

Limits of Social Data

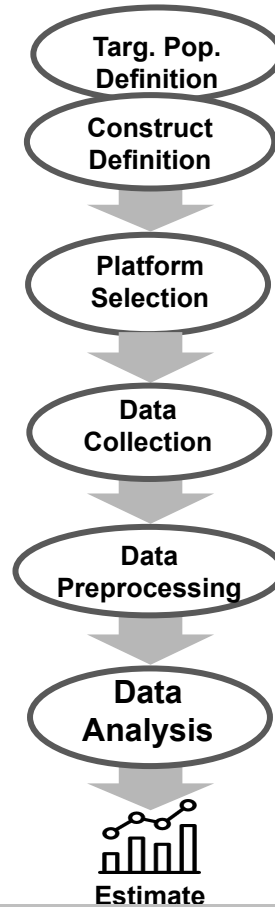
Indira Sen

University of Konstanz
Measurement and Representation Biases (MRB) in
Digital Trace Data-based Studies

Agenda

- ❖ Recap
- ❖ Reading: Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social data: Biases, methodological pitfalls, and ethical boundaries." *Frontiers in big data* 2 (2019): 13.
<https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full>
- ❖ Discussion: We will start the discussions based on the different roles:
<https://colinraffel.com/blog/role-playing-seminar.html>

A typical research pipeline with digital traces



[“TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms”](#) Sen et al., 2021, Public Opinion Quarterly and <https://arxiv.org/pdf/1907.08228>

Use case: Detecting the flu with digital traces

Google Flu (trends)

nature

Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

**What proportion of US-Americans have the flu?
What is the approval rating of A. Merkel?
Are anti-immigration sentiments on the rise?**

Ginsberg, Jeremy, et al. "[Detecting influenza epidemics using search engine query data.](https://doi.org/10.1038/nature07634)" *Nature* 457.7232 (2009): 1012-1014.

The fail

The image shows two overlapping screenshots of web pages. The top screenshot is from the Wired website, featuring a navigation bar with categories like 'BACKCHANNEL', 'BUSINESS', 'CULTURE', 'GEAR', 'IDEAS', 'SCIENCE', and 'SECURITY'. Below the navigation, there's a section for 'CORONAVIRUS' with sub-links for 'FAQS', 'BEST FACE MASKS', 'ESSENTIALS', 'HOW TO DISINFECT EVERYTHING', 'SYMPTOMS AND TESTING', 'NEWSLETTER', and 'LATEST NEWS'. The main article title is 'What We Can Learn From the Epic Failure of Google Flu Trends' by David Lazer and Ryan Kennedy, dated 10.01.2015 07:00 AM. The article text begins with 'GFT seeming searches into...'. The bottom screenshot is from the Nature website, showing the journal's header 'nature International weekly journal of science' and a navigation menu. The article title is 'When Google got flu wrong' by Declan Butler, dated 13 February 2013. The article text begins with 'US outbreak foxes a leading web-based method for tracking seasonal flu.'

“In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. Nature reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2).”

Example study

- How would a researcher study influenza prevalence in a national population using digital traces?



