«Critical Social Media Analysis using Mixed Methods»

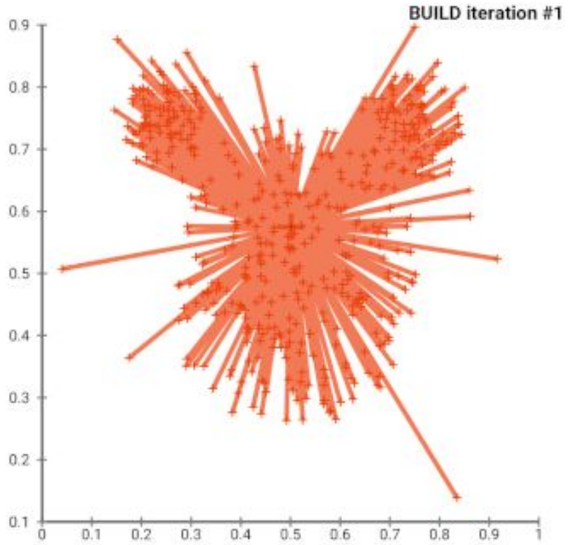
Combining ML and interpretivist approaches to social media data analysis

Dr. Simon David Hirsbrunner, Michael Tebbe
Human-Centered Computing, Institute of Computer Science
Freie Universität Berlin

Session IV, 10 Dec 2020

Recap last session

- Clustering in „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Tv Discourse“.
- Clustering in general
- K-Medoids
  -> Clustering has an interpretative component



Sanders, Abraham, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S Erickson, und Kristin P Bennett. „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Twitter Discourse“. Preprint. Health Informatics, 1. September 2020. https://doi.org/10.1101/2020.08.28.20183863.

Plan for today

- Making the case for combining ML and interpretivist approaches
  - Interdisciplinary cooperation
  - Opportunities for social sciences and humanities
  - Opportunities for computer and data science
- How to combine approaches
  - Perspectives from computer science
  - Perspectives from an interpretivist stance
- Assignments
- Group work (Discord)

Seminar progress / today

Seminar

Theory / Methodology
Critical Data Science

Research questions	**Collection**	**Exploration**	**Machine Learning** language models / clustering / visualization		**Paper writing**
			Interpretative analysis	**Focused analysis and selected aspects**	
online videos + user debates **Material**	digital traces of social interaction **Data**		cleaned datasets, categorized data **Data aggregation**	categories and theoretical elements **Theory building**	

interdisciplinary collaboration and combining approaches

- Inter- trans- and postdisciplinary research
- Parallel, master-servant or blended constellations of interdisciplinary work
- True interdisciplinary work takes time, sometimes fails, requires development of interdisciplinary expertise
- So why do it anyway?

(on modes of interdisciplinary collaboration, see: Barry et al. 2008)

Combining approaches: opportunities for the social sciences and humanities? (I)

Faster and better through technology?

- not a valid argument in itself
- ‘interpretativist approaches’ are already highly mediated by (taylored) technology

Big Data?

- volume, velocity, variety, veracity
- exhaustiveness, resolution, indexicality, relationality, flexibility

(boyd and Crawford, 2012; Kitchin 2013; 2014; Mayer-Schonberger and Cukier, 2013)

- but why should we consider this data?

Combining approaches: opportunities for the social sciences and humanities? (II)

Social life and culture is mediated through digital dat and technologies

Methodological considerations

- Consider the many mediations and distortions in digital life
- Formalized exploration, filtering and ordering of research material

New perspectives

- Instrumentarium enables discovery and addressing new research questions
- ML is a new lense / perspective to look at data

(Baumer et al. 2017)

Image source: https://www.tableau.com/about/blog/2020/4/how-data-culture-can-help-unify-time-crisis

Combining approaches: opportunities for computer and data science?

