## Introduction to Computational Social Science

van Atteveldt, W., Strycharz, J., Trilling, K., & Welbers, K. (2019). Toward open computational communication science: a practical road map for reusable data and code. *International Journal of Communication*, 13, 3935–3954.

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, 12(2–3), 81–92.

boyd, D., & Crawford, K. (2012). Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.

Brady, H.E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22(1), 297–323.

Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), 259–271.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., & Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1), 325–346.

Di Maggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society,* 2(2), 1–5.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning and causal inference work together. *PS: Political Science & Politics*, 48(1), 80–83.

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., González-Bailón, S., Lamberso, PJ., Pan, J., Peng, T.-Q., Shen, C.C., Smaldino, P.E., van Atteveldt, W., Waldherr, A., Zhang, J., & Zhu, J.J.H. (2019). Computational communication science: methodological catalyzer for a maturing discipline. *International Journal of Communication*, 13, 3912–3934.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & van Alstyne, M. (2009). Computational Social Science. *Science,* 323(5915), 721–723.

Niemann-Lenz, J., Bruns, S., Hefner, D., Knop-Huelss, K., Possler, D., Reich, S., Reinecke, L., Scheper, J., & Klimmt, C. (2019). Crafting a strategic roadmap for computational methods in communication science: learnings from the CCS 2018 conference in Hanover. *International Journal of Communication*, 13, 3885-3893.

Peng, T-Q., Liang, H., & Zhu, J.H. (2019). Introducing computational social science for Asia-Pacific communication research. *Asian Journal of Communication*, 29(3), 205–216.

Possler, D., Bruns, S., & Niemann-Lenz, J. (2019). Data is the new oil – but how do we drill it? Pathways to access and acquire large data sets in communication science. *International Journal of Communication*, 13, 3894–3911.

Shah, D.V., Cappella, J.N., & Neuman, R. (2015). Big data, digital media, and computational social science: possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, 69(1), 6–13.

Wallach, H. (2016). Computational social science: toward a collaborative future. In R. Alvarez (Ed.), *Computational social science: discovery and prediction* (Analytical methods for social research, pp. 307–316). Cambridge: Cambridge University Press.

Windsor, L.C. (2020). Advancing interdisciplinary work in computational communication science. *Political Communication*. Advance online publication.

**Introduction to Automated Content Analysis**

Anandarajan, M., Hill, C., & Nolan, T. (2019). *Practical text analysis. Maximizing the value of text data*. Cham: Springer.

van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). *Studying political decision-making with automatic text analysis*. Oxford Research Encyclopedias.

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 verfügbar via: https://kenbenoit.net)

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.

Evans, J.A., & Aceves, P. (2016). Machine translation: mining text for social theory. *Annual Review of Sociology*, 42, 21–50.

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.

Günther, E., & Quandt, T. (2016). Word counts and topic models. Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1), 75–88.

Krippendorff, K. (2019). *Content analysis. An introduction to its methodology*. Chapter 11: Computer AIDs. Fourth Edition. Los Angeles: SAGE.

Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2), 254–277.

Manning, C.D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Nguyen, D., Liakata, M., DeDeo, S. Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2019). *How we do things with words: Analyzing text as social and cultural data*. Preprint verfügbar via: https://arxiv.org/abs/1907.01468

Riffe, D., Lacy, S., & Fico, F.G. (2005). *Analyzing media messages. Using quantitative content analysis in research*. