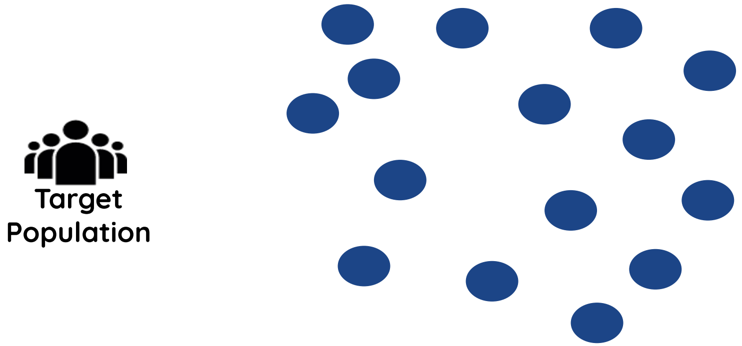
Conclusion

Indira Sen

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

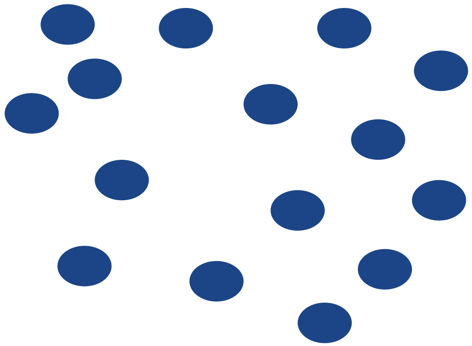
A typical research pipeline with digital trace data for measuring social phenomena

A typical research design with digital traces



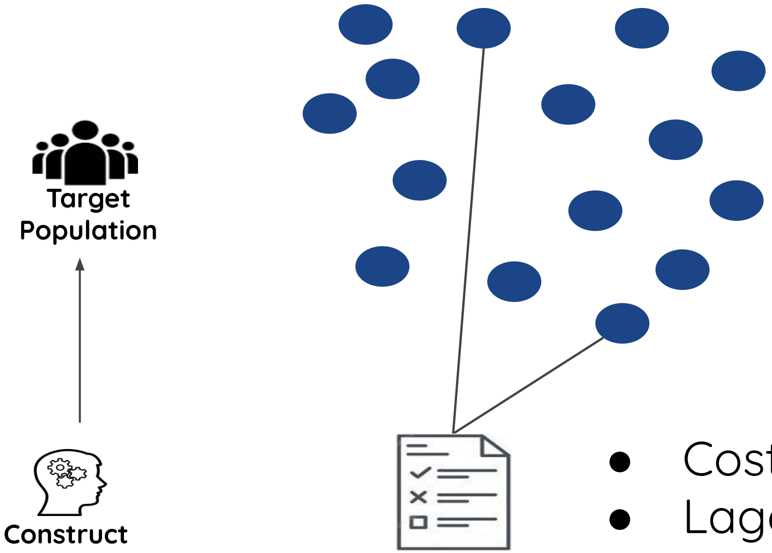
A typical research design with digital traces

Target Population All US adults



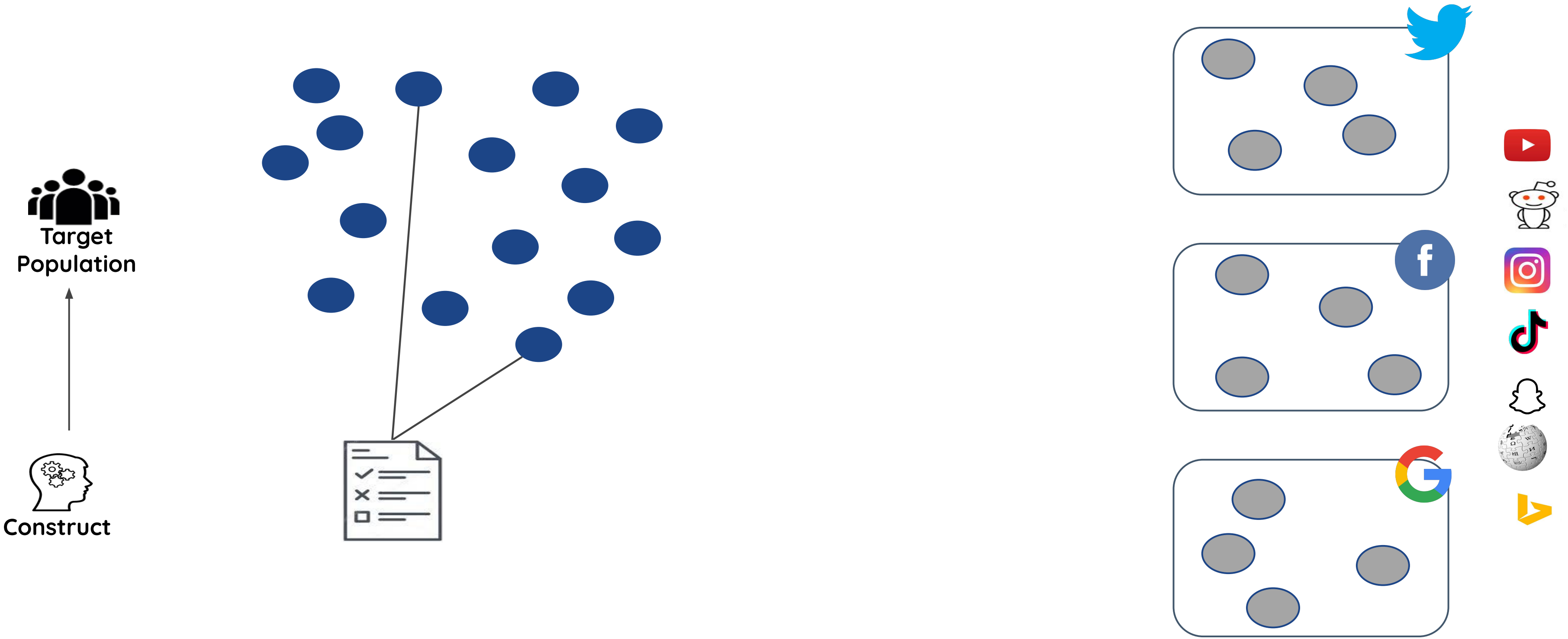
Construct Approval of D. Trump as president

A typical research design with digital traces

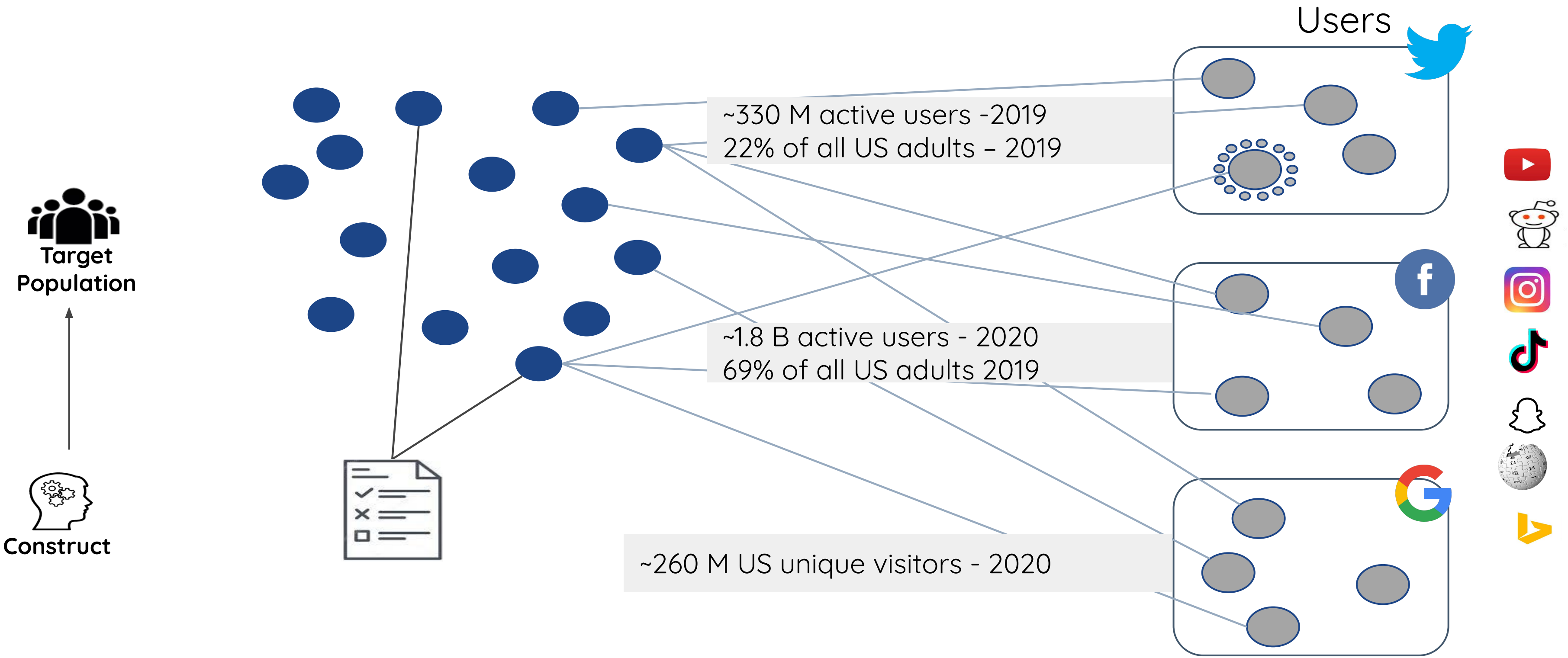


- Costly
- Lagging (after important events)
- Small samples (max. in the thousands)

