

gesis

Leibniz Institute
for the Social Sciences



Potentials and Pitfalls of Social Media Data

Indira Sen & Katrin Weller

9. Recap & Conclusions



Photo: Katrin Weller

Workshop goals

- provide an overview on current approaches in research **based on digital traces from social media**
- outline different steps in the research process when working with social media data, and provide **practical examples** for data collection, cleaning and analysis
- offer a **structured approach** to think about potential pitfalls and error sources in social media research, that can help to design, present, talk about research approaches

Workshop goals

- provide an overview on current approaches in research **based on digital traces from social media**
- outline different steps in the research process when working with social media data, and provide **practical examples** for data collection, cleaning and analysis
- offer a **structured approach** to think about potential pitfalls and error sources in social media research, that can help to design, present, talk about research approaches

Social media research...

- offers new opportunities for studying humans **attitudes, behaviour, characteristics**
- can complement other research methods
- is happening across disciplines
- may focus on specific platforms as model organisms
- may provide insights into our “online lives”

Social media research...

- is depending on access to platform data
- faces many challenges of ephemeral data:
 - platforms change
 - user behaviour may change
 - data access options change

Workshop goals

- provide an overview on current approaches in research based on digital traces from social media
- outline different steps in the research process when working with social media data, and provide **practical examples** for data collection, cleaning and analysis
- offer a **structured approach** to think about potential pitfalls and error sources in social media research, that can help to design, present, talk about research approaches

Hands-on examples

- During several hands-on examples we have illustrated practical details of social media data collection, preprocessing and analysis.
- Reusable and editable examples via notebooks.



hands-on
examples

Hands-on examples

Collection

- Different ways to access social media data - example reddit API.
- Access options may influence what data can be studied.
- Also important: data selection criteria (e.g. which subreddit, which time period).

Hands-on examples

(Pre-)processing

- We looked at different ways, in which textual data can be prepared for analysis.
- typical steps include tokenization, removal of stopwords, lemmatization.
- Different tools and programmatic resources can support data cleaning.
- Choices for processing steps and tools can influence the research results.

Hands-on examples

Analysis

- For our exemplary case of analysing reddit data, we chose **sentiment** analysis, augmentation with **toxicity** scores, and **topic modelling**.
- Different approaches may be needed for different types of data (e.g. multimedia, networks).

Workshop goals

- provide an overview on current approaches in research **based on digital traces from social media**
- outline different steps in the research process when working with social media data, and provide **practical examples** for data collection, cleaning and analysis
- offer a **structured approach** to think about potential pitfalls and error sources in social media research, that can help to design, present, talk about research approaches

Identifying, mitigating and documenting pitfalls

- Typical pitfalls exist for all phases of social media research.
- We present a structured way to think about these potential pitfalls along the typical research workflow.
- Our framework is inspired by similar approaches in survey research.

Identifying, mitigating and documenting pitfalls

Identifying errors

- We distinguish representation and measurement errors as two main sources for pitfalls.
- We have illustrated errors along the research workflow based on exemplary research designs.
- We have also seen a hands on example on stance vs. sentiment detection.

Identifying, mitigating and documenting pitfalls

mitigating and documenting errors

- We have looked at potential remedies for selected errors, including augmentation and reduction strategies, and reweighting.
- We have worked with specification sheets to document pitfalls.

Look on the bright side

- Social media data has several advantages compared to surveys:
 - rich, large-scale, multimodal, high-resolution
 - spans long time frames
 - avoid response and recall biases in surveys
 - data access can be immediate after an important event instead of a lag
- Error frameworks can be generative in trying to understand gaps in study designs, where we need transparency, and brainstorming alternatives

Missing pieces - not covered in this workshop

- Research ethics
- Additional types of data (multimedia)
- Ways to measure errors
- Combination of survey and social media data

Some of these have been addressed in our Meet the Experts season on CSS and Digital Behavioural Data:

<https://www.gesis.org/en/services/sharing-knowledge/consulting-and-guidelines/meet-the-experts>

Missing pieces

Please also see GESIS' additional training activities for specialized topics (workshops, summer school etc.)

Cannot find something that matches your need?

Please email us -

we are currently still working on the courses for next year.

Open Questions?

- Indira: indira.sen@gesis.org, @indiiigosky
- Katrin: katrin.weller@gesis.org, @kwelle

Thanks for participating - and greetings from Cologne



Photo: Katrin Weller

Resources from this workshop

Materials will be shared via Github:

https://github.com/Indiiigo/social_media_data_research_2022

This includes:

- **Slides**
- List of **references** at the end of the slides
- **Notebooks** with examples of code that can be executed

References

Digital Traces and works exploring them

- Salganik, Matthew J. Bit by Bit: Social Research in the Digital Age. Princeton, NJ: *Princeton University Press. Open review edition* (2017).
- O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Fourth international AAAI conference on weblogs and social media* (2010).
- Gilbert, Eric, and Karrie Karahalios. "Predicting tie strength with social media." *Proceedings of the SIGCHI conference on human factors in computing systems* (2009).
- Merullo, Jack, et al. "Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts." *arXiv preprint arXiv:1909.03343* (2019).
- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21.3 (2013): 267-297.
- Japek, Lilli, et al. "Big data in survey research: AAPOR task force report." *Public Opinion Quarterly* 79.4 (2015): 839-880.
- Stier, Sebastian, et al. "Integrating survey data and digital trace data: key issues in developing an emerging field." (2019): 0894439319843669.

