pseudo-relevance feedback.

6.3.1. Building the Lexicon

Figure 6.2 depicts the steps we take to construct the lexicon. We start by selecting the set of terms that discriminate crisis-related messages (L_0). Next, we refine this set by performing a series of curation steps filtering out both contextual and general terms as decided by crowdworkers ($L_{1...3}$). Finally, we filter out terms that frequently co-occur to maximize recall for a limited sized lexicon ($\mathrm{topdiv}(\cdot)$).

Candidate Generation Step (L_0)

Term selection. Our candidate terms are word unigrams and bigrams. We start with tweets from the positive and negative classes described in Section 6.2. We remove URLs and user mentions (*@username*). After tokenizing, we discard tokens that are too short (2 characters or less), too long (16 characters or more, typically corresponding to concatenated strings of words), or that correspond to punctuation, numbers, or stopwords. The remaining words are stemmed using Porter's stemmer.[8] Word unigram and bigrams are then extracted, and kept if they appear in at least 0.5% of the tweets.

Term scoring. Each term is then scored by two well-known statistical tests: chi-squared (χ^2) and point-wise mutual information (PMI), used in the past for lexicon creation [163]. Details are in Appendix A.1.

[8] http://tartarus.org/ martin/PorterStemmer/

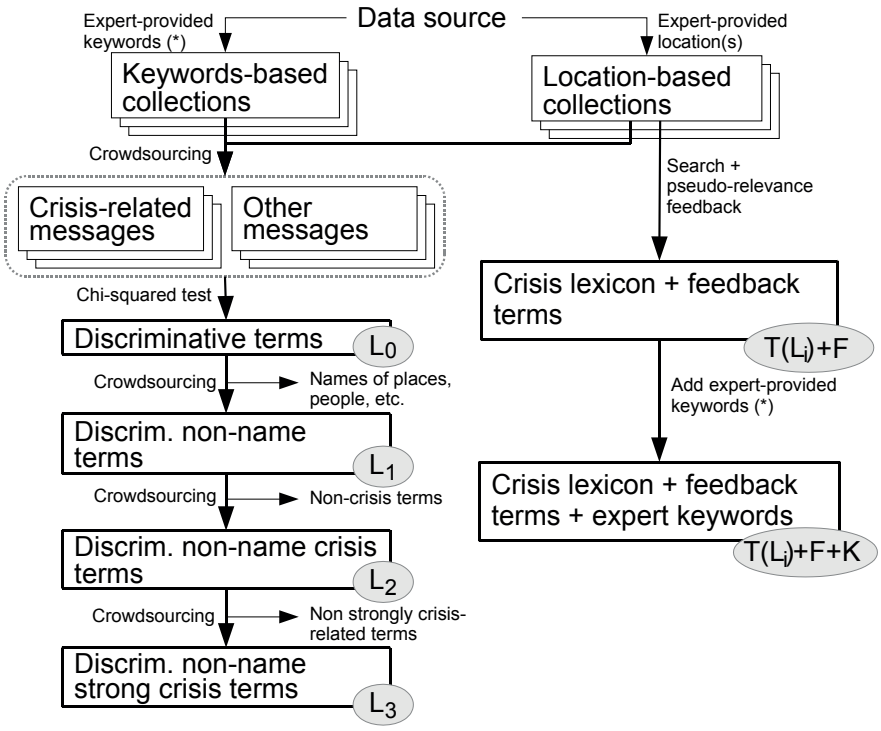


Figure 6.2.: Steps in the lexicon construction (left), and in the evaluation of the lexicon combination with pseudo-relevance feedback and expert-provided keywords (right). $T(\cdot)$ selects the highest-scoring terms: $\mathrm{top}(\cdot)$, or the highest-scoring terms ensuring diversity: $\mathrm{topdiv}(\cdot)$

We refer to the result of a statistical test of discriminative value for a term t on a crisis c as its *discriminative score* $\mathrm{discr}(c, t)$. We rank terms according to this score, divide them in n-quantiles of one term each, and score each term t belonging to the k-th quantile according to the quantile probability ($\frac{k}{n}$). We can use this score directly, or combine it with the term's frequency in the crisis-related tweets (γ) by multiplying it with the probability of the quantile to which t belongs when the ranking is done according to γ instead of $\mathrm{discr}(c, t)$. We map scores to quantiles to give equal weight to the term's $\mathrm{discr}(c, t)$ and its frequency. The outcome is a per-crisis score of a term $s(c, t)$.

For our lexicon to be general, we look for terms that work well across a variety of crises. We tested multiple aggregations of scores across crises including median, mean, and harmonic mean. The best result was obtained when computing the mean crisis score of a term across crises, and then multiplying it by a sigmoid function to favor terms that appear in (at least 0.5%

Indicate if the term is specific to a particular disaster: it contains the name of a place, the name of a person, or the name of a disaster:
A. YES, it contains a place name or it refers to the name of a region, city, etc.
– "Jersey flood"; "California people"; "okc tornado"
B. YES, it contains a person name or it refers to the name of a politician, etc.
– "Obama"; "Kevin donate"; "John hurt"
C. YES, it contains a reference to the name given to a disaster
– "Sandy hurricane"; "abfloods"; "yycfloods"
D. NO.
– "tornado"; "hurricanes"; "help rebuild"; "firefighter"; "rise"; "flame"; "every"

Indicate if the term is more likely to appear in Twitter during hazards:
A. YES, it is likely to appear more often during hazards/disasters.
– "tornado"; "donate help"; "people killed"; "state emergency"
B. NO, but could appear frequently during hazards/disasters as well.
– "power"; "water"; "nursing"; "recover"
C. NO, it shouldn't appear more often during hazards/disasters.
– "children"; "latest"; "south"; "voted"

Figure 6.3.: Crowdtask for filtering name terms (top) and identifying strong and weak crisis-related terms (bottom).

of the tweets of) several crises:

$$s_{\text{agg}}(t) = \frac{1}{1 + e^{-\frac{|C_t|}{2}}} \frac{1}{|C_t|} \sum_{c \in C_t} s(c, t) \tag{6.1}$$

Where C_t is the set of crises in which t appears. If C_t is large enough the sigmoid function converges to 1 (> 0.9 when $|C_t| > 4$), while when the term appears to be discriminative in only one crisis, this factor is around 0.6.

Curation Steps ($L_{1...3}$). After identifying and scoring the set of candidate terms L_0, we perform a series of curation steps depicted in Figure 6.2 which yield increasingly filtered sets L_1 through L_3, which we detail next.

Removal of names (L_1). We remove terms that name contextual elements unique to a crisis. Such terms mainly fall within three categories: (a) the names of affected areas; (b) the names of individuals involved in the disaster; and (c) the names used to refer to a disaster. We ask evaluators if a term contains such proper nouns, which filtered out about 25% of the terms. The task description is in Figure 6.3 (top).

Removal of non-crisis terms (L_2 and L_3). Next, we filter out those words that are not specific to disasters. We consider three levels of crisis relevance:

(1) *strongly crisis-specific*: the term is likely to appear more often during disasters;
(2) *weakly crisis-specific*: the term *could* appear frequently during disasters; and
(3) *not crisis-specific*: the term should not appear more often during disasters.

We ask evaluators to label each term with one of these categories. This task is depicted in Figure 6.3 (bottom). Of the terms that pass the previous filtering step (L_1), around 50% of them are filtered out by *weak* filtering (L_2) and around 65% by *strong* filtering (L_3).

Top-terms selection step. Twitter's API allows us to track up to $K = 400$ keywords, making this the maximum size of our lexicon. To use this allocation effectively, we test two strategies: $\mathrm{top}(\cdot)$ and $\mathrm{topdiv}(\cdot)$. The first strategy selects the top terms according to their crisis score. The second also selects the top terms according to crisis scores, but removes terms with lower crisis scores that frequently co-occur with higher score terms, as they match on a similar set of tweets. To find such a subset of terms, we compute the *independent set* on the term co-occurrence graph thresholded at a given level.[9] Given a set of queries (keywords- and location-based) and a collection of relevant tweets for each query, we build a graph G in which nodes are terms weighted by their crisis score, and between each pair of terms that co-occur in more than 50% of the tweets, we draw an unweighted edge. Then, we determine the *maximum weighted independent set (MWIS)* of G, which represents a subset of terms with high scores that rarely co-occur. Intuitively, this improves recall (since the lexicon has a limited number of terms).

The maximum independent set problem is NP-complete [303]. To this end, we compared the approximation method in [31] with a simple greedy algorithm (GMWIS) that keeps the most discriminative terms that rarely co-occur. Since the latter obtained slightly higher recall scores, we focus the discussion on those results obtained with GMWIS.

6.3.2. Applying the Lexicon

Pseudo-relevance feedback. We adapt the *generic* lexicon with terms specific to the targeted crisis. To identify such terms we employ pseudo-relevance feedback (PRF) mechanisms with the following framework:

(i) Given a lexicon lex containing at most 400 terms, retrieve crisis relevant tweets in the first Δ_t hours of the event. We refer to these tweets as pseudo-relevant.
(i) From these tweets, extract and sort the terms (unigrams and bigrams) – which do not already belong to the lexicon – by their PRF score (explained below). Return the top k terms to be added to the lexicon.

A similar methodology has showed effectiveness in other Twitter-related search tasks [92].

[9] The idea of mapping terms co-occurrences on a graph is inspired from [339, 328].

PRF term scoring. PRF terms are usually scored according to their distribution in the feedback tweets, or according to the comparison of the distribution in the feedback tweets and the entire collection [330]. Due to having only the extracted PRF tweets, the scoring strategies we implement fall within the former category:

- **Frequency-based** scoring ranks PRF terms according to their frequency in the feedback tweets: $s_{prf}(t) = fr(t)$.
- **Label propagation-based** scoring propagates the scores from the query terms to PRF terms based on their co-occurrence in the feedback tweets:

$$s_{prf}(t) = \frac{\sum_{q \in lex} co(q,t) s_{agg}(q)}{\sum_{q \in lex} co(q,t)} \tag{6.2}$$

  where $co(q,t)$ is the number of co-occurrences between query term q and PRF term t, and $s_{agg}(q)$ the crisis score of q as defined in Equation 6.1.

PRF term selection. To select the top PRF terms we test again the two strategies described in (§6.3.1): $\mathrm{top}(\cdot)$ and $\mathrm{topdiv}(\cdot)$. For $\mathrm{topdiv}(\cdot)$, we compute the maximum weighted independent set based on the co-occurrence graph formed only by PRF terms.

Terms sampling. We note that some of the selected terms might actually be harmful [47]. A workaround is to resample the terms based on their co-occurrence with sub-samples of the original query [62]. The main hypotheses are that feedback documents form clusters according to the query terms that matched them, and that good PRF terms occur in multiple such clusters [330]. Yet, in contrast with [330], we cannot make assumptions about terms distribution in the whole collection, since we only have the pseudo-relevant tweets, and given the short nature of tweets we do not attempt to model their language. In contrast, we use the sigmoid function to favor the PRF terms that co-occur with multiple query terms:

$$s_{prf}(t) / (1 + e^{-\frac{|T_{prf}(t)|}{2}}) \tag{6.3}$$

where $T_{prf}(t)$ is the number of terms co-occurring with term t and $fr(t)$ is t's frequency in PRF documents.

Hashtags. Hashtags are topical markers for tweets [310], used to learn about events and join the conversation [293]. During crises, specific hashtags emerge from the start, with some of them quickly fading away, while others end up being widely adopted [259]. Kamath et al. [164] found that hashtags can reach their usage peak many hours after their initial use. Thus, even

if they are scarce in the beginning, if widely adopted later on, hashtags improve recall. On the other hand, if they are not adopted, they end up having little impact on the retrieved data. Therefore, we lower the selection barrier for hashtags by employing a dedicated PRF-step: we add the top k hashtags (appearing in at least 3 tweets) to the query according to their frequency in the PRF documents, similar to [319].

6.4. Experimental Evaluation

We compare against two standard practices: sampling using a manually pre-selected set of keywords, and sampling using a geographical region. The goal of the lexicon is to sample a large set of crisis-related messages; this is what we evaluate first (§6.4.1). Next, we see if our method introduces biases in the data collection compared to existing methods (§6.4.2).

In both cases, we perform *cross-validation* across disasters:

(1) leave one disaster data set out;
(2) build the crisis lexicon ($L_{0...3}$) using data from the remaining disasters;
(3) evaluate on the excluded disaster data set;
(4) repeat the process for each of the 6 disasters, averaging the results.

6.4.1. Precision and Recall

We evaluate the precision and recall of different strategies for sampling crisis-related messages. We also incorporate other metrics, particularly those that emphasize recall, as described in §6.2.3.

Lexicon generation. First, we identify the best versions of our lexicon along the analyzed metrics. There are several design choices that we exhaustively explore:

- The term scoring method (§6.3.1): χ^2, PMI, χ^2 + γ, PMI + γ, and γ.
- The curation steps executed (§6.3.1): no curation (L_0), removing names (L_1), keeping weak and strong crisis terms (L_2) and keeping strong crisis terms only (L_3).
- Whether to select the top scoring terms: $\mathrm{top}(\cdot)$, or the top scoring terms removing co-occurring terms: $\mathrm{topdiv}(\cdot)$.

This yields 40 configurations that we test along the two existing methods, i.e., keyword-based

and geo-based sampling. Figure 6.4 highlights the *skyline* configurations, i.e., the configurations for which there is no other configuration that simultaneously leads to higher recall and higher precision. Further, given that the points with similar properties tend to cluster along the skyline, we keep on the skyline only those points with the highest precision when they are within 5 percentage points from each other in terms of both precision and recall.

We notice that *different methods have different precision-recall trade-offs*. The term-scoring method appears to influence these trade-offs the most. Specifically, the scoring methods that penalize more a term's appearance in non-crisis tweets lead to high precision at the cost of recall (e.g., PMI); those methods that put more weight on the absolute frequency of terms in the crisis tweets lead to high recall at the cost of precision (e.g. γ). χ^2 and the combination of PMI and χ^2 with γ lead to better precision-recall trade-offs, or, in other words, they lead to higher F_k scores.

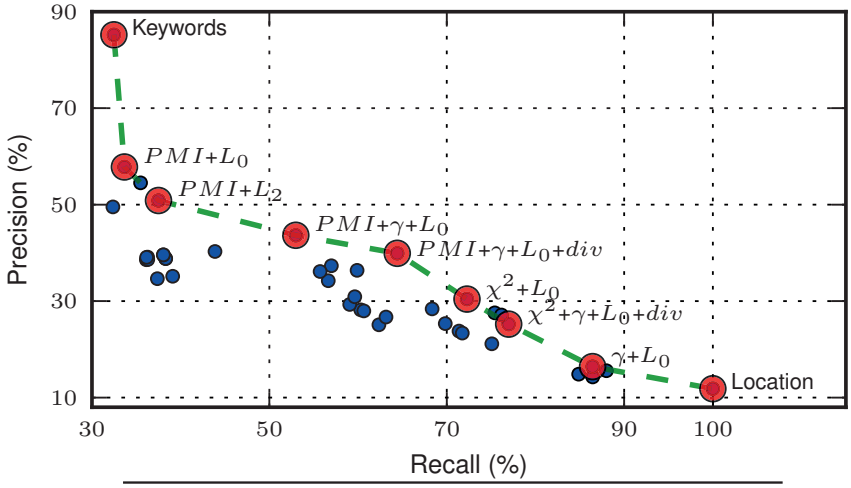
We curate the initial list of terms to improve precision (by removing terms that are too general) and recall (by removing terms that are too specific). Yet, curating the lexicon by removing proper nouns (L_1) lowers both the recall and precision. This effect is less pronounced when we remove terms with lower crisis scores that often co-occur with more discriminative terms ($\mathrm{topdiv}(\cdot)$). The next curation steps (L_2 and L_3) also alleviate this effect leading to higher precision overall. However, keeping only strong crisis-related terms (L_3) heavily impacts recall (the points clustered around 40% recall and precision in Figure 6.4).

Lexicon expansion. With the parameter combinations from the skyline in Figure 6.4 (7 options), we test the performance of our lexicon when using various pseudo-relevance feedback (PRF) mechanisms (§6.3.2). We explore the following design choices:

- PRF term scoring (§6.3.2): frequency (Fr) and label propagation (Lp).
- Whether to select the top scoring terms: $\mathrm{top}(\cdot)$, or the top scoring terms removing co-occurring terms: $\mathrm{topdiv}(\cdot)$.
- Whether to favor terms that co-occur with more query terms (§6.3.2): sp, or not: $\neg\,\mathrm{sp}$.
- Whether to use *only* a hashtag ($\#$) dedicated PRF, combine it with the PRF for terms (as defined by the previous choices), or use the later alone (§6.3.2).

We also combine lexicons by first running PRF with L_i, select the PRF terms, and then add them to L_j, where L_i, L_j are lexicons obtained with the skyline configurations of Figure 6.4; combination denoted $(L_i)L_j$. This yields about 700 configurations to test. For these tests we set the number of PRF terms to 30, and PRF interval to $\Delta_t = 3$ hours. We assume the data

[10]For brevity, in the rest of the chapter we refer to the lexicons corresponding to these configurations by this code.



Code[10]	Config.		Prec.	Rec.	GMean	F1	F2
-	Keywords		**85.2**	32.5	56.1	45.5	36.6
-	Location		11.8	**100.0**	0.0	20.5	37.6
1	PMI	L_0 top	**57.8**	33.7	57.6	42.0	36.5
2	PMI	L_2 top	**50.9**	37.5	60.5	41.1	38.6
3	PMI+γ	L_0 top	43.7	53.0	72.7	**46.9**	50.0
4	PMI+γ	L_0 topdiv	39.9	64.4	80.2	**48.0**	56.2
5	χ^2	L_0 top	30.4	72.3	85.0	41.6	**55.0**
6	$\chi^2+\gamma$	L_0 topdiv	25.2	77.0	87.7	36.6	**52.3**
7	γ	L_0 top	16.4	**86.4**	**93.0**	26.5	43.6

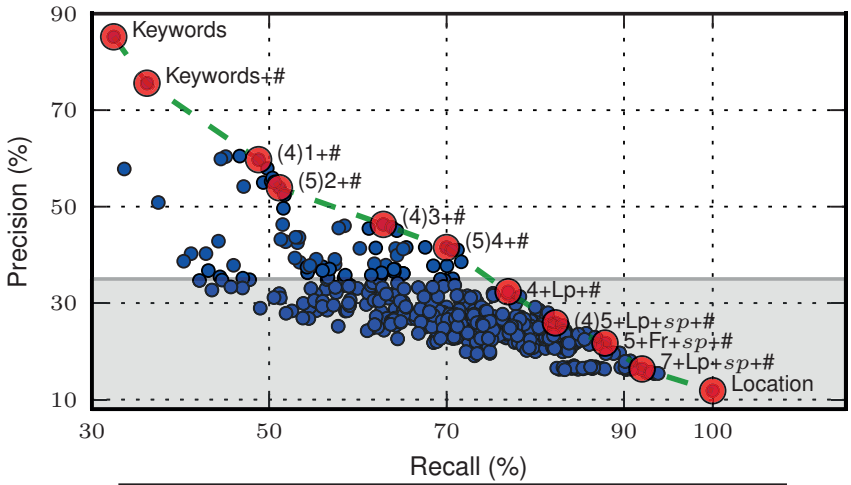
Figure 6.4.: Averaged performance of existing methods and our lexicon. Among 40 tested (small dots), the table includes the skyline configurations (large dots).

collection, and the PRF, start simultaneously with the keywords-based collection. Results are in Figure 6.5.

