Examples of Research with Web and Social Media Data

Indira Sen

University of Konstanz

Research Projects in Computational Studies of Social Phenomena

https://github.com/Indiiigo/css_research_projects/tree/main

Agenda

❖ Repurposing digital traces for (Computational) Social Science research

❖ An overview of different types of CSS studies

❖ Typical CSS methods

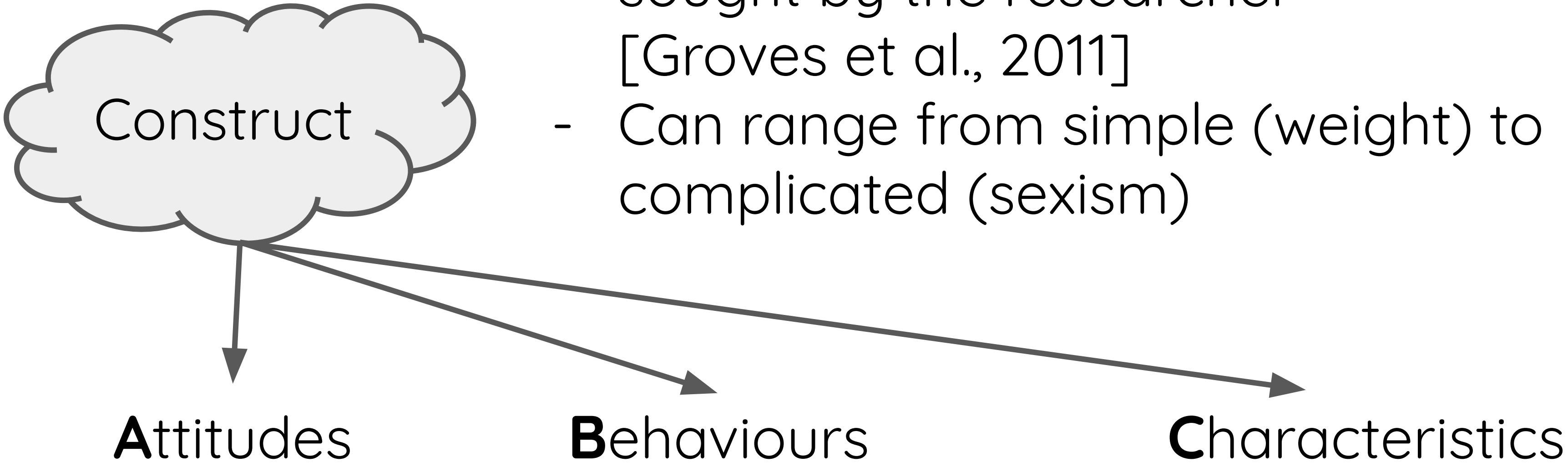
❖ Resources

recap

Understanding Social Phenomena

Understanding Social Phenomena

- Elements of **information** that are sought by the researcher [Groves et al., 2011]
- Can range from simple (weight) to complicated (sexism)

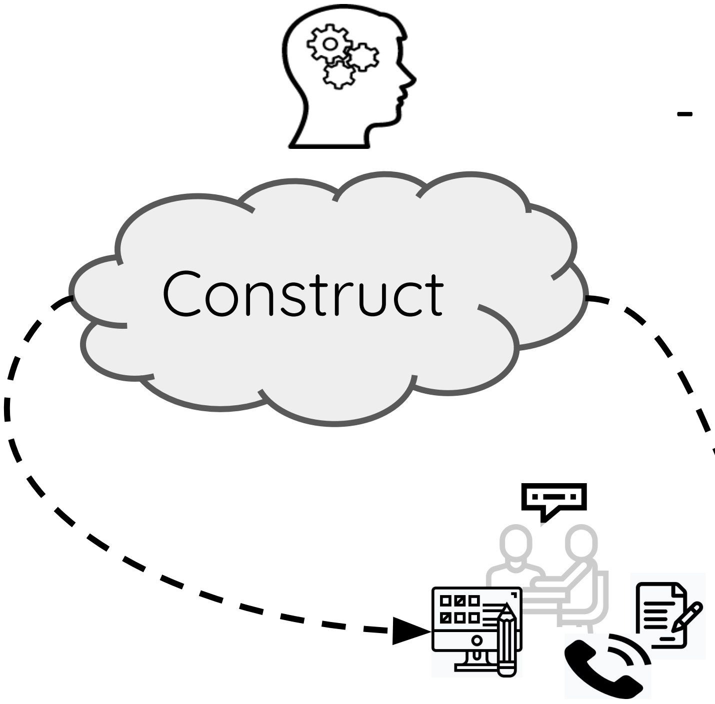


Understanding Social Phenomena

- The set of **people** to be studied [Groves et al., 2011]
- Can be easily defined (all the preschool students in a city) to more difficult (all refugees)
- Usually a national population, can also be any “system population” (-> platform study vs. “Social Sensing”)