References

Challenges of Digital Traces

- Olteanu, Alexandra, et al. "Social data: Biases, methodological pitfalls, and ethical boundaries." *Frontiers in Big Data* 2 (2019): 13.
- Howison, James, Andrea Wiggins, and Kevin Crowston. "Validity issues in the use of social network analysis with digital trace data." *Journal of the Association for Information Systems* 12.12 (2011): 2.
- Tufekci, Zeynep. "Big questions for social media big data: Representativeness, validity and other methodological pitfalls." *Eighth International AAAI Conference on Weblogs and Social Media* (2014).
- Metaxas, Panagiotis T., Eni Mustafaraj, and Dani Gayo-Avello. "How (not) to predict elections." *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE (2011).
- Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency* (2018).
- Snijders, C., Matzat, U., & Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet. *International Journal of Internet Science*, 7, 1-5. Retrieved from http://www.ijis.net/ijis7_1/ijis7_1_editorial.html
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203-1205.
- Boyd, Danah, & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>

References

Better transparency and documentation

- Hsieh, Yuli Patrick & Murphy, Joe. "Total twitter error." *Total survey error in practice* (2017): 23-46.
- Gebru, Timnit, et al. "Datasheets for datasets." *arXiv preprint arXiv:1803.09010* (2018).
- Mitchell, Margaret, et al. "Model cards for model reporting." *Proceedings of the conference on fairness, accountability, and transparency* (2019).

Research ethics

- Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1), 205630511876336. <https://doi.org/10.1177/2056305118763366>
- Franzke, A., Bechmann, A., Zimmer, M., Ess, C. and the Association of Internet Researchers (2020). *Internet Research: Ethical Guidelines 3.0*. <https://aoir.org/reports/ethics3.pdf>
- Zimmer, M., & Kinder-Kurlanda, K. (Hrsg.). (2017). *Internet research ethics for the social age: New challenges, cases, and contexts*. Peter Lang.

Data Access

- Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544–1566. <https://doi.org/10.1080/1369118X.2019.1637447>
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. <http://arxiv.org/abs/1306.5204>

References

Potential Remedies

→ Validity

- Joseph, Kenneth, et al. "Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-Based Measures." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. No. 01 (2019).
- Diaz, Fernando, et al. "Online and social media data as an imperfect continuous panel survey." *PloS one* 11.1 (2016).

→ Platform Selection

- Malik, Momin M., and Jürgen Pfeffer. "Identifying platform effects in social media data." *Tenth International AAAI Conference on Web and Social Media* (2016).
- Weber, Ingmar. "Demographic research with non-representative internet data." *International Journal of Manpower* 36.1 (2015): 13-25.

→ Data Collection

- Ruiz, Eduardo J., Vagelis Hristidis, and Panagiotis G. Ipeirotis. "Efficient filtering on hidden document streams." *Eighth International AAAI Conference on Weblogs and Social Media* (2014).

References

Potential Remedies

→ Data Collection (contd...)

- Linder, Fridolin. "Improved data collection from online sources using query expansion and active learning." *Available at SSRN 3026393* (2017).
- Wang, Zijian, et al. "Demographic inference and representative population estimates from multilingual social media data." *The World Wide Web Conference* (2019).

→ Data Preprocessing

- Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26.2 (2018): 168-189.
- González-Bailón, Sandra, and Georgios Paltoglou. "Traces of public opinion in online communication: A comparison of methods and data sources." *The ANNALS of the American Academy of Political and Social Science* 659.1 (2015): 95-107.
- Karimi, Fariba, et al. "Inferring gender from names on the web: A comparative evaluation of gender detection methods." *Proceedings of the 25th International Conference Companion on World Wide Web* (2016).

References

Potential Remedies

→ Data Preprocessing (contd...)

- Keyes, Os. "The misgendering machines: Trans/HCI implications of automatic gender recognition." *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018): 1-22.

→ Data Analysis

- Ahmed, Saifuddin, Kokil Jaidka, and Marko M. Skoric. "Tweets and votes: A four-country comparison of volumetric and sentiment analysis approaches." *Tenth International AAAI Conference on Web and Social Media* (2016).
- Li, Lingling, et al. "On weighting approaches for missing data." *Statistical methods in medical research* 22.1 (2013): 14-30.

→ Presidential Approval

- Gallup, George, ed. *The gallup poll: public opinion 1993*. Rowman & Littlefield, 1994.
- O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Fourth international AAAI conference on weblogs and social media* (2010).
- Pasek, Josh, et al. "Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces?." *Social Science Computer Review* (2019): 0894439318822007.

References

Documentation approaches and critical reflections

- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018).
- Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." Transactions of the Association for Computational Linguistics 6 (2018): 587-604.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In Proceedings of the conference on fairness, accountability, and transparency, pp. 220-229. 2019.
- Jacobs, Abigail Z., and Hanna Wallach. "Measurement and fairness." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 375-385. 2021.
- Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social data: Biases, methodological pitfalls, and ethical boundaries." Frontiers in Big Data 2 (2019): 13.
- Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33-44. 2020.
- Hsieh, Y. P., & Murphy, J. (2017). *Total twitter error* (Vol. 74, pp. 23-46). Hoboken, NJ, USA: John Wiley & Sons,.
- Amaya, Ashley, Paul P. Biemer, and David Kinyon. "Total error in a big data world: Adapting the TSE framework to big data." *Journal of Survey Statistics and Methodology* 8, no. 1 (2020): 89-119.

Images Used in this Tutorial

Images used in TED-On diagram

→ designed by Becris, EliasBikbulatov and Pixel perfect from www.flaticon.com