We notice that PRF boosts recall, but has little impact on precision. Further, the lexicon combinations with the #-dedicated PRF lead to better precision-recall trade-offs when L_i has high recall and L_j has high precision.

Expert-defined terms. To analyze how the expert-defined crisis-specific terms and the lexicon complement each other, we add the former to the queries corresponding to the top skyline configurations depicted in Figure 6.5.

As shown on Table 6.3, such combinations generally lead to improvements over both the keywords and the lexicon (e.g., up to 40 percentage points recall over the crisis-specific keywords). The only metric we do not improve on is the *precision* of the keyword collection, yet this is an upper bound for precision as the expert-edited keywords are chosen to be specific only to



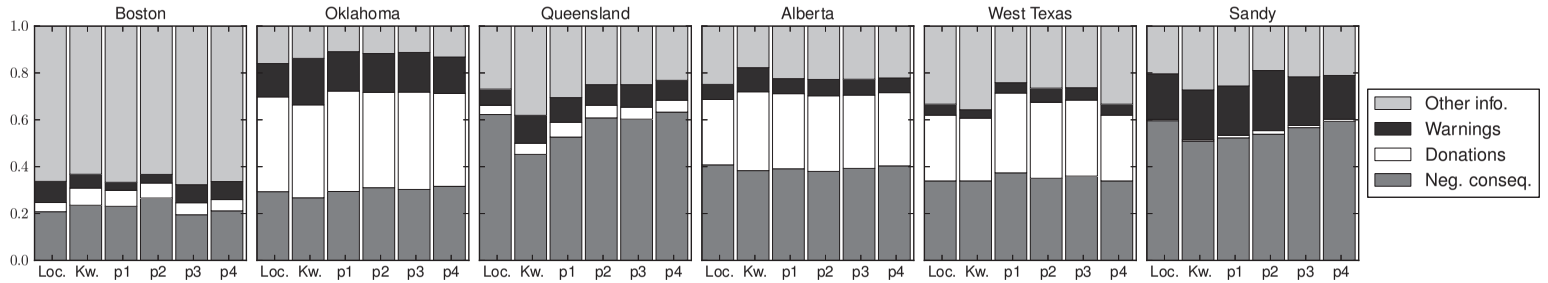
Code[10]	Config.				Prec.	Rec.	GMean	F1	F2
-	Kw.	-	#	-	75.5	36.2	59.4	47.9	40.0
p1	(4)1	-	#	-	**59.7**	48.8	69.6	**53.2**	50.4
p2	(5)2	-	#	-	53.9	51.2	70.9	50.6	50.6
p3	(4)3	-	#	-	46.3	62.9	79.2	52.3	57.8
p4	(5)4	-	#	-	41.5	**70.0**	**83.6**	50.6	**60.0**
-	4	Lp	#	¬ sp	32.4	76.9	87.6	43.1	56.8
-	(4)5	-	#	-	25.9	82.3	90.7	37.9	54.5
-	5	Fr	#	sp	21.7	87.9	93.7	33.2	51.0
-	7	Lp	#	sp	16.4	95.9	95.7	26.6	44.5

Figure 6.5.: Averaged performance of existing methods and our lexicon with PRF. From about 700 tested (small dots), the table includes the skyline configurations (large dots). The gray area marks the configurations with precision below 35% and places the corresponding skyline points at the end of the table. $(L_i)L_j$ means that we run PRF with L_i and then add the PRF terms to L_j, where L_i is a lexicon code from Figure 6.4.

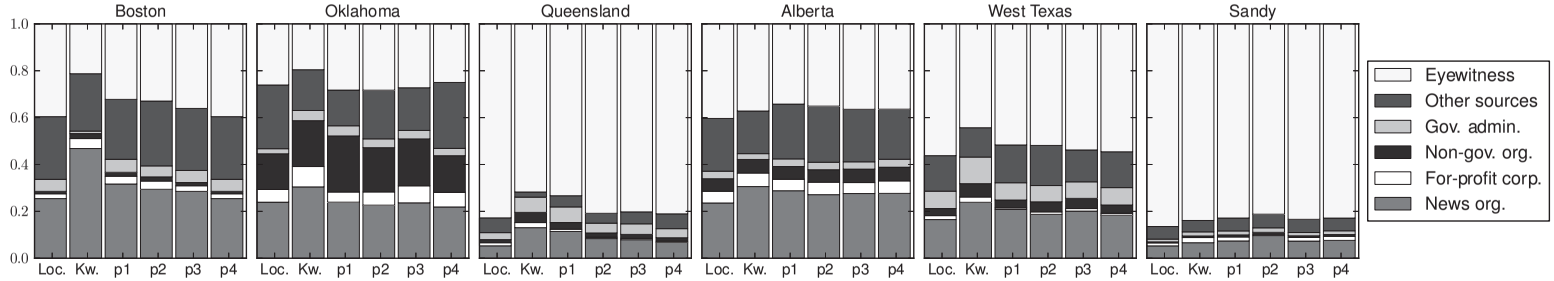
a given disaster. Furthermore, though the precision decreases, the combination leads to better precision-recall trade-offs, as it improves over the F-score metrics. p2 leads to the highest gains over the lexicon-based approach and over the F1-score of the keyword-based approach—meaning that the samples obtained with p2 and those obtain with the crisis-specific keywords overlap the least.

Performance over time. Finally, to analyze the performance variation over time, we test two design decisions: running PRF only one time at the beginning of the crisis (one-time PRF), or re-running PRF after every 24 hours (online PRF). We measure the average performance's variation across the first three days from the start of the keyword-collections.[11] Figure 6.6 shows

[11] We restrict this analysis to the first three days for two reasons: all collections span across at least three days, and, typically, the largest volumes of tweets happen in the first days of the event (i.e. around the peak in the Twitter



(a) Tweet distribution per type of information for each sampling method. The average BC coefficient between the distribution of the message types in the location-based collection and in data sampled from it with our lexicon and the keywords: $BC(keywords) = 0.994$, $BC(p1) = 0.995$, $BC(p2) = 0.996$, $BC(p3) = 0.998$, $BC(p4) = 0.999$



(b) Tweet distribution per type of source for each sampling method. The average BC coefficient between the distribution of the message source in the location-based collection and in data sampled from it with our lexicon and the keywords: $BC(keywords) = 0.984$, $BC(p1) = 0.993$, $BC(p2) = 0.996$, $BC(p3) = 0.997$, $BC(p4) = 0.999$

Config.	Prec.	Rec.	Gmean	F1	F2
p1	60.8 (-24.4/1.1)	55.7 (23.1/6.9)	74.2 (18.2/4.5)	56.1 (11.7/4.1)	57.3 (19.5/5.6)
p2	56.9 (-28.3/**3.1**)	60.7 (28.4/**8.4**)	77.7 (21.7/**6.0**)	**57.7** (**12.2**/**6.8**)	59.2 (22.7/**7.8**)
p3	47.7 (-37.4/1.6)	66.6 (34.1/3.7)	81.5 (25.5/2.2)	54.8 (9.3/2.7)	61.0 (24.5/3.3)
p4	42.3 (-42.8/1.0)	**73.5** (**41.0**/3.5)	**85.7** (**29.6**/1.8)	52.4 (6.9/2.3)	**62.7** (**26.2**/2.7)

Table 6.3.: Average performance of our lexicon when combined with crisis-specific keywords. We also report (the improvement over such keywords/the improvement over the method without these keywords) as percentage points.

the performance of the lexicon with both one-time PRF and online PRF in terms of recall and F1-score relative to the crisis-specific keywords, which is the reference values. We omit the corresponding precision plots, but note that an increase in recall with no improvement in F1-score indicates a loss in precision.

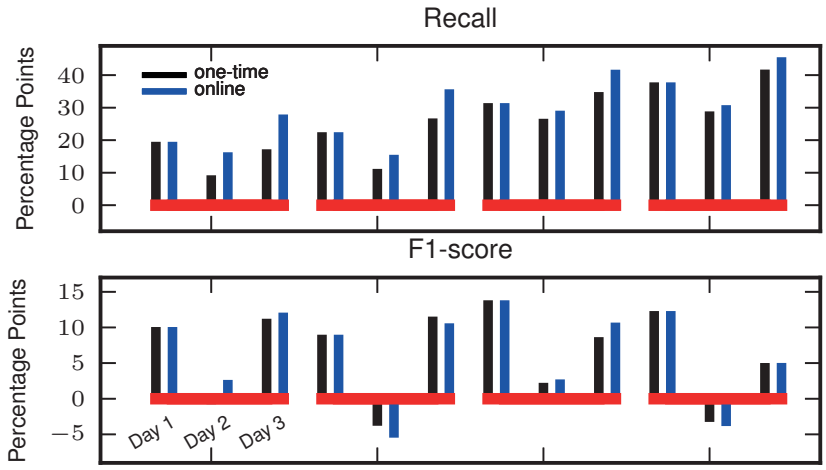
In our experiments, the lexicon based approaches do better on average (in the range of 20 to 40 percentage points for recall and 9 to 13 percentage points for F1-score) towards the beginning of the crisis compared to the crisis specific keywords. Then, we see a drop in the performance relative to the keywords which might be due to more users conforming to keywords use as the event gets global coverage, followed by an increase when the event loses coverage. Finally, although employing online PRF leads to better recall values later on in the crisis, it's improvement in terms of F1-score over one-time PRF is only marginal.

6.4.2. Distribution of message types

We measure changes in the distribution of tweets of different types, as sampling by keywords may introduce *biases* that favor one class of tweets at the expense of another. We evaluate by asking crowdworkers to categorize tweets, and then measure the divergence between the distribution of tweets into categories across the sampling methods. We repeat this twice using three categorizations: informativeness, information type and information source (details in Appendix A.2).

First we check if any sampling method biases the collection towards the tweets deemed informative by crowdworkers. With one exception, we find only marginal differences across crises; looking at crisis-relevant tweets, we find that between the lexicon and the crisis-specific key-

messages).



	Recall			F1-score		
	Day 1	Day 2	Day 3	Day 1	Day 2	Day 3
Keywords	35.3	**39.0**	31.6	49.4	**53.4**	41.7

Figure 6.6.: Relative performance over time of our lexicon with one-time PRF and online PRF re: crisis-specific keywords. The table contains the reference performance by the keywords—represented by the (red) horizontal line.

words there is a difference of less than 10 percentage points regarding the proportion of informative tweets. The (reference) location-based samples have lower proportions of informative tweets than the lexicon and keywords-based samples. The exception is Hurricane Sandy, for which the p2 configuration collects more informative tweets (about 18 percentage points) than the keywords sample.

Figures 6.6(a) and 6.6(b) depict the tweets distribution according to the type and source of information. We also show the Bhattacharyya coefficient (BC) which quantifies the overlap between the reference location-based collection, and the lexicon and the keyword-based samples in terms of information type and source—higher BC values indicate higher similarity.

We notice large variations in tweet distributions according to both the information type and source across crises; yet it has little to no impact on the sampling methods' ability to preserve the distributions. Generally, high-precision methods diverge more from the reference sample, with the keyword-based samples being the least representative, e.g., it collects more tweets coming from news organizations and fewer eyewitness reports (Figure 6.6(b)). In contrast, our lexicon better preserves the reference distribution, with a BC close to 1.

6.5. Conclusions

We have described a methodology for constructing an effective, general lexicon for monitoring domain specific events, which we have extensively evaluated for crisis events. Our experiments demonstrate a range of precision and recall operating points previously not well understood when using only keyword or location-based sampling. This work provides researchers an informed strategy for assembling a set of relevant tweets. This is a fundamental technology for automatic linguistic analysis tools such as temporal summarization [127].

The impact of these results goes beyond an algorithmic understanding. We showed that the amount of data that it is currently mined represents only a fraction of the data posted during disasters. We believe that such lexicons can support others interested in increasing the recall of their data collections, but who may not have the ability to finely tune their lexicons.

6.5.1. Future Work

There are many directions in which to take the work we described in this Chapter.

Social Media in Crises. First, users are often interested in classifications more finely grained than 'relevant' or 'non-relevant': e.g., emergency responders may be interested in personal or property loss tweets, each of which will admit its own lexicon.

Improving Data Collection. Second, though our techniques to improve the quality of the data collections are in principle language-independent and domain-independent, we want to build lexicons which demonstrate this. Further, when using a lexicon to collect data through an API, if the API is more limited or less limited, or limited in a different way, our results may have to be adapted. Finally, it would be desirable to keep human effort to a minimum—mostly because one may want to build a specialized lexicon in a short time—and, thus, more efforts are needed to develop methods that simplify the manual steps of the process.

6.5.2. Reproducibility & Data Release

To ensure and support the reproducibility and replicability of this case study, the crisis lexicon, the list of keywords, geographical regions, etc. along with the labeled data sets as sets of (tweet-ids, label, and metadata) are available for research purposes at `http://crisislex.org/`.

7. Methods Assessment: A Study of Item Recommendation

In this Chapter, we explore the sensitivity of social data methods to data biases and variability, by focusing on recommendation systems—one of the most popular (and long-standing) applications that leverages online social behavioral traces, which is today inescapable in a wide range of web applications. Particularly, the rise of online social networks created new prediction opportunities for recommendation systems: instead of relying on past rating history through the use of collaborative filtering (CF), they can leverage the social relations among users as a predictor of user tastes similarity.

Alas, little effort has been put into (i) understanding when and why—e.g., for which users and what items—the *social affinity* (how well connected users are in the social network) is a better predictor of user preferences than the *interest affinity* among them as algorithmically determined by CF; and (ii) and how to better evaluate recommendations depending on the recommendation context such as the type of users a recommendation application targets. This oversight is explained in part by the lack of a systematic collection of data sets including both the explicit social network among users and the collaborative annotated items. To fill this gap, we conduct an extensive empirical analysis on six real-world publicly available data sets from four distinct recommendation sites. We dissect the impact of user and item attributes, such as the density of social ties or item rating patterns, on the performance of recommendation strategies relying on either the social ties or past rating similarity. Our findings represent practical guidelines that can inform methods evaluation and can assist in future deployments and mixing schemes.

7.1. Background

The recommendation systems are inescapable in a wide range of web applications, e.g. Amazon or Netflix, to provide users with books or movies that match their interest. Accurate recommendations generate returns of investments up to 30% due to increased sales [180]. Many such

systems rely on collaborative filtering (CF) approaches that recommend items based on user rating history. Concomitantly, the rising popularity of social networks has provided new opportunities to filter out relevant content for users. For instance, recommendation services like Epinions, Last.fm or BeerAdvocate are enhanced with virtual social networks.

As a result, existing works have proposed both pure social recommenders (SR)[1] that only leverage the social ties among users [216], and hybrid approaches that either augment the CF recommendation engine with social guidelines [283, 211], or incorporate CF mechanisms into a social recommendation engine [154].

A common practice in evaluating such approaches is to resort to *(i)* one [302, 334, 223, 156, 154, 181, 210], sometimes two [283, 211] data sets and *(ii)* global averages for the metrics of choice. Alas, this has made it difficult to draw generalizable conclusions on the effectiveness of leveraging the social ties for recommendations compared with CF across data sets of different nature.

Furthermore, the use of global metrics[2] to evaluate and compare the recommendation approaches may be inconclusive as they provide little insight into *when* and *why* the approaches succeed or fail [94]. Although the impact of the parameters of a recommendation strategy has been often inspected [56, 283, 156, 301, 209, 154, 34], little systematic effort has been devoted into understanding how various user or item attributes are affecting the performance [10], and none of such analyses, to our knowledge, have included SR approaches.

7.1.1. Contributions

Orthogonal to designing better hybrid approaches that combine SR and CF features, our goal is to gain insight into the relative benefits of each of these approaches that, in turn, can guide future deployments and mixing schemes. To do so, we perform an extensive empirical analysis that dissects the recommendation performance, measured by precision and coverage, and does a fine grained comparison across various user and item classes on six *publicly available* data sets including both the ratings information and the social network among users (Section 7.3). All data sets are medium to large-scale and exhibit various properties regarding user social ties and items ratings. We focus on the two ends of the problem spectrum, which places on the one side the *interest affinity* among users (respectively items), as algorithmically determined by CF from user rating history, and at the other side the *social affinity* as inferred from users social

[1] For readability, in this chapter we use *social* to refer to both trust and social ties.
[2] Metrics that are computed or averaged over all predictions.

network by *pure* SR (Section 7.2). Our analysis addresses two main questions:

(1) Are global metrics able to reflect the performance of a given recommendation strategy across various settings? Our analysis shows that one cannot rely on global metrics to assess a given recommender performance not only across all data sets but also within each data set, across different classes of users or items. Even a slight change in the global average might hide important changes in the performance distribution across a data set demographics. One may thus need to understand and optimize the performance on a specific demographic subset depending on the application specifics—e.g., for a beer recommendation service, it might be more important to be accurate in the recommendations made to experienced and, likely, harder to please users [217].

(2) Are there user or item attributes that hint at the CF (interest affinity) performance with respect to SR (social affinity)? In our results, we find that when the basis of formulating connections among users stems from *plain* friendship, rather than from sharing interests, SR leads to less precise recommendations. Further, items likeability (the rating they received on average) and user selectiveness (the rating they give on average) are good predictors of the recommendation performance: relying on *social affinity* leads to more precise predictions for highly liked items, while for indulgent users (that typically give high ratings) leveraging the *interest affinity* for items similarity is best. More results are discussed in (Section 7.3).

7.1.2. Related Work

Collaborative Filtering (CF) has been widely used by major commercial applications such as Amazon, Movielens, or Netflix [180, 207, 9]. These methods leverage users rating history and predict the rating of a target item i and a source user u by looking at the ratings on the target item given by similar users to u, *user-based approaches* [118], or at what ratings items similar to the target item have received from the source user, *item-based approaches* [277]. Yet, relying solely on collaborative filtering is known to be ineffective when dealing with large numbers of items, given the sparsity of the user-item ratings matrix. *Cold start* users and items are particularly affected, CF often failing to make predictions in such cases (i.e., leading to a low *coverage*). While collaborative filtering approaches fall within two main classes, *neighborhood-* and *model-based* approaches [78], for this case-study we focus on the former as due to its' tendency to better capture local associations in the data we consider it more suitable for juxtaposing different ways of measuring the affinity among users or items (as we detail in §7.2.1).

Social recommender systems (SR) In contrast, SR systems leverage users social ties to make predictions [340, 216, 256, 117, 223], assuming that these reflect common tastes or interests.

SR systems deal better with *cold start* users, as they require users only to be connected to other users in the social network, and, thus, do not have to wait for users to grow a rating history to make predictions. Alas, while these systems tend to achieve better coverage, they can also suffer due to sparse ratings and sparse trust relations. Thus, in order to consider the ratings of users that are not directly connected, various approaches propagate the trust among their users [340, 117, 216, 223]. Yet, in these cases the recommender might end up considering ratings of weakly trusted users, thus affecting the precision [154].

Social-enhanced collaborative approaches incorporate social factors to the collaborative framework by tailoring the rating similarity based on the social ties [181]; making predictions based on friend ratings weighted by the level of trust, and integrating them in the CF framework [209]; adding social regularization factors to matrix factorization recommendation techniques by constraining a user inferred taste (her feature vector) with the average taste of her friends, and the similarity with each of them [211], thus making her feature vector depend on those of her friends [156], or by accounting for the social ties heterogeneity [283].

In contrast, *collaborative-enhanced social approaches* implement a social-based framework that falls back on CF when trusted users did not rate the target item. TrustWalker enhances a social-based approach with item-based CF [154], and employs a random walk model that first tries to exploit the social network by looking for the ratings on the target item at trusted nodes (*trust-based approach*). Yet, as the random walk advances, if a rating on this item is not found, the likelihood to return the rating of a similar item (*item-based approach*) increases. TrustWalker acts in extreme settings as a pure SR approach when the random walk never stops for similar items, and as pure item-based CF when the walk never starts (navigating the same problem spectrum as us).

7.2. Problem Definition

Typically, a recommender task is to predict ratings for unseen items to users. To do so, a set of items I, a set of users U, and a set of items $I_u \subseteq I$ rated by each user u with a rating $r_{u,i}$ on a Likert scale from 1 to 5 is considered. If the recommender system exploits the social ties among users, for each user u a set of friends F_u is assumed. This chapter looks at the predictive capability of social ties (SR) compared to the one of items or users rating similarity (CF) for items recommendation.

7.2.1. Comparison Framework

We conduct our study using a comparison framework that implements a recommendation template under which, to make a recommendation for user u on target item i, two main steps are performed[3]: (1) identify the set of similar users (respective items) with u (respective i) and (2) compute weighted aggregates of their ratings on i (respectively from u) according to the similarity with u (respective i). On top of it, we implement the main building blocks of SR and CF as used for comparison in literature [11, 154, 155, 223, 181, 168]. Specifically, we implement (a) item- and user-based CF variants as often used as reference point by previous work [154, 223, 277, 181], and (b) a SR approach that aggregates the ratings similarly with CF, yet, instead of deriving users affinity based on how similar they rated items in the past, it does so based on their social ties. Next we describe each approach and motivate our choices.

Collaborative Filtering (CF) approaches are usually grouped in two main classes: *neighborhood-* and *model-based* [78]. Model-based variants have received lot of attention as their accuracy was considered superior, yet neighborhood-based CF, although simpler, remains competitive [70]. Further, they exploit different patterns in data, none of them consistently out-performing the other: model-based CF is typically effective at estimating the overall model related to all items simultaneously, while neighborhood-based CF better captures local associations in data [33]. This trait makes neighborhood-based CF suitable for our purpose to compare the predictive capability of *interest affinity* (inferred based on implicit similarity links as determined by CF) and *social affinity* (computed based on explicit social links among users). Further, neighborhood-based CF offers a simple and intuitive template for recommendation to easily implement a pure SR-based approach on top of it and fairly compare the two under the same setting.

We use common variants of the two main types of neighborhood-based CF: user- and item-based CF. Briefly, for each user u (respectively item i) a neighborhood UN_u (respective IN_i) of users (items) similar with u (respective i) is built and their ratings on the target item i (respective from active user u) are aggregated as:

$$p_{u,i} = \frac{\sum_{v \in UN_u} sim(u,v) r_{v,i}}{\sum_{v \in UN_u} sim(u,v)} \tag{7.1}$$

for user-based CF, where $sim(u,v)$ is the similarity between users u and v, as estimated by the Pearson correlation of the ratings given by u and v on the same items[4]; respective, $p_{u,i} =$

[3] As in neighborhood-based CF [132].

[4] Note that we also consider only positive correlations [154].

$\frac{\sum_{j \in IN_i} sim(i,j) r_{u,j}}{\sum_{j \in IN_i} sim(i,j)}$ for item-based CF, where $s(i,j)$ is the Pearson correlation of the ratings received by i and j from the same users.

Social Recommendation (SR) In contrast to CF[5], when the ratings received by target item i are aggregated according to Equation (7.1), SR weights them based on the social affinity between the active user (i.e., the user for which we want to make a prediction) and the users that have rated item i in the past.

Social Affinity (relatedness) of two nodes in a social graph can be estimated using random walks (RWs) [208], which have been used for both friend [24, 204] and item recommendations [332, 154, 102]. In short, for each prediction, we run RWs on the social graph that start at user u needing a recommendation on item i, and stops when they either reach a user v that have rated the target item i, or have performed a maximum number of steps k_{max}[6]. We denote a RW stopping condition with $s_{v,i,k}$, which is $true$ if $i \in I_v$ or $k >= k_{max}$, meaning that the RW stops at v. Then, the social affinity between u and user v that rated the target item i is the probability to reach v using different paths and number of steps:

$$P(X_{u,i} = v) = \frac{\sum_k P(X_{u,i,k} = v)}{\sum_{w \in U} \sum_k P(X_{u,i,k} = w)}, \tag{7.2}$$

where the random variable $X_{u,i}$ represents the nodes that rated item i and can be reached at any step of the RW starting at node u, while $X_{u,i,k}$ represents only the subset of nodes reachable at step k:

$$P(X_{u,i,k} = v) = \sum_{w \in U} P(X_{u,i,k-1} = w) P(X_w = v) \tag{7.3}$$

where $P(X_{u,i,0}) = 1$ and X_w the random variable to pick a friend of node w. For unweighted graphs (as those used in our evaluation), we have:

$$P(X_w = v) = \frac{1}{|F_w|} \tag{7.4}$$

Thus, the probability to step on node $v \in F_w$ at step $k+1$ after being at node w at step k is:

[5] For brevity, when referring to both user-based and item-based CF, we use only CF.

[6] Set to 6 based on the "six-degree of separation" assumption [224] that most of the nodes are reachable within 6 hopes [154].

$$P(X_{u,i,k+1} = v | X_{u,i,k} = w, \overline{s}_{w,i,k}) = P(X_w = v) \tag{7.5}$$

where $X_{u,i,k}$ is the random variable for nodes that can be reached at step k when looking for i, $\overline{s}_{w,i,k}$ is the negation of $s_{w,i,k}$, and $P(X_{u_i,k+1} = v | X_{u,i,k} = u, s_{w,i,k}) = 0$ to complete the probability distribution. To also complete the specification of the probability distribution in Equation (7.3), we define a final state $\perp$, to which the RW goes when it terminates:

$$P(X_{u,i,k} = \perp) = 1 - \sum_{v \in U} P(X_{u,i,k} = v) \tag{7.6}$$

To determine if we performed enough RWs to make an admissible prediction, after each RW we compute the variance $\sigma^2 = \frac{\sum_{j=1..T}(r_j - \overline{r})}{T}$ in the results of all the walks [154], where T is the number of successful walks[7], r_j is the result returned by the j-th RW, and $\overline{r}$ is the mean of the results return by the RWs. If the variance σ^2 converges to a constant (i.e., the variance after $j+1$ walks varies with less than $\epsilon = 0.0001$ from the variance after j walks), or the total number of (successful and unsuccessful) walks reaches the maximum number of walks $T_{max} = 1000$, we stop from running more RWs. Then, to make a prediction, in Eq. (7.1), we replace the similarity between active user u and user v which have rated item i with their relatedness in the social network:

$$p_{u,i} = \sum_{\{v \in U | i \in R_v\}} P(X_{u,i} = v) r_{v,i} \tag{7.7}$$

7.3. Empirical Analysis

In this section we perform an extensive analysis that juxtaposes the SR (social affinity) and CF (interest affinity) as predictors for item recommendation, structured in three parts. First, we present a comprehensive characterization of the data sets. Second, we apply global metrics to evaluate the recommendation strategies, and examine if they capture the performance variation across various settings. Finally, we do a fine grained analysis of the impact of user and item properties on the performance, organized as a set of questions about CF and SR properties. These questions are largely inspired by admitted properties of CF or SR, such as, CF performs better on users for which it has more information [116, 15, 45], the recommendation accuracy

[7] A random walk is successful if it encounters a user that have rated the target item.

Data set	Users	Items	Ratings	Social Links	Links Type
Ciao	12,375	99,762	284,086	237,350	direct
Epinions1	49,290	139,738	664,824	487,181	direct
Epinions2	22,166	296,277	922,267	355,813	direct
Epinions	132,000	755,760	13,668,319	841,372	direct
Flixster	786,936	48,794	8,196,077	7,058,819	symmetric
Douban	129,490	58,541	16,830,839	1,692,952	symmetric

Table 7.1.: Data sets Figures

decreases towards the long-tail items (i.e., less popular items) [296], or SRs are superior on *cold start* users [154, 216].

7.3.1. Metrics and Experimental Setup

To evaluate the recommendation performance, we use the well-known *leave one out* strategy. Specifically, we remove from the data set only the rating we want to predict and leave the other ratings and social network unchanged. Then, we compare CF and SR along two popular metrics:

1. The *coverage* measures a recommendation strategy ability to make predictions, and it is the number of ratings the system succeeded to make divided by the total number of ratings that it tried to predict.
2. The *Root Mean Square Error (RMSE)* captures the average error between the predictions and the real ratings, measuring the recommendation precision:

$$RMSE = \sqrt{\frac{1}{N}\sum(r_{u,i} - p_{u,i})^2} \tag{7.8}$$

where N is the number of predictions, $r_{u,i}$ the real rating given by u to item i, while $p_{u,i}$ is the prediction. Note that the smaller the RMSE is, the more precise the recommendations are.

Albeit the RMSE ability to gauge the performance for pervasive top-k recommendations is debated [66], it best fits our purpose to measure performance shifts across classes of items/users. The accuracy metrics deemed suitable to evaluate top-k performance, are biased towards the performance on preferred items (i.e., high ratings) [133]. Moreover, many recommendation systems that leverage the social ties optimize for RMSE [333] (as it is perhaps the most popular metric [281]), making our analysis convenient to compare with.

Two approaches are used to report RMSE and coverage values for a set of users/items:

(i) compute the RMSE (respectively coverage) over all the predictions to users (or for items) in the set; or
(ii) compute the RMSE (respectively coverage) for each item/user separately and average the results over all users (respectively items) in the set.

While the first measures the overall performance on estimating the ratings, the second weights each user (respectively item) equally measuring how good the predictions are on average for each user (respectively item) in the set. We measured both, yet, when the two variants lead to similar conclusions we show only results with the second one; both are included otherwise. Finally, when measuring how a certain user (respectively item) property impacts the results, we group the users (respectively items) by logarithmically binning them regarding the property value, and then compute the performance for each bin[8].

7.3.2. Data sets Characterization

We conduct our analysis on 6 real world publicly available data sets including both ratings and a social network (figures are summarized in Table 7.1):

Epinions is a popular product review site where people rate products and build lists of trusted users whose reviews they find useful. We use two rating data sets from Epinions: one that is collected by the authors of [215] around 2006 (noted *epinions1*), and one that is collected in May 2011 by the authors of [302] (noted *epinions2*). In addition to product ratings, in Epinions, users can also rate product reviews. We also use a data set, made available by Epinions.com to the authors of [215] containing ratings on product reviews, instead of ratings on products (noted *epinions*). In all data sets the ratings are on a scale from 1 to 5.

Douban is a Chinese product review site that represents one of the largest online communities in China. As in Epinions, users rate and review products in order to receive recommendations. In addition, at the date of crawling, it provided a Facebook-like social networking service [211].

Ciao defines itself as a *multi-million-strong online community* in which users critically review and rate millions of products. It provides the same functionality as Epinions (i.e., users can both rate products and indicate the trusted users) [301].

Flixster is a large social movie rating service that allows users to create Facebook-like friendship relations and share ratings [156], which are from 0.5 to 5 (with a step of 0.5). To ensure

[8] We use logarithmic binning (in base 4) to account for the fact that some values in the degree, popularity, or activity distributions are frequent while others are not. A linear binning leads to bins with few or no points.

Data set	Ratings Per User	Ratings Per Item	Avg. Degree	Mean Rating	Median Rating
Ciao	22.9	*2.8*	**19.1**	4.16	4
Epinions1	13.4	4.7	9.8	3.99	4
Epinions2	41.6	3.1	16	3.97	4
Epinions	103.5	18.0	*6.3*	**4.67**	**5**
Flixster	*10.4*	167.9	8.9	*3.8*	4
Douban	**129.9**	**287.5**	13.0	3.84	4

Table 7.2.: Data set Statistics. Bold marks the highest value per column, while italic the lowest.

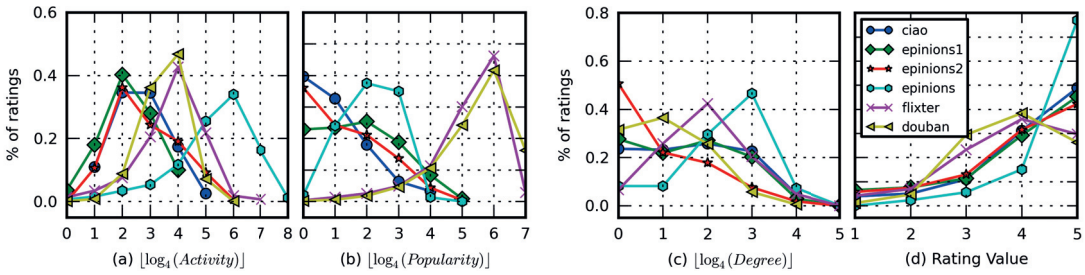


Figure 7.1.: Distribution of ratings as function of: (a) user activity; (b) item popularity; (c) user degree; (d) rating value

uniformity across the analyzed data sets, we round the ratings to the next integer so as to obtain ratings on a 1 to 5 scale.

Data Statistics. We want to understand the properties of the data sets we analyze, the resemblance among them, as they might explain the performance variations across them. Table 7.2 highlights basic statistics for each data set.

Rating Distributions. Figure 7.1 shows the rating distributions across user and item properties, and the rating value. In Figure 7.1(a) we notice similar patterns across data sets with only little variation (for larger data sets, the level of user activity at which the peak number of ratings is

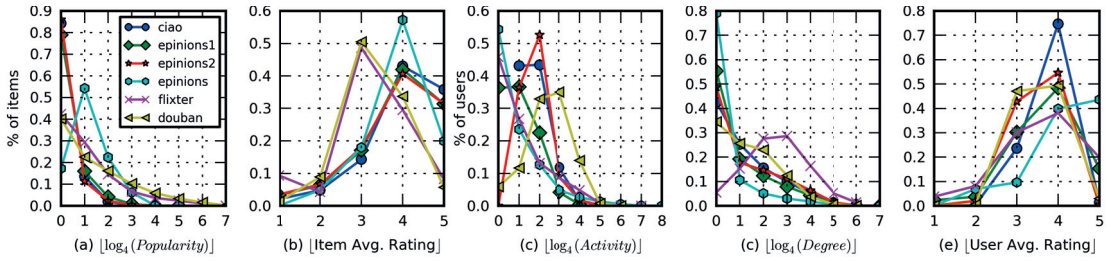


Figure 7.2.: The distribution of items as a function of (a) item popularity and (b) average rating per item, and the distribution of users as a function of (c) user activity, (d) user (out-)degree and (e) average rating per user.

produced is shifted towards higher ranges). In contrast, the rating distribution according to item popularity, Figure 7.1(b), varies greatly: while in some data sets (*ciao*, *epinions1*, *epinions2*) the highest fraction of ratings is given to unpopular items, in others (the largest ones) this is accounted for popular items. Figure 7.1(c) also shows that while in *flixster* and *epinions* most ratings are given by moderately social connected users, in other data sets a higher number of ratings is credited to lower degree users. Looking at rating distributions according to the rating value, Figure 7.1(d), we see that in all data sets the values are skewed towards higher ranges (peaking around 5).

Item Distributions. We observe similar patterns across all data sets: Figure 7.2(a) illustrates that with only one exception (*epinions*) the cold start items (with only few ratings) represent a significant fraction of all items. Figure 7.2(b) shows that in all data sets most of the items received on average a rating of 3 or 4.

User Distributions. SR is believed to address *cold start* users, as it does not require them to rate items for making predictions, but only to be connected in the social network. Given that in some data sets the number of *cold start* users is significant (roughly 50% [154]), improving on this set of users might significantly impact the overall performance. Thus, on average such approaches were found to outperform CF [154, 216]. Yet, when the percentage of cold start users is not significant, this might not be the case. Figure 7.2(c) shows that while in some data sets (*epinions*, *flixster*) cold start users are a significant percentage, this is clearly not the case in others (*douban*, *ciao*). Additionally, regardless of their fraction, cold start users always produce a minor fraction of ratings (see Figure 7.1). In Figure 7.2(d), we notice that, except *flixster*, the number of low degree users is larger than the number of cold start users, which in turn might affect SR overall performance. Finally, Figure 7.2(e) shows that, on average, users tend to give higher rating values.

Correlations. We also checked the correlation among item and user properties (item popularity, user activity and degree, and the average rating received by an item or given by a user). Given that, in general, we found low or no correlation, we report only on statistically significant ($p < 0.01$) moderate Pearson correlations ($|r| \geq 0.2$). We found moderate and positive correlations among users degree and their level of activity in *ciao* ($r = 0.59$), *epinions1* ($r = 0.45$) and *epinions*($r = 0.36$). In *flixster* ($r = 0.43$), *douban* ($r = 0.35$) and *epinions* ($r = 0.30$) there is a positive correlation between items popularity and the ratings they got, i.e, popular items tend to obtain higher ratings. Item popularity also correlates negatively with users level of activity in *flixster* and *douban* ($r = -0.20$ in both data sets), i.e., active users are more inclined to rate unpopular items. While in *douban* there is a negative correlation ($r = -0.29$) between users

Data set	User CF	Item CF	Social
Ciao	**1.144** (0.410)	1.285 (0.318)	1.252 (**0.626**)
Epinions1	**1.186** (0.512)	1.428 (0.463)	1.362 (**0.663**)
Epinions2	**1.164** (**0.483**)	1.361 (0.395)	1.406 (0.365)
Epinions	**0.466** (0.930)	0.602 (0.579)	0.559 (**0.951**)
Flixster	1.013 (0.969)	**0.889** (**0.991**)	1.349 (0.985)
Douban	**0.784** (0.996)	0.809 (**0.997**)	1.037 (0.894)

Table 7.3.: Overall performance. In each cell we report RMSE (Coverage) computed over all the ratings in the data set. Bold highlights the best value on each row.

level of activity and the ratings they give on average, indicating that active users are more likely to give lower ratings; in *epinions* popular items tend to get higher ratings ($r = 0.31$).

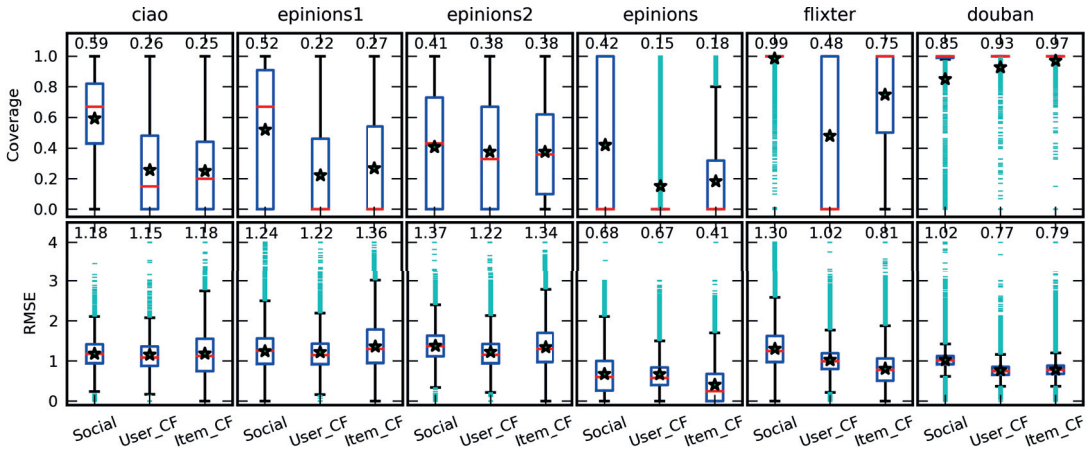
We will see in the next sections how these varying data properties explain the different performance numbers obtained when aggregating the results differently (e.g., user-oriented vs. item-oriented evaluation) within and across data sets.