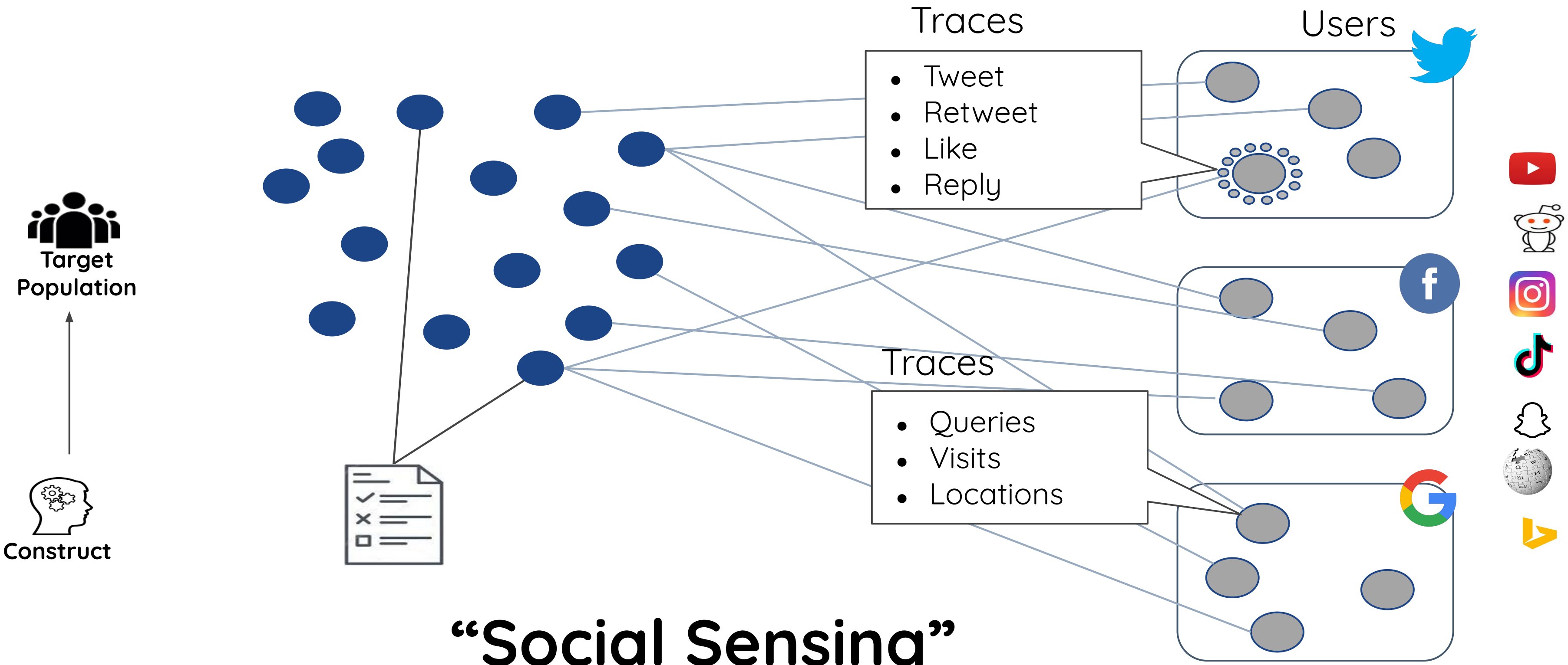
A typical research design with digital traces



A typical research design with digital traces

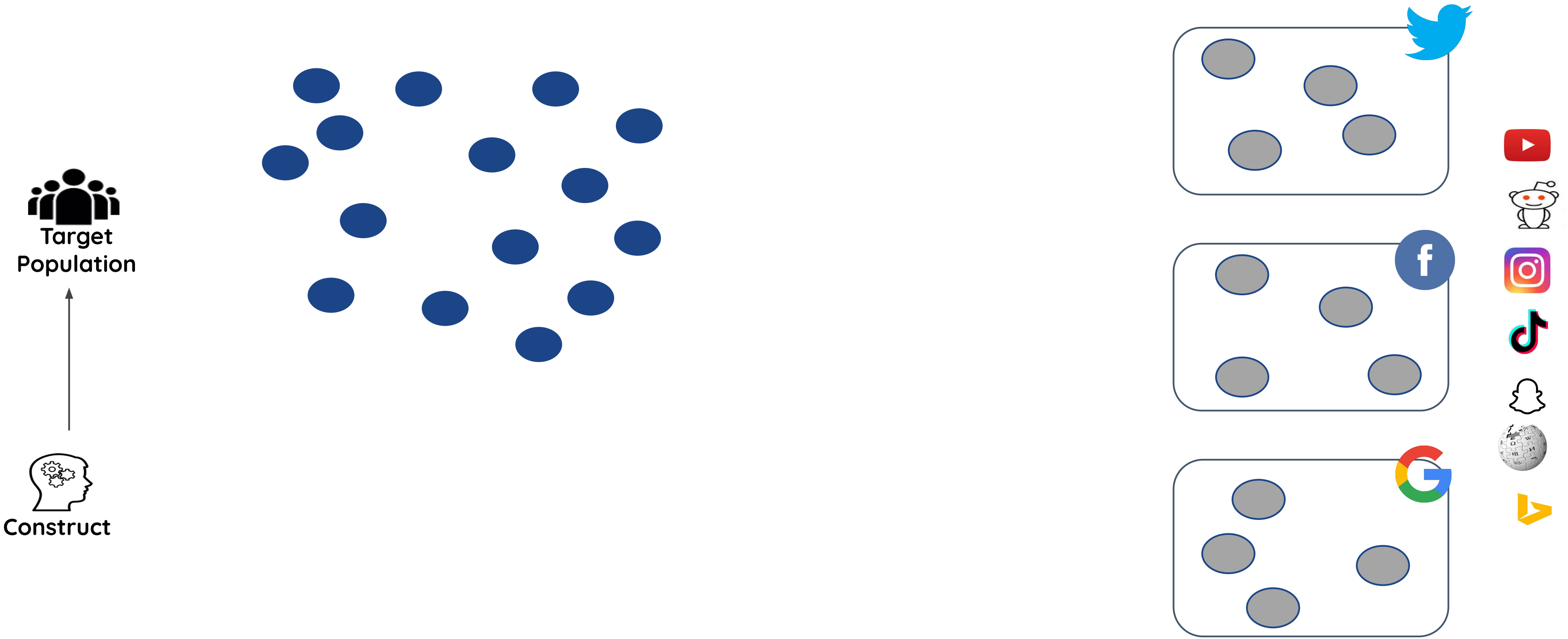


A typical research design with digital traces



"Social Sensing"

A typical research design with digital traces



A typical research design with digital traces

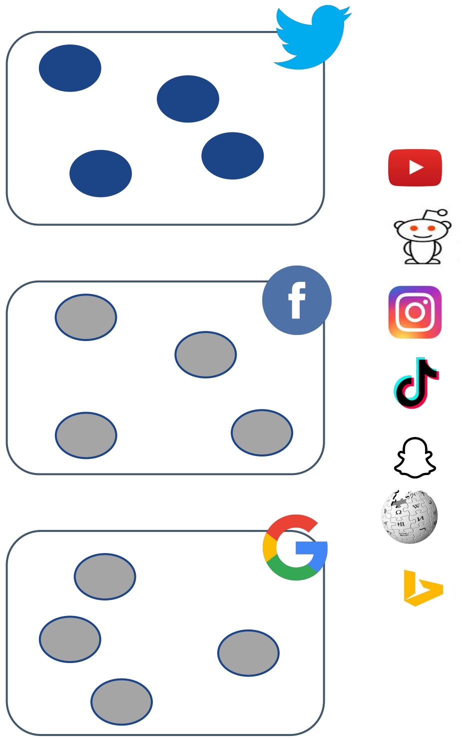
Target Population

All German-speaking Twitter Users

Construct

Susceptibility to Fake News

“Platform Study”



Examples of research with digital traces

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

Munmun De Choudhury[1] Andrés Monroy-Hernández[1] Gloria Mark[2]
[1] Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
{munmund, andresmh}@microsoft.com
[2] Department 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 warfare, and the impact on civic engagement.