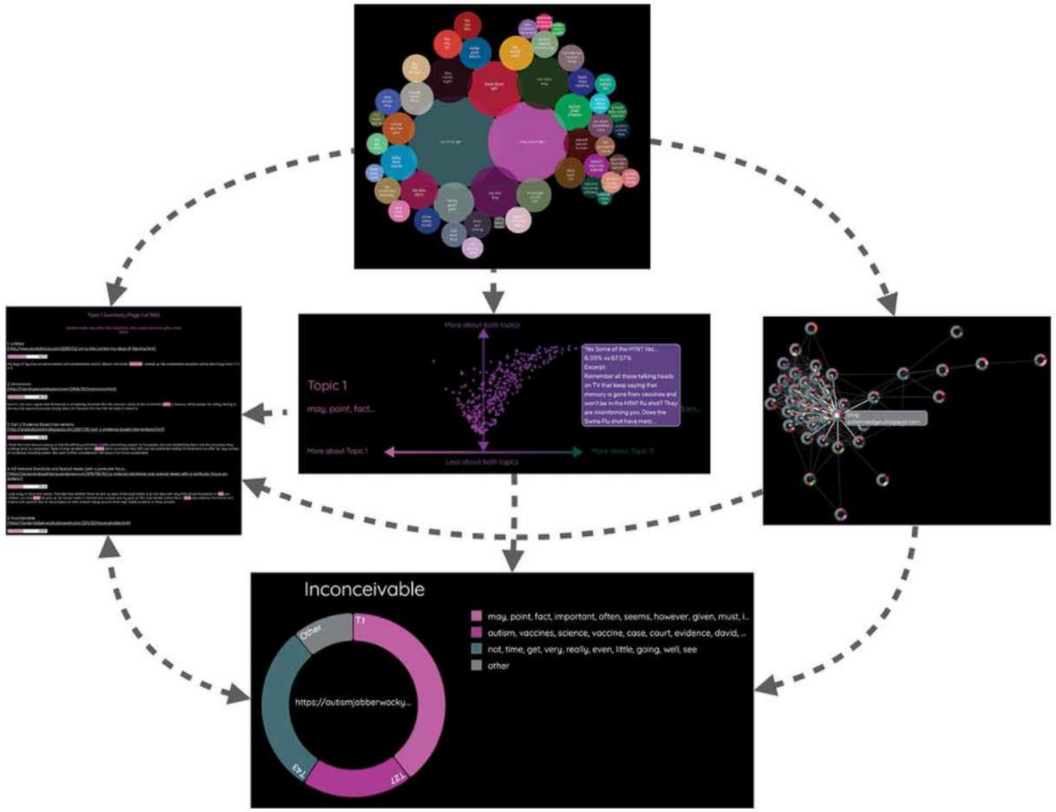
- Insights informed by state-of-the art theory about society (social sciences), culture (humanities) or other domains
- Discover biases in data (e.g. discussion on fairness in ML) (Lepri et al. 2018; Selbst et al. 2019)
- Create more meaningful data (Brooks et al., 2013) and real-world insights
- More user- and human-centered design of technology

→ Human-Centered Machine Learning (Gillies et al. 2016)

Short break: 5 Minutes

How to combine Social Sciences and ML?computer science perspective

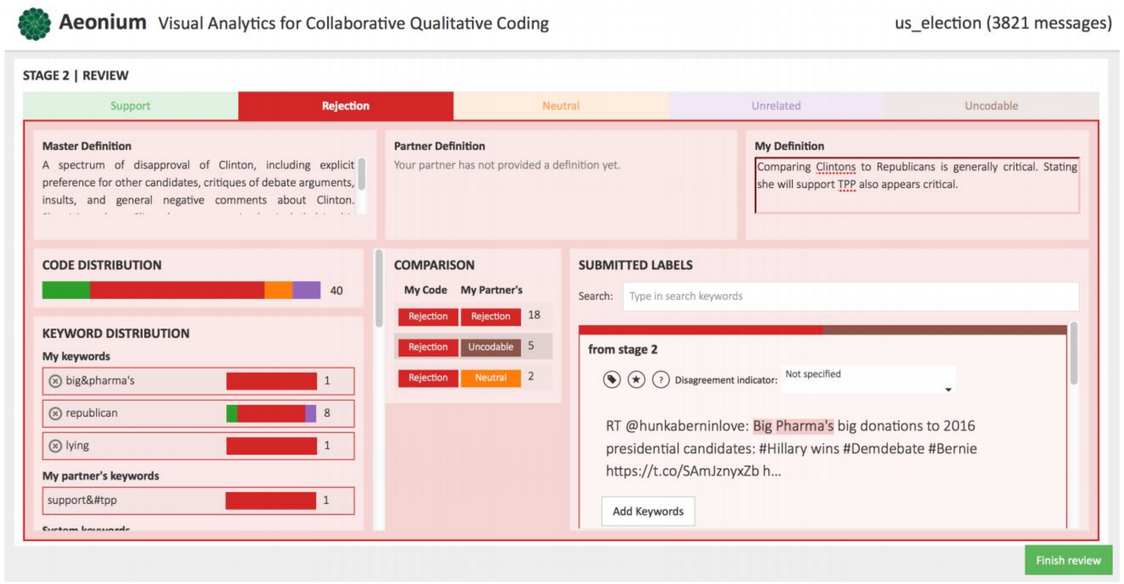
- Topicalizer (Baumer et al. 2020)
- Co-designed with interpretivist scholars
- Visual analytics tool