Construct Definition

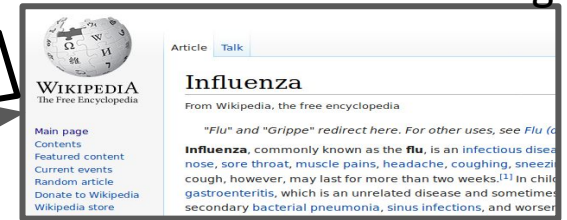
Platform Selection

SIGNALS

search queries related to flu



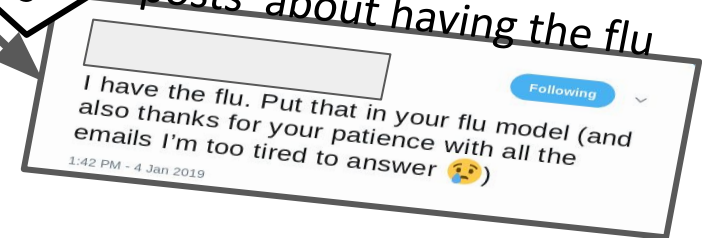
flu related information usage



SIGNALS

SIGNALS

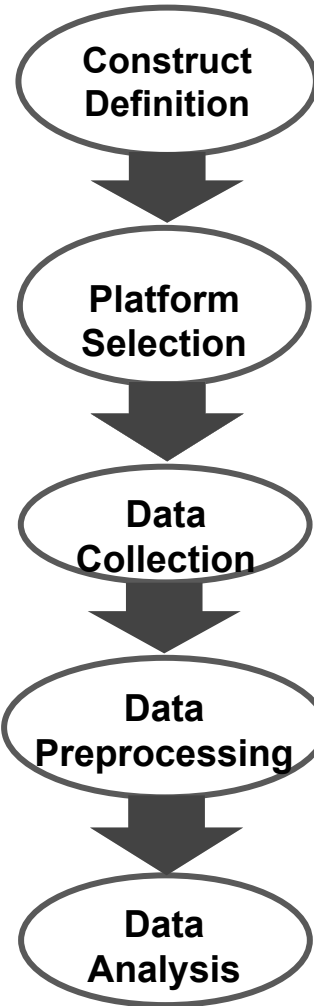
posts about having the flu



Preprocessing

Data Analysis





How to read and review papers?

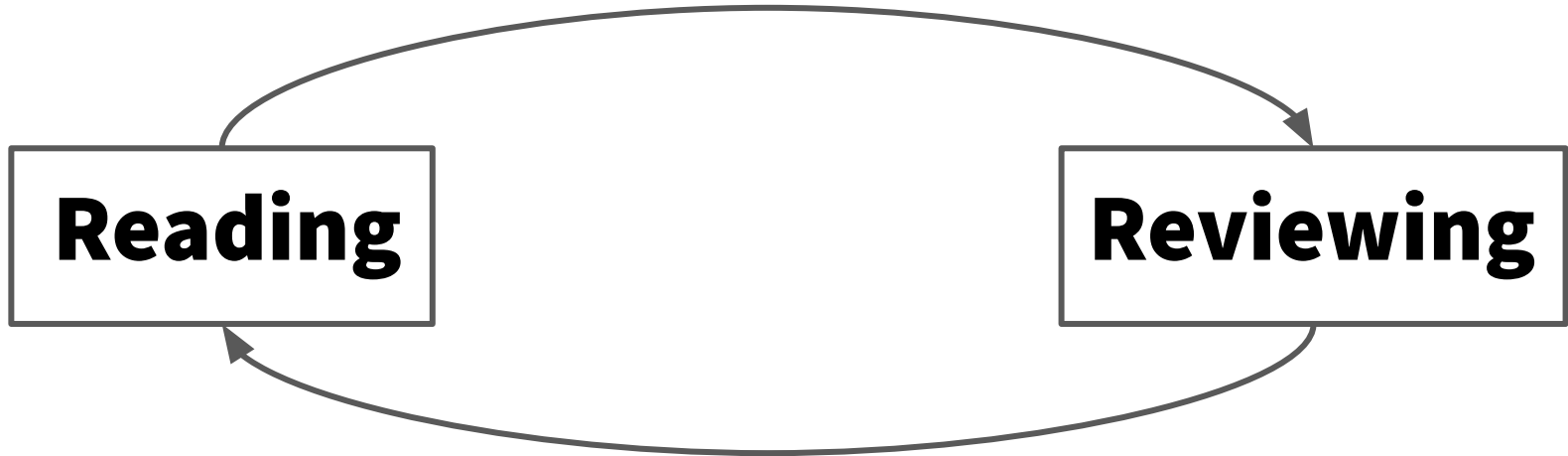
1. Keshav, Srinivasan. "[How to read a paper.](#)" ACM SIGCOMM Computer Communication Review 37.3 (2007): 83-84.
2. Pain, Elisabeth "[How to review a paper](#)"

How to review papers

Many different ways, this is what I do:

1. Read the abstract
2. Skim the introduction, results, and conclusions
3. Then read the entire paper more carefully
4. Make notes as you go: comments and questions
5. Then write the first review draft:
 - a. Summary of the paper; focus on its contributions
 - b. Strengths [does it address an important gap? Who does it 'help'?]
 - c. Weaknesses [not limitations, but e.g., does the paper *do* what it says it *does*?]

Bullet points are all you need for your first draft!