Author Keywords
affect; desensitization; social media; crisis informatics.

ACM Classification Keywords
H.5.3. Group and Organization Interfaces; Asynchronous interaction; Web-based interaction.

as it can lead to cognitive performance decline, impairment [20], and is a stressor of the onset of P traumatic stress disorder), an anxiety disorder ass harmful physiological outcomes [30].

The Mexican Drug War is an example of the typ conflict that has exposed people to persistent acts Since the war started in, many Mexican cities h rapid increase of shootings and homicides that, or affect innocent civilians. Furthermore, the c triggered an increase of criminal activities such as and kidnappings affecting the general population generalized violence in some Mexican cities, co constrained information reporting on news m contributed to the emergence of citizen alert netw platforms like Twitter and Facebook to i collectively grieve, critique, and express frustratio violence in the streets [25].

Previous research in crisis informatics has demo role of social media as a lens to understand h copes with crises and how communities leverage for civic engagement and social support [21,39,2 this paper, we use social media to examine the e affective reactions to persistent violence, and whether affective desensitization may be mar social media, focusing on the Mexican Drug War step to this investigation, we focus particularly

scientific reports

Check for updates

OPEN

Validating daily social media macroscopes of emotions

Max Pellert[1,2,3,4], Hannah Metzler[1,2,3,5], Michael Matzenberger[6] & David Garcia[1,2,3]

Measuring sentiment in social media text has become an important practice in studying emotions at the macroscopic level. However, this approach can suffer from methodological issues like sampling biases and measurement errors. To date, it has not been validated if social media sentiment can actually measure the temporal dynamics of mood and emotions aggregated at the level of communities. We ran a large-scale survey at an online newspaper to gather daily mood self-reports from its users, and compare these with aggregated results of sentiment analysis of user discussions. We find strong correlations between text analysis results and levels of self-reported mood, as well as between inter-day changes of both measurements. We replicate these results using sentiment data from Twitter. We show that a combination of supervised text analysis methods based on novel deep learning architectures and unsupervised dictionary-based methods have high agreement with the time series of aggregated mood measured with self-reports. Our findings indicate that macro level dynamics of mood expressed on an online platform can be tracked with social media text, especially in situations of high mood variability.

User generated text from social media has become an important data source to analyze expressed mood and emotions at large scales and high temporal resolutions, for example to study seasonal mood oscillations[1], emotional responses to traumatic events[2], the effect of pollution on happiness[3], and the role of climate change in suicide and depression[4]. Despite these promising applications, using social media text to measure emotion aggregates can suffer a series of methodological issues typical of studies of this kind of *found data*[5–7]. Common validity threats are measurement error in sentiment analysis tools and the performative behavior of social media users due to platform effects or community norms. Sampling biases can generate a mismatch between users that produce text and a target group that might include silent individuals.

The validation of sentiment analysis methods has focused on micro level measurement accuracy at the individual post level[8]. Recent work has assessed the measurement validity also at the individual person level, using historical records of text from a user. This has revealed low to moderate correlations between aggregates of sentiment produced by an individual over a period of time and emotion questionnaires[9,10]. At the group level, static measurements of social media sentiment are only moderately correlated with affective well-being and life-satisfaction across regions[11]. These earlier findings highlight the limits of static aggregations of sentiment to measure concepts like life satisfaction that are only slowly changing over time. However, it is still an open question if analyses of social media text can shed light on *faster* phenomena, for example core emotional experiences, when we stick to aggregating individual signals to a community of interest and observe variation over time.

Here, we address this research gap by testing whether social media text sentiment tracks the macro level dynamics of emotions with daily resolution in an online community. We study the convergence validity of two approaches to study *emotions* at scale: *sentiment* aggregates from social media text and *mood* self-report frequencies in a survey. For 20 days, we collected 268,128 emotion self-reports through a survey in an Austrian online newspaper. During the same period, we retrieved text data from user discussions on the same platform, including 452,013 posts in our analysis using our pre-existing Austrian social media monitor[12]. To replicate our results with a second dataset, we conducted a pre-registered analysis of 635,185 tweets by Austrian Twitter users. We applied two off-the-shelf German sentiment analysis tools on the text data: a state-of-the-art supervised tool based on deep learning (German Sentiment, GS[13]) and a popular dictionary method based on expert word lists (Linguistic Inquiry and Word Count, LIWC[14]). Our results strongly support the assumption that social media sentiment can reflect both mean levels and changes of self-reported emotions in explicit daily surveys. We additionally analyze

nature human behaviour

Article https://doi.org/10.1038/s41562-023-01691-w

From alternative conceptions of honesty to alternative facts in communications by US politicians

Jana Lasser[1,2], Segun T. Aroyehun[1,3], Fabio Carrella[4], Almog Simchon[4], David Garcia[1,2,3] & Stephan Lewandowsky[4,5,6]

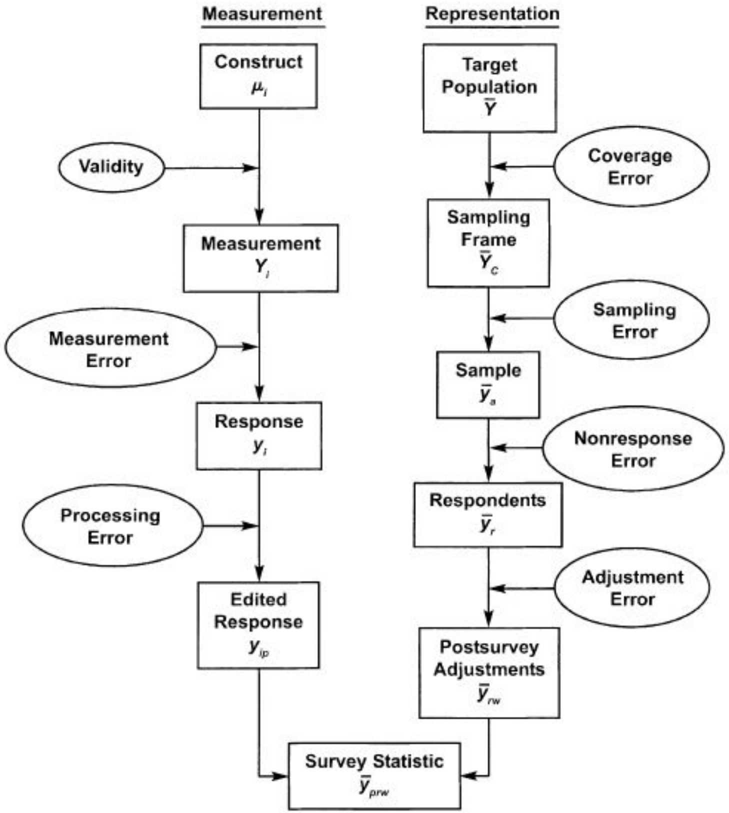
The spread of online misinformation on social media is increasingly perceived as a problem for societal cohesion and democracy. The role of political leaders in this process has attracted less research attention, even though politicians who 'speak their mind' are perceived by segments of the public as authentic and honest even if their statements are unsupported by evidence. By analysing communications by members of the US Congress on Twitter between 2011 and 2022, we show that politicians' conception of honesty has undergone a distinct shift, with authentic belief speaking that may be decoupled from evidence becoming more prominent and more differentiated from explicitly evidence-based fact speaking. We show that for Republicans—but not Democrats—an increase in belief speaking of 10% is associated with a decrease of 12.8 points of quality (NewsGuard scoring system) in the sources shared in a tweet. In contrast, an increase in fact-speaking language is associated with an increase in quality of sources for both parties. Our study is observational and cannot support causal inferences. However, our results are consistent with the hypothesis that the current dissemination of misinformation in political discourse is linked to an alternative understanding of truth and honesty that emphasizes invocation of subjective belief at the expense of reliance on evidence.

that people consume that later turns out to be false. Misinformation can be spread unintentionally, when communicators mistakenly believe some item of information to be true, or it can be spread intentionally, for example, in pursuit of a political agenda. Intentionally disseminated misinformation is often referred to as 'disinformation'. The psychological and cognitive consequences of disinformation are indistinguishable from those of unintentional misinformation, and we therefore use the latter term throughout.