Eric P. S. Baumer, Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul & Micki McGee (2020) Topicalizer: reframing core concepts in machine learning visualization by co-designing for interpretivist scholarship, Human–Computer Interaction, 35:5-6, 452-480, DOI: 10.1080/07370024.2020.1734460

How to combine Social Sciences and ML?computer science perspective

- Aeonum: A visual analytics tool for qualitative coding
- highlighting ambiguity between qualitative coders (Chen et al., 2018)



Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. ACM Trans. Interact. Intell. Syst. 8, 2, Article 9 (July 2018), 20 pages. DOI:https://doi.org/10.1145/3185515

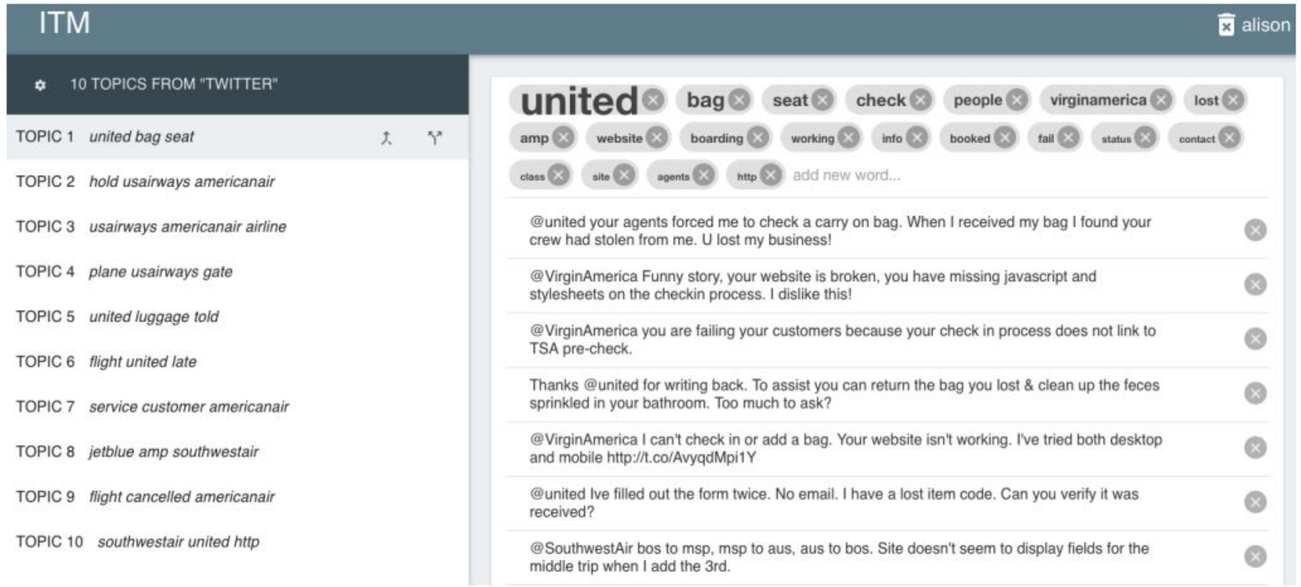
How to combine Social Sciences and ML? computer science perspective

- Human-in-the-loop topic modeling
- RQ: How do humans employ refinements to improve a topic model?