Target Population

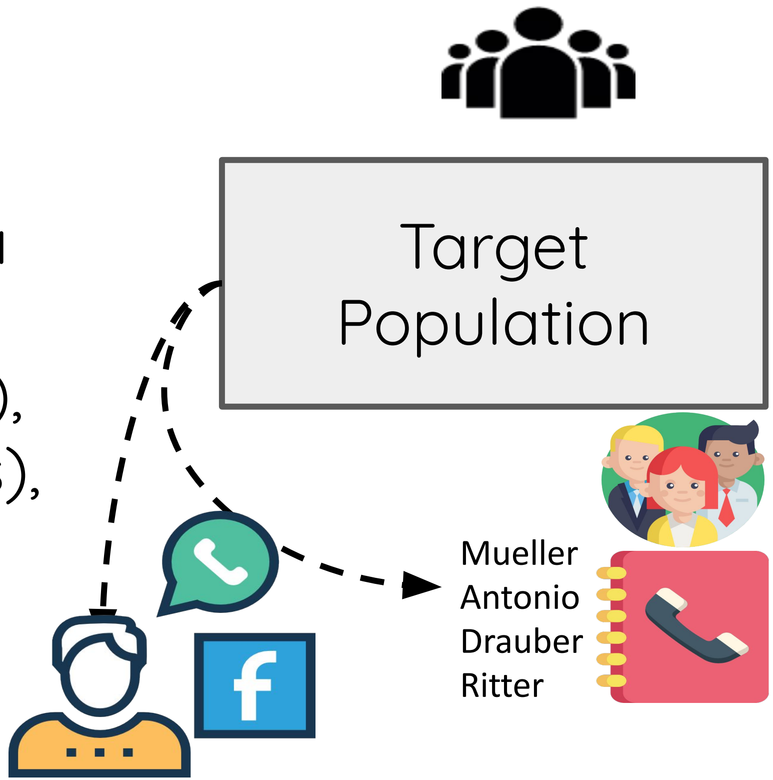
Understanding Social Phenomena



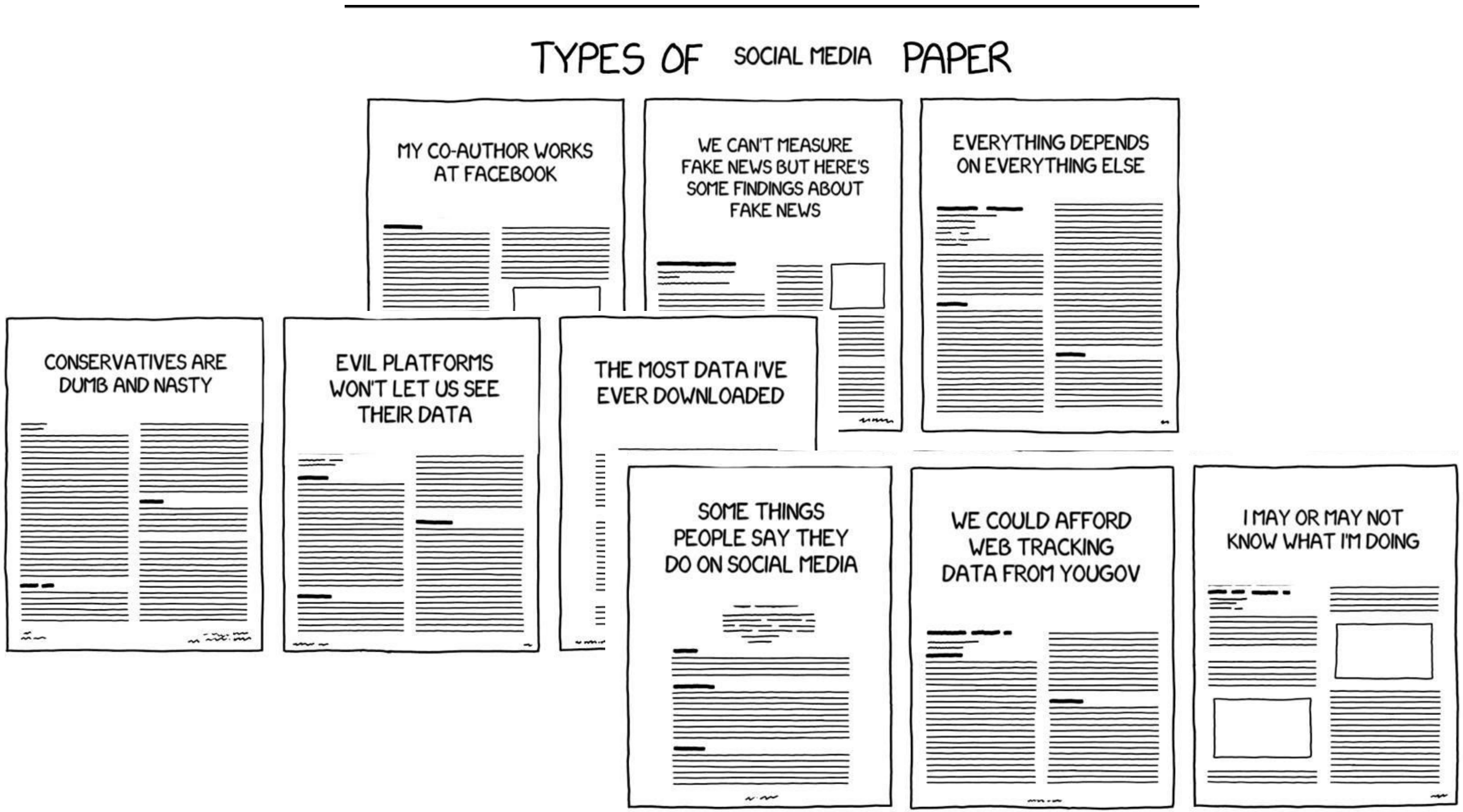
- Intangible constructs can be understood through tangible **traces**: responses to survey questions or digital actions and interactions

Understanding Social Phenomena

- **Members** of the target population (TP) are usually represented as elements in a record. This can be offline records (census, phonebook), online platforms (profiles, IPs), or other digital forms (bank accounts, fitness trackers)



Prototypical CSS Studies Leveraging Digital Trace Data



From: Fabrizio Gilardi
(@fgilardi)

‘Readymade’ and ‘Custommade’ data

Readymade

Custommade

“*Fountain* by Marcel Duchamp and *David* by Michaelangelo.

Fountain is an example of a **readymade**, where an artist sees something that already exists in the world then creatively repurposes it for art. *David* is an example of art that was intentionally created; it is a **custommade**. Social research in the digital age will involve both readymades and custommades.”

https://www.bitbybitbook.com/en/1st-ed/introduction/themes/

‘Readymade’ and ‘Custommade’ data

Readymade: repurposing of existing sources that were originally created for purposes other than your research question, often by companies (platform data), governments(register data), or other organizations (church attendance records). Similar to **Found** data

Custommade: a researcher started with a specific question and then create the data needed to answer that question (surveys).

‘Readymade’ and ‘Custommade’ data

Examples of Research with Readymade data: repurposes existing data, like web or social media data (but could also be other types of content — books, newspaper articles...)