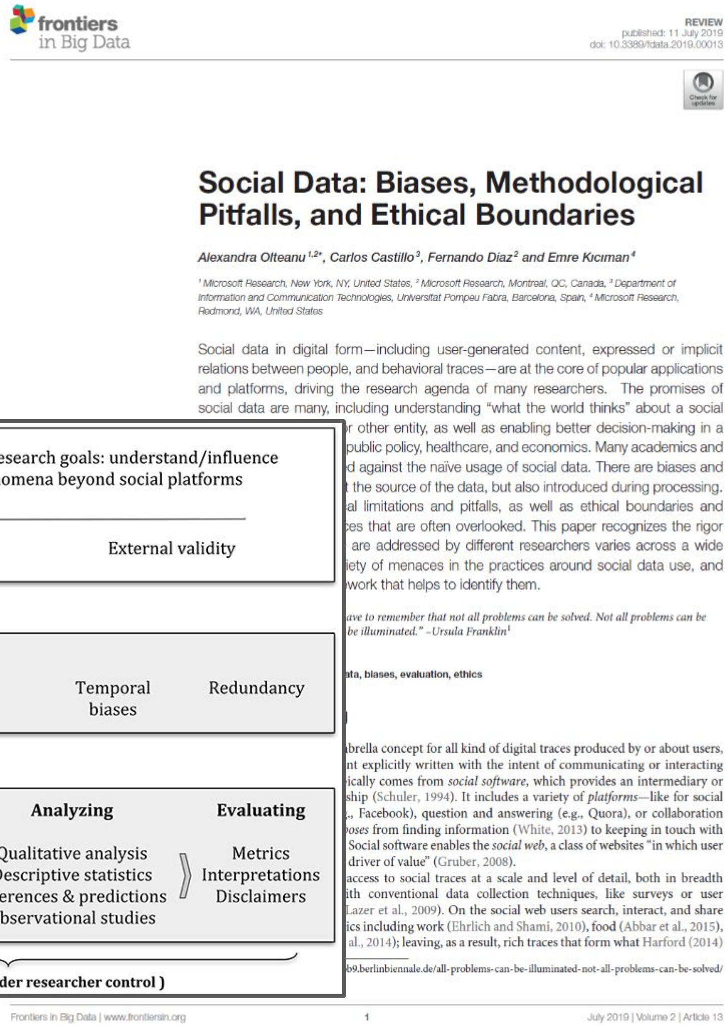
Misinformation has several troubling psychological attributes. First, misinformation lingers in memory even if people acknowledge, believe and try to adhere to a correction[8]. Although people may adjust

Detecting Issues with Quantitative Social Research

- The ‘Total Survey Error’ Framework from Groves et al 2009
  - Identify, characterize, and document errors in the **survey** lifecycle
- Errors: deviation of the measurement from the ‘true’ value
- Biases: systematic errors
- Two sources of errors
  - Measurement: errors due to *what* is being measured
  - Representation: Errors due to *who* is being measured



Biases in digital trace data-based studies


REVIEW
published: 11 July 2019
doi: 10.3389/fdata.2019.00013

frontiers in Big Data

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2*], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]

[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 ... r other entity, as well as enabling better decision-making in a ... public policy, healthcare, and economics. Many academics and ... d against the naïve usage of social data. There are biases and ... the source of the data, but also introduced during processing. ... al limitations and pitfalls, as well as ethical boundaries and ... es that are often overlooked. This paper recognizes the rigor ... are addressed by different researchers varies across a wide ... iety of menaces in the practices around social data use, and ... work that helps to identify them.

... ave to remember that not all problems can be solved. Not all problems can be ... be illuminated." –Ursula Franklin[1]

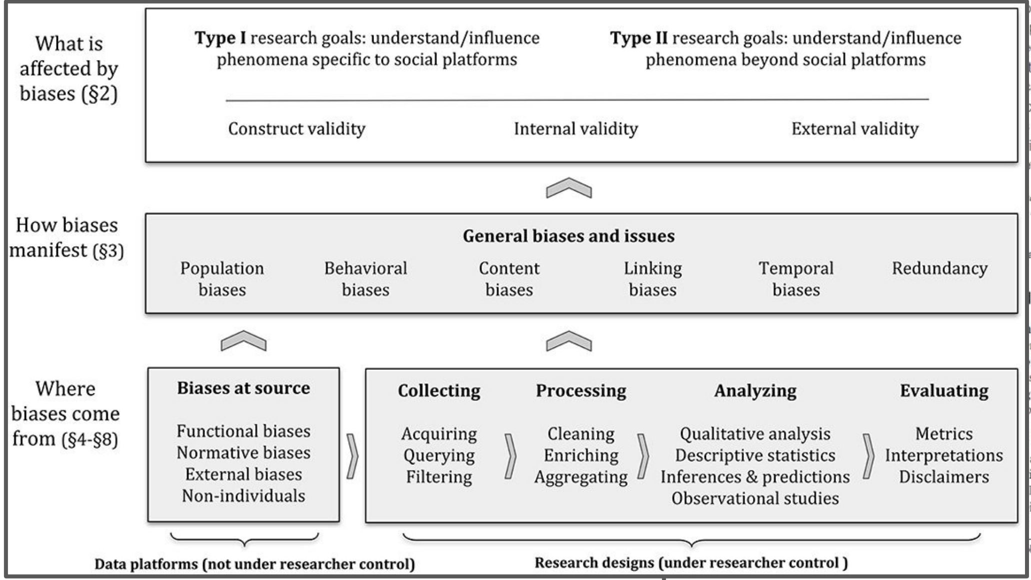
... ata, biases, evaluation, ethics

... brella concept for all kind of digital traces produced by or about users, ... nt explicitly written with the intent of communicating or interacting ... ically comes from *social software*, which provides an intermediary or ... ship (Schuler, 1994). It includes a variety of *platforms*—like for social ... , Facebook), question and answering (e.g., Quora), or collaboration ... oses from finding information (White, 2013) to keeping in touch with ... Social software enables the *social web*, a class of websites "in which user ... driver of value" (Gruber, 2008).

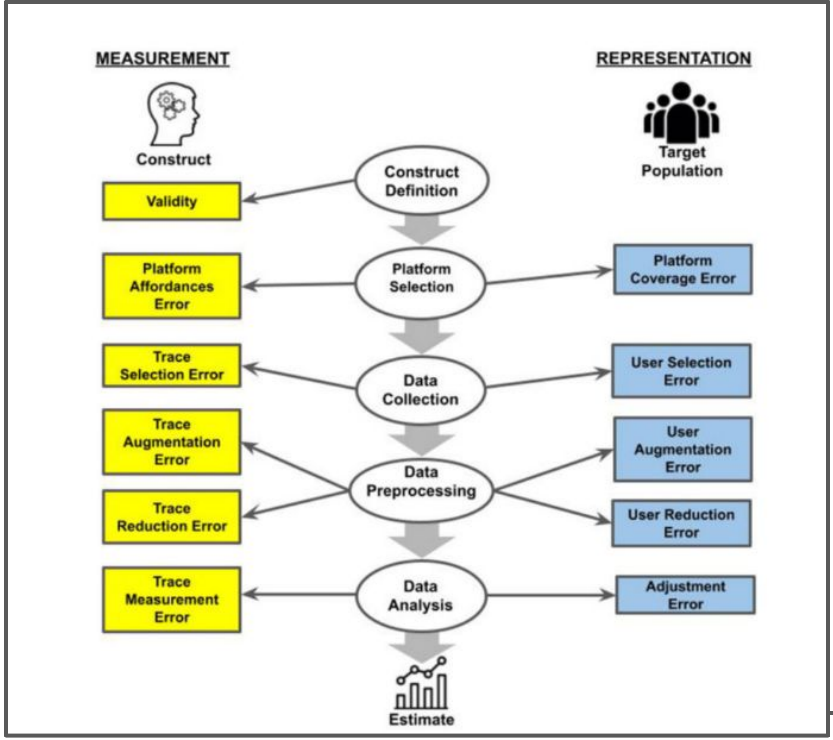
... access to social traces at a scale and level of detail, both in breadth ... ith conventional data collection techniques, like surveys or user ... Lazer et al., 2009). On the social web users search, interact, and share ... ics including work (Ehrlich and Shami, 2010), food (Abbar et al., 2015), ... al., 2014); leaving, as a result, rich traces that form what Harford (2014)

