«Text as Data Methods in R –
Applications for Automated Analyses of News Content»

Information about instructor:

Valerie Hase, M.Sc., M.A.

v.hase@ikmz.uzh.ch, Raum AND-3-26

Information about seminar:

Part of the Master "Internet & Society" (Research Areas in Internet & Society)

takes place via in-person weekly meetings (Tuesdays, 10:15-12:00, HS2021)

in room BIN-1-D.25

Link to course in UZH course catalogue

Link to course in OLAT

Link to corresponding bookdown tutorial

Summary of seminar

Automated content analyses, i.e., (semi-)automated analysis of text, are becoming an increasingly important method in communication science. In this course, students will learn how to apply different methods of automated content analysis using R and conduct their own analyses within a small project. Please note that interested participants should have a basic interest in working with the programming language R as well as in learning new quantitative methods. However, previous knowledge of R is not required.

Learning goals

- Students will learn basic knowledge of the programing language R and will be able to conduct their own analyses by writing corresponding functions.
- Students will learn basic knowledge of the method "automated content analysis" and will be able to conduct their own automated content analysis by writing corresponding functions.
- Students will be able to apply this knowledge to independently conduct automated content analyses using data on news media coverage. This means that given predefined research questions/hypotheses, they develop a suitable operationalization via R, conduct descriptive analyses/statistical tests to answer these questions/test hypotheses, and critically interpret results.
- Students will be able to critically discuss current German- and English-language research on the topic of automated content analyses, summarize its strengths and weaknesses, and evaluate it.
- Students will be able to engage in factual and competent discussions. They are able to communicate their acquired knowledge alone and in groups in a comprehensible and structured manner, both orally and in writing. They are also able to work in groups, resolve potential conflicts, and organize themselves.

Workload & Assessment

This is a 6 ECTS seminar which equals 180 hours of work.

Please know that the time spent in in-person meetings in class is only a fraction of the actual workload needed to pass the class. The bulk of the required work needs to be done between sessions.

Workload

Attendance in seminar sessions	26h
Readings	24h
Working through R tutorials & doing related assignments	70h
Seminar paper	60h

Grading

Assignments in R (group or individual work)	30%
Seminar paper (individual work)	70%

Assignments in R (group or individual work)

During the semester, you will receive a total of three exercises in R which you can work on either in groups or individually. You will upload your solutions via OLAT. In the script, please specify the name of all students who worked on the corresponding task. The grade for each task will make up 10% of your final grade, leading to all tasks contributing a total of 30% to your final grade.

Task in R (I)
Please submit until: 22.10.2021 (folder closes as midnight) via OLAT, folder: "Task in R (1)".

Task in R (II)
Please submit until: 19.11.2021 (folder closes as midnight) via OLAT, folder: "Task in R (2)".

Task in R (III)
Please submit until: 17.12.2021 (folder closes as midnight) via OLAT, folder: "Task in R (3)".

Seminar paper (individual work)

Since this seminar is a bit more practice- and programming oriented, you will do a seminar paper in a slightly different form than your usual term paper. In short, the seminar paper will consist of two parts:

(1) A theoretical part where you need to answer a set of pre-defined questions on automated content analyses using relevant literature (for instance by describing a specific method and discussing its (dis-)advantages

(2) An empirical part where you need answer a set of pre-defined questions with a given data set and write-up your results using both literature in the field and R.

Seminar paper
Please submit until: 07.01.2022 (folder closes as midnight) via OLAT, folder: "Seminar paper".

Overview Seminar "Text as Data Methods in R" (HS2021)

Date	Phase	Session	Work by students	Input by instructor
21.09.2021	**Introduction + organizational information**	Session 1: *Introduction*		Basic information on the seminar; expectations; structure and assessment
28.09.2021	**Introduction to R**	Session 2: *Introduction to R (I)*	Work through necessary tutorials for corresponding session before in-person meeting	Summary of tutorials & moderating Q&A
05.10.2021		Session 3: *Introduction to R (II)*	Work through necessary tutorials for corresponding session before in-person meeting	Summary of tutorials & moderating Q&A
12.10.2021		Session 4: *Introduction to R (III)*	Work through necessary tutorials for corresponding session before in-person meeting	Summary of tutorials & moderating Q&A
19.10.2021		Session 5: *Introduction to R (IV)*	Work through necessary tutorials for corresponding session before in-person meeting **Assignment deadline**: Task in R (I) until 22.10.2021	Summary of tutorials & moderating Q&A
26.10.2021	**Theoretical Introduction: Automated Content Analysis**	Session 6: *Theoretical Introduction: Automated Content Analysis (I)*	Prepare necessary readings: *Benoit (2019), Grimmer & Stewart (2013), Welbers (2017)*	Input: Basic logic of automated content analysis
02.11.2021		Session 7: *Theoretical Introduction: Automated Content Analysis (II)*	Prepare necessary readings: *Denny & Spirling (2018), Hase (in press)*	Input: Preprocessing, rule-based approaches & validation
09.11.2021		Session 8: *Theoretical Introduction: Automated Content Analysis (III)*	Prepare necessary readings: *Blei (2012), Maier et al. (2018), Roberts et al. (2014)*	Input: Topic modeling