[As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians]

WebSci ’19, June 30–July 3, 2019, Boston, MA, USA

As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians

Armin Mertens
Cologne Center for Comparative Politics
Cologne, Germany
mertens@wiso.uni-koeln.de

Franziska Pradel
Cologne Center for Comparative Politics
Cologne, Germany
pradel@wiso.uni-koeln.de

Ayjeren Rozyjumayeva
Faculty of Management, Economics and Social Sciences
Cologne, Germany
rozyjumayeva@wiso.uni-koeln.de

Jens Wäckerle
Cologne Center for Comparative Politics
Cologne, Germany
jens.waeckerle@uni-koeln.de

ABSTRACT

This study investigates gender bias in political interactions on digital platforms by considering how politicians present themselves on Twitter and how they are approached by others. Incorporating social identity theory, we use dictionary analyses to detect biases in individual tweets connected to the German federal elections in 2017. Besides sentiment analysis, we introduce a new measure of personal- vs. job-related content in text data, that is validated with structural topic models. Our results indicate that politicians’ communication on Twitter is driven by party identity rather than gender. However, we find systematic gender differences in tweets directed at politicians: female politicians are significantly more likely to be reduced to their gender rather than to their profession compared to male politicians.

1 INTRODUCTION

> “Why do you even care for the #BER [Berlin airport] opening? Because you have to choose your clothes for the opening ceremony??”[1]

This tweet directed at a female German politician resembles an often observed and reported phenomenon in our digital society: digital communication is driven by gender stereotypes rather than job-related content. This, in turn, can lead to bias and discrimination towards female professionals. Building on insights from research on social identities [60, 61] as well as gender biases and gender roles in politics [1, 9, 52], we expect to find gender biased communication on Twitter coming from — but also directed at — politicians. One of the key elements of social identity theory is group membership and

‘Readymade’ and ‘Custommade’ data

Examples of Research with Readymade data: repurposes existing data, like web or social media data (but could also be other types of content — books, newspaper articles...)

Examples of Research with Custommade data: creates surveys or survey experiments to test perceptions of politicians based on their gender

[As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians]()

WebSci ’19, June 30–July 3, 2019, Boston, MA, USA

As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians

Armin Mertens
Cologne Center for Comparative Politics
Cologne, Germany
mertens@wiso.uni-koeln.de

Franziska Pradel
Cologne Center for Comparative Politics
Cologne, Germany
pradel@wiso.uni-koeln.de

Ayjeren Rozyjumayeva
Faculty of Management, Economics and Social Sciences
rozyju

Jens Wäckerle
Cologne Center for Comparative Politics

ABSTRACT

This study investigates
ital platforms by consi
on Twitter and how the
social identity theory,
in individual tweets c
in 2017. Besides sentim
of personal- vs. job-rel
with structural topic m
communication on Twi
gender. However, we fi
directed at politicians:
likely to be reduced to
compared to male polit

Article

The Price of Power: Power Seeking and Backlash Against Female Politicians

Personality and Social Psychology Bulletin
36(7) 923–936
© 2010 by the Society for Personality and Social Psychology, Inc
Reprints and permission: sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167210371949
http://pspb.sagepub.com
SAGE

Tyler G. Okimoto[1] and Victoria L. Brescoll[1]

Abstract

Two experimental studies examined the effect of power-seeking intentions on backlash toward women in political office. It was hypothesized that a female politician’s career progress may be hindered by the belief that she seeks power, as this desire may violate prescribed communal expectations for women and thereby elicit interpersonal penalties. Results suggested that voting preferences for female candidates were negatively influenced by her power-seeking intentions (actual or perceived) but that preferences for male candidates were unaffected by power-seeking intentions. These differential reactions were partly explained by the perceived lack of communality implied by women’s power-seeking intentions, resulting in lower perceived competence and feelings of moral outrage. The presence of moral-emotional reactions suggests that backlash arises from the violation of communal prescriptions rather than normative deviations more generally. These findings illuminate one potential source of gender bias in politics.

Keywords

gender stereotypes, backlash, power, politics, intention, moral outrage

Received June 5, 2009; revision ac

[The Price of Power: Power Seeking and Backlash Against Female Politicians]()

Combining ‘Readymade’ and ‘Custommade’ data

Combines readymade data (mobile phone records) and custommade data (surveys)

ECONOMICS

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual’s past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

Reliable, quantitative data on the economic characteristics of a country’s population are essential for sound economic policy and research. The geographic distribution of poverty and wealth is used to make decisions about resource allocation and provides a foundation for the study of inequality and the determinants of economic growth (*1*, *2*). In developing countries, however, the scarcity of reliable quantitative data represents a major challenge to policy-makers and researchers. In much of Africa, for instance, national statistics on economic production may be off by as much as 50% (*3*). Spatially disaggregated data, which are necessary for small-area statistics and which are used by both the private and public sector, often do not exist (*4*, *5*).

In wealthy nations, novel sources of passively collected data are enabling new approaches to

unemployment (*9*), electoral outcomes (*10*), and economic development (*8*). Although most comparable sources of big data are scarce in the world’s poorest nations, mobile phones are a notable exception: They are used by 3.4 billion individuals worldwide and are becoming increasingly ubiquitous in developing regions (*11*).

Here we examine the extent to which anonymized data from mobile phone networks can be used to predict the poverty and wealth of individual subscribers, as well as to create high-resolution maps of the geographic distribution of wealth. That this may prove fruitful is motivated by the fact that mobile phone data capture rich information, not only on the frequency and timing of communication events (*12*) but also reflecting the intricate structure of an individual’s social network (*13*, *14*), patterns of travel and location choice (*15*–*17*), and histories of consumption and

individual’s socioeconomic characteristics. This distinction is a scientific one, which also has several important implications: First, it allows for the method to be used in contexts for which recent census or household survey data are unavailable. Second, when an authoritative source of data does exist, it can be used to more objectively validate or refute the model’s predictions. This limits the likelihood that the model is overfit on data from a single source, which is otherwise difficult to control, even with careful cross-validation (*20*). Third, our approach allows for a broad class of potential applications that require inferences about specific individuals instead of census tracts. As we discuss in the supplementary materials (section 6), future iterations of this approach could help to improve the targeting of humanitarian aid and social welfare, disseminate information to vulnerable populations, and measure the effects of policy interventions.

For this study, we used an anonymized database containing records of billions of interactions on Rwanda’s largest mobile phone network and supplemented this with follow-up phone surveys of a geographically stratified random sample of 856 individual subscribers. Upon contacting and surveying each of these individuals, we received informed consent to merge their survey responses with the mobile phone transaction database. The surveys solicited no personally identifying information but contained questions on asset ownership, housing characteristics, and several other basic welfare indicators. From these data, we constructed a composite wealth index using the first principal component of several survey responses related to wealth (*21*, *22*) (supplementary materials section 1D). For each of the 856 respondents, we thus have ~75 survey responses, as well as the historical records of thousands of phone-based interactions such as calls and text messages (Table 1).