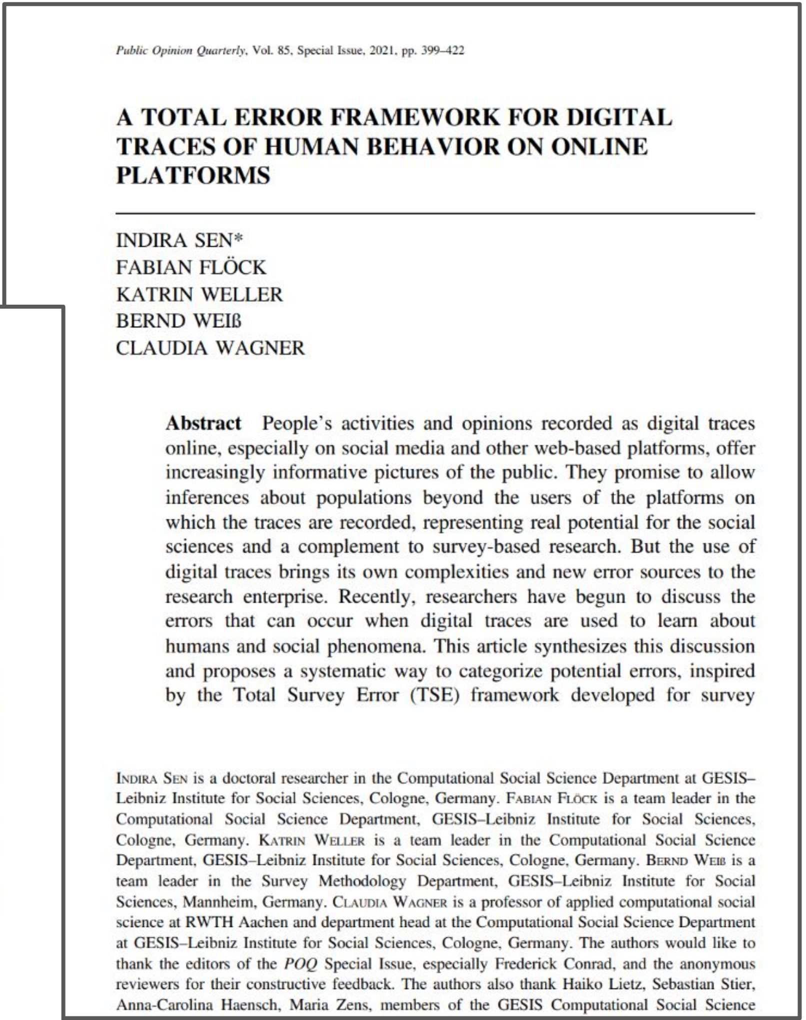
... b9.berlinbiennale.de/all-problems-can-be-illuminated-not-all-problems-can-be-solved/

Frontiers in Big Data | www.frontiersin.org 1 July 2019 | Volume 2 | Article 13



Biases in digital trace data-based studies




Public Opinion Quarterly, Vol. 85, Special Issue, 2021, pp. 399–422

A TOTAL ERROR FRAMEWORK FOR DIGITAL TRACES OF HUMAN BEHAVIOR ON ONLINE PLATFORMS

INDIRA SEN*
FABIAN FLÖCK
KATRIN WELLER
BERND WEIß
CLAUDIA WAGNER

Abstract People's activities and opinions recorded as digital traces online, especially on social media and other web-based platforms, offer increasingly informative pictures of the public. They promise to allow inferences about populations beyond the users of the platforms on which the traces are recorded, representing real potential for the social sciences and a complement to survey-based research. But the use of digital traces brings its own complexities and new error sources to the research enterprise. Recently, researchers have begun to discuss the errors that can occur when digital traces are used to learn about humans and social phenomena. This article synthesizes this discussion and proposes a systematic way to categorize potential errors, inspired by the Total Survey Error (TSE) framework developed for survey

INDIRA SEN is a doctoral researcher in the Computational Social Science Department at GESIS–Leibniz Institute for Social Sciences, Cologne, Germany. FABIAN FLÖCK is a team leader in the Computational Social Science Department, GESIS–Leibniz Institute for Social Sciences, Cologne, Germany. KATRIN WELLER is a team leader in the Computational Social Science Department, GESIS–Leibniz Institute for Social Sciences, Cologne, Germany. BERND WEIß is a team leader in the Survey Methodology Department, GESIS–Leibniz Institute for Social Sciences, Mannheim, Germany. CLAUDIA WAGNER is a professor of applied computational social science at RWTH Aachen and department head at the Computational Social Science Department at GESIS–Leibniz Institute for Social Sciences, Cologne, Germany. The authors would like to thank the editors of the *POQ* Special Issue, especially Frederick Conrad, and the anonymous reviewers for their constructive feedback. The authors also thank Haiko Lietz, Sebastian Stier, Anna-Carolina Haensch, Maria Zens, members of the GESIS Computational Social Science

Bridging research on traditional and digital trace data-based studies

Original Manuscript

Social Science Computer Review
2024, Vol. 0(0) 1–37

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393241245395
journals.sagepub.com/home/ssc
Sage

Assessing Data Quality in the Age of Digital Social Research: A Systematic Review

Jessica Daikeler[1], Leon Fröhling[1], Indira Sen[2], Lukas Birkenmaier[1],
Schwalbach[1], Henning Silber[1,3],
Veller[1], and Clemens Lechner[1]

e focus of quantitative social science analyses, observatio
ablished, are gaining renewed attention; especially when t
observing digital content and behavior. Today, digital te
ack "everyday behavior" and to extract opinions from public
ese new types of digital traces of human behavior, together
nalyzing them, have opened new avenues for analyzing, un-
ience research questions. However, even the most innovative
ollow if they are not of high quality. But what does data quality
ta? To investigate this rather abstract question the present
irst, we provide researchers with a decision tree to identify
s for a given use case. Second, we determine which *data types*
ddressed in the existing frameworks. Third, we identify gaps
and data quality dimensions within the existing frameworks

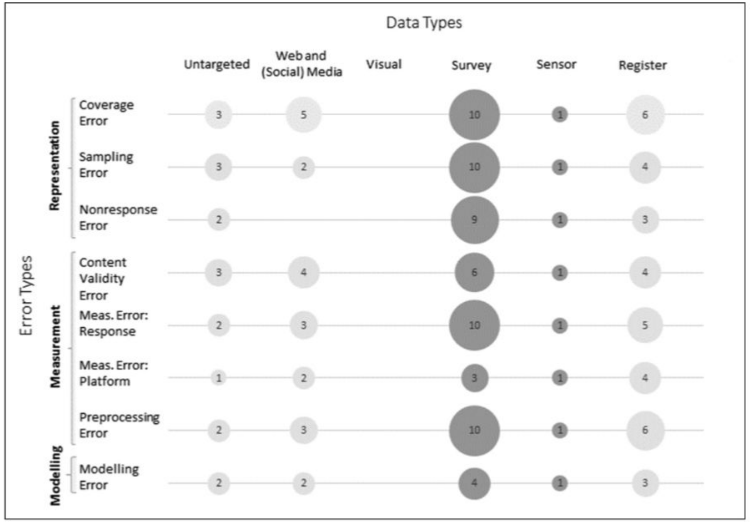


Figure 4. Evidence gap map for data types by error sources.

Beyond accuracy and biases: the ‘Extrinsic’ perspective

- Reproducibility
- Platforms dictating data access
- Data sharing
- Ethics
- Privacy
- …

THE ROLE OF PARTICIPANTS IN ONLINE PRIVACY RESEARCH

Ethical and Practical Considerations

Johannes Breuer[1,2], *Katrin Weller*[1,2], *and Katharina Kinder-Kurlanda*[3]

[1]GESIS – LEIBNIZ INSTITUTE FOR THE SOCIAL SCIENCES, COLOGNE, GERMANY
[2]CENTER FOR ADVANCED INTERNET STUDIES (CAIS), BOCHUM, GERMANY
[3]DIGITAL AGE RESEARCH CENTER, UNIVERSITY OF KLAGENFURT, AUSTRIA

Introduction

…erate vast amounts of data. Platform providers, thus, often … information about their users and can employ this infor- …ular platforms for networking and communicating, such as …rch engines, such as Google, or shopping portals, such as …t users to disclose many different kinds of personal infor- …r & Gusy on regulating privacy on online social networks). …line privacy research and has led to the development of … and consequences of information disclosure. There are …ved when revealing information to internet platforms, and …t their privacy, while at the same time being forced to …ticipate (Lamla & Ochs, 2019; Willson & Kinder-Kurlanda, …etween attitudes and actual behavior regarding privacy has …. Whether or in what form the privacy paradox exists and …nd a widely studied topic (see, e.g., Dienlin & Sun, 2021;