Let's read and review a paper together

“Narco” Emotions: Affect and Desensitization in Social Media during the Mexican Drug War

Munmun De Choudhury¹ Andrés Monroy-Hernández¹ Gloria Mark²

¹Microsoft Research

One Microsoft Way

Redmond, WA 98052 USA

{munmund, andresmh}@microsoft.com

²Department of Informatics

University of California, Irvine

Irvine, CA 92697 USA

gmark@uci.edu

ABSTRACT

Social media platforms have emerged as prominent information sharing ecosystems in the context of a variety of recent crises, ranging from mass emergencies, to wars and political conflicts. We study affective responses in social media and how they might indicate desensitization to violence experienced in communities embroiled in an armed conflict. Specifically, we examine three established affect measures: negative affect, activation, and dominance as observed on Twitter in relation to a number of statistics on protracted violence in four major cities afflicted by the Mexican Drug War. During a two year period (Aug 2010-Dec 2012), while violence was on the rise in these regions, our findings show a decline in negative emotional expression as well as a rise in emotional arousal and dominance in Twitter posts: aspects known to be psychological markers of desensitization. We discuss the implications of our work for behavioral health, facilitating rehabilitation efforts in communities enmeshed in an acute and persistent urban

as it can lead to cognitive performance decline, attentional impairment [20], and is a stressor of the onset of PTSD (post-traumatic stress disorder), an anxiety disorder associated with harmful physiological outcomes [30].

The Mexican Drug War is an example of the type of armed conflict that has exposed people to persistent acts of violence. Since the war started in, many Mexican cities have seen a rapid increase of shootings and homicides that, on occasions, affect innocent civilians. Furthermore, the conflict has triggered an increase of criminal activities such as extortions, and kidnappings affecting the general population [24]. This generalized violence in some Mexican cities, coupled with constrained information reporting on news media, have contributed to the emergence of citizen alert networks using platforms like Twitter and Facebook to inform and collectively grieve, critique, and express frustration about the violence in the streets [25].

Previous research in crisis informatics has demonstrated the role of social media as a lens to understand how society

Reading and discussing the paper

- Summary
- Strengths
- Weaknesses
- Building on this work

Limits of Social Data II

Indira Sen

University of Konstanz
Measurement and Representation Biases (MRB) in
Digital Trace Data-based Studies

Agenda

- ❖ Recap
- ❖ Reading: Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social data: Biases, methodological pitfalls, and ethical boundaries." *Frontiers in big data* 2 (2019): 13.
<https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full>
- ❖ Discussion
- ❖ Next time and logistics

Recap

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu^{1,2*}, Carlos Castillo³, Fernando Diaz² and Emre Kiciman⁴

¹ Microsoft Research, New York, NY, United States, ² Microsoft Research, Montreal, QC, Canada, ³ Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, ⁴ Microsoft Research, Redmond, WA, United States

Social data in digital form—including user-generated content, expressed or implicit relations between people, and behavioral traces—are at the core of popular applications and platforms, driving the research agenda of many researchers. The promises of social data are many, including understanding “what the world thinks” about a social issue, brand, celebrity, or other entity, as well as enabling better decision-making in a variety of fields including public policy, healthcare, and economics. Many academics and practitioners have warned against the naïve usage of social data. There are biases and inaccuracies occurring at the source of the data, but also introduced during processing. There are methodological limitations and pitfalls, as well as ethical boundaries and unexpected consequences that are often overlooked. This paper recognizes the rigor with which these issues are addressed by different researchers varies across a wide range. We identify a variety of menaces in the practices around social data use, and organize them in a framework that helps to identify them.

“For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated.” –Ursula Franklin¹

Keywords: social media, user data, biases, evaluation, ethics

Situation: [*The current context and background*]

Complication: [*The challenges in the current situation*]

Resolution: [*proposal for resolving the challenge*]

Hulett J (2021) [How to tell a business story using the McKinsey situation-complication-resolution \(SCR\) framework.](#)

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu^{1,2*}, Carlos Castillo³, Fernando Diaz² and Emre Kiciman⁴

¹Microsoft Research, New York, NY, United States, ²Microsoft Research, Montreal, QC, Canada, ³Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, ⁴Microsoft Research, Redmond, WA, United States

Social data in digital form—including user-generated content, expressed or implicit relations between people, and behavioral traces—are at the core of popular applications and platforms, driving the research agenda of many researchers. The promises of social data are many, including understanding “what the world thinks” about a social issue, brand, celebrity, or other entity, as well as enabling better decision-making in a variety of fields including public policy, healthcare, and economics. Many academics and practitioners have warned against the naïve usage of social data. There are biases and inaccuracies occurring at the source of the data, but also introduced during processing. There are methodological limitations and pitfalls, as well as ethical boundaries and unexpected consequences that are often overlooked. This paper recognizes the rigor with which these issues are addressed by different researchers varies across a wide range. We identify a variety of menaces in the practices around social data use, and organize them in a framework that helps to identify them.