7.3.3. Overall Performance Characterization

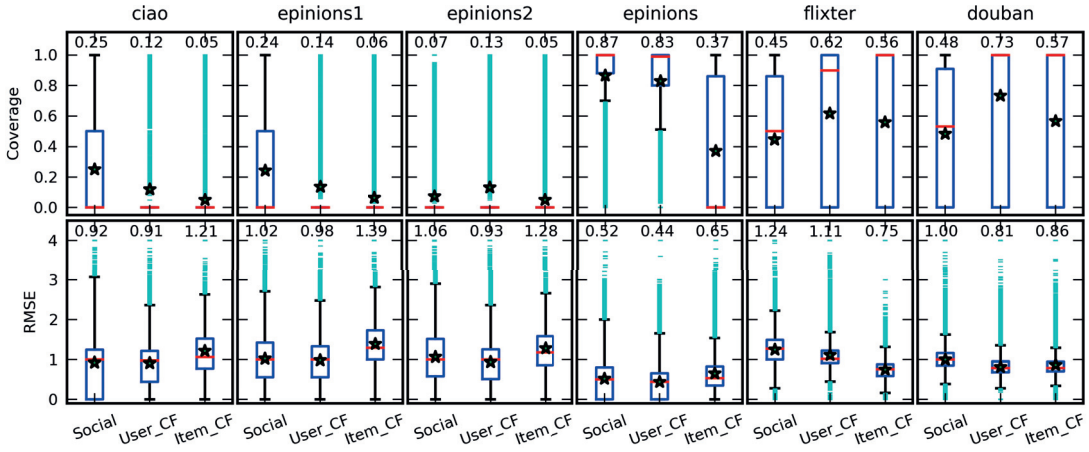
A common practice in recommender systems evaluation is to show how their performance varies with approach-dependent parameters. Yet, even when there are correlations between the parameter values and performance level, it is difficult to know, for instance, if the improvements hold for the entire population, or only for some subgroups. Thus, we want to observe if there is a trivial relationship between the experimental results obtained through globally computed metrics that summarize the performance, typically used to evaluate recommendation systems [154, 156, 223, 283], and the averaged performance at user (respectively item) level.

Table 7.3 reports the globally computed metrics (*rating-oriented evaluation*) per data set and approach. For error rates, with only one exception (i.e., *flixster*), user-based CF performs best across all the data sets. In terms of coverage, there is no clear winner: SR performs best for *ciao*, *epinions1* and *epinions*, while user-based CF for *douban* and *epinions2*, and item-based CF for *flixster*. Next, we check if these results are also confirmed by the *user (respectively item)-oriented evaluations* (§7.3.1) which measures how well an approach does on average per user (respectively item). In Figure 7.3 the boxplots show the shape of the average performance distribution for users (respectively items), its central value, and variability.

User-oriented evaluation. Figure 7.3(a) shows the per-user performance variation across data sets. Though it mostly confirms the overall results (in terms of winners) for most data sets, there are exceptions in which SR, respectively item-CF, fares better than the globally computed



(a) Per-user distributions



(b) Per-item distributions

Figure 7.3.: Results Distribution: The boxplots divide the data, except outliers (the blue lines), in four equal buckets. A data point displays the performance on a particular user (respectively item). The redline splitting the boxplot is the median, while the star is the average performance (also plotted above each boxplot).

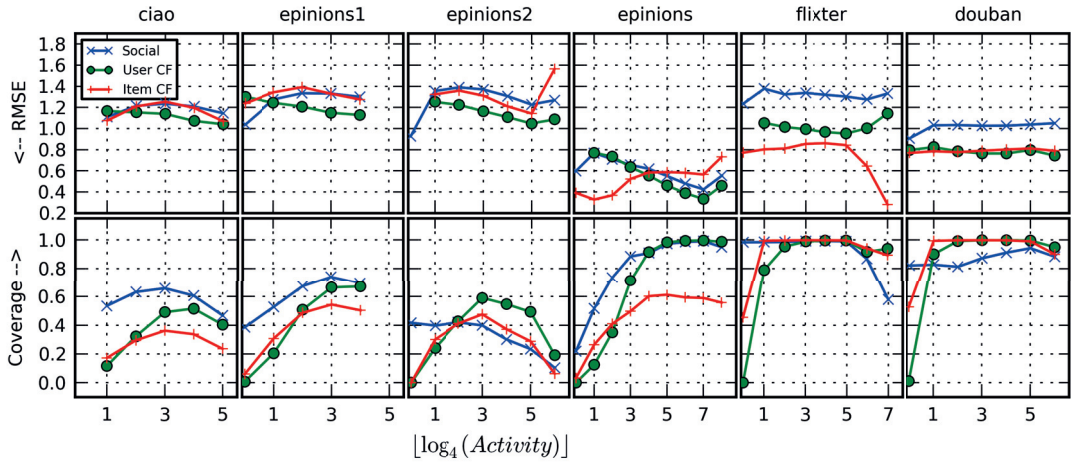


Figure 7.4.: Performance as a function of user activity: (top) average RMSE per user; (bottom) average coverage per user.

metrics indicate in Table 7.3: e.g., the coverage on *flixster*, where the fraction of unsocial users is lower than that of cold start users, and RMSE on *epinions*, where there is a higher fraction of items with similar ratings, than of users giving similar ratings.

Item-oriented evaluation. Similarly, barring the coverage on *flixster* and *douban*, Figure 7.3(b) also confirms (in terms of winners) the figures in Table 7.3. Yet, we notice that except *epinions* and *flixster*, in all the other data sets both the distributions and the average coverage values are significantly shifted towards lower ranges regarding the user-oriented evaluation, which is explained in part by the much higher fraction of unpopular items than of cold start users that these data sets exhibit.

This demonstrates that it is difficult to rely on global metrics to assess or explain a given recommender performance, a finer granularity has to be applied; and that indeed no general conclusion can be drawn regarding the relative superiority of a given recommendation method over another, not only across data sets but also within each data set.

7.3.4. In-Depth Performance Characterization

We aim to understand the benefit of each approach under a variety of settings. In this regard, we address a set of questions about the properties of CF and SR, some of which are well embedded in the conventional wisdom:

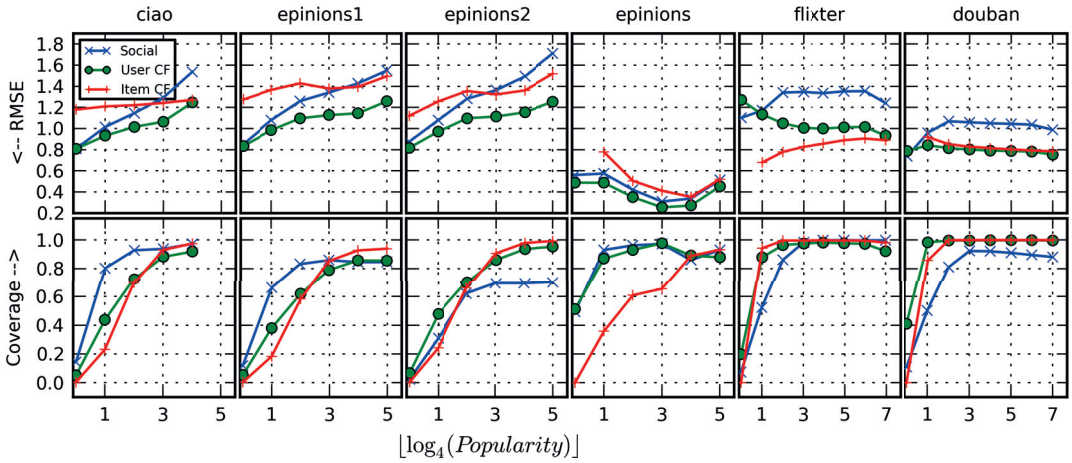


Figure 7.5.: Performance as a function of item popularity: (top) average RMSE per item; (bottom) average coverage per item.

Does CF fare better for users (respectively items) with more ratings? The belief is that CF does better when a user has rated more items [116, 45]. To test it, we analyze how CF performs as users are more active (have rated more items). Figure 7.4 shows that users' level of activity impacts the ability to make predictions (the coverage) similarly across all approaches: being more active helps only until some threshold after which rating more items either does not help (*epinions*, *douban*) or can even be harmful (*epinions2*). Further, while rating more items tends to help user-based CF to make precise prediction (in *epinions* and *flixster* after slightly improving for a while, the error increases again), item-based CF has a more inconsistent pattern. Looking at the relative performance of CF regarding SR (barring cold start users, i.e., the first bin on the $\log_4$ scale), we notice that users level of activity impacts user-based CF and SR similarly in terms of both coverage and RMSE. Exceptions are the coverage results on the data sets that exhibit no correlation among users social degree and their level of activity (*douban*, *flixter*).

As with more ratings per user, the belief is that more ratings per item help CF [296]. To challenge it, we look how CF performs with the number of ratings per item. Figure 7.5 shows that the average coverage per item is improving as items are more popular only until some threshold when they plateau. In addition, as a rule-of-thumb we also notice that items with about over 2^6 items tend attain a coverage of 80% or more. In contrast, for *ciao*, *epinions1*, *epinions2* (data sets with a small number of ratings per item, Table 7.2) the predictions are less precise as the items are more popular, *invalidating* the belief. Checking the relative performance of CF regarding SR, we notice that more ratings per item helps CF to increase its precision regarding SR. The only exception is *epinions* (to easily spot the patterns, follow on y-axis the distance

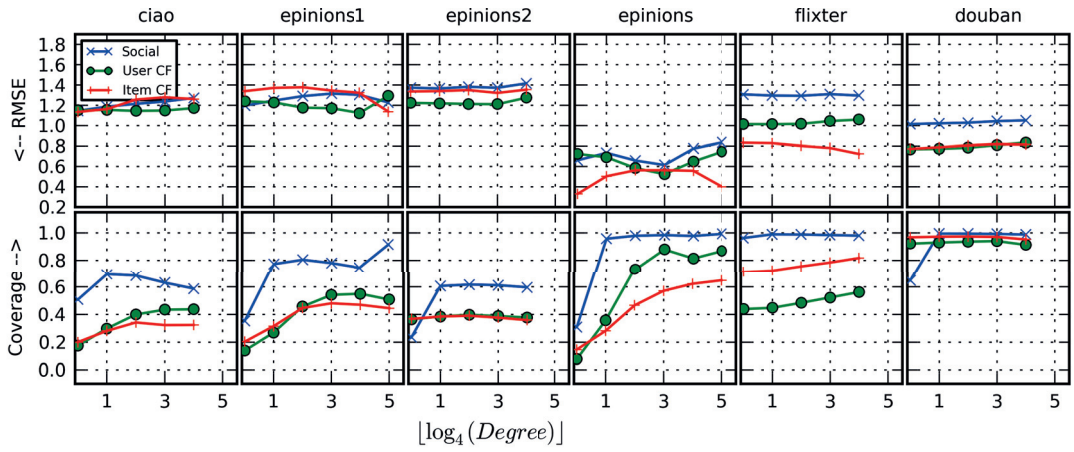


Figure 7.6.: Performance as a function of node degree: (top) average RMSE per user; (bottom) average coverage per user.

between points corresponding to the same bin but with distinct approaches).

Does SR fare better for cold start users? The belief is that SR deals better with cold start users [154] (with less than 5 items rated [117]) as it only requires them to be connected to other users to make predictions. Indeed, Figure 7.4 shows that SR achieves better coverage for these users (leftmost bins) across data sets. Yet, this is not always the case when it comes to precision (RMSE). For instance, we observe that for *flixster* and *douban* (when the social ties stem from friendship), CF attains a better precision for all users, including cold start ones.

Does SR fare better for users with more social connections? Intuitively, more social information available should help SR. To check this, we study how SR performs across users with various social degrees. Figure 7.6 shows that higher degrees help improve the coverage only until users are moderately connected (have at least 5 connections), after which linking to more users seems to bring little or no benefit for SR, even declining on *ciao*. Neither SR's precision improves as users are more socially active: it either slightly decreases, or plateau. This means that having too many friends might also introduce noise. This hints that many social ties might not reflect as much friendship, similarity or trust. However, on most data sets higher degrees tend to have a weak to no impact on SR's precision. Further, as with the level of activity, barring the low degree users, the social degree impacts user-based CF and SR in a similar way, in particular for those data sets in which the degree correlates with the level of activity.

Is CF doing better on low degree nodes? Since CF does not leverage the social links to make predictions, it should not be affected by their absence, and, thus, should perform better on

unsocial (low-degree) users. Yet, Figure 7.6 shows that CF succeeds to obtain a better coverage on unsocial users only for *douban* and *epinions2*. For RMSE, while on some data sets CF does better on *unsocial* users, when there is a correlation between user degrees and how many items they rate (*ciao*, *epinions1*, and *epinions*), it performs comparable with SR.

Is the Precision of SR Lower Relative to CF on Facebook-like Networks ? The process of creating connections primarily based on "plain" friendship (Facebook-like) does not necessarily correlate with one's opinions as it is orthogonal to a product recommendation task. Yet, when the basis of forming connections is to connect with people whose opinions one shares, there might be more agreement in how users rate the same items. Indeed, this distinction is clearly visible in our results (Figure 6 to 8): while SR fares comparable with CF in terms of RMSE in Epinions data sets and *ciao*, in Facebook-like *flixster* and *douban* CF significantly outperforms SR. In addition, being more socially active has little to no impact on the results obtained for *flixster* and *douban* (Figure 7.6). Thus, this indicates that the underlying nature of the network and whether or not the connections are related or orthogonal to the recommendation task is an important factor as well.

Is the performance independent of users selectiveness or items likeability? Only few studies hint at the relation between user selectiveness [217] or items likeability and recommendation performance. Yet, in Figure 7.7 we notice consistent patterns across data sets, in particular, for RMSE. In all data sets item-based CF is less precise when items are either liked (received high ratings), or disliked (received low ratings), while SR and user-based CF are less precise for users that are either very selective (giving mostly low ratings) or indulgent (offering mostly high ratings). Also note how similarly both the user and item average rating impacts the precision across all data sets (i.e., leading to similar curves for all data sets). This is surprising as it indicates that the users (respectively items) average rating is predictive for the recommendation approach precision. It is also worth noting that user-based CF and SR precision (although with slightly different values) follow almost identical curves. Yet, as Figure 7.7 illustrates, for coverage the patterns are not consistent across all data sets.

7.4. Conclusions

We conducted an in-depth empirical analysis on six publicly available data sets to study the respective merits of the *interest affinity*, as derived by CF, and the *social affinity*, reflecting how well connected users are in the social graph, for items recommendations. We focused on the building blocks of the analyzed strategies, without aiming to exhaustively inspect all

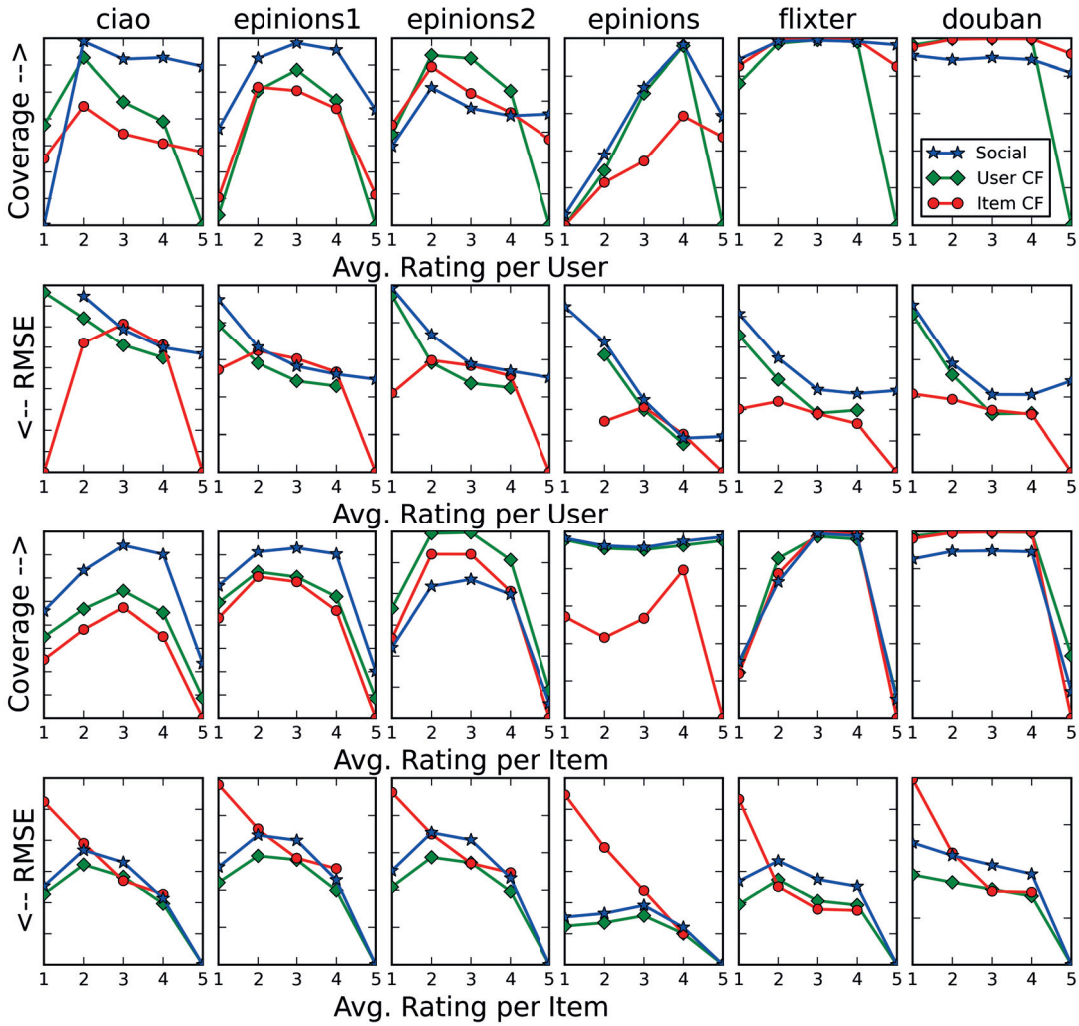


Figure 7.7.: Performance as a function of average rating value per item and per user.

possible implementations, as we argue that their understanding can better guide more complex deployments. Our study conveys that the level of user activity, item popularity or the density and nature of the underlying social network are as many characteristics that can impact the performance of recommendation systems. One needs to understand the data set demographics and optimize the performance based on the specifics of each application. We make a case for hybrid approaches, which dynamically adapt as the system evolves and the properties of user and item change over time.

7.4.1. Reproducibility

As we emphasised earlier, this case study is based on 6 publicly available data sets (see [215, 302, 211, 301, 156] for more details about the data sets), implements basic recommendation approaches and relies on popular evaluation metrics. This enables the reproducibility of our study, and also makes it convenient to contrast with both past and future related studies.