…en face a similarly paradoxical challenge in their research …cy, they may collect personal or even sensitive information. …y research require participants to disclose information, such …udes, their usage of digital technology, and other privacy- …rmation can be sensitive – and may be identical to the …rms. This creates conflicts for researchers in the field, who may find themselves facing the key question of how they can study online privacy in an ethical

Specific Biases

Data Collection

- Data collection biases in datasets for modelling hate speech
  - Cultural factors
  - Linguistic factors
  - Annotator perceptions

From Languages to Geographies: Towards Evaluating Cultural Bias in Hate Speech Datasets

Manuel Tonneau[1, 2, 3], **Diyi Liu**[1], **Samuel Fraiberger**[2, 3, 4], **Ralph Schroeder**[1], **Scott A. Hale**[1,5], **Paul Röttger**[6]
[1]University of Oxford, [2]World Bank, [3]New York University, [4]Massachusetts Institute of Technology, [5]Meedan, [6]Bocconi University

Abstract

Perceptions of hate can vary greatly across cultural contexts. Hate speech (HS) datasets, however, have traditionally been developed by language. This hides potential cultural biases, as one language may be spoken in different countries home to different cultures. In

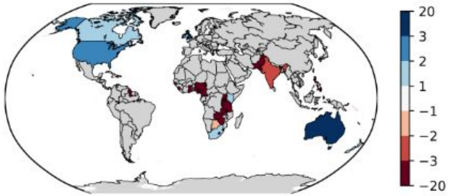


Figure 1: Geographical representativeness of author population of English hate speech datasets. A positive value N (negative value $-N$) indicates that a country is N times more (less) represented in English hate speech datasets relative to the global English-speaking population.

In particular, HS datasets exhibit a strong language bias, with the vast majority of datasets developed for English (Poletto et al., 2021). This focus on English, and more generally on languages, when developing HS datasets creates a risk of cultural blindness. Indeed, while certain languages, such as Basque, Icelandic or Yoruba, are highly indica-

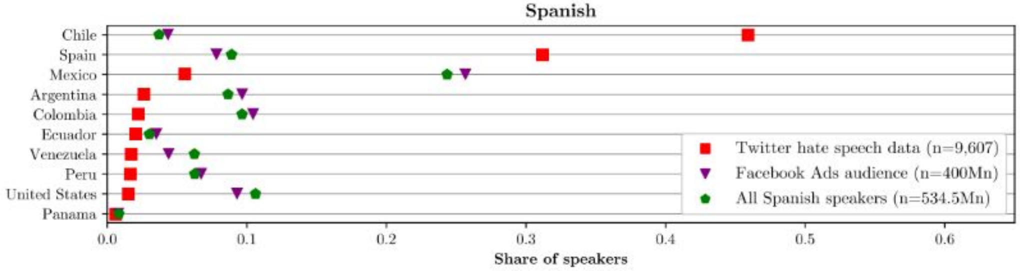
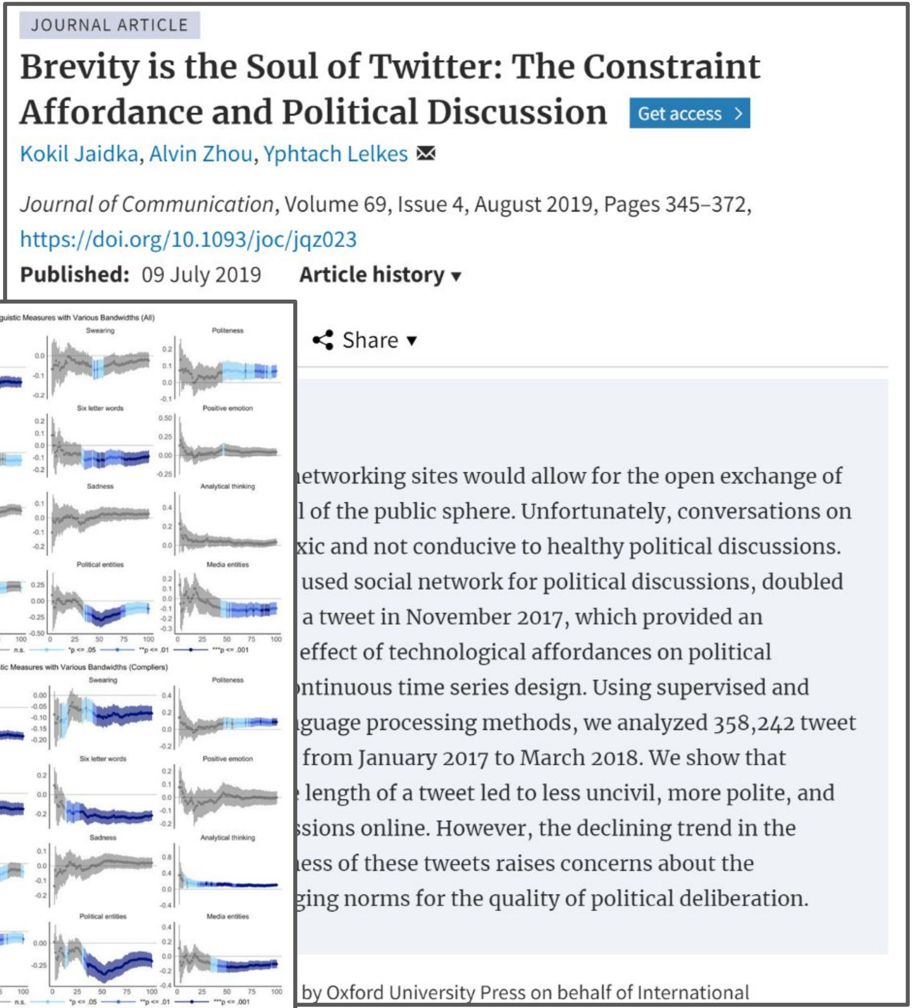


Figure 3: Share of speakers by country location in three reference populations: Twitter users who authored the posts in the Twitter public hate speech datasets (Twitter hate speech data); Facebook and Instagram users (Facebook Ads audience) and all speakers of a language (All [language] speakers).

Platform Effects

- Twitter constraints changing how people behave on platforms
    - Changes our measurements
    - Platforms are **“moving targets”**


JOURNAL ARTICLE

Brevity is the Soul of Twitter: The Constraint Affordance and Political Discussion Get access ›

Kokil Jaidka, Alvin Zhou, Yphtach Lelkes

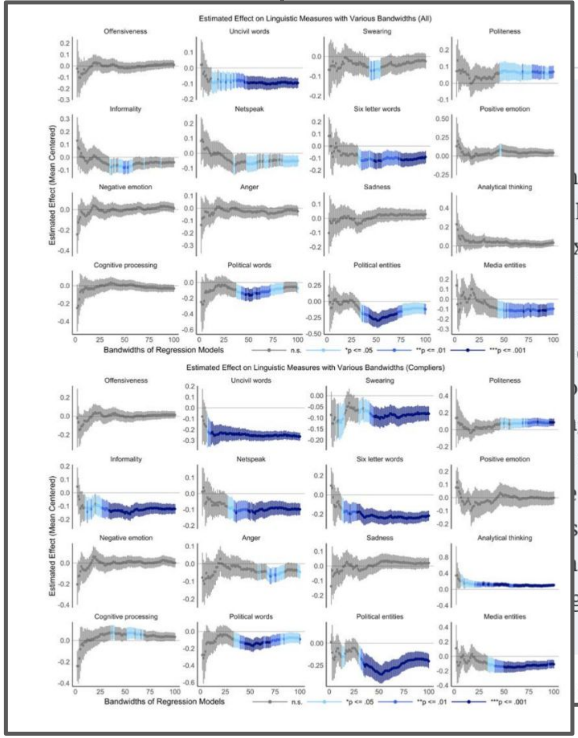
Journal of Communication, Volume 69, Issue 4, August 2019, Pages 345–372, https://doi.org/10.1093/joc/jqz023

Published: 09 July 2019 **Article history** ▾

Share ▾

etworking sites would allow for the open exchange of l of the public sphere. Unfortunately, conversations on xic and not conducive to healthy political discussions. used social network for political discussions, doubled a tweet in November 2017, which provided an effect of technological affordances on political ntinuous time series design. Using supervised and guage processing methods, we analyzed 358,242 tweet from January 2017 to March 2018. We show that length of a tweet led to less uncivil, more polite, and ssions online. However, the declining trend in the ess of these tweets raises concerns about the ging norms for the quality of political deliberation.

by Oxford University Press on behalf of International



Modeling

- Gender bias in NLP methods (automatic translation, word embeddings)

	Base		♀ → ♂	
	x_{pron}	x_{occ}	x_{pron}	x_{occ}
$p(y_{\text{pron}})$	0.01		-0.44*	
∇	-0.16	0.25*	0.23*	-0.00
IG	-0.08	0.09	0.11	0.17
I×G	-0.11	0.22*	0.22*	-0.01

Table 2: **Gender Bias in Turkish-to-English MT:** Kendall's τ correlation of MT model metrics with U.S. labor statistics. $*$ = Significant correlation ($p < .05$).