“For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated.” –Ursula Franklin¹

Keywords: social media, user data, biases, evaluation, ethics

Situation: Social Data (content from social media, web platforms, and other social softwares) is an alternative to traditional sources of data

Complication: While this data has several potentials (up to date, ‘non-reactive’, large, etc), it also presents several challenges (population biases, lack of ecological validity)


Resolution: trying to understand these challenges better systematically

Hulett J (2021) [How to tell a business story using the McKinsey situation-complication-resolution \(SCR\) framework.](#)

What did we cover?

- ❖ The potentials of ‘social data’ / digital trace data
 - Larger samples
 - Samples that are not WEIRD
 - ...
- ❖ Growing evidence of pitfalls
 - Google flu
 - We still need theory
 - Validity issues
- ❖ General challenges

General Challenges

- 
- **Population Biases:** differences in demographics
 - **Behavioral Biases:** differences in user behavior
 - **Content Biases:** lexical, syntactic, and semantic biases in user content
 - **Linking Biases:** differences in network connections, interactions
 - **Temporal Biases:** changing biases over time
 - **Redundancy:** duplicates, near duplicates

What did 'you' think about the paper?

- ❖ **Private investigator:** authors are generally from CS and have worked in industry research labs which colors their perspective
- ❖ **Social impact assessor:** somewhat lacking grounding in theory
- ❖ **Industry Practitioner:** Social Science Bias Prevention Tool

This time

- ❖ Zooming into the specific parts of the pipeline
- ❖ Ethical Boundaries

Tutorial Series: Limits of Social Data:

<https://www.aolteanu.com/SocialDataLimitsTutorial/slidesKDD17.html>

Challenges During Data Analysis

- **Data Source:** biases at the source of social data
- **Data Collection:** biases due to data collection
- **Data Processing:** biases due to data preprocessing
- **Data Analysis:** validity threats due to methods selection and usage

Data Source

- **Functional:** biases due to platform affordances and algorithms
- **Normative:** biases due to community norms
- **External:** biases due to phenomena outside social platforms
- **Non-individuals:** e.g., organizations, automated agents

Data Collection

- **Acquisition:** biases due to, e.g., API limits
- **Querying:** biases due to, e.g., query formulation
- **Filtering:** biases due to removal of data "deemed" irrelevant

Data Processing

- **Cleaning:** biases due to, e.g., default values
- **Enrichment:** biases from manual or automated annotations
- **Aggregation:** e.g., grouping, organizing, or structuring data

Data Analysis

- **Qualitative Analyses:** lack generalizability, interpret. biases
- **Descriptive Statistics:** confounding bias, obfuscated measurements
- **Prediction & Inferences:** data representation, perform. variations
- **Observational studies:** peer effects, select. bias, ignorability

Evaluation

- **Metrics:** e.g., reliability, lack of domain insights

Ethical Boundaries

- **General Concepts and Principles**
- **Individual Autonomy:** ensure informed consent
- **Beneficence and Nonmaleficence:** actions are beneficial and do not cause harm
- **Justice:** risks and benefits justly shared

Data Source



- **Functional:** biases due to platform affordances and algorithms
- **Normative:** biases due to community norms
- **External:** biases due to phenomena outside social platforms
- **Non-individuals:** e.g., organizations, automated agents

Data Collection



- **Acquisition:** biases due to, e.g., API limits
- **Querying:** biases due to, e.g., query formulation
- **Filtering:** biases due to removal of data “deemed” irrelevant

Data Processing



- **Cleaning:** biases due to, e.g., default values
- **Enrichment:** biases from manual or automated annotations
- **Aggregation:** e.g., grouping, organizing, or structuring data

Data Analysis



- **Qualitative Analyses:** lack generalizability, interpret. biases
- **Descriptive Statistics:** confounding bias, obfuscated measurements
- **Prediction & Inferences:** data representation, perform. variations
- **Observational studies:** peer effects, select. bias, ignorability