We use the merged data from this sample of 856 phone survey respondents to show that a

[Predicting poverty and wealth from mobile phone metadata](#)

What are some things we create when doing social science?

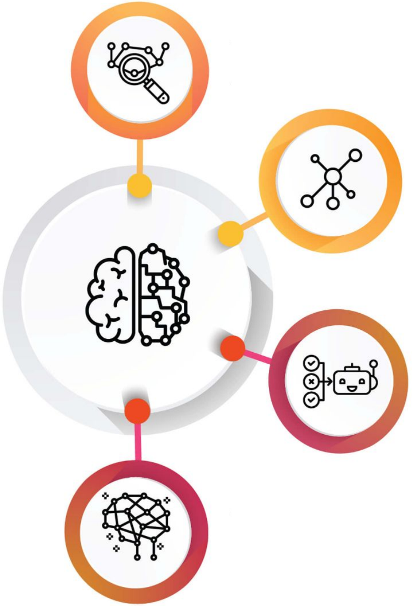
Prototypical Study Leveraging Digital Trace Data

1. **Analysis** (quantitative, qualitative, or mixed)
    a. *Social Sensing:* measuring events in a **broader target (offline) population** with platform data -> elections, markets, diseases, …

    b. *Platform studies:* Studying phenomenon **on the platform** -> content moderation, sharing behaviour, network effects, …

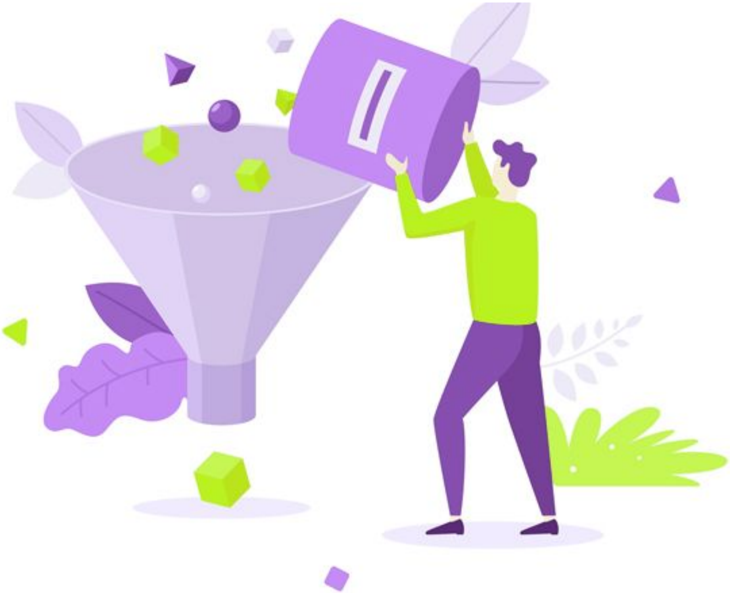
Prototypical Study Leveraging Digital Trace Data

2. **Method development/benchmarking**: Building, improving, evaluating methods for the specific social media context
    - Measuring stance towards X
    - hate speech, misinformation
    - Finding communities
    - User annotation
    - ….



Prototypical Study Leveraging Digital Trace Data

3. **Datasets**: Collection of users and/or posts that represent general social or platform phenomena
    - Misinformation
    - Hateful tweet
    - political candidates
    - …



Prototypical Study Leveraging Digital Trace Data

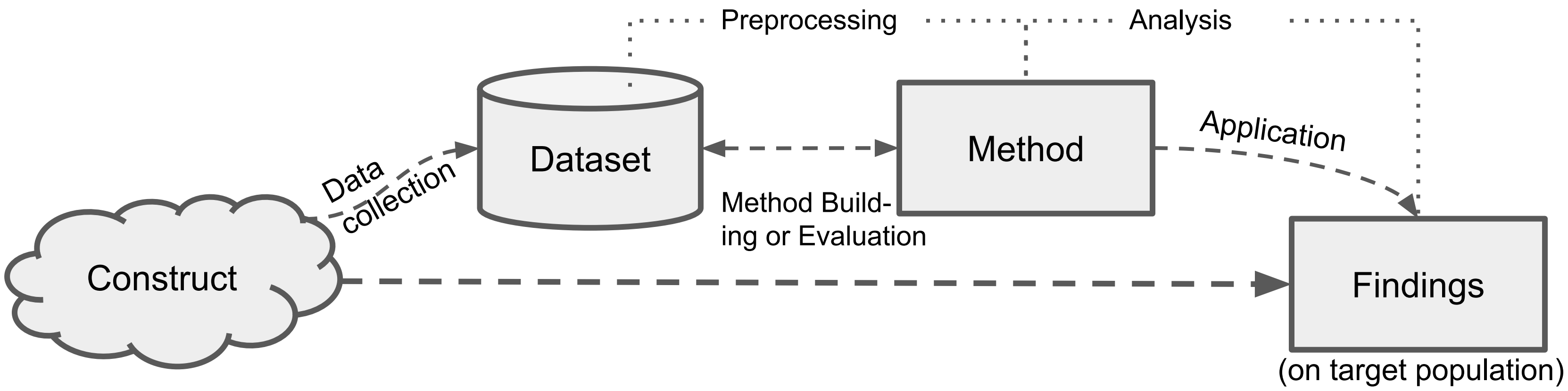
These prototypes (analyses, methods, and datasets) are usually interconnected

- Methods ← are either trained or tested on datasets

- Analyses ← rely on methods and provide insights based on particular datasets

…but studies need not have all components!

Prototypical Pipeline - Artifacts and Steps



Let’s do a quick activity

- Pick any social science paper you read recently. It doesn’t *have* to use digital traces

- What type of contributions does it make?
    - Dataset
    - Methods
    - Analysis

- Discuss

Let’s talk about some of the peculiarities of research with digital trace data

Prototypical Pipeline: From Construct to Measurement

From Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, January). The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 706-706).