8. Conclusions

"*There are a lot of small data problems that occur in big data. They don't disappear because you've got lots of the stuff. They get worse.*"
—Prof. David Spiegelhalter, 2014, quoted in [130]

8.1. Retrospective

Social data promise to provide us with many fascinating insights about human phenomena [194] and exciting applications [173, 219]. Unfortunately, there are limits to what can be discerned from social data about real-world phenomena that have yet to be addressed. In this thesis we examine four instances of such limits that demonstrate the need for more efforts in defining common standards to collect and analyze social data, but also to devise and evaluate the methods that work with it.

Social Media Biases. First, we have devised a methodology for comparing news agendas that allowed us to quantify the biases in the coverage of domain-specific news events in social and mainstream media. We showed that social media significantly deviates from mainstream media and tends to focus on actions by individuals, original investigative journalism, and legal actions involving governments. It typically pays more attention than mainstream media to news events considered ordinary, predictable, and of low-magnitude. Thus, while there is some overlap between the two type media, social media (or at least Twitter) is far from being a good proxy for online mainstream news media.

Data Collection Biases. Second, we have performed an in-depth analysis of 26 data sets of social media messages posted during multiple crisis events. Our results suggest that some intrinsic characteristics of each event (e.g. being instantaneous or progressive) tend to produce consistent effects on the types of information available in the messages and their sources. However, we also uncover substantial variability across data sets (e.g. with the number of eyewitness accounts varying from over 50% during one event to less than 1% during another), revealing

pitfalls related to the generalization of findings from one data set to another, even across those data sets that are seemingly similar. Our study makes a case for running analyses and evaluate frameworks or methods on multiple data sets.

Leveraging Domain Knowledge to Improve Data Collection. Further, we have shown that using a generic, domain-specific lexicon to collect social media messages during domain-specific events leads to collections with higher recall than obtained with crisis-specific keywords manually chosen by domain experts, while it also better preserves the original distribution of message types. The impact of these results goes beyond an algorithmic understanding. We also show that the amount of data that it is currently mined represents only a fraction of the relevant data.

Methods Evaluation. Finally, we conducted an in-depth empirical analysis on multiple data sets to study the respective merits of different social cues (explicit social links vs. implicit interest affinity) to aid item recommendations. We show that one needs to understand the data set demographics and optimize the performance based on each application specificities, as the performance of different recommendation strategies varies not only across data sets, but also within each data set, across different classes of users or items. Thus, our results make a case for more extensive, fine-grained evaluations, not only across data sets, but also across data demographics.

8.2. Prospective

The research debating the challenges of relying on online social data to model human behavior has been more focused on asking questions rather than trying to answer them. In Chapter 2, we conducted a comprehensive survey that focuses on a variety of the challenges from issues with the current working data sets to the employed methods to capture and leverage these data sets, as well as ethical aspects, rather than reviewing existing solutions or tentatives to addressing them (if they exists). We believe that reviewing existing solutions is an equally important investigation to pursue to further guide the community towards the most pressing issues that need to be addressed. Among the all the surveyed issues, below we highlight a few that we consider important to be addressed:

Ethics, Standards, and Algorithms Discrimination. First, the ethical issues when working with such data are often overlooked [342]. Consistent standards across the community about how to handle such data are needed: e.g. when and how can we disclose user identifiers in papers? what about disclosing their social media posts as they are?

The way in which we evaluate and validate methods and observations is also important. Ruths and Pfeffer [274] put together an evaluation guideline to help reduce biased and flows in social data including showing results on multiple platforms, distinct data sets from the same and different platforms. Yet, as we discussed in Chapter 7, especially when looking at a problem that is confined within a certain context, looking at the performance variation across data demographics is also necessary. Thus, effort should be put in developing minimum evaluation requirements per classes of problems.

In the same context, there is a need to further scrutinize the implications of algorithmic decisions across applications and types of data. First, not all mistakes are equal. Thus, when choosing among methods that show different performance trade-offs, one should consider the cost of various types of mistakes that might occur—automatically classifying someone as unfitted for a job due unintentional racial profiling has much grave consequences, than the cost of interviewing more candidates. In [32], Barocas and Selbst discuss how various design choices at different steps in the analysis pipeline (see Chapter 3 for an overview) can lead to discriminatory decisions. Thus, future research should carefully investigate the differences, especially the errors, in the results of different methodological choices .

Social Issues. Further, in spite of existing limitations, we believe that if handled with care online social data can be instrumental for policy makers and advocacy groups as it can help them estimate both existing biases towards protected classes (typically including minorities), as well as the impact of various relevant policies, programmes or campaigns.

For instance, while the growing number of discussions about minority (a group that is subordinate to a more dominant group in society) issues—including gender [3], income [1], or race [4]—is good news, empirical evidence suggests that they are held mainly among the discriminated group: women dominate the debate on gender [5], while African-Americans dominate the one on race [2]. This suggests that, although social media has led to a paradigm shift for advocacy by increasing the effectiveness, the speed and the outreach of social campaigns, many—even the online campaigns—still fail to reach far beyond the communities for which they advocate. Knowing the extent to which each stakeholder group contributes to the debate is helpful in learning how to alter the message to appeal to them. Yet, although important, studies that look at various online social data to understand the public opinion and the different narratives around minority groups issues across stakeholders are scant. We are interested in this problem, and to fill this gap, we have started efforts in this direction [250].

Parting Thoughts. Eliminating all data biases, noise or other limitations is unlikely, and sometimes even undesirable. Additionally, the solutions for various limitations might pull in

opposite directions (e.g. to solve privacy related issues one might need to compromise various performance metrics). Yet, as mentioned earlier, not all mistakes are equal. It is, thus, important to assess and categorize existing limitations based on their effects, in order to understand what type of trade-offs should be made.

A. Supporting Material

A.1. Statistical Tests for Terms

For each term t we compute the following contingency table:

	related	not related
t	$n(\mathrm{t}, \mathrm{rel})$	$n(\mathrm{t}, \neg\,\mathrm{rel})$
$\neg$ t	$n(\neg\,\mathrm{t}, \mathrm{rel})$	$n(\neg\,\mathrm{t}, \neg\,\mathrm{rel})$

where $n(\mathrm{t}, c)$ is the number of tweets belonging to class c in which term t appears, $n(\neg\,\mathrm{t}, \mathrm{rel})$ the number of tweets in which term t does not appear and $c \in \{\mathrm{rel}, \neg\,\mathrm{rel}\}$. Then, similarly with [163], we use two popular statistical measures to estimate how strong the association between a term and the crisis-related tweets is (the *discriminative score*): Chi-square (χ^2) and Pointwise Mutual Information (PMI).

χ^2-based crisis score. The statistical measure χ^2 tests whether a term t occurrence is independent of the tweet being about a disaster or not; and is defined as follows:

$$\chi^2 = \sum_{x \in \{\mathrm{t}, \neg\,\mathrm{t}\}} \sum_{c \in \{\mathrm{rel}, \neg\,\mathrm{rel}\}} \frac{(n(x,c) - E[n(x,c)])^2}{E[n(x,c)]}$$

where $E[n(x,c)]$ is the expected value for $n(x,c)$.

Although χ^2 estimates the discriminative power of a term t towards one of the classes, it does not indicate if t is discriminative for the crisis-related tweets. So we ignore the χ^2 when t appears more often in the non-crisis-related tweets and define the crisis score as follows:

$$cs_{\chi^2}(t) = \begin{cases} \chi^2 & \text{if } n(\mathrm{t}, \mathrm{rel}) > n(\mathrm{t}, \neg\,\mathrm{rel}) \\ 0 & \text{otherwise} \end{cases}$$

PMI-based crisis score. PMI measure the relatedness between term t and a certain class c and

it is defined as [59]:

$$\text{PMI}(t, c) = \log_2 \frac{P(t, c)}{P(t)P(c)}$$

where $P(t, c)$ is the joint probability of t and c, and $P(t)$ and $P(c)$ are the marginal probability of t and c.

Even if PMI measures how strongly associated term t and class c are, it does not account for how strongly associated t and the other class are. So we compute the crisis score as the difference between the association strength with crisis-related tweets and the association strength with non-crisis-related tweets [163]:

$$cs_{\text{PMI}}(\text{t}) = \text{PMI}(\text{t}, \text{rel}) - \text{PMI}(\text{t}, \neg\,\text{rel}) = \log_2 \frac{p(\text{t} \mid \text{rel})}{p(\text{t} \mid \neg\,\text{rel})}$$

where $p(\text{t} \mid \text{rel})$ and $p(\text{t} \mid \neg\,\text{rel})$ are the probabilities of t to appear in crisis-related, respectively non-crisis-related tweets:

$$p(\text{t} \mid \text{rel}) = \frac{n(\text{t}, \text{rel})}{n(\text{t}, \text{rel}) + n(\neg\,\text{t}, \text{rel})}$$

$$p(\text{t} \mid \neg\,\text{rel}) = \frac{n(\text{t}, \neg\,\text{rel})}{n(\text{t}, \neg\,\text{rel}) + n(\neg\,\text{t}, \neg\,\text{rel})}$$

This yields positive scores when t has a higher probability of appearing in crisis tweets than in non-crisis tweets, and negative otherwise. Therefore, we consider only positive values.

A.2. Message Types Categorization

We label crisis-relevant tweets distribution along two main categorizations: information type, and information source. For each, we present workers a tweet and ask them to label it with the likeliest category (see Figure A.1). For quality control, one of every 10 tweets presented to a worker was labeled by one of the authors and was chosen to be an obvious case.

Indicate if the tweet is informative for decision makers and emergency responders:
"*RT @Boston_Police: Despite various reports, there has not been an arrest* "
Choose the best one: The tweet is ...
A. Informative about negative consequences of the bombings
B. Informative about donations or volunteering
C. Informative about advice, warnings and/or preparation
D. Other informative messages related to the bombings
E. Not informative: messages of gratitude, prayer, jokes, etc.
F. Not understandable because it is not readable, too short, etc.

Indicate the information source for tweets posted during a crisis situation:
"*family & friends are bruised & slightly damaged but ALIVE. now i can rest..*"
Choose the best one: This information seems to come from ...
A. News organizations or journalists: TV, radio, news organizations, or journalists
B. Eyewitness: people directly witnessing the event
C. Government: local or national administration departments
D. Non-governmental organizations (not for profit)
E. Companies, business, or for-profit corporations (except news organizations)
F. Other sources: e.g, friends or relatives of eyewitnesses
G. Not sure

Figure A.1.: Crowd-tasks for categorizing tweets according to informativeness and type (top), and source (bottom).

Table A.1.: Abbreviated GDELT themes and taxonomies used to select articles covering climate change subjects.

Bootstrap Theme: ENV_CLIMATECHANGE
Themes (22): NATURAL_DISASTER_: -HURRICANES, -STORMS, -BUSHFIRE, -BUSHFIRES, -BUSH_FIRE, -BUSH_FIRES, -BURIED_HOMES, -CYCLONES, -DESERTIFICATION, -INTENSE_RAINFALL, -EXTREME_WEATHER, WIND_STORMS, -TROPICAL_STORMS, -FOREST_BURNED, -VIOLENT_TORNADOES, -DROUGHTS; ENV_: -OVERFISH, -DEFORESTATION, -CARBONCAPTURE; MOVEMENT_ENVIRONMENTAL, MAN-MADE_DISASTER_DISRUPTION_OF_POWER, SOC_MASSMIGRATION
Taxonomies (17): TAX_FNCACT_: -CLIMATOLOGIST, -EARTH_SCIENTIST, -ENVIRONMENTAL_SCIENTIST, -ATMOSPHERIC_SCIENTIST, -WEATHERCASTER, -OIL_BARONS, -OCEANOGRAPHER, -OCEANOGRAPHERS, -SCRUTINEER; -ECOLOGIST; TAX_ETHNICITY_: -INUPIAT, -CHUKCHI, -KIRIBATIS, -KIRIBATI, -MARSHALLESE, -CHIPEWYAN; TAX_POLITICAL_PARTY_ECOLOGIST_GREENS

A.3. Climate Change Themes and Keywords

- Table A.1 containing the list of themes and taxonomies used to locate climate change related news in GDELT.
- Table A.2 containing the list of keywords used to collect Twitter data about climate change.

Table A.2.: Keywords used for sampling Twitter data

Bootstrap terms (9): climate change, global warming, climatechange, globalwarming, climate_change, global_warming, climate-change, global-warming, #climate
Terms (230): algore, australian climate, #peoplesclimate climate, climate makes, #emissions, climate rally, climate fighting, cooling warming, climate forward, climate major, climate stand, climate least, caused climate, climate going, climate economic, climate continues, climate rising, climate takes, climate warning, climate young, climate national, climate fracking, climate comes, climate far, michaelemann, caused warming, climate important, climate renewable, climate mitigate, climate david, climate working, climate environmental, climate james, #ecology, serious warming, #climate2014, climate cooling, breaking climate, methane, climate dangerous, climate international, climate give, climate leading, climate getting, #pjnet climate, climate making, climate daily, co2, climate planning, actually climate, climate explains, climate united, freezing warming, called climate, climate flooding, climate pretty, climate tackling, climate likely, climate driving, climate high, climate said, climate never, climate threatened, climate denying, climate taking, changing climate, climate left, climate growing, made warming, climate interesting, climate saying, #adaptation, #climatemarch, climate critical, climate ignoring, climate linked, climate ready, climate military, #actonclimate climate, affecting climate, climate strong, abrupt climate, climate cut, climate told, emissions, #cop19, climate finally, climate keep, climate free, #climateaction, climate natural, causing climate, climate warming, climate talking, climate put, climate recent, climate late, climate coming, climate thought, climate paul, claims climate, climate huge, climate needed, climate political, according climate, addressing climate, #green climate, climate little, climate concerned, climate seriously, bring climate, climate serious, climate seen, climate sustainable, climate public, climate made, climate flat, arctic climate, change international, change sustainable, climateprogress, change environmental, #unfccc, un_climatetalks, change denying, #green change, polluters, cfigueres, change fracking, catastrophic change, change natural, #eu2030, change renewable, #peoplesclimate change, manmade, climatereality, change warning, australian change, #co2, #actonclimate change, ghg, arctic change, change tackling, billmckibben, peoples_climate, dana1981, ginaepa, permafrost, #oceans, nrdc, acidification ocean, epa plants, arctic ice, epa plan, obama plants, assessment national, gas greenhouse, gas study, carbon, epa obama, rise sea, arctic sea, antarctic ice, ipcc report, droughts, march nyc, ice melting, caused humans, glaciers, #peoplesclimate people, ecowatch, ice melt, epa rule, ice sea, #weather hurricane, gases greenhouse, fossil fuel, rising sea, emission, extreme weather, #auspol #nuclear, #pollution, level rise, events weather, ice scientists, level sea, 400 ppm, #nuclear #thorium, #ipcc report, #peoplesclimate action, #actonclimate action, ice sheet, events extreme, #epa, 400ppm, #peoplesclimate march, epa, #auspol #thorium, pollution, antarctic collapse, ecology, greenland sheet, #peoplesclimate nyc, #forests, fossil subsidies, caps melting, #actonclimate president, fossil fuels, arctic loss, antarctic sheet, converting oxygen, dioxide, gina mccarthy, fracking study, antarctic ship, #acidification, hansen james, nanotubes, guardianeco, monoxide, #actonclimate plan, clean plan, pollutants, antarctic scientists, arctic scientists, rising seas, divestment fossil, greg hunt, agw, environmental protection

A.4. News Values Annotation

Instructions given to crowdsource annotators were the following:

> Help researchers to categorize Climate Change related events that happen in 2013 and 2014. These events have received a good deal of attention in Online News Media, on Twitter, or both. Please follow the provided links, and categorize the events according to the following dimensions:

We create several tasks for this purpose. The question and examples provided to annotators are the following.

Negativity: is this bad news?

A. This is bad news, e.g. *Sinkhole swallows resort in Florida*
B. This is neutral news, e.g. *UN Climate Change Conference in Bonn*
C. This is good news, e.g. *Fresh water reserves found under ocean floor*

Conflict: are there two persons or groups in antagonism?

A. This depicts a conflict between two opposing persons/groups, e.g. *EPA fines Shell for Arctic Air Violations*
B. This does not depict a conflict between two opposing persons/groups, e.g. *Yarnell Hill Fire*

Extraordinary:[1] is this something out of the ordinary or rare, or is it something that normally happens?

A. This is an ordinary event, e.g. *Texas family to install solar panels*
B. This is an extraordinary event, e.g. *Epic Drought in West is Literally Moving Mountains*

Predictability: could a member of the public have known this was going to happen, or not?

A. A member of the public could not have know this will happen, e.g. *Google buys solar-powered drone maker Titan Aerospace*
B. A member of the public could have know this will happen, e.g. *People March for Climate in NYC*

Magnitude: does this event affects a large number of people, has important consequences, or is of global interest (high magnitude); or it is hyper-local with no or limited consequences and involves a low number of people (low magnitude)?

[1] This news value is referred to in the literature as "unexpectedness," but to avoid confusion we avoid the term in this paper as one common meaning of the word "unexpected" is "unpredictable," which can be confused with a different news value.

A. The magnitude of this event is high, e.g. *Air Pollution Linked to 1.2 Million Deaths in China*
B. The magnitude of this event is moderate, e.g. *California bans plastic bags, legislation signed*
C. The magnitude of this event is low, e.g. *FoxNews tells scientist not to talk about climate change*

Reference to elite persons: does it involve someone rich, powerful or famous?

A. This involves someone rich, powerful, or famous, e.g. *Obama's Climate Change plan in Congress*
B. This does not involve someone rich, powerful, or famous, e.g. *Species disappearing far faster than before*

A.5. Crisis Data Sets Characteristics

- Table A.3 containing the list of keywords used to collect data from each crisis.
- Tables A.4 and A.5 depicting temporal distributions of tweets on each crisis for each type, and for each source.

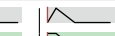
Table A.3.: List of keywords used to collect data for each of the crises in this study.