Evaluation



- **Metrics:** e.g., reliability, lack of domain insights
- **Interpretation:** e.g., contextual validity, generalizability
- **Disclaimers:** e.g., lack of negative results and reproducibility

Tutorial Series: Limits of Social Data:

<https://www.aolteanu.com/SocialDataLimitsTutorial/slidesKDD17.html>

Data Source

Functional Biases: Biases that are a result of platform-specific mechanism or affordances, that is, the possible actions within each system or environment

Normative Biases: Biases that are a result of written norms or expectations about unwritten norms describing acceptable patterns of behavior on a given platform

External Sources of Bias

Non-Individual Accounts

[the meaning of retweets or likes] “could range from affirmation to denunciation to sarcasm to approval to disgust”

Data Source

Functional Biases: Biases that are a result of platform-specific mechanism or affordances, that is, the possible actions within each system or environment

Normative Biases: Biases that are a result of written norms or expectations about unwritten norms describing acceptable patterns of behavior on a given platform

External Source

Non-Individual

See for example, norms on Reddit: Chandrasekharan et al., "[The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales](#)"

[the meaning of retweets or likes] "could range from affirmation to denunciation to sarcasm to approval to disgust"

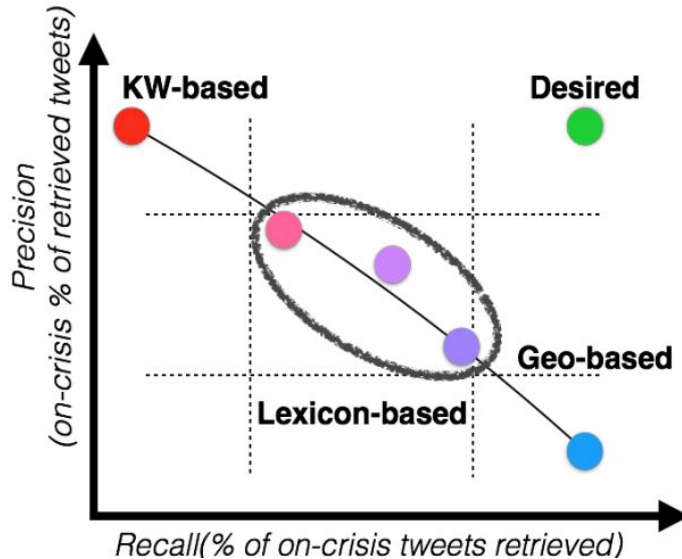
See data brokers like: <https://socialscience.one/>

But also the potential impact of reliance on platforms: Wu and Taneja, "[Platform enclosure of human behavior and its measurement: Using behavioral trace data against platform episteme](#)"

The choice of keywords in keyword based samples shapes the resulting datasets

The choice of keywords “**equivalent to specifying the boundaries of a data collection**”: working with the wrong list of keywords might cause relevant data to be missed”

González-Bailón, Sandra, et al. "[Assessing the bias in samples of large online networks.](#)" *Social Networks* 38 (2014): 16-27.



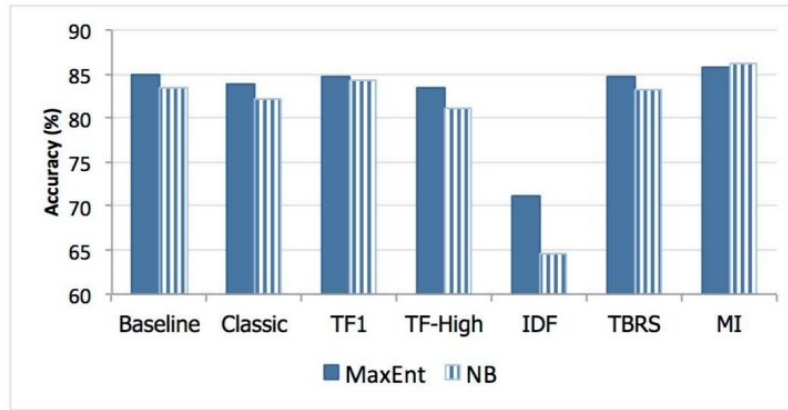
- Keyword-based sampling:
#sandy, #bostonbombings, #qldflood
- Location-based sampling:
tweets geo-tagged in disaster area

Olteanu, Alexandra, et al. "[Crisislex: A lexicon for collecting and filtering microblogged communications in crises.](#)" *Proceedings of the international AAAI conference on web and social media*. Vol. 8. No. 1. 2014.