Prototypical Pipeline: Data Collection

collect data potentially containing tangible signals regarding our construct

Content

videos

text

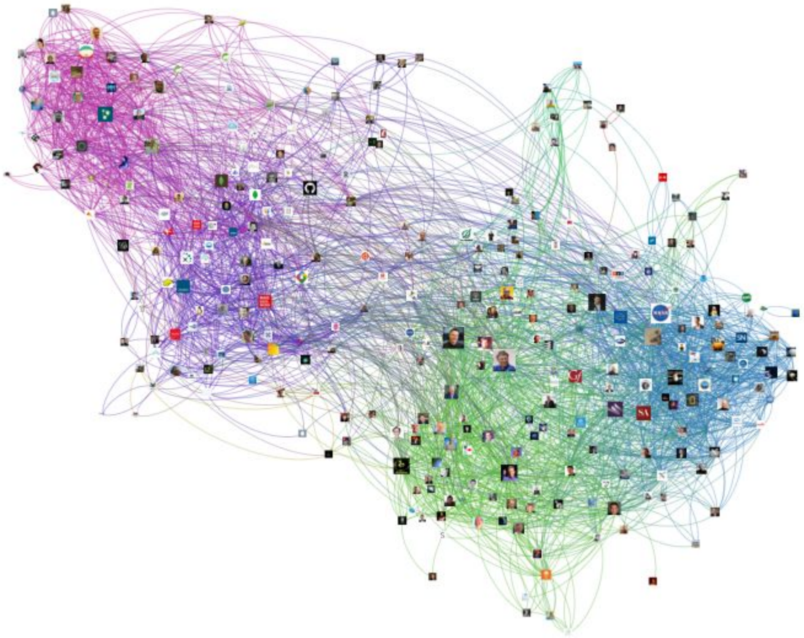
images

Prototypical Pipeline: Data Collection

collect data potentially containing tangible signals regarding our construct

Other data:

- Connections (friends, followers, links/URLs)
- Actions (likes, favs, comments)
- Aggregate data (views, shares, top searches)
- User profiles (with metadata)



Prototypical Pipeline: Data Collection

Source types:

- APIs (data collection), via tools or code libraries
- Web scraping
- Web tracking
- Data donations
- Reuse existing datasets (data archives)
- Official sale
- Use 3rd party services (e.g. Pushsift, Crimson Hexagon)

APIs as an example

- Typically not designed for research purposes (therefore not tailored to research needs)

- Several platforms offer APIs as access points

- Often you have to register, sometimes also accept terms of services

APIs as an example: Twitter [Past]

- Different access options for different purposes:
  - Twitter Developer: https://developer.twitter.com/
  - APIs: https://developer.twitter.com/en/docs
  - Premium "Firehose" access vs 1% sample. See also: https://developer.twitter.com/en/enterprise
  - Then: Academic API access
  - Now: DSA?
- Limitations on volume and functions
  - Free APIs cover 7 days tweets; Premium APIs exist for 30-day search and full archive search.

APIs are forever changing

- Facebook and Twitter/X completely closed down many of its APIs and it is now much harder to get Facebook data
- Constant need to keep up with changing APIs, sometimes certain data is just not available anymore