-> Interpretability

-> Control

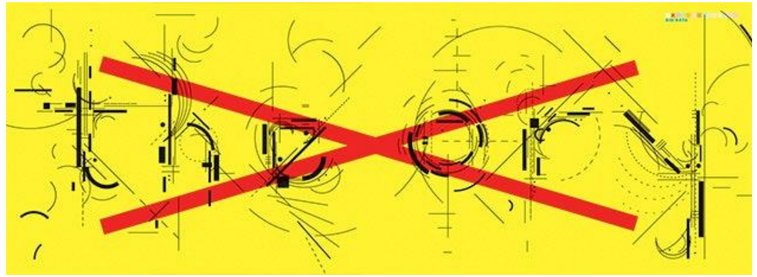
-> Latency



Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In 23rd International Conference on Intelligent User Interfaces (IUI '18). Association for Computing Machinery, New York, NY, USA, 293–304. DOI:https://doi.org/10.1145/3172944.3172965

Is data enough? or ‘the end of theory’

“Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.”
(Anderson 2008)



Assumptions of data-driven research

- **Big Data can capture a whole domain** and provide full resolution;
- there is **no need for** a priori theory, models or hypotheses;
- through the application of agnostic data analytics the **data can speak for themselves** free of human bias or framing, and any patterns and relationships within **Big Data are inherently meaningful and truthful**;
- **meaning transcends context or domain-specific knowledge**, thus can be interpreted by anyone who can decode a statistic or data visualization.
(Kitchin 2014)

Image source: https://www.wired.com/2008/06/pb-theory/

Big Data critique

“In reality, working with Big Data is still **subjective**, and what it quantifies does not necessarily have a closer claim on objective truth – **particularly when considering messages from social media sites.**”

“Big Data enables the practice of **apophenia**: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions.”

“**Interpretation is at the center of data analysis.** Regardless of the size of a data, it is subject to limitation and bias. Without those biases and limitations being understood and outlined, misinterpretation is the result. Data analysis is most effective when researchers take account of the complex methodological processes that underlie the analysis of that data.”

(boyd and Crawford 2012)

“Data (re-)constructivism”

> “**reality**, as well as our **knowledge** thereof, are **social products** and hence incapable of being understood independent of the social actors (including the researchers) that construct (...) and make sense of that reality” (Orlikowski and Baroudi 1991, cited in Baumer et al. 2017)

Social constructivism
Thomas theorem: “If men define situations as real, they are real in their consequences” (Thomas and Thomas 1928)
E.g. social labels, bank runs, algorithmic predictions

Data (re)constructivism?
- Meaning of data is context- and situation-specific (Leonelli 2015)
  “raw data is an oxymoron” (Gitelman 2013)
- Gap between data and reality to be filled by interpretation
  Data analysis → re-contextualize / re-engineer the original meaning(s) of data
- Especially relevant for social (media) data

Reflexivity and “constant comparison”

- **Formulate open research questions.** No testing of hypothesis, but develop innovative questions and ways to make sense of social media data.

- **Constant comparison** and revision of data, categories and questions
  Asking questions like: “What are these data a study of?”, “What do the data suggest and pronounce?”, “From whose point of view” and “What theoretical category does this specific datum indicate?” (Charmaz 2006; Glaser 1978; Glaser and Strauss 1999)

- **Adapt your categories / identify higher level issues** during the process and put them at the center of analysis (e.g. kinds of ambiguity, debunk debunking algorithms)

- **Don’t hide process.** Reflect on changes, failures and learnings during the whole research process (and document some of them in your paper).
  E.g. alterations of stop lists, choice of cluster numbers, replacement of datasets, disagreements in your group

What’s up next session?

Focused session:
political engagement around
climate change on YouTube

Assignments for next week

1 Reading assignment

- Read Paper: Uldam, Julie. “Online Civic Cultures: Debating Climate Change Activism on YouTube.” International Journal of Communication 7 (2013): 1185–1204. (available on Whiteboard: CSMA / Resources / 7)

- Evaluate in a **commentary** of 150 words whether and how ML methods could to add to the insights from the paper above.