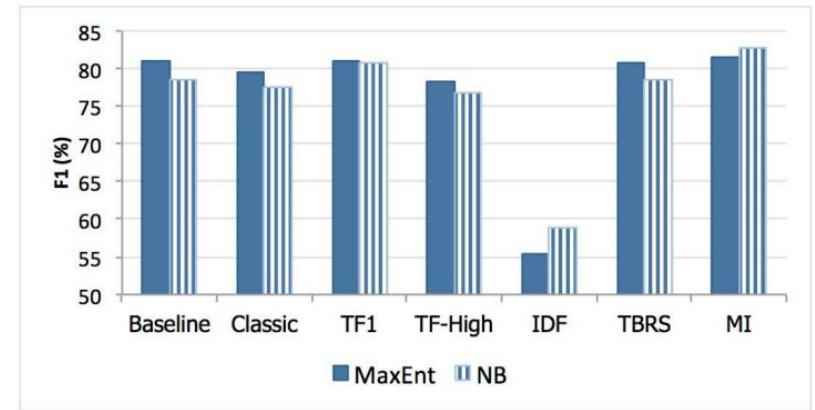
Figure based on

Text filtering operations may limit certain analyses.

Does removing stop-words helps or hampers the effectiveness of Twitter sentiment classification methods?



(a) Average Accuracy



(b) Average F-measure

Figure 4: Average Accuracy and F-measure of MaxEnt and NB classifiers using different stoplists

Data Processing

Data Cleaning

Data Enrichment

Data Aggregation

Data Processing

Data Cleaning

Data **Enrichment**

Data Aggregation

**Manual
annotation
often yields
subjective
and noisy
labels**

**Data enrichment
involves adding
annotations to
data items.**

Factors that affect the quality of human-annotations:

- **unreliable annotations**
- **poor annotation guidelines**
- **poor category design** → categories that are too broad, too narrow, or too vague
- **insufficient information** to make a reliable assessment
- **ethical concerns** → how and what information can or cannot be presented to the annotators
- **skewed difficulty** → some categories are easier to identify than others e.g., baby vs. people in their 40s

Data Processing

Data Cleaning

Data **Enrichment**

Data Aggregation

**Manual
annotation
often yields
subjective
and noisy
results**

**enrichment
involves adding
annotations to
data items.**

One common enrichment
technique for textual
data: [Perspective API](#)

Factors that affect the quality of human-annotations:

- **unreliable annotations**
- **poor annotation guidelines**
- **poor category design** → categories that are too broad, too narrow, or too vague
- **insufficient information** to make a reliable assessment
- **ethical concerns** → how and what information can or cannot be presented to the annotators
- **skewed difficulty** → some categories are easier to identify than others e.g., baby vs. people in their 40s