Computational research in the post-API age

Deen Freelon

University of North Carolina at Chapel Hill

Forthcoming in *Political Communication*

Keywords: API, computational, Facebook, Twitter, social media

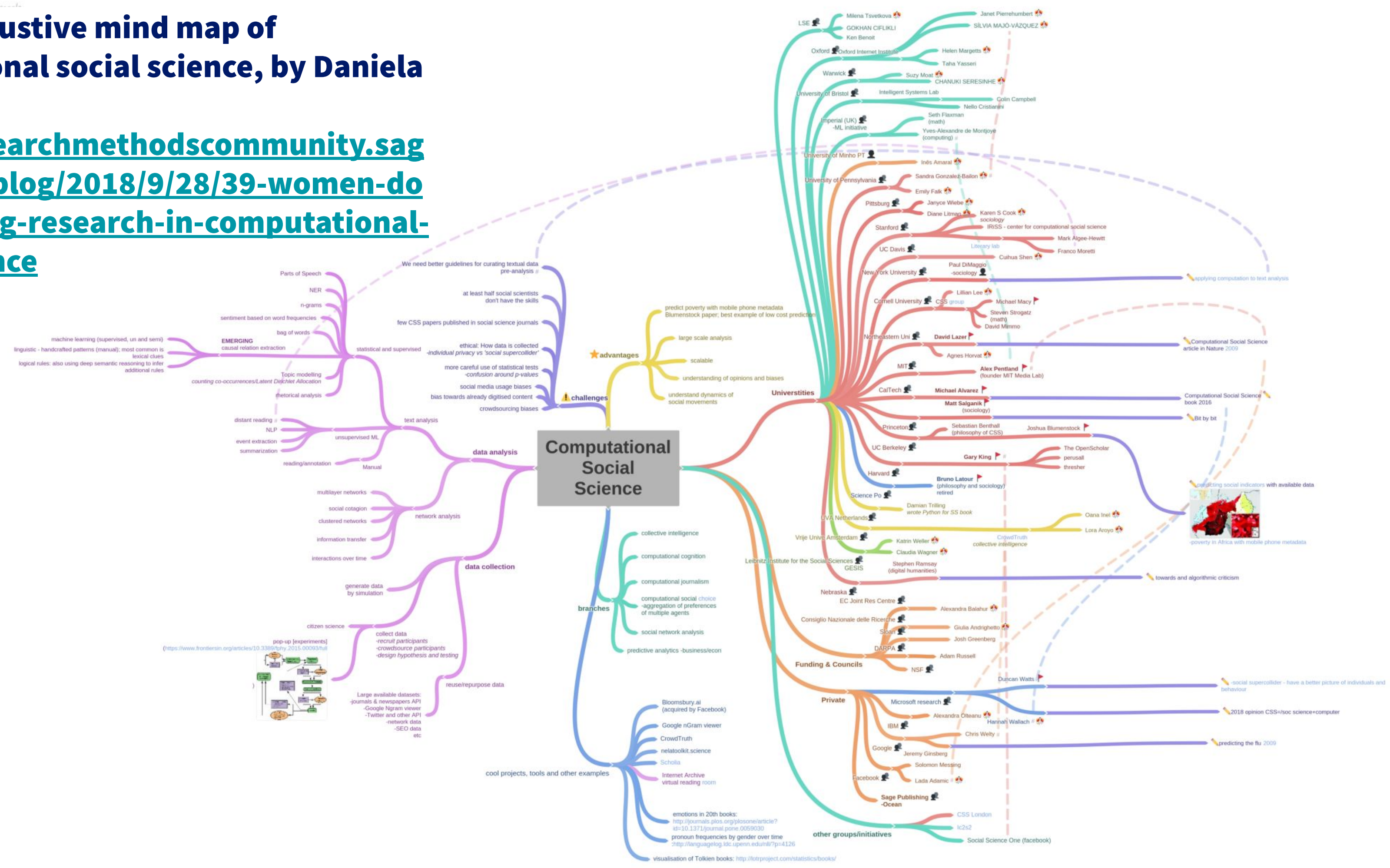
2018-08-20

On April 4, 2018, the post-API age reached a milestone. On that day, Facebook closed access to its Pages API, which had allowed researchers to extract all posts, comments, and associated metadata from public Facebook pages (Schroepfer, 2018). This decision followed the company's April 2015 closure of its public search API, which provided searchable access to all public posts within a rolling two-week window (Facebook, n.d.). The closure of the Pages API eliminated all terms of service (TOS)-compliant

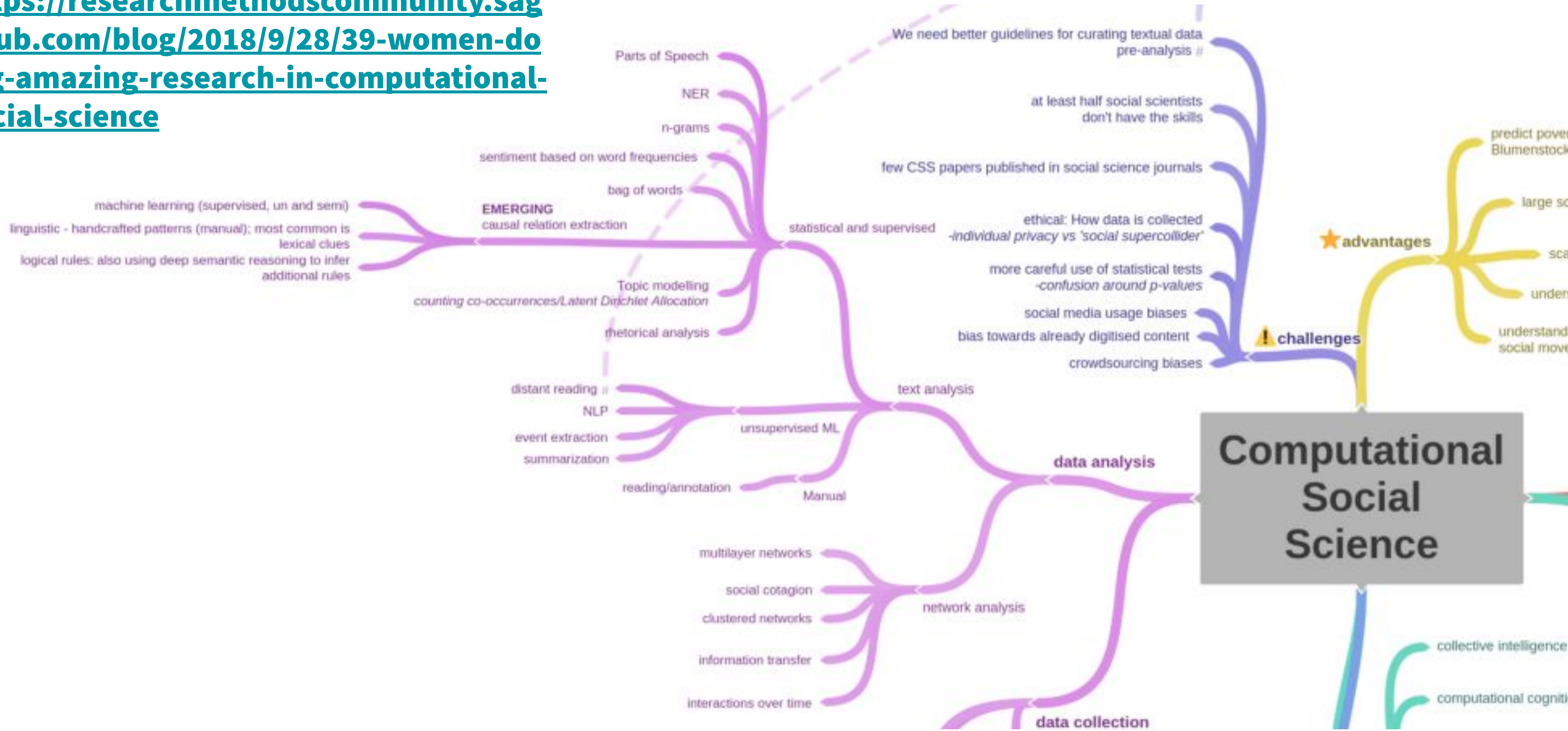
[Computational research in the post-API age]()

What about methods?

A non-exhaustive mind map of computational social science, by Daniela Duca: https://researchmethodscommunity.sagepub.com/blog/2018/9/28/39-women-doing-amazing-research-in-computational-social-science



A non-exhaustive mind map of computational social science, by Daniela Duca: https://researchmethodscommunity.sagepub.com/blog/2018/9/28/39-women-doing-amazing-research-in-computational-social-science



A rough list of (inter-connected) methods

- Network analysis
- Simulations
- NLP / Text-as-data
- Content analysis
- Human Computer Interaction (‘social computing’)
- Causal inference
- Surveys + digital traces

(Social) Network Analysis

Studying social structures with networks and graph theory

Natural fit for web and social media data

scientific reports

Check for updates

OPEN

Inequality and inequity in network-based ranking and recommendation algorithms

Lisette Espín-Noboa[1,2,3], Claudia Wagner[1,4,5], Markus Strohmaier[1,4,6] & Fariba Karimi[1]

Though algorithms promise many benefits including efficiency, objectivity and accuracy, they may also introduce or amplify biases. Here we study two well-known algorithms, namely PageRank and Who-to-Follow (WTF), and show to what extent their ranks produce *inequality* and *inequity* when applied to directed social networks. To this end, we propose a directed network model with preferential attachment and homophily (DPAH) and demonstrate the influence of network structure on the rank distributions of these algorithms. Our main findings suggest that (i) inequality is positively correlated with inequity, (ii) inequality is driven by the interplay between preferential attachment, homophily, node activity and edge density, and (iii) inequity is driven by the interplay between homophily and minority size. In particular, these two algorithms *reduce*, *replicate* and *amplify* the representation of minorities in top ranks when majorities are homophilic, neutral and heterophilic, respectively. Moreover, when this representation is reduced, minorities may improve their visibility in the rank by connecting strategically in the network. For instance, by increasing their out-degree or homophily when majorities are also homophilic. These findings shed light on the social and algorithmic mechanisms that hinder equality and equity in network-based ranking and