- **Submit on Github** (reply to issue) until 16 Dec 12h00 (noon)

- **Read the commentaries** from your peers before the session on Thursday

2 Seminar project preparation

- Continue to work on your **group project idea** and add elements to the GSheet: https://docs.google.com/spreadsheets/d/1DdkST3KZV4x9D5nGsHgevIASmu_rFkK0Bx2r4AeBGPE/edit?usp=sharing

- **If you haven’t found a topic or group yet,** you can a) either join an existing group, b) propose a new topic and ask others to join or c) pick one of the example topic indicated by instructors. **Use the GSheet in these decisions.**

Github issue for assignment: https://github.com/FUB-HCC/seminar_critical-social-media-analysis/issues/27

Discussion and group work

Discussion on Discord (voice channels)

- Discuss in what ways your research project can combine ML with an interpretivist / critical perspective.

- Continue to work on your project idea and add (further) elements in the GSheet: https://docs.google.com/spreadsheets/d/1DdkST3KZV4x9D5nGsHgevIASmu_rFkK0Bx2r4AeBGPE/edit?usp=sharing (Instructors have started to provide feedback in the last row of the sheet on the right)

Further info

- Instructors will drop by and can give you feedback. You can also ask instructors to join your discussion)
- Indications about the research project and seminar paper can be found here: https://github.com/FUB-HCC/seminar_critical-social-media-analysis/issues/15

Important note a first short paper about the research project (concept, 750 words) has to be submitted until 7 January 2021. In order to prevent work during the winter break, you should already start to work for this assignment. Let us know if you have questions in this regard.

Literature

Anderson, Chris. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” Wired. Accessed December 10, 2020. https://www.wired.com/2008/06/pb-theory/.

Barry, Andrew, Georgina Born, and Gisa Weszkalnys. “Logics of Interdisciplinarity.” Economy and Society 37, no. 1 (February 2008): 20–49. https://doi.org/10.1080/03085140701760841.

Baumer, Eric P. S., Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul, and Micki McGee. “Topicalizer: Reframing Core Concepts in Machine Learning Visualization by Co-Designing for Interpretivist Scholarship.” Human–Computer Interaction 35, no. 5–6 (November 1, 2020): 452–80. https://doi.org/10.1080/07370024.2020.1734460.

Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan, und Geri K. Gay. „Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?“ Journal of the Association for Information Science and Technology 68, Nr. 6 (Juni 2017): 1397–1410. https://cpb-us-e1.wpmucdn.com/blogs.cornell.edu/dist/c/3483/files/2017/02/Muller2016-Machine-2bp3h65.pdf

boyd, danah, and Kate Crawford. “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” Information, Communication & Society 15, no. 5 (June 2012): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Charmaz, Kathy. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. London; Thousand Oaks, Calif.: Sage Publications, 2006.

Chen, Nan-Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. “Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity.” ACM Transactions on Interactive Intelligent Systems 9, no. 4 (2018): 21.

Gillies, Marco, Bongshin Lee, Nicolas d’Alessandro, Joëlle Tilmanne, Todd Kulesza, Baptiste Caramiaux, Rebecca Fiebrink, et al. “Human-Centred Machine Learning.” In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA ’16, 3558–65. San Jose, California, USA: ACM Press, 2016. https://doi.org/10.1145/2851581.2856492.

Gitelman, Lisa. “Raw Data” Is an Oxymoron. Cambridge Mass.: MIT Press, 2013.

Glaser, Barney G. Theoretical Sensitivity: Advances in the Methodology of Grounded Theory, 1978.

Glaser, Barney, and Anselm Strauss. The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago: Aldine Transaction, 1999.

Kitchin, Rob. “Big Data, New Epistemologies and Paradigm Shifts.” Big Data & Society 1, no. 1 (July 10, 2014): 205395171452848. https://doi.org/10.1177/2053951714528481.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. “Fairness and Abstraction in Sociotechnical Systems.” In Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19, 59–68. Atlanta, GA, USA: ACM Press, 2019. https://doi.org/10.1145/3287560.3287598.

Smith, Alison, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. “Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System.” In Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI 18, 293–304. Tokyo, Japan: ACM Press, 2018. https://doi.org/10.1145/3172944.3172965.

M. Brooks, K. Kuksenok, M. K. Torkildson, D. Perry, J. J. Robinson, T. J. Scott, O. Anicello, A. Zukowski, P. Harris, C. Aragon. Statistical Affect Detection in Collaborative Chat. CSCW 2013