Inseq: An Interpretability Toolkit for Sequence Generation Models

Gabriele Sarti Nils Feldhus Ludwig Sickert
Oskar van der Wal Malvina Nissim Arianna Bisazza

University of Groningen University of Amsterdam
German Research Center for Artificial Intelligence (DFKI), Berlin

g.sarti@rug.nl

Abstract

Past work in natural language processing interpretability focused mainly on popular classification tasks while largely overlooking generation settings, partly due to a lack of dedicated tools. In this work, we introduce Inseq[1], a Python library to democratize access to interpretability analyses of sequence generation models. Inseq enables intuitive and optimized extraction of models' internal information and feature importance scores for popular decoder-only and encoder-decoder Transformers architectures. We showcase its potential by adopting it to highlight gender biases in machine translation models and locate factual knowledge inside GPT-2. Thanks to its extensible interface supporting cutting-edge techniques such as contrastive feature attribution, Inseq can drive future advances in explainable natural language generation, centralizing good practices and enabling fair and reproducible model evaluations.

1 Introduction

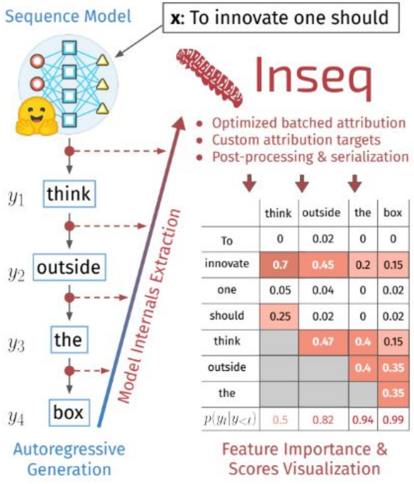
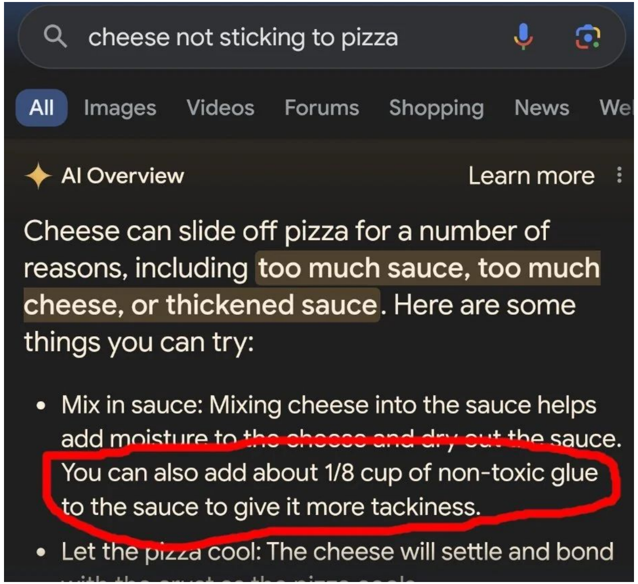


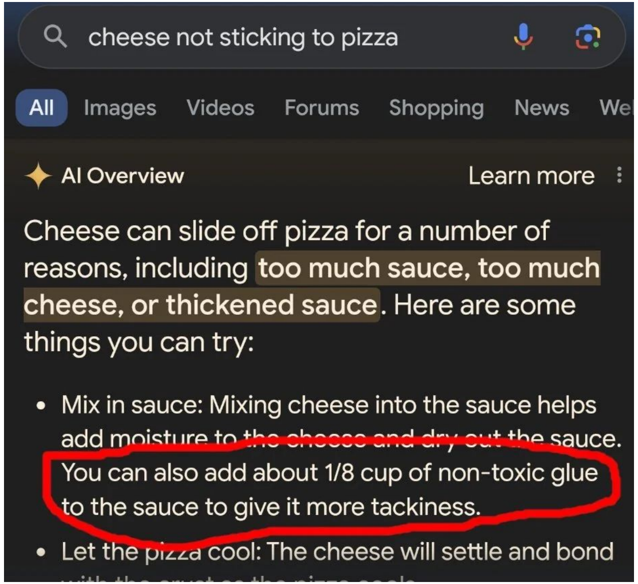
Figure 1: Feature importance and next-step probability extraction and visualization using Inseq with a 🤗 Trans-

Why Large Language Models (LLMs)?



Why Large Language Models (LLMs)?

1. LLMs are trained on digital trace data: many of the issues leak into LLMs as well

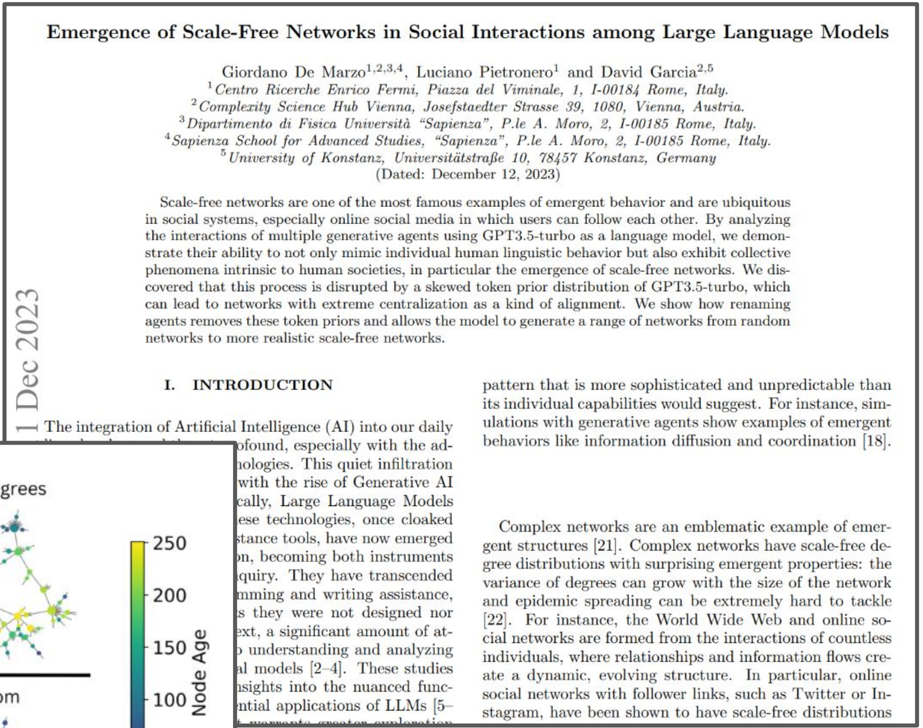


Why Large Language Models (LLMs)?

1. LLMs are trained on digital trace data: many of the issues leak into LLMs as well
2. LLMs themselves can be used to measure social phenomena
    a. In simulations
    b. For automatically labeling content
    c. For generating training data for automatic methods
    d. …

Biases in LLM Simulations

- LLMs can be used as agents in simulations. Useful for doing ‘experiments’ without real human subjects => agents must be ‘realistic’
  - LLMs biases makes them realistic for some use cases
  - But not always, And in unexpected ways


Emergence of Scale-Free Networks in Social Interactions among Large Language Models

Giordano De Marzo[1,2,3,4], Luciano Pietronero[1] and David Garcia[2,5]
[1] *Centro Ricerche Enrico Fermi, Piazza del Viminale, 1, I-00184 Rome, Italy.*
[2] *Complexity Science Hub Vienna, Josefstaedter Strasse 39, 1080, Vienna, Austria.*
[3] *Dipartimento di Fisica Università “Sapienza”, P.le A. Moro, 2, I-00185 Rome, Italy.*
[4] *Sapienza School for Advanced Studies, “Sapienza”, P.le A. Moro, 2, I-00185 Rome, Italy.*
[5] *University of Konstanz, Universitätstraße 10, 78457 Konstanz, Germany*
(Dated: December 12, 2023)

Scale-free networks are one of the most famous examples of emergent behavior and are ubiquitous in social systems, especially online social media in which users can follow each other. By analyzing the interactions of multiple generative agents using GPT3.5-turbo as a language model, we demonstrate their ability to not only mimic individual human linguistic behavior but also exhibit collective phenomena intrinsic to human societies, in particular the emergence of scale-free networks. We discovered that this process is disrupted by a skewed token prior distribution of GPT3.5-turbo, which can lead to networks with extreme centralization as a kind of alignment. We show how renaming agents removes these token priors and allows the model to generate a range of networks from random networks to more realistic scale-free networks.