[Inequality and inequity in network-based ranking and recommendation algorithms]

Simulations

Fully or partially relies on 'simulated' data or simulations of people's behaviors

Agent-based models are one example

Institutions and Cultural Diversity: Effects of Democratic and Propaganda Processes on Local Convergence and Global Diversity

Roberto Ulloa, Celina Kacperski, Fernando Sancho

Published: April 8, 2016 • https://doi.org/10.1371/journal.pone.0153334

Article | Authors | Metrics | Comments | Media Coverage

Abstract

In a connected world where people influence each other, what can cause a globalized monoculture, and which measures help to preserve the coexistence of cultures? Previous research has shown that factors such as homophily, population size, geography, mass media, and type of social influence play important roles. In the present paper, we investigate for the first time the impact that institutions have on cultural diversity. In our first three studies, we extend existing agent-based models and explore the effects of institutional influence and agent loyalty. We find that higher institutional influence increases cultural diversity, while individuals' loyalty to their institutions has a small, preserving effect. In three further studies, we test how bottom-up and top-down processes of institutional influence impact our model. We find that bottom-up democratic practices, such as referenda, tend to produce convergence towards homogeneity, while top-down information dissemination practices, such as propaganda, further increase diversity. In our last model—an integration of bottom-up and top-down processes into a feedback loop of information—we find that when democratic processes are rare, the effects of ropaganda are amplified, i.e., more diversity emerges; however, when democratic processes re common, they are able to neutralize or reverse this propaganda effect. Importantly, our nodels allow for control over the full spectrum of diversity, so that a manipulation of our

[Institutions and Cultural Diversity: Effects of Democratic and Propaganda Processes on Local Convergence and Global Diversity]()

NLP / Text-as-Data

Lexicon-based methods, supervised (e.g., classifiers) and unsupervised ML (e.g., topic modeling)

Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia

Chan Young Park,* Xinru Yan,* Anjalie Field,* Yulia Tsvetkov
Language Technologies Institute, Carnegie Mellon University
{chanyoun, anjalief, ytsvetko}@cs.cmu.edu, xinruyan@alumni.cmu.edu

t reflect both the
ative and influence
examined descrip-
l affective analysis,
nique that seeks to
mensions of *power*,
n extension of this
ch is enabled by a
ilingual model. We
s differ across lan-
culty of generaliz-
. We then demon-
alyzing Wikipedia
community across
panish. Our results
GBT community is

English Wikipedia:
He *accepted* the option of injections of what was then called stilboestrol.

Spanish Wikipedia:
Finalmente escogió las inyecciones de estrógenos.
Finally he chose estrogen injections.

Russian Wikipedia:
Учёный предпочёл инъекции стильбэстрола
The scientist preferred stilbestrol injections.

Figure 1: Example from Alan Turing's biography page on Wikipedia in different languages. Verb choice in different languages can have subtly different connotations.

Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings

Dorottya Demszky[1] Nikhil Garg[1] Rob Voigt[1] James Zou[1]
Matthew Gentzkow[1] Jesse Shapiro[2] Dan Jurafsky[1]
[1]Stanford University [2]Brown University
{ddemszky, nkgarg, robvoigt, jamesz, gentzkow, jurafsky}@stanford.edu
jesse_shapiro_1@brown.edu

Abstract

We provide an NLP framework to uncover four linguistic dimensions of political polarization in social media: topic choice, framing, affect and illocutionary force. We quantify these aspects with existing lexical methods, and propose clustering of tweet embeddings as a means to identify salient topics for analysis across events; human evaluations show that our approach generates more cohesive topics than traditional LDA-based models. We apply our methods to study 4.4M tweets on 21 mass shootings. We provide evidence that the discussion of these events is highly polarized politically and that this polarization is primar-

2016) and Facebook (Bakshy et al., 2015). Prior NLP work has shown, e.g., that polarized messages are more likely to be shared (Zafar et al., 2016) and that certain topics are more polarizing (Balasubramanyan et al., 2012); however, we lack a more broad understanding of the many ways that polarization can be instantiated linguistically.

This work builds a more comprehensive framework for studying linguistic aspects of polarization in social media, by looking at topic choice,

[Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings]()

[Multilingual Contextual Affective Analysisof LGBT People Portrayals in Wikipedia]()

Content Analysis

Automated ('computer-assisted'), semi-automated, fully manual

4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community

Michael S. Bernstein[1], Andrés Monroy-Hernández[1], Drew Harry[1], Paul André[2], Katrina Panovich[1] and Greg Vargas[1]

[1]Massachusetts Institute of Technology
Cambridge, MA 02139
{msbernst, amonroy, dharry, kp, gvargas}@mit.edu

[2]University of Southampton
SO17 1BJ, United Kingdom
pa2@ecs.soton.ac.uk

en, Kraut, and Kiesler forthcoming; Grudin 2002). However, we have a limited understanding of how an anonymous nd ephemeral community design might actually play out — specially at large scale.