Data Analysis

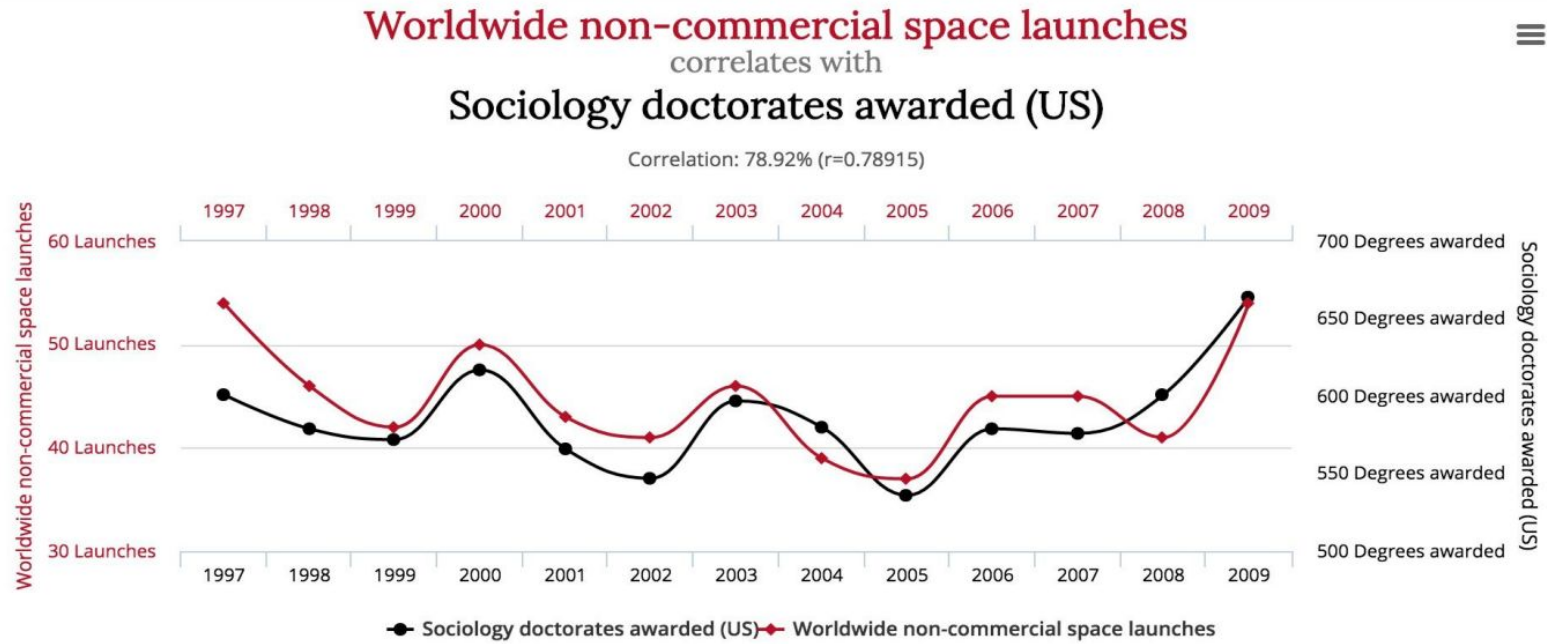
Qualitative Analyses

Descriptive Statistics

Predictions and Inferences

Observational Studies

Correlation analyses are sensitive to bias, confounders, and randomness



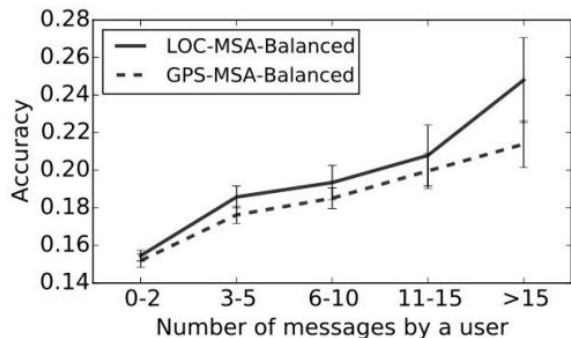
Tutorial Series: Limits of Social Data: Data sources: Federal Aviation Administration and National Science Foundation

<https://www.aolteanu.com/SocialDataLimitsTutorial/slidesKDD17.html>

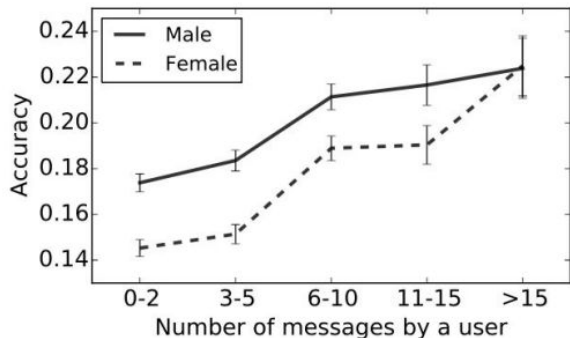
Try more on <https://www.tylervigen.com/spurious-correlations>

There are performance variations across and within dataset

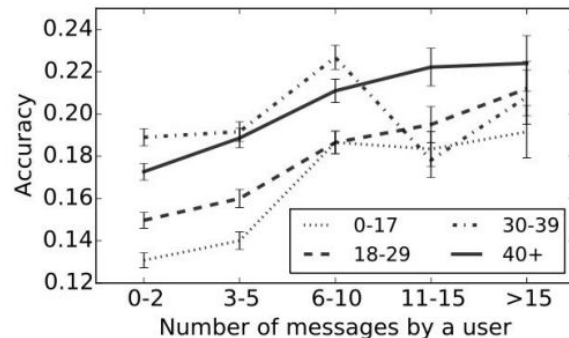
Even when a model achieves good overall accuracy, **errors may be unevenly concentrated on certain groups of users.**



(a) Classification accuracy by sampling technique



(b) Classification accuracy by user gender



(c) Classification accuracy by imputed age

Classification accuracies from figure in Pavalanathan, Umashanthi, and Jacob Eisenstein. "[Confounds and Consequences in Geotagged Twitter Data.](#)" *EMNLP* 2015.