16.11.2021	**Automated Content Analysis in R**	No in-person meeting (session used for working through next tutorials) **Assignment deadline**: Task in R (II) until 19.11.2021		
23.11.2021		Session 9: *Introduction to automated content analysis with* R *(I)*	Work through necessary tutorials for corresponding session before in-person meeting	Summary of tutorials & moderating Q&A
30.11.2021		Session 10: *Introduction to automated content analysis with* R *(II)*	Work through necessary tutorials for corresponding session before in-person meeting	Summary of tutorials & moderating Q&A
07.12.2021		Session 11: *Introduction to automated content analysis with* R *(III)*	Work through necessary tutorials for corresponding session before in-person meeting	Summary of tutorials & moderating Q&A
14.12.2021		Session 12: *Introduction to automated content analysis with* R *(IV)*	Work through necessary tutorials for corresponding session before in-person meeting **Assignment deadline**: Task in R (III) until 17.12.2021	Summary of tutorials & moderating Q&A
21.12.2021	**Wrap-Up & information on seminar paper**	Session 13: Wrap-Up & Information on seminar paper	**Assignment deadline**: Seminar paper until 07.01.2022	Wrap-Up of seminar, further information on seminar paper

Necessary readings

Session 6 (26.10.2021): *Theoretical Introduction: Automated Content Analysis (I)*

Benoit, K. (2019). Text as data: An overview. Forthcoming in Cuirini, L., & Franzese, R. (Eds.), Handbook of Research Methods in Political Science and International Relations. Thousand Oaks: Sage. (Preprint via: https://kenbenoit.net/pdfs/28%20Benoit%20Text%20as%20Data%20draft%202.pdf)

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. Communication Methods and Measures, 11(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Session 7 (02.11.2021): *Theoretical Introduction: Automated Content Analysis (II)*

Denny, M., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis, 26(2), 168-189. https://doi:10.1017/pan.2017.44

Hase, V. (in press). Automated Content Analysis. In F. Oehmer, S. H. Kessler, E. Humprecht, K. Sommer, & L. Castro Herrero (Eds.), Handbook of Standardized Content Analysis: Applied Designs to Research Fields of Communication Science. VS Springer.

Session 8 (09.11.2021): *Theoretical Introduction: Automated Content Analysis (III)*

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. Communication Methods and Measures, 12(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses: Structural topic models for survey responses. American Journal of Political Science, 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Further readings

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. Discourse Studies, 21(1), 3–21. https://doi.org/10.1177/1461445618814032

van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). Studying Political Decision Making with Automatic Text Analysis. In W. van Atteveldt, K. Welbers, & M. van der Velden, Oxford Research Encyclopedia of Politics. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.957

Boumans, J. W., & Trilling, D. (2016). Taking Stock of the Toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. Digital Journalism, 4(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). Big Data & Society. https://doi.org/10.1177/2053951715602908

Evans, M. S. (2014). A Computational Approach to Qualitative Analysis in Large Textual Datasets. PLoS ONE, 9(2), e87908. https://doi.org/10.1371/journal.pone.0087908

Günther, E., & Quandt, T. (2016). Word Counts and Topic Models: Automated text analysis methods for digital journalism research. Digital Journalism, 4(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. Political Analysis, 23(2), 254–277. https://doi.org/10.1093/pan/mpu019

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. Poetics, 41(6), 545–569. https://doi.org/10.1016/j.poetic.2013.10.001

Nicholls, T., & Culpepper, P.D. (2021). Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities. *Political Communication*, 38(1-2), 159-181. https://doi.org/10.1080/10584609.2020.18127

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. American Journal of Political Science, 54(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Wilkerson, J., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annual Review of Political Science, 20(1), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Zamith, R., & Lewis, S. C. (2015). Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis. The ANNALS of the American Academy of Political and Social Science, 659(1), 307–318. https://doi.org/10.1177/0002716215570576

Other tutorials: Introduction to R

Bail, C.: SICSS Boot Camp (Beta) https://compsocialscience.github.io/summer-institute/boot_camp

Breuer, K., & Jünger, S. (2020). "Introduction to R for Data Analysis", GESIS Summer School in Survey Methodology 2020. https://github.com/jobreu/r-intro-gesis-2020

Cotton, R. (2013). Learning R (First Edition). O'Reilly.

Grolemund G., & Wickham, H.: R for Data Science. https://r4ds.had.co.nz/index.html

Haim, M. wegweisR. https://wegweisr.haim.it/

Long, J., Teetor, P., & Safari, and O. M. C. (2019). R Cookbook, 2nd Edition. https://rc2e.com/

Phillips N.D. (2018). YaRrr ! The Pirate's Guide to R. https://bookdown.org/ndphillips/YaRrr/

De Vries, A., & Meys, J. (2015). R for dummies (2nd edition). John Wiley & Sons, Inc.

Unkel, J. (2020). Methodische Vertiefung: Computational Methods mit R und R Studio. https://bookdown.org/joone/ComputationalMethods/

Wickham, H., & Grolemund, G. (2016). R for data science: Import, tidy, transform, visualize, and model data (First edition). O'Reilly. https://r4ds.had.co.nz/

Other tutorials: Automated Content Analysis in R

van Atteveldt, W. Text Analysis in R workshop at University of Vienna http://vanatteveldt.com/vienna-r-text-analysis

Bail, C. Day 3: Automated Text Analysis. https://compsocialscience.github.io/summer-institute/curriculum#day_3

Bernauer J, & Traber D. Quantitative Analysis of Political Text. https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/quantitative-analysis-of-political-text/

Silge, J., & Robinson, D. Text mining with R: A tidy approach. https://www.tidytextmining.com/

Puschmann, C., & Haim, M. (no date). Automated Content Analysis with R. https://content-analysis-with-r.com

Unkel, J. (2020). Methodische Vertiefung: Computational Methods mit R und R Studio. https://bookdown.org/joone/ComputationalMethods/

Watanabe, K., & Müller, S. Quanteda Tutorials. https://tutorials.quanteda.io/

Wiedemann, Gregor; Niekler, Andreas (2017): Hands-on: a five day text mining course for humanists and social scientists in R. Proceedings of the 1st Workshop Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin. https://tm4ss.github.io/docs/