In this paper we analyze one such large-scale, anonyhous, and ephemeral community: the imageboard website chan. We focus on 4chan's first and most popular board, he "random" board known as /b/. Our goal is to use /b/ s a lens to understand the concepts of *anonymity* and *phemerality* online. /b/ implements these concepts in more xtreme ways than most other online communities. First, osts are fully anonymous by default and very rarely contain seudonyms or other identity signals. This lack of identity hakes traditional reputation systems unworkable. Second, nstead of archiving conversations, /b/ deletes them when ewer content arrives — often within minutes — which

Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks

Carsten Schwemmer & Sebastian Jungkunz

To cite this article: Carsten Schwemmer & Sebastian Jungkunz (2019) Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks, Political Research Exchange, 1:1, 1-23, DOI: 10.1080/2474736X.2019.1646102

To link to this article: https://doi.org/10.1080/2474736X.2019.1646102

[4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community]

[Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks]

Human Computer Interaction / Social Computing

Often focused on explaining platforms ('platform studies')

Audits of platforms and recommendation systems

Content moderation

You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech

ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology
UMASHANTHI PAVALANATHAN, Georgia Institute of Technology
ANIRUDH SRINIVASAN, Georgia Institute of Technology
ADAM GLYNN, Emory University
JACOB EISENSTEIN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

In 2015, Reddit closed several subreddits—foremost among them r/fatpeoplehate and r/CoonTown—due to violations of Reddit's anti-harassment policy. However, the effectiveness of banning as a moderation approach remains unclear: banning might diminish hateful behavior, or it may relocate such behavior to different parts of the site. We study the ban of r/fatpeoplehate and r/CoonTown in terms of its effect on both participating users and affected subreddits. Working from over 100M Reddit posts and comments, we generate hate speech lexicons to examine variations in hate speech usage via causal inference methods. We find that the *ban worked for Reddit*. More accounts than expected discontinued using the site; those that stayed drastically decreased their hate speech usage—by at least 80%. Though many subreddits saw an influx of r/fatpeoplehate and r/CoonTown "migrants," those subreddits saw no significant changes in hate speech usage. In other words, other subreddits did not inherit the problem. We conclude by reflecting on the apparent success of the ban, discussing implications for online moderation, Reddit and internet communities more broadly.

[You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech]

Causal Inference

Experiments, quasi-experimental designs, matching

Polit Behav (2017) 39:629–649
DOI 10.1007/s11109-016-9373-5

CrossMark

ORIGINAL PAPER

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

ber 2016
s Media New York 2016

experiment which examines the impact of group norm pro-
ioning on racist online harassment. Racist online harassment
ities it targets, and the open, unopposed expression of racism
gitimize racist viewpoints and prime ethnocentrism. I employ
d to reduce the use of anti-black racist slurs by white men on
nple of Twitter users who have harassed other users and use
ts") to sanction the harassers. By varying the identity of the
(white man) and out-group (black man) and by varying the
wers each bot has, I find that subjects who were sanctioned by
nale significantly reduced their use of a racist slur. This paper
lab experiments to a naturalistic setting using an objective,
This

A Social Media Study on the Effects of Psychiatric Medication Use

Koustuv Saha,[†] Benjamin Sugar,[†] John Torous,[‡] Bruno Abrahao,[*]
Emre Kıcıman,[§] Munmun De Choudhury[†]
[†]Georgia Tech, [‡]Harvard Medical School, [*]NYU Shanghai, [§]Microsoft Research
[†]{koustuv.saha,bsugar,munmund}@gatech.edu, [‡]jtorous@bidmc.harvard.edu, [*]bd58@nyu.edu, [§]emrek@microsoft.com

Abstract

Understanding the effects of psychiatric medications during mental health treatment constitutes an active area of inquiry. While clinical trials help evaluate the effects of these medications, many trials suffer from a lack of generalizability to broader populations. We leverage social media data to examine psychopathological effects subject to self-reported usage of psychiatric medication. Using a list of common approved and regulated psychiatric drugs and a Twitter dataset of 300M posts from 30K individuals, we develop machine learning models to first assess effects relating to mood, cognition, depression, anxiety, psychosis, and suicidal ideation. Then, based on a stratified propensity score based causal analysis,

practice, their effects vary across individuals, and often do not achieve the intended result. Without any biological markers to match patients with the most appropriate medication, the selection of drug treatments is based primarily on trial-and-error (Cipriani et al. 2018; Trivedi et al. 2006). Unsurprisingly, frustration with treatment and side effects often causes treatment discontinuation (Bull et al. 2002).

Consequently, literature in precision psychiatry has emphasized the need to understand the psychiatric effects of these medications (Cipriani et al. 2009). Presently, most knowledge of drug reactions comes from clinical trials and reports of adverse events; e.g., the FDA's Adverse Event Reporting System (*open.fda.gov/data/faers*) clinical
atabase. However, these trials can be biased, being
ucted and funded by pharmaceutical companies, an

[A Social Media Study on the Effects of Psychiatric Medication Use](#)

[Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment](#)

Surveys + Digital Traces

Relatively new, combines the ‘best of both worlds’

However, also has several challenges

- Data linking
- Skewed samples
- Data cleaning
- …

Is There a Link between Climate Change Scepticism and Populism? An Analysis of Web Tracking and Survey Data from Europe and the US

Pu Yan [a], Ralph Schroeder [a*], Sebastian Stier [b]

Abstract

Climate change and populism are two major phenomena in contemporary politics. Recent successes of populist parties and politicians, especially in Europe and in the United States, have given rise to extensive debates in the academic literature and beyond. Yet the link between populism and climate change scepticism (henceforth ‘scepticism’), has so far received little attention. This paper examines the link between scepticism and populism by means of a unique dataset: a survey and detailed web browsing histories of participants from France, Germany, Italy, Spain, the US, and the UK. The web tracking data includes more than 150 million website visits over a period of three months and the survey contains questions about political attitudes and attitudes towards climate change. We analyse the 8893 websites which contained ‘climate change’ and its cognates in the URLs and link these website visits and the content of these websites to the political attitudes and climate change orientation of those who visited them. The contribution is both methodological (linking surveys and web tracking data, including cross-country comparison) and substantive (uncovering links between populists, their climate change orientations, and the content of related websites

s the value of computational methods in

rovides insights into the link between populists

[Is There a Link between Climate Change Scepticism and Populism? An Analysis of Web Tracking and Survey Data from Europe and the US]

A rough list of (inter-connected) methods

- Network analysis
- Simulations
- NLP / Text-as-data
- Content analysis
- Human Computer Interaction (‘social computing’)
- Causal inference
- Surveys + digital traces
- …?

~~Let’s do a another quick activity~~ Homework

- Go back to the paper that you had picked.

- What type of method did the authors use? Is it appropriate for what is being studied?

- Discuss

Some Resources

- AwesomeCSS github page: https://github.com/gesiscss/awesome-computational-social-science?tab=readme-ov-file#online-courses-and-material
- Bit by Bit: (you can read it online for free!): https://www.bitbybitbook.com/en/1st-ed/preface/
- The Summer Institute in CSS: https://sicss.io/
- NLP + CSS tutorials: https://nlp-css-201-tutorials.github.io/nlp-css-201-tutorials/