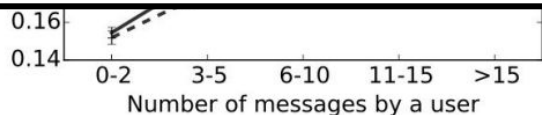
13

There are performance variations across and within dataset

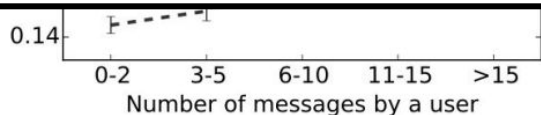
Even when a model achieves good overall accuracy, **errors may be unevenly concentrated on certain groups of users.**

See also for the consequences of differential error rates:

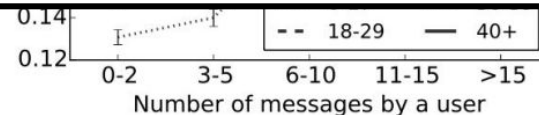
<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>



(a) Classification accuracy by sampling technique



(b) Classification accuracy by user gender



(c) Classification accuracy by imputed age

Classification accuracies from figure in Pavalanathan, Umashanthi, and Jacob Eisenstein. "[Confounds and Consequences in Geotagged Twitter Data](#)," EMNLP 2015.

13

Evaluation

Metrics Selection

Results Assessment
and Interpretation

Disclaimers and
Reproducibility

**Analyses
should go
beyond studies
confined to a
single dataset
or method**

Be wary of claims of generalizability: one must clearly state in what *context* or under what *assumptions*: the observations may be generalizable to other domains, platforms, groups of users, etc.

- results of **different methods** to collect, measure or process the data should be routinely juxtaposed
- if possible, results on **different datasets** should be reported

While longitudinal, comparative, multi-datasets or multi-platforms research exists, they tend to be the exception, rather than the rule.

Ethical Issues

Concepts and Principles

Individual Autonomy

Beneficence and
Nonmaleficence

Justice

Institutional review boards (IRBs)

Required by **funding agencies**: NSF in the US and the EC in Europe

Required by **CS conferences** e.g. CHI for many years, others more recently

Some benefits:

1. Set **common standards within an institution**, but not across institutions
2. Provide researchers a framework to **think critically** about consequences
3. **Demonstrate to others that careful decisions have been made** for a study

Giving back!

Researchers using social media should remember: we are often entering a conversation — grabbing all we can and leaving violates social norms

- Can we provide a valuable service to the communities we are studying?
- Can we share anonymized data back with to the community?
- What can we do to ensure a balance between the benefits we obtain as a researcher and the benefits the public receives from our research?

How would you improve on this work?

Next Time

- Measurement and Representation Errors
 - Groves, Robert M., and Lars Lyberg. "[Total survey error: Past, present, and future.](#)" Public Opinion Quarterly 74.5 (2010)
 - Sen, Indira, et al. "[A total error framework for digital traces of human behavior on online platforms.](#)" Public Opinion Quarterly 85.S1 (2021)

- Pick your roles by Monday:

Course Timeline and Deadlines

24.05.24: register
your paper and
timeslot

5.06.24-7.09.24: present your chosen
paper

15.09.24: submit
final report on
your paper

08.05.24: graded
discussions start
(pick your roles)

08.05.24-7.09.24: participate in discussions of
papers by students and guests

Updated Schedule and when you present

Jun 5 student presentation

Jun 12 guest presentation: [Max Pellert](#)

Jun 19 student presentation

Jun 26 guest presentation: [Nils Feldhus](#)

Jul 3 guest presentation: [Katrin Weller](#)

Jul 10 guest presentation: [Giordano de Marzo](#)

Jul 17 student presentation

Jul 24 guest presentation: [Jessica Daikeler](#)

Jul 31 student presentation

Aug 7 student presentation