Year	Crisis Name	Keywords	#Kw
2012	Italy earthquakes	earthquake italy; quake italy; #modena; #sanfelice; san felice; modena terremoto; modena earthquake; modena quake; #norditalia; terremoto italia; #terremoto;	11
2012	Colorado wildfires	#cofire; #boulderfire; #colorado; #wildfire; #waldocanyonfire; #waldofire; #waldocanyon; colorado springs; #highparkfire; #flagstafffire; #littlesandfire; #treasurefire; #statelinefire; #springerfire; #lastchancefire; #fourmilefire; #4milefire; #fourmilecanyonfire; #boulderfire; #bisonfire; colorado wildfire; colorado wildfires; colorado fire; colorado fires; boulder fire; boulder fires; boulder wildfires; boulder wildfires;	28
2012	Philipinnes Floods	#rescueph; #reliefph; #floodsph; #prayforthephilippine; manila flood; manila floods; philippine floods; philippine flood; #floodph; #phalert; #habagat;	11
2012	Venezuela refinery explosion	paraguana refinery; venezuela refinery; paraguaná refinery; #paraguana; #paraguaná; amuay refinery; venezuelan refinery; #amuay; paraguaná refinería; paraguana refineria; amuay refineria; amuay refinería; #falcon; #falcón; refinería venezuela; refineria venezuela; refinería paraguana;	17
2012	Costa Rica earthquake	#temblorcr; #terremotocr; #costarica; #terremoto; costa rica quake; costa rica earthquake; costa rica temblor; costa rica terremoto; #creq; costa rican quake; costa rican earthquake; #quake; #earthquake;	13
2012	Guatemala earthquake	#sismo; #guatemala; tiemblaenguate; temblorgt; terremotogt; temblor guatemala; terremoto guatemala; sismo guatemala; earthquake guatemala; quake guatemala; #sanmarcos; #terremotoguatemala; #tremorguatemala;	13
2012	Typhoon Pablo	#pabloph; #reliefph; #bopha; #typhoonpablo; #typhoonbopha; typhoon bopha; typhoon pablo; #bopha; #pablo; #typhoon; #walangpasok; #mindanao; #visayas; #hinatuan; #rescueph; #pablosafetytips; #cdo; #strongerph;	18
2013	Brazil nightclub fire	#forçasantamaria; boate kiss; #boatekiss; #santamaria; #tragédiaemsm; #tragediaemsm; #todosdesejamforçasasantamaria; #brazilfire; #brazil fire; brazil nightclub; #brasildesejaforçasasvitimasdesantamaria; #prayforsantamaria; #prayforbrazil;	13
2013	Queensland Floods	#qldflood; #bigwet; queensland flood; australia flood; #qldfloods; queensland floods; australia floods; queensland flooding; qld flood; qld floods; qld flooding; australia flooding;	12
2013	Russian Meteor	#метеорит; #meteor; #meteorite; russia meteor; russian meteor; #russianmeteor; #chelyabinsk; #челябинск;	8
2013	Boston Bombings	boston explosion; boston explosions; boston blast; boston blasts; boston tragedies; boston tragedy; prayforboston; boston attack; boston attacks; boston terrorist; boston terrorists; boston tragic; bostonmarathon; boston marathon; boston explosive; boston bomb; boston bombing; dzhokhar; tsarnaev; marathon attack; marathon explosion; marathon explosions; marathon tragedies; marathon tragedy; marathon blasts; marathon blast; marathon attacks; marathon bomb; marathon bombing; marathon explosive;	30
2013	Savar building collapse	#savar; #bangladesh; bangladesh collapse; #ranaplaza; savar bangladesh; savar collapse; rana plaza;	7
2013	West Texas Explosion	#westexplosion; west explosion; waco explosion; texas explosion; texas fertilizer; prayfortexas; prayforwest; waco tx; west tx; west texas; waco texas; #west; #waco; westexplosion; west explosion; waco explosion; tx explosion; fertilizer explosion; prayfortexas; prayforwest; westtx; wacotx; west texas; waco texas; west tx; waco tx; texas fertilizer; west fertilizer; waco fertilizer;	29
2013	Alberta Floods	alberta flood; #abflood; canada flood; alberta flooding; alberta floods; canada flooding; canada floods; #yycflood; #yycfloods; #yycflooding; calgary flood; calgary flooding; calgary floods;	13
2013	Singapore Haze	#sghaze; singapore haze; #hazyday; blamethehaze; mustbehaze; #sg #haze; singapore #hazy;	7
2013	Lac-Megantic train crash	#lacmegantic; #lacmégantic; #lacmég; #lacmeg; #tragedielacmegantic; #tragédielacmégantic; #mégantic; lac mégantic; lac megantic; quebec train explosion; quebec train derailment; quebec train crash; quebec oil train; canada train oil; canada train oil; canadian train oil;	16
2013	Spain train crash	compostela train; spain train; tren compostela; españa tren; #santiagocompostela; #accidentesantiago;	6
2013	Manila Floods	baha manila; #maringph; #rescueph; #reliefph; #floodsph; #prayforthephilippine; manila flood; manila floods; philippine floods; philippine flood; #floodph; #phalert; #safenow; #trafficph; #habagat; #maring; #maringupdates;	17
2013	Colorado Floods	#cofloodrelief; colorado floods; colorado flooding; #coloradoflood; #coflood; #opcoflood; #boulderflood; #longmont;	8
2013	Australia wildfires	#nswfires; #nswbushfire; #nswbushfires; #nswrfs; #sydneybushfire; #sydneyfire; #sydneyfires; #sydneybushfires; nsw #bushfire; #redoctober; australia #bushfire; #faulconbridge; #nswrfs; #bushfire sydney; nsw fire; #prayforaustralia; #prayfornsw; australia fire; sydney fire; nsw fires; australia fires; sydney fires; prayfornsw;	23
2013	Bohol earthquake	#phquake; #pheq; #phtrenchquake; philippines earthquake; philippines quake; ph earthquake; ph quake; #phtrenchquake; #prayforthephilippines; #rescueph; #reliefph; #tabangbohol; #tabangcebu; #bohol; #cebu; prayforvisayas; prayforbohol; #lindol;	18
2013	Glasgow helicopter crash	#prayerforglasgow; #helicopter; glasgow helicopter; #clutha; helicopter crash;	5
2013	LA Airport Shootings	lax shooting; lax shootings; lax shooter; lax suspect; #laxshooting; lax airport; #lax; airport shooting; airport shootings; #losangeles airport; lax victims;	11
2013	NYC train crash	#newyork derailment; ny derailment; nyc derailment; #metronorth derailment; #spuyten duyvil; #nyctrain; new york derailment; metro north derailment; #metronorth derailment; ny train crash; nyc train crash; newyork train crash; york train crash; #metronorth train crash; metro north crash; ny train derailed; york train derailed; nyc train derailed;	18
2013	Sardinia Floods	sardinia floods; sardinia flooding; cyclone cleopatra; #cyclonecleopatra; #sardinia; sardegna alluvione; #cleopatra alluvione; #sardegna;	8
2013	Typhoon Yolanda	#typhoonyolanda; #yolandaph; #yolanda; #haiyan; #tracingph; #floodph; #safenow; #rescueph; #reliefph; typhoon yolanda; typhoon haiyan; typhoon philippines; #typhoonhaiyan; #typhoonaid; #philippines; #typhoon; #supertyphoon; #redcrossphilippines; #yolandaactionweekend; rescue ph; typhoon ph; super typhoon;	22

Table A.4.: Temporal distribution of tweets across information sources (top) and types (bottom) for the progressive events we analyzed. The 3 most frequent sources, respectively, information types, per crisis are highlighted in green. The red vertical bar indicates the peak volume for all tweets related to each event. (Best seen in color.)

Name	Alberta floods'13	Australia wildfires'13	Colorado floods'13	Colorado wildfires'12	Manila floods'13	Philipinnes floods'12	Queensland floods'13	Sardinia floods'13	Singapore haze'13	Typhoon Pablo'12	Typhoon Yolanda'13
Overall											
Eyewitness											
Government											
NGOs											
Business											
Media											
Outsiders											
Affected Ind.											
Infrast. & Utilities											
Donat. & Volun.											
Caution & Advice											
Sympathy											
Other Useful Info.											

Table A.5.: Temporal distribution of tweets across information sources (top) and types (bottom) for the instantaneous events we analyzed. The 3 most frequent sources, respectively, information types, per crisis are highlighted in green. The red vertical bar indicates the peak volume for all tweets related to each event. (Best seen in color.)

Name	Bohol earthq.'13	Boston bombings'13	Brazil nightclub fire'13	Costa Rica earthq.'12	Glasgow helicopter crash'13	Guatemala earthq.'12	Italy earthq.'12	LA airport shootings'13	Lac-Megantic train crash'13	NYC train crash'13	Russian meteor'13	Savar building collapse'13	Spain train crash'13	Venezuela refinery explosion'12	West Texas explosion'13
Overall															
Eyewitness															
Government															
NGOs															
Business															
Media															
Outsiders															
Affected Ind.															
Infrast. & Utilities															
Donat. & Volun.															
Caution & Advice															
Sympathy															
Other Useful Info.															

Bibliography

[1] Black lives matter; a tale of two covers. [14-april-2015]. http://www.nbcnews.com/news/nbcblk/black-lives-matter-tale-two-covers-n339796.

[2] Can we talk about race? a few rules of engagement. [14-april-2015]. http://articles.baltimoresun.com/2006-08-01/news/0608010135_1_racial-inequality-political-change-problem-of-racial.

[3] Gender equality won't happen unless men speak up. [14-april-2015]. http://edition.cnn.com/2013/04/17/business/sandberg-gender-equality/.

[4] Seven signs you are clueless about income inequality. [14-april-2015]. http://fortune.com/2015/03/20/anand-giridharadas-ted-inequality/.

[5] What's missing from the debate about women leaders in the nhs? men. [14-april-2015]. http://www.theguardian.com/healthcare-network/2014/jan/08/female-managers-gender-equality-nhs.

[6] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through Twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206. ACM, 2015.

[7] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3):392–402, 2011.

[8] David S Adams. Policies, programs, and problems of the local Red Cross disaster relief in the 1960s. Technical report, Univ. of Delaware, Disaster Research Center, 1970.

[9] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 2005.

[10] Gediminas Adomavicius and Jingjing Zhang. Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems (TMIS)*, 3(1):3, 2012.

[11] Gediminas Adomavicius and Jingjing Zhang. Stability of recommendation algorithms. *ACM*

Transactions on Information Systems (TOIS), 30(4):23, 2012.

[12] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–26. ACM, 2006.

[13] James Allan. Incremental relevance feedback for information filtering. In *Proceedings of the 19th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 270–278. ACM, 1996.

[14] Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 897–908. ACM, 2013.

[15] Xavier Amatriain. Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, 14(2):37–48, 2013.

[16] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 16(07), 2008.

[17] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, 2011.

[18] Anne Archambault and Jonathan Grudin. A longitudinal study of Facebook, Linkedin, & Twitter use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2741–2750. ACM, 2012.

[19] Jaime Arguello, Fernando Diaz, and Jean-François Paiement. Vertical selection in the presence of unlabeled verticals. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 691–698. ACM, 2010.

[20] Bryan Arva, John Beieler, Bejamin Fisher, Gustavo Lara, Philip A Schrodt, Wonjun Song, Marsha Sowell, and Sam Stehle. Improving forecasts of international events of interest. In *EPSA 2013 Annual General Conference Paper*, volume 78, 2013.

[21] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. Tweedr: Mining Twitter to inform disaster response. *Proceedings of the 11th International Conference on Information Cystems for Crisis Response and Management*, 2014.

[22] Sitaram Asur, Bernardo Huberman, et al. Predicting the future with social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 492–499. IEEE, 2010.

[23] Anne Aula, Rehan M Khan, and Zhiwei Guan. How does search behavior change as search be-

comes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 35–44. ACM, 2010.

[24] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 635–644. ACM, 2011.

[25] Ricardo Baeza-Yates. Wisdom of crowds or wisdom of a few? *Web Engineering*, page 573.

[26] Ricardo Baeza-Yates and Yoelle Maarek. Usage data in web search: benefits and limitations. In *Scientific and Statistical Database Management*, pages 495–506. Springer, 2012.

[27] Ricardo A Baeza-Yates. Big data or right data? In *AMW*, 2013.

[28] Mossaab Bagdouri and Douglas W Oard. On predicting deletions of microblog posts. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1707–1710. ACM, 2015.

[29] Paul Baker. Representations of Islam in British broadsheet and tabloid newspapers 1999–2005. *Journal of Language and Politics*, 9(2):310–338, 2010.

[30] Eytan Bakshy, Brian Karrer, and Lada A Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, pages 325–334. ACM, 2009.

[31] Reuven Bar-Yehuda and Shimon Even. A local-ratio theorem for approximating the weighted vertex cover problem. *North-Holland Mathematics Studies*, 109:27–45, 1985.

[32] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Social Science Research Network Working Paper Series*, 2014.

[33] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[34] Alejandro Bellogín, Iván Cantador, Fernando Díez, Pablo Castells, and Enrique Chavarriaga. An empirical comparison of social, collaborative filtering, and hybrid recommenders. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):14, 2013.

[35] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam Conference (CEAS)*, volume 6, page 12, 2010.

[36] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[37] Pablo J Boczkowski. The divergent online news preferences of journalists and readers. *Communications of the ACM*, 53(11):24–25, 2010.

[38] Tom Boellstorff. Making big data, in theory. *First Monday*, 18(10), 2013.

[39] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, 2012.

[40] Maxwell T Boykoff and S Ravi Rajan. Signals and noise. *EMBO reports*, 8(3):207–211, 2007.

[41] Axel Bruns. Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday*, 18(10), 2013.

[42] Axel Bruns, Jean E Burgess, Kate Crawford, and Frances Shaw. #qldfloods and @qpsmedia: Crisis communication on Twitter in the 2011 South East Queensland floods. Technical report, ARC Centre, Queensland Univ. of Technology, 2012.

[43] Axel Bruns and Yuxian Eugene Liang. Tools and methods for capturing Twitter data during natural disasters. *First Monday*, 17(4), 2012.

[44] Axel Bruns and Stefan Stieglitz. Twitter data: what do they represent? *it-Information Technology*, 56(5):240–245, 2014.

[45] Robin Burke. Integrating knowledge-based and collaborative-filtering recommender systems. In *Proceedings of the Workshop on AI and Electronic Commerce*, pages 69–72, 1999.

[46] Sam Burnett and Nick Feamster. Encore: Lightweight measurement of web censorship with cross-origin requests. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. ACM, 2015.

[47] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–250. ACM, 2008.

[48] Cornelia Caragea, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H Tapia, Lee Giles, Bernard J Jansen, et al. Classifying text messages for the Haiti earthquake. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and management*, 2011.

[49] Lowell Juilliard Carr. Disaster and the sequence-pattern concept of social change. *American Journal of Sociology*, pages 207–218, 1932.

[50] Andy Carvin. *Distant Witness*. CUNY Journalism Press, 2013.

[51] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 211–223. ACM, 2014.

[52] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684, 2011.

[53] Pew Research Center. Demographics of key social networking platforms. [19-sept-2015]. http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/.

[54] Pew Research Center. The demographics of social media users. [19-sept-2015]. http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/.

[55] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, everybody Tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 1–14, New York, NY, USA, 2007. ACM.

[56] Wei Chen, Wynne Hsu, and Mong Li Lee. Making recommendations from multiple domains. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 892–900. ACM, 2013.

[57] Sophie Chou. Race and the machine: Re-examining race and ethnicity in data mining. 2015. http://www.sophiechou.com/papers/chou_racepaper.pdf.

[58] Sophie Chou, William Li, and Ramesh Sridharan. Democratizing data science. 2014.

[59] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[60] Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It not easy! In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[61] William W Cohen. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123, 1995.

[62] Kevyn Collins-Thompson and Jamie Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 303–310. ACM, 2007.

[63] Scott Counts, Munmun De Choudhury, Jana Diesner, Eric Gilbert, Marta Gonzalez, Brian Keegan, Mor Naaman, and Hanna Wallach. Computational social science: CSCW in the social media era. In *Proceedings of the companion publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 105–108. ACM, 2014.

[64] Kate Crawford and Megan Finn. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, pages 1–12, 2014.

[65] Kate Crawford and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College. Law School. Boston College Law Review*, 55(1):93, 2014.

[66] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM Conference on Recommender systems*, pages 39–46. ACM, 2010.

[67] W Bruce Croft and David J Harper. Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35(4):285–295, 1979.

[68] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github: Transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1277–1286. ACM, 2012.

[69] Sauvik Das and Adam Kramer. Self-censorship on Facebook. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[70] Luis M De Campos, Juan M Fernández-Luna, Juan F Huete, and Miguel A Rueda-Morales. Measuring predictive capability in collaborative filtering. In *Proceedings of the third ACM Conference on Recommender systems*, pages 313–316. ACM, 2009.

[71] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1431–1442. ACM, 2013.

[72] Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. Unfolding the event landscape on Twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 241–244. ACM, 2012.

[73] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[74] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1365–1376. ACM, 2014.

[75] Arnaud De Myttenaere, Bénédicte Le Grand, Boris Golden, and Fabrice Rossi. Reducing offline evaluation bias in recommendation systems. *arXiv preprint arXiv:1407.0822*, 2014.

[76] Claes H De Vreese. News framing: Theory and typology. *Information design journal+ document design*, 13(1):51–62, 2005.

[77] Sebastian Denef, Petra S Bayerl, and Nico A Kaptein. Social media and the police: tweeting practices of british police forces during the August 2011 riots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3471–3480. ACM, 2013.

[78] Christian Desrosiers and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook*, pages 107–144. Springer, 2011.

[79] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.

[80] Nicholas Diakopoulos, Dag Elgesem, Andrew Salway, Amy Zhang, and Knut Hofland. Compare clouds: Visualizing text corpora to compare media frames. In *Proceedings of IUI Workshop on Visual Text Analytics*, 2015.

[81] Nicholas Diakopoulos, Amy X Zhang, Dag Elgesem, and Andrew Salway. Identifying and analyzing moral evaluation frames in climate change blog discourse. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[82] Nicholas Diakopoulos, Amy X. Zhang, and Andrew Salway. Visual analytics of media frames in online news and blogs. In *Proceedings of IEEE InfoVis Workshop on Text Visualization*, 2013.

[83] Fernando Diaz. Experimentation standards for crisis informatics. *SIGIR Forum*, 48(2):22–30, 2014.

[84] Fernando Diaz, Michael Gamon, Jake Hofman, Emre Kiciman, and David Rothschild. Online and social media data as an imperfect continuous panel survey. *Unpublished manuscript, Microsoft Research*, 2015.

[85] Anton Dimitrov, Alexandra Olteanu, Luke Mcdowell, and Karl Aberer. Topick: Accurate topic distillation for user streams. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*, pages 882–885. IEEE Computer Society, 2012.

[86] Simon D Donner and Jeremy McDaniels. The influence of national temperature fluctuations on opinions about climate change in the US since 1990. *Climatic change*, 2013.

[87] Virgile Landeiro Dos Reis and Aron Culotta. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[88] Chris Drummond. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop*, 2009.

[89] Susan Dumais, Robin Jeffries, Daniel M Russell, Diane Tang, and Jaime Teevan. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*, pages 349–372. Springer,

2014.

[90] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[91] Cynthia Dwork and Deirdre K Mulligan. It's not privacy, and it's not fair. *Stanford Law Review Online*, 66:35, 2013.

[92] Miles Efron, Jana Deisner, Peter Organisciak, Garrick Sherman, and Ana Lucic. The Univ. of Illinois Graduate School of Library and Information Science at TREC 2012. In *Proc. of TREC*, 2012.

[93] Kate Ehrlich and N Sadat Shami. Microblogging inside and outside the workplace. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[94] Michael Ekstrand and John Riedl. When recommenders fail: predicting recommender failure for algorithm selection and combination. In *Proceedings of the sixth ACM Conference on Recommender systems*, pages 233–236. ACM, 2012.

[95] Dag Elgesem, Lubos Steskal, and Nicholas Diakopoulos. Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication*, 9(2):169–188, 2015.

[96] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2):415–441, 2006.

[97] Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173, 2007.

[98] Frank Esser and Thomas Hanitzsch, editors. *The Handbook of Comparative Communication Research*. ICA Handbooks. Routledge, April 2012.

[99] Henry W. Fischer. *Response to disaster: fact versus fiction & its perpetuation—the sociology of disaster*. Univ. Press of America, 1998.

[100] National Commission for the Proptection of Human Subjects of Biomedical and MD. Behavioral Research, Bethesda. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. ERIC Clearinghouse, 1978.

[101] Adam Fourney, Ryen W White, and Eric Horvitz. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 737–746. ACM, 2015.

[102] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *Knowledge and data engineering, ieee transactions on*, 19(3):355–369, 2007.

[103] Julia D. Fraustino, Brooke Liu, and Yan Jin. Social media use during disasters: A review of the knowledge base and gaps. Technical report, Science and Technology Directorate, U.S. Department of Homeland Security, 2012.

[104] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, (3):10–14, 2011.

[105] Ruth García-Gavilanes, Yelena Mejova, and Daniele Quercia. Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1511–1522. ACM, 2014.

[106] Venkata Rama Kiran Garimella, Ingmar Weber, and Sonya Dal Cin. From "i love you babe" to "leave me alone"-romantic relationship breakups on twitter. In *Social Informatics*, pages 199–215. Springer, 2014.

[107] Daniel Gayo-Avello. ï wanted to predict elections with Twitter and all i got was this lousy paper¨-a balanced survey on election prediction using Twitter data. *arXiv preprint arXiv:1204.6441*, 2012.

[108] Daniel Gayo-Avello. No, you cannot predict elections with twitter. *Internet Computing, IEEE*, 16(6):91–94, 2012.

[109] Daniel Gayo Avello, Panagiotis T Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[110] Eni Mustafaraj Markus Strohmaier Harald Schoen Gayo-Avello, Panagiotis Takis Metaxas, Daniel Peter Gloor, Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, and Peter Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

[111] Carolin Gerlitz and Bernhard Rieder. Mining one percent of Twitter: collections, baselines, sampling. *M/C Journal*, 16(2), 2013.

[112] Paolo Giardullo. Does 'bigger' mean 'better'? pitfalls and shortcuts associated with big data for social research. *Quality & Quantity*, pages 1–19, 2015.

[113] Jim Giles. Making the links. *Nature*, 488(7412):448–450, 2012.

[114] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*,

457(7232):1012–1014, 2009.

[115] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan Watts. The structural virality of online diffusion. *Management Science*, 2015.

[116] Nadav Golbandi, Yehuda Koren, and Ronny Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 595–604. ACM, 2011.

[117] J. Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland, 2005.

[118] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[119] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[120] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in communication networks sampled from Twitter. *Social Networks*, 38:16–27, July 2014.

[121] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27, 2014.

[122] Daniel L Goroff. Balancing privacy versus accuracy in research protocols. *Science*, 347(6221):479–480, 2015.

[123] Rebecca Gray, Nicole B Ellison, Jessica Vitak, and Cliff Lampe. Who wants to know?: question-asking and answering practices among Facebook users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1213–1224. ACM, 2013.

[124] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the Electronic Society*, pages 71–80. ACM, 2005.

[125] Tom Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web semantics: science, services and agents on the World Wide Web*, 6(1):4–13, 2008.

[126] Pedro Calais Guerra, Wagner Meira Jr, and Claire Cardie. Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pages 443–452. ACM, 2014.

[127] Qi Guo, Fernando Diaz, and Elad Yom-Tov. Updating users about time critical events. In *Ad-*

vances in Information Retrieval, pages 483–494. Springer, 2013.

[128] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 305–318. ACM, 2014.

[129] Tony Harcup and Deirdre O'neill. What is news? Galtung and Ruge revisited. *Journalism studies*, 2(2):261–280, 2001.

[130] Tim Harford. Big data: A big mistake? *Significance*, 11(5):14–19, 2014.

[131] Eszter Hargittai. Whose space? differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*, 13(1):276–297, 2007.

[132] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237. ACM, 1999.

[133] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[134] Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. Share, like, recommend: Decoding the social media news consumer. *Journalism Studies*, 13(5-6):815–824, 2012.

[135] Luis E Hestres. Preaching to the choir: Internet-mediated advocacy, issue public mobilization, and climate change. *New Media & Society*, 2013.

[136] Kashmir Hill. Facebook added 'research' to user agreement 4 months after emotion manipulation study. *Tech*, 2014.

[137] Mary Hodder. Why amazon didn't just have a glitch. *TechCrunch Blog*, 2009.

[138] Lichan Hong, Gregorio Convertino, and Ed H Chi. Language matters in Twitter: A large scale study. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2011.

[139] Daniel Hoornweg. *Cities and climate change: responding to an urgent agenda*. World Bank Publications, 2011.

[140] Eric Horvitz and Deirdre Mulligan. Data, privacy, and the greater good. *Science*, 349(6245):253–255, 2015.

[141] William Housley, Rob Procter, Adam Edwards, Peter Burnap, Matthew Williams, Luke Sloan, Omer Rana, Jeffrey Morgan, Alex Voss, and Anita Greenhill. Big and broad social data and the

sociological imagination: a collaborative response. *Big Data & Society*, 1(2), 2014.

[142] Dirk Hovy, Barbara Plank, and Anders Søgaard. When POS datasets don't add up: Combatting sample bias. *Proceedings of Conference on Language Resources and Evaluation (LREC)*, 2014.

[143] Zhicong Huang, Alexandra Olteanu, and Karl Aberer. Credibleweb: a platform for web credibility evaluation. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1887–1892. ACM, 2013.

[144] Amanda L. Hughes. Participatory design for the social media needs of emergency public information officers. In *Proceedings of Information Systems for Crisis Response and management*, 2014.

[145] Amanda L Hughes, Lise AA St Denis, Leysia Palen, and Kenneth M Anderson. Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1505–1514. ACM, 2014.

[146] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *Proceedings of Information Systems for Crisis Response and Management*, 2009.

[147] David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[148] Luke Hutton and Tristan Henderson. " i didn't sign up for this!": Informed consent in social network research. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

[149] Luke Hutton and Tristan Henderson. Towards reproducibility in online social network research. *IEEE Transactions on Emerging Topics in Computing*, 2015.

[150] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: a survey. *ACM Computing Surveys (CSUR)*, 47(4):67, 2015.

[151] Muhammad Imran, Carlos Castillo, Ji Lucas, M Patrick, and Jakob Rogstadius. Coordinating human and machine intelligence to classify microblog communications in crises. In *Proceedings of Information Systems for Crisis Response and management (ISCRAM)*, 2014.

[152] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 1021–1024, 2013.

[153] Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *Proceedings of Information Systems for Crisis Response and Management*, 2013.

[154] Mohsen Jamali and Martin Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 397–406. ACM, 2009.

[155] Mohsen Jamali and Martin Ester. Using a trust network to improve top-n recommendation. In *Proceedings of the third ACM Conference on Recommender systems*, pages 181–188. ACM, 2009.

[156] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM Conference on Recommender systems*, pages 135–142. ACM, 2010.

[157] J Jashinsky, SH Burton, CL Hanson, J West, C Giraud-Carrier, MD Barnes, and T Argyle. Tracking suicide risk factors through twitter in the us. *Crisis*, 35(1):51, 2014.

[158] Wei Jin, Hung Hay Ho, and Rohini K Srihari. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1195–1204. ACM, 2009.

[159] Adam N Joinson. Looking at, looking up or keeping up with people?: motives and use of Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1027–1036. ACM, 2008.

[160] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1% s don't make a whole: Comparing simultaneous samples from Twitter's streaming api. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 75–83. Springer, 2014.

[161] David Jurgens, Tyler Finethy, Caitrin Armstrong, and Derek Ruths. Everyone's invited: A new paradigm for evaluation on non-transferable datasets. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[162] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[163] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *EMNLP-CoNLL*, pages 1075–1083, 2007.

[164] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 667–678, 2013.

[165] Nattiya Kanhabua and Wolfgang Nejdl. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 1335–1342, 2013.

[166] Renato Kempter, Valentina Sintsova, Claudiu Musat, and Pearl Pu. EmotionWatch: Visualizing fine-grained emotions in event-related tweets. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[167] Erin Kenneally. How to throw the race to the bottom: revisiting signals for ethical and legal research using online data. *ACM SIGCAS Computers and Society*, 45(1):4–10, 2015.

[168] Anne-Marie Kermarrec, Vincent Leroy, Afshin Moin, and Christopher Thraves. Application of random walks to decentralized recommender systems. In *Principles of Distributed Systems*, pages 48–63. Springer, 2010.

[169] Norbert L Kerr. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998.

[170] Emre Kıcıman. OMG, i have to tweet that! a study of factors that influence tweet rates. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[171] Emre Kıcıman. Towards learning a knowledge base of actions from experiential microblogs. In *AAAI Spring Symposium on Knowledge Representation and Reasoning*. AAAI, 2015.

[172] Emre Kıcıman, Scott Counts, Michael Gamon, Munmun De Choudhury, and Bo Thiesson. Discussion graphs: Putting social media analysis in context. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[173] Emre Kıcıman and Matthew Richardson. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 547–556. ACM, 2015.

[174] Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 243–248. ACM, 2014.

[175] Gary King. Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721, 2011.

[176] Lauren Kirchner. When discrimination is baked into algorithms. *The Atlantic*, 2015.

[177] Andrei P Kirilenko, Tatiana Molodtsova, and Svetlana O Stepchenkova. People as sensors: Mass media and local temperature influence climate change discussion on Twitter. *Global Environmental Change*, 30:92–100, 2015.

[178] Andrei P Kirilenko and Svetlana O Stepchenkova. Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change*, 26:171–182, 2014.

[179] Jon Kleinberg. Small-world phenomena and the dynamics of information. *Advances in neural information processing systems*, 1:431–438, 2002.

[180] Joseph Konstan, John Riedl, et al. Recommended for you. *Spectrum, IEEE*, 49(10):54–61, 2012.

[181] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and development in information Research and Development in Information Retrievall*, pages 195–202. ACM, 2009.

[182] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[183] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[184] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 85–94. ACM, 2015.

[185] Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. Whom should i follow?: identifying relevant users during crises. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 139–147. ACM, 2013.

[186] Jim A. Kuypers. *Framing Analysis*, pages 181+. Lexington Press, 2009.

[187] Haewoon Kwak and Jisun An. A first look at global news coverage of disasters by using the gdelt dataset. In *Social Informatics*, pages 300–308. Springer, 2014.

[188] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World wide web*, pages 591–600. ACM, 2010.

[189] Sang Jib Kwon, Eunil Park, and Ki Joon Kim. What drives successful social networking services? a comparative analysis of user acceptance of Facebook and Twitter. *The Social Science Journal*, 51(4):534–544, 2014.

[190] Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.

[191] Cliff Lampe, Nicole B Ellison, and Charles Steinfield. Changes in use and perception of Facebook. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 721–730. ACM, 2008.

[192] Carsten Lanquillon and Ingrid Renz. Adaptive information filtering: Detecting changes in text streams. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 538–544. ACM, 1999.

[193] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google flu: traps in big data analysis. *Science*, 343(14 March), 2014.

[194] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[195] Alex Leavitt and Joshua A Clark. Upvoting hurricane Sandy: event-based news production processes on a social news site. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1495–1504. ACM, 2014.

[196] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5), 2013.

[197] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web*, pages 251–260. ACM, 2012.

[198] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and Twitter social networks. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2010.

[199] Kristina Lerman and Tad Hogg. Leveraging position bias to improve peer recommendation. *PLoS ONE*, 9(6), 2014.

[200] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506. ACM, 2009.

[201] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. of Experimental Social Psychology*, 2013.

[202] Linna Li, Michael F Goodchild, and Bo Xu. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40(2):61–77, 2013.

[203] Hai Liang and King-wa Fu. Testing propositions derived from twitter studies: Generalization and replication in computational social science. *PloS one*, 10(8):e0134270, 2015.

[204] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal*

of the American society for information science and technology, 2007.

[205] Bang Hui Lim, Dongyuan Lu, Tao Chen, and Min-Yen Kan. # mytweet via instagram: Exploring user behaviour across multiple social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 113–120. ACM, 2015.

[206] Kuan-Yu Lin and Hsi-Peng Lu. Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in Human Behavior*, 27(3):1152–1161, 2011.

[207] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[208] László Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

[209] Hao Ma, Irwin King, and Michael R Lyu. Learning to recommend with social trust ensemble. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 203–210. ACM, 2009.

[210] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 931–940. ACM, 2008.

[211] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. Recommender systems with social regularization. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 287–296. ACM, 2011.

[212] Jim Maddock, Kate Starbird, and Robert Mason. Using historical Twitter data for research: Ethical challenges of tweet deletions. In *Proc. of CSCW 5 Workshop on Ethics at the 2015 Conference on Computer Supported Cooperative Work*, 2015.

[213] Walid Magdy and Tamer Elsayed. Adaptive method for following dynamic topics on Twitter. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[214] Annette Markham and Elizabeth Buchanan. Ethical decision-making and internet research: Version 2.0. *AOIR Executive Committee*, 2012.

[215] Paolo Massa and Paolo Avesani. Trust-aware bootstrapping of recommender systems. In *Proceedings of ECAI Workshop on Recommender Systems*, volume 28, page 29, 2006.

[216] Paolo Massa and Paolo Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender systems*, pages 17–24. ACM, 2007.

[217] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, pages 897–908, 2013.

[218] Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. On building a reusable Twitter corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1114. ACM, 2012.

[219] Patrick Meier. *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*. CRC Press, 2014.

[220] Yelena Mejova, Ingmar Weber, and Michael W Macy. *Twitter: A Digital Socioscope*. Cambridge University Press, 2015.

[221] Panagiotis Metaxas and Eni Mustafaraj. The rise and the fall of a citizen reporter. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 248–257. ACM, 2013.

[222] Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–655. ACL, 2012.

[223] Simon Meyffret, Lionel Médini, and Frédérique Laforest. Trust-based local and social recommendation. In *Proceedings of the 4th ACM RecSys Workshop on Recommender Systems and the Social Web*, pages 53–60. ACM, 2012.

[224] Stanley Milgram. The small world problem. *Psychology today*, 1967.

[225] Claire Cain Miller. Algorithms and bias: Q. and a. with Cynthia Dwork. *Hidden in the data*, 2015.

[226] Claire Cain Miller. When algorithms discriminate. *Hidden Bias*, 2015.

[227] Tehila Minkus, Kelvin Liu, and Keith W Ross. Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the 24th International Conference on World Wide Web*, pages 776–786, 2015.

[228] Alan Mislove, Sune Lehmnan, Yong-Yeol Ahn Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 554–557. AAAI Press, 2011.

[229] Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Improving pseudo-relevance feedback via tweet selection. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 439–448. ACM, 2013.

[230] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro

Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4), 2013.

[231] T Molodtsova, A Kirilenko, and S Stepchenkova. Utilizing the social media data to validate'climate change'indices. In *AGU Fall Meeting Abstracts*, 2013.

[232] Andrés Monroy-Hernández, Emre Kiciman, Munmun De Choudhury, Scott Counts, et al. The new war correspondents: The rise of civic media curation in urban warfare. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1443–1452. ACM, 2013.

[233] Michael Moricz, Yerbolat Dosbayev, and Mikhail Berlyant. Pymk: friend recommendation at myspace. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 999–1002. ACM, 2010.

[234] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. When is it biased?: assessing the representativeness of Twitter's streaming api. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 555–556, 2014.

[235] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from Twitter's streaming API with Twitter's Firehose. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2013.

[236] Robert Munro and Christopher D Manning. Short message communications: users, topics, and in-language processing. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, page 4. ACM, 2012.

[237] Arvind Narayanan and Bendert Zevenbergen. No Encore for Encore? ethical questions for web-based censorship measurement. *Ethical Questions for Web-Based Censorship Measurement (September 24, 2015)*, 2015.

[238] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153. IEEE, 2012.

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

[240] UN OCHA. World humanitarian data and trends. Technical report, December 2013.

[241] UN OCHA. Hashtag standards for emergencies. Technical report, November 2014.

[242] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of International AAAI Conference on Weblogs and Social Media*, 11(122-129):1–2, 2010.

[243] Hüseyin Oktay, Brian J Taylor, and David D Jensen. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, pages 1–9. ACM, 2010.

[244] Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, and Karl Aberer. Comparing events coverage in online news and social media: The case of climate change. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[245] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2014.

[246] Alexandra Olteanu, Anne-Marie Kermarrec, and Karl Aberer. Comparing the predictive capability of social and interest affinity for recommendations. In *Web Information Systems Engineering*, pages 276–292. Springer, 2014.

[247] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web credibility: features exploration and credibility prediction. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, pages 557–568. Springer-Verlag, 2013.

[248] Alexandra Olteanu and Guillaume Pierre. Towards robust and scalable peer-to-peer social networks. In *Proceedings of the Fifth Workshop on Social Network Systems*, SNS '12, pages 10:1–10:6, New York, NY, USA, 2012. ACM.

[249] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 994–1009. ACM, 2015.

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

[251] Raphael Ottoni, Diego Las Casas, João Paulo Pesce, Wagner Meira Jr, Christo Wilson, Alan Mislove, and Virgilio Almeida. Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[252] Leysia Palen and Sophia B Liu. Citizen communications in crisis: anticipating a future of ICT-supported public participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 727–736. ACM, 2007.

[253] Zizi Papacharissi and Maria de Fatima Oliveira. News frames terrorism: A comparative analysis of frames employed in terrorism coverage in US and UK newspapers. *The International Journal of Press/Politics*, 13(1):52–74, 2008.

[254] Warren Pearce, Kim Holmberg, Iina Hellsten, and Brigitte Nerlich. Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 report. *PloS one*, 9(4), 2014.

[255] Ronald W Perry and Enrico Louis Quarantelli. *What is a disaster?: New answers to old questions*. Xlibris Corporation, 2005.

[256] Georgios Pitsilis and Svein J. Knapskog. Social trust as a solution to address sparsity-inherent problems of recommender systems. In *Recommender Systems and the Social Web*, 2009.

[257] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. Do all birds tweet the same?: characterizing Twitter around the world. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1025–1030. ACM, 2011.

[258] Lindsay Poirier. Reforming mis-care in big data analysis. 2015.

[259] Liza Potts, Joyce Seitzinger, Dave Jones, and Angela Harrison. Tweeting disaster: hashtag constructions and collisions. In *Proceedings of the 29th ACM International Conference on Design of Communication*, pages 235–240. ACM, 2011.

[260] Andrew J Prelog. *Social Change and Disaster Annotated Bibliography*. PhD thesis, Department of Economics, Colorado State Univ., 2010.

[261] Pearl Pu, Li Chen, and Rong Hu. Evaluating recommender systems from the user perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4-5):317–355, 2012.

[262] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1), 2013.

[263] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 25–34. ACM, 2011.

[264] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. Smelly maps: The digital life of urban smellscapes. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[265] Kira Radinsky, Krysta Svore, Susan Dumais, Jaime Teevan, Alex Bocharov, and Eric Horvitz. Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st International Conference on World Wide Web*, pages 599–608. ACM, 2012.

[266] Filip Radlinski, Paul N Bennett, and Emine Yilmaz. Detecting duplicate web documents using clickthrough data. In *Proceedings of the fourth ACM International Conference on Web Search*

and Data Mining, pages 147–156. ACM, 2011.

[267] Christian Reuter and Simon Scholl. Technical limitations for designing applications for social media. In *Mensch & Computer*, pages 131–140, 2014.

[268] Matthew Richardson. Learning about the world through long-term query logs. *ACM Transactions on the Web (TWEB)*, 2(4):21, 2008.

[269] Simone Rödder and Mike S Schäfer. Repercussion and resistance. An empirical study on the interrelation between science and mass media. *Communications*, 35(3):249–267, 2010.

[270] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of International Conference on World Wide Web*, 2011.

[271] Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. Representation and communication: Challenges in interpreting large social media datasets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 357–362. ACM, 2013.

[272] Eduardo Ruiz, Vagelis Hristidis, and Panagiotis G Ipeirotis. Efficient filtering on hidden document streams. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*, 2014.

[273] Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, pages 513–522. ACM, 2012.

[274] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[275] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geotagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[276] Justin Sampson, Fred Morstatter, Ross Maciejewski, and Huan Liu. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 237–245. ACM, 2015.

[277] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295, 2001.

[278] Andreas Schmidt, Ana Ivanova, and Mike S Schäfer. Media attention for climate change around the world. *Global Environ. Change*, 2013.

[279] H Andrew Schwartz, Greg Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, and Mar-

garet Kern. Extracting human temporal orientation in Facebook language. In *Proceedings of the The 2015 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies, NAACL*, 2015.

[280] Alexandra Segerberg and W Lance Bennett. Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests. *The Communication Review*, 14(3):197–215, 2011.

[281] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[282] Frances Shaw, Jean Burgess, Kate Crawford, Axel Bruns, et al. Sharing news, making sense, saying thanks: Patterns of talk on Twitter during the Queensland floods. *Australian Journal of Communication*, 40(1):23, 2013.

[283] Yelong Shen and Ruoming Jin. Learning personal+ social latent factor model for social recommendation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1303–1311. ACM, 2012.

[284] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Alan Gary Smith, and Wenbo Wang. Chapter title: Twitris-a system for collective social intelligence. *Encyclopedia of Social Network Analysis and Mining*, 2014.

[285] Galit Shmueli. To explain or to predict? *Statistical science*, pages 289–310, 2010.

[286] Fabrizio Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1):1–174, 2010.

[287] Meredith M Skeels and Jonathan Grudin. When social networks cross boundaries: a case study of workplace use of Facebook and Linkedin. In *Proceedings of the ACM 2009 International Conference on Supporting group work*, pages 95–104. ACM, 2009.

[288] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. ACL, 2008.

[289] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *The semantic web: research and applications*, pages 624–639. Springer, 2007.

[290] Nirupama Dharmavaram Sreenivasan, Chei Sian Lee, and Dion Hoe-Lian Goh. Tweet me home: Exploring information use on Twitter in crisis situations. In *Online Communities and Social Computing*, pages 120–129. Springer, 2011.

[291] Kate Starbird, Grace Muzny, and Leysia Palen. Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions. In *Proceedings of*

Information Systems for Crisis Response and Management, 2012.

[292] Kate Starbird and Leysia Palen. Pass it on?: Retweeting in mass emergency. In *Proceedings of Information Systems for Crisis Response and Management*, 2010.

[293] Kate Starbird and Leysia Palen. Voluntweeters: Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1071–1080. ACM, 2011.

[294] Kate Starbird and Leysia Palen. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the acm 2012 Conference on computer supported cooperative work*, pages 7–16. ACM, 2012.

[295] Kate Starbird, Leysia Palen, Amanda L Hughes, and Sarah Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pages 241–250. ACM, 2010.

[296] Harald Steck. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM Conference on Recommender Systems*, pages 125–132. ACM, 2011.

[297] James G. Stovall. *Journalism: who, what, when, where, why and how*. Pearson, November 2004.

[298] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

[299] Melanie Swan. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2):85–99, 2013.

[300] Edson C Tandoc Jr. Why web analytics click: Factors affecting the ways journalists use audience metrics. *Journalism Studies*, (ahead-of-print):1–18, 2014.

[301] Jiliang Tang, Huiji Gao, and Huan Liu. mTrust: discerning multi-faceted trust in a connected world. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, pages 93–102. ACM, 2012.

[302] Jiliang Tang, Huiji Gao, Huan Liu, and Atish Das Sarma. eTrust: Understanding trust evolution in an online world. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 253–261. ACM, 2012.

[303] Robert Endre Tarjan and Anthony E Trojanowski. Finding a maximum independent set. *SIAM Journal on Computing*, 6(3):537–546, 1977.

[304] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 35–44. ACM, 2011.

[305] Loren Terveen and Will Hill. Beyond recommender systems: Helping people help each other. *HCI in the New Millennium*, 1:487–509, 2001.

[306] Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu, Ryo Hayasaka, Ryuzo Isochi, and Zian Wang. Trusting tweets: The Fukushima disaster and information source credibility on Twitter. In *Proceedings of Information Systems for Crisis Response and Management*, 2012.

[307] Kathleen J Tierney. Disaster preparedness and response: Research findings and guidance from the social science literature. Technical report, Univ. of Delaware, Disaster Research Center, 1993.

[308] Marie Truelove, Maria Vasardani, and Stephan Winter. Towards credibility of micro-blogs: characterising witness accounts. *GeoJournal*, pages 1–21, 2014.

[309] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 565–574. ACM, 2011.

[310] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, pages 643–652. ACM, 2012.

[311] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[312] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[313] José Van Dijck. 'you have one identity': performing the self on Facebook and LinkedIn. *Media, Culture & Society*, 35(2):199–215, 2013.

[314] Sarah Vieweg. *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications*. PhD thesis, Univ. of Colorado at Boulder, 2012.

[315] Sarah Vieweg, Oliver L Haimson, Michael Massimi, Kenton O'Hara, and Elizabeth F Churchill. Between the lines: Reevaluating the online/offline binary. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2337–2340. ACM, 2015.

[316] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM, 2010.

[317] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution

of user interaction in Facebook. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA, 2009. ACM.

[318] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 2014.

[319] Xinyue Wang, Laurissa Tokarchuk, Félix Cuadrado, and Stefan Poslad. Exploiting hashtags for adaptive microblog crawling. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 311–315. ACM, 2013.

[320] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. I regretted the minute i pressed share: A qualitative study of regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 10. ACM, 2011.

[321] Spencer R Weart. *The discovery of global warming: revised and expanded edition*. Harvard Univ. Press, 2008.

[322] Ingmar Weber and Carlos Castillo. The demographics of web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 523–530. ACM, 2010.

[323] Katrin Weller, GE Gorman, and GE Gorman. Accepting the challenges of social media research. *Online Information Review*, 39(3), 2015.

[324] Katrin Weller and Katharina E Kinder-Kurlanda. Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research? In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[325] Robert West, Ryen W White, and Eric Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1399–1410, 2013.

[326] Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. ACM, 2013.

[327] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*, EuroSys '09, pages 205–218. ACM, 2009.

[328] Ping Wu, Ji-Rong Wen, Huan Liu, and Wei-Ying Ma. Query selection techniques for efficient crawling of structured web sources. In *Proceedings of the 22nd International Conference on Data Engineering*, pages 47–47. IEEE, 2006.

[329] Clayton Wukich and Ines A. Mergel. Closing the Citizen-Government communication gap: Content, audience, and network analysis of government tweets. *Social Science Research Network Working Paper Series*.

[330] Jinxi Xu and W Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1):79–112, 2000.

[331] Xin Yan, Raymond Y.K. Lau, Dawei Song, Xue Li, and Jian Ma. Toward a semantic granularity model for domain-specific information retrieval. *ACM Trans. Inf. Syst.*, 29(3):15:1–15:46, July 2011.

[332] Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, 2011.

[333] Xiwang Yang, Harald Steck, Yang Guo, and Yong Liu. On top-k recommendation using social networks. In *Proceedings of the sixth ACM Conference on Recommender systems*, pages 67–74. ACM, 2012.

[334] Xiwang Yang, Harald Steck, and Yong Liu. Circle-based recommendation in online social networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1267–1275. ACM, 2012.

[335] David S Yeager, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, page nfr020, 2011.

[336] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 2012.

[337] Elad Yom-Tov and Evgeniy Gabrilovich. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6), 2013.

[338] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the International Conference on World Wide Web Companion Publication*, 2014.

[339] Petros Zerfos, Junghoo Cho, and Alexandros Ntoulas. Downloading textual hidden web content through keyword queries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 100–109. IEEE, 2005.

[340] C.-N. Ziegler. *Towards Decentralized Recommender Systems*. PhD thesis, Albert-Ludwigs-

Universitat Freiburg, 2005.

[341] Michael Zimmer. "but the data is already public": on the ethics of research in Facebook. *Ethics and information technology*, 12(4):313–325, 2010.

[342] Michael Zimmer and Nicholas John Proferes. A topology of twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3):250–261, 2014.

Alexandra Olteanu

Ph.D. Candidate

École Polytechnique Fédérale de Lausanne (EPFL)
EPFL-IC-LSIR, BC147, Station 14
1015 Lausanne, Switzerland

+41 21 69 37559 (office)
+41 789 07 88 35 (mobile)
alexandra.olteanu@epfl.ch
http://people.epfl.ch/alexandra.olteanu

Research Interests

Social Media, Social Computing, Crisis Computing, Data Biases and Quality, Social Systems and Tools, Data Science for Social Good

Education

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland Sept. 2011 - Jan. 2016
Ph.D. Candidate in Computer Science
Distributed Information Systems Laboratory, under the supervision of Prof. Karl Aberer
Thesis: *Probing the Limits of Social Data: Biases, Methods and Domain Knowledge*

Vrije University of Amsterdam, Netherlands & 2009 - 2011
University Politehnica of Bucharest, Romania
Double-degree Research M.Sc. in Parallel and Distributed Computer Systems
Thesis: *P2P Social Networking*, under the supervision of Prof. Guillaume Pierre

University Politehnica of Bucharest, Romania 2005 - 2009
B.Sc. in Computer Science. Specialization in Systems Programming.
Thesis: *Rescheduling and Error Recovery Algorithm for GRID Environments*, under the direction of Prof. Valentin Cristea & Prof. Florin Pop.

Research & Work Experience

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland September 2011 - present
Research Assistant in the Distributed Information Systems Laboratory, working with Prof. Karl Aberer *Lausanne, Switzerland*

My work focuses on helping users to better gauge and assess web content. Specifically, I'm identifying and leveraging distinctive traits of data (e.g., source, type, credibility) within particular domains (e.g., climate change, crises) and media (e.g., Twitter, websites) to build or guide the development of dedicated tools (e.g., data collection and filtering, credibility prediction).

Microsoft Research, US June 2015 - September 2015
Research Intern in the Context, Learning, and User Experience for Search (CLUES) Group, working with Emre Kiciman *Redmond, US*

I worked on the experimental design and evaluation methodology of the QuantifiedAll system which aims to extract information about the outcomes of people's actions from social media in order to help others with e.g. decision-making, or goal achievement.

Qatar Computing Research Institute (QCRI), Qatar September 2013 - February 2014
Research Associate Intern in the Social Computing Group, working with Carlos Castillo *Doha, Qatar*

We showed that using a generic crisis-lexicon to collect migroblogged posts during mass emergencies leads to collections with higher recall than obtained with crisis-specific keywords manually chosen by experts; it also helps to preserve the original distribution of message types.

Vrije University of Amsterdam (VU), Netherlands January - August 2011
Research Master Project, working with Prof. Guillaume Pierre *Amsterdam, Netherlands*

We showed that building robust and scalable online social networks on top of P2P systems is feasible from an architectural standpoint. In this regard, I implemented a dedicated simulation framework that scales to millions of nodes.

"Politehnica" University of Bucharest, Romania October 2009 - August 2010
Research Assistant, working with Prof. Costin Boiangiu *Bucharest, Romania*

I worked on designing image segmentation algorithms that improve the input of Optical Character Recognition (OCR) systems. Part of my work became part of a course on analysis and extraction of information in documents. My work was supported by CCS Content Conversion Specialists GmbH.

WebWise.ro, Romania July - August 2009
Software Engineer (project-based) *Slatina, Romania*

I implemented, using a classic LAMP stack, an application-specific content management system consisting of two types of interfaces for different levels of user privileges.

"Politehnica" University of Bucharest, Romania December 2008 - August 2010
Research student collaborator, working with Prof. Valentin Cristea & Prof. Florin Pop *Bucharest, Romania*

Research projects on resource management in large-scale distributed systems. I evaluated a set of scheduling heuristics, employed data mining techniques to uncover system behavior patterns and implemented various extensions to the MONARC II simulation framework.

Refereed Publications

[AAAI SS'16 EA] **Characterizing the Demographics Behind the #BlackLivesMatter Movement.** Alexandra Olteanu, Ingmar Weber, and Daniel Gatica-Perez. *In AAAI Spring Symposia on Observational Studies through Social Media and Other Human-Generated Content (SS'16 OSSM), Stanford, US, March 2016.*

[ICWSM'15] **Comparing Events Coverage in Online News and Social Media: The Case of Climate Change.** Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, Karl Aberer. *In Proc. of 9th International AAAI Conference on Web and Social Media (ICWSM'15), Oxford, UK, May 2015.*

[CSCW'15] **What to Expect When the Unexpected Happens: Social Media Communications Across Crises.** Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. *In Proc. of 18th ACM Computer Supported Cooperative Work and Social Computing (CSCW'15), Vancouver, BC, Canada, March 2015.*

[WISE'14] **Comparing the Predictive Capability of Social and Interest Affinity for Recommendations.** Alexandra Olteanu, Anne-Marie Kermarrec, and Karl Aberer. *In Proc. of 15th International Conference on Web Information Systems Engineering (WISE'14), Thessaloniki, Greece, October 2014 (Best paper award).*

[ICWSM'14] **CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises.** Alexandra Olteanu, Carlos Castillo, Fernando Diaz and Sarah Vieweg. *In Proc. of 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14), Ann Arbor, US, June 2014.*

[CHI'13 EA] **CredibleWeb: A Platform for Web Credibility Evaluation**. Zhicong Huang, Alexandra Olteanu, and Karl Aberer. *In Proc. of CHI'13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, May 2013.*

[ECIR'13] **Web Credibility: Features Exploration and Credibility Prediction.** Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. *In Proc. of the 34th European Conference on Information Retrieval (ECIR'13), Moscow, Russia, March 2013.*

[ICDM'12 Demo] **Topick: Accurate Topic Distillation from Users Streams.** Anton Dimitrov, Alexandra Olteanu, Luke McDowell, and Karl Aberer. *In Proc. of 13th IEEE International Conference on Data Mining (ICDM'12), Demo Session, Brussels, Belgium, December 2012.*

[CIKM'12] **A Decentralized Recommender System for Effective Web Credibility Assessment.** Thanasis Papaioannou, Jean-Eudes Ranvier, Alexandra Olteanu, Karl Aberer. *In Proc. of the 21th ACM International Conference on Information and Knowledge Management (CIKM'12), Hawaii, USA, November 2012.*

[SNS'12] **Towards Robust and Scalable Peer-to-Peer Social Networks.** Alexandra Olteanu, Guillaume Pierre. *In Proc. of the 5th EuroSys' Workshop on Social Network Systems (SNS'12), Bern, Switzerland, May 2012 (Best paper award).*

[Elsevier CMA'12] **A Dynamic Rescheduling Algorithm for Resource Management in Large Scale Dependable Distributed Systems.** Alexandra Olteanu, Florin Pop, Ciprian Dobre, Valentin Cristea. *Computers & Mathematics with Applications 63(9), Elsevier, Volume 63(9), May 2012.*

[GHC'10 Poster] **Adaptive Scheduling Approach Used for Rescheduling in Distributed Systems.** Alexandra Olteanu, Florin Pop. *Grace Hopper Celebration of Women in Computing Conference, Atlanta, USA, November 2010.*

[GHC'10 Poster] **Adaptive Binarization Method with Variable Window Size.** Alexandra Olteanu, Alexandru Stefanescu, Costin-Anton Boiangiu. *Grace Hopper Celebration of Women in Computing Conference , Atlanta, USA, November 2010.*

Awards & Grants

ACM-W Scholarship 2015
for 2015 ACM Conference on Computer-Supported Cooperative Work and Social Computing

Student Travel Grant 2015
for 8th and 9th International AAAI Conference on Web and Social Media -2014

Best Paper Award 2014
at 15th International Conference on Web Information Systems Engineering (WISE'14)

Best Paper Award 2012
at 5th EuroSys' Workshop on Social Network Systems (SNS'12)

Google Global Community Scholarship 2010
for Grace Hopper Celebration of Women in Computing

"Dinu Patriciu" Private Scholarship 2010
for excellent academic results

Research scholarship 2009
in cooperation with CCS Content Conversion Specialists GmbH

3rd place Image Processing Session 2010
3rd place Managerial Communication Session 2006
Student Scientific Communications at "Politehnica" University of Bucharest

Media Coverage & Impact

CrisisLex Taxonomies Available in GDELT Global Knowledge Graph
GDELT Project, July 2015
http://blog.gdeltproject.org/crisislex-taxonomies-now-available-in-gkg/

Why You Shouldn't Only Get Your Climate Change News from the Mainstream Media
The Washington Post, April 2015
http://www.washingtonpost.com/news/energy-environment/wp/2015/04/21/how-to-get-twitter-riled-up-about-global-warming/

Earthquakes, Hurricanes and Other Disasters on Twitter
La Stampa (in Italian), March 2015
http://www.lastampa.it/2015/03/26/multimedia/tecnologia/terremoti-uragani-e-altre-calamitcos-li-racconta-twitter-jQGdKvF6Dq61Q3axYkfmUI/pagina.html

Crisis Tweets Study Identifies What to Expect During Emergencies
Emergency Management, February 2015
http://www.emergencymgmt.com/training/Crisis-Tweets-Study-Expect-During-Emergencies.html

The Role of Twitter During Emergencies and Disasters
Sky TG25 (in Italian), February 2015
http://tg24.sky.it/tg24/mondo/2015/02/18/twitter_ruolo_disastri_studio_differenti_tipi_di_informazioni.html

Could This Be The Most Comprehensive Study of Crisis Tweets Yet?
iRevolutions Blog, February 2015
http://irevolution.net/2015/02/16/comprehensive-crisis-tweets-study/

World Humanitarian Data and Trends 2014
United Nations Office for the Coordination of Humanitarian Affairs, December 2014
http://www.unocha.org/data-and-trends-2014/

Teaching

EPFL

Teaching Assistant

- *Probabilities and Statistics* (bachelor level) — Spring 2015
- *Analysis I* (bachelor level) — Fall 2014
- *Algorithms* (bachelor level) — Fall 2012

Advising Activities

- *Twitter as an Indicator of Food Security* (M.Sc. Thesis, Busser Alexander) — Spring 2015
- *The Impact of Twitter on Stock Markets* (Semester Project, Victor Kristof) — Fall 2014
- *An Empirical Analysis of the Evolution of Climate Change Discussions in Twitter* (Semester Project, Antoine Bastien) — Fall 2014
- *Debating Climate Change: How to Identify Credible Information and Sources from Twitter* (M.Sc. Thesis, Ashish Bindal) — Spring 2014
- *Robust Web Content Evaluation Systems* (M.Sc. Thesis, Bogdan Stoica) — Fall 2013
- *A Platform for Web Content Evaluation* (Semester Project, Zhicong Huang) — Fall 2012
- *Assess Web Credibility* (Semester Project, Stanislav Peshterliev) — Spring 2012
- *Inferring User Expertise from Online Social Network Content* (Semester Project, Anton Dimitrov) — Spring 2012

"Politehnica" University of Bucharest

Advising Activities

- *Data Transmissions for Multimedia* (class projects, master level) — Fall 2009
- *Analysis and Extraction of Information in Documents* (class projects, master level) — Spring 2010
- *Computational Geometry* (class projects, master level) — Spring 2010
- *Summer school for developing image processing applications* (instructor) — Summer 2010

Professional Service

Program Committee: ICWSM'16

External Reviewer: CHI'16, CSCW'16, WISE'12/'13/'14, VLDB'13, ICDE'13, IWSOS'13, ISWC'12, ICDCS'12

Skills

Programming: Python (current projects); C++, Java, C (past projects); PHP, CSS, HTML (occasionally)

Miscellaneous

Languages: English (fluent), Romanian (native), French, Spanish, Italian (basic)

Hobbies: Snowboarding, cooking, swimming. In a past life I also used to paint clocks and mirror frames.

Last updated: January 6th, 2016
http://static.alexandra.olteanu.eu/aolteanu-cv.pdf