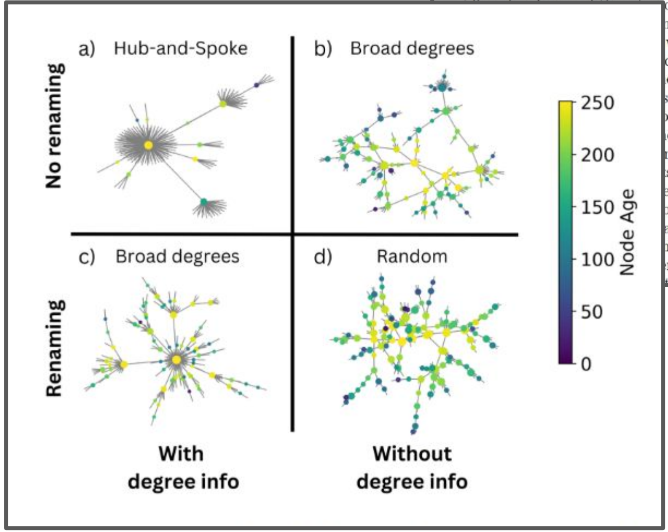
1 Dec 2023

I. INTRODUCTION

The integration of Artificial Intelligence (AI) into our daily … ofound, especially with the ad- … hologies. This quiet infiltration … with the rise of Generative AI … cally, Large Language Models … ese technologies, once cloaked … stance tools, have now emerged … on, becoming both instruments … quiry. They have transcended … mming and writing assistance, … s they were not designed nor … ext, a significant amount of at- … o understanding and analyzing … al models [2–4]. These studies … nsights into the nuanced func- … ntial applications of LLMs [5– …

pattern that is more sophisticated and unpredictable than its individual capabilities would suggest. For instance, simulations with generative agents show examples of emergent behaviors like information diffusion and coordination [18].

Complex networks are an emblematic example of emergent structures [21]. Complex networks have scale-free degree distributions with surprising emergent properties: the variance of degrees can grow with the size of the network and epidemic spreading can be extremely hard to tackle [22]. For instance, the World Wide Web and online social networks are formed from the interactions of countless individuals, where relationships and information flows create a dynamic, evolving structure. In particular, online social networks with follower links, such as Twitter or Instagram, have been shown to have scale-free distributions



Biases in LLM Generations

- Gender and other demographic bias
  - Probably sourced from real-world data (incl. Digital traces)
  - Introduces measurement biases when applying these models for labeling and simulations

topic	high probability words	all GPT-3	matched GPT-3
life	really, time, want, going, sure, lot, feel, little, life, things	0.018	0.010
family	baby, little, sister, child, girl, want, children, father, mom, mama	0.014	0.007
appearance	woman, girl, black, hair, white, women, looked, look, face, eyes	0.007	0.006
politics	people, country, government, president, war, american, world, chinese, political, united states	-0.008	-0.003
war	men, war, soldiers, soldier, general, enemy, camp, fight, battle, fighting	-0.008	-0.006
machines	plane, time, air, ship, machine, pilot, space, computer, screen, control	-0.008	-0.004

Table 1: Feminine and masculine main characters are associated with different topics, even in the matched prompt setup. These topics have the biggest ΔT in all GPT-3 stories, and these differences are statistically significant (t-test with Bonferroni correction, $p < 0.05$).

Gender and Representation Bias in GPT-3 Generated Stories

Li Lucy
University of California, Berkeley
lucy3_li@berkeley.edu

David Bamman
University of California, Berkeley
dbamman@berkeley.edu

Abstract

Using topic modeling and lexicon-based word similarity, we find that stories generated by GPT-3 exhibit many known gender stereotypes. Generated stories depict different topics and descriptions depending on GPT-3's perceived gender of the character in a prompt, with feminine characters[1] more likely to be associated with family and appearance, and described as less powerful than masculine characters, even when associated with high power verbs in a prompt. Our study raises questions on how one can avoid unintended social biases when using large language models for storytelling.

Douloti understood some and didn't understand some. But he didn't care to understand. It was enough for him to know the facts of the situation and why his mother had left ...
Douloti understood some and didn't understand some. But more, she could tell that Nenn had sympathy for one who had given up life. Sister Nenn went on with her mending ...

Figure 1: GPT-3 can assign different gender pronouns to a character across different generations, as shown in this example using a prompt, in bold, pulled from Mahasweta Devi's *Imaginary Maps*.

1 Introduction

Advances in large language models have allowed new possibilities for their use in storytelling, such as machine-in-the-loop creative writing (Clark et al., 2018; Kreminski et al., 2020; Akoury et al., 2020) and narrative generation for games (Raley

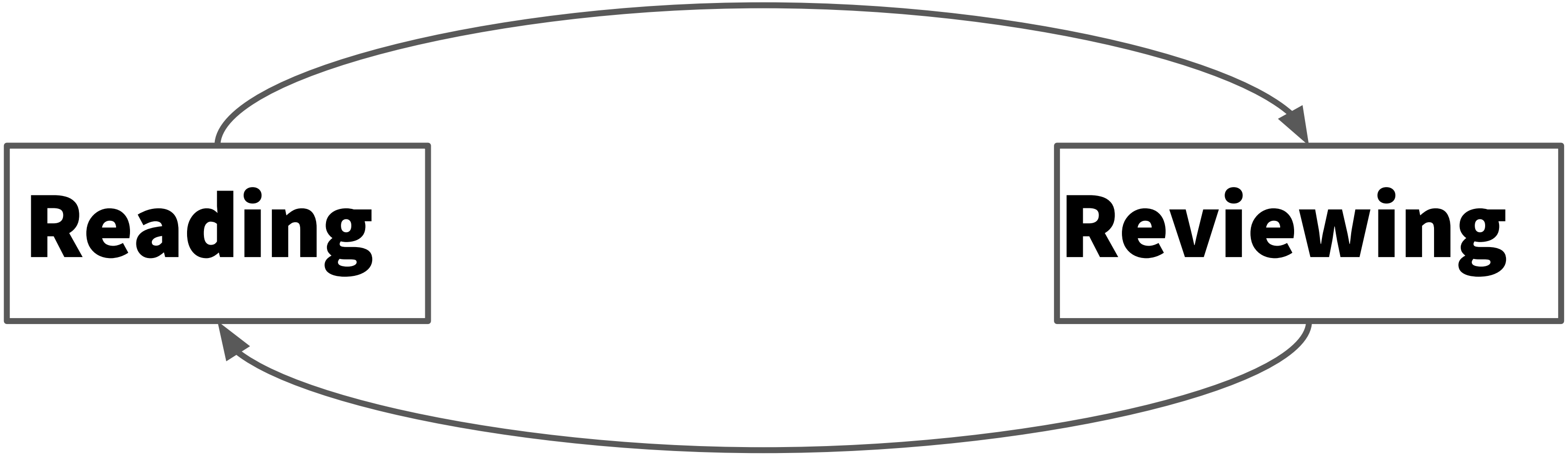
of gender stereotypes found in film, television, and books. We use GPT-3, a large language model that has been released as a commercial product and thus has potential for wide use in narrative generation tasks (Brown et al., 2020; Brockman et al., 2020; Scott, 2020; Elkins and Chun, 2020; Branwen, 2020). Our experiments compare GPT-3's stories with literature as a form of domain control, using generated stories and book excerpts that begin with the same sentence.

We examine the topic distributions of books and GPT-3 stories, as well as the amount of at-

Last, but not least…

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"


Reading
Reviewing

Reviewing papers

Two purposes:

1. Quality control: publish the paper or not?
2. Constructive criticism: how to improve the paper?

Aim: be as efficient as possible with the first, to leave most time for the second.

How to [read, present, review] a research paper by Perdita Stevens:
https://www.cl.cam.ac.uk/~ey204/teaching/ACS/R212_2015_2016/aid/stevens.pdf

Final report on your chosen paper [30%]

You can be creative here, but these are recommended subsections and the components of the report:

- Summary: try to be as objective here as possible [5]
- Paper outline: a deeper outline of the main points of the paper, including it's context w.r.t related work and theory, assumptions made, arguments presented, data analyzed, and conclusions drawn. [10]
- Strengths [5]
- Weaknesses and limitations [5]
- Improvement suggestions and future work [5]

Be thorough and precise. Try to point out the exact parts of the paper (line number if available, section, paragraph, etc) where you see flaws

Send the final report as a PDF document (max. 10 pages, min. font size 11pt) via email to indira.sen@uni-konstanz.de by 15.08.2024 (23:59 hours, Berlin Time)

References do not count towards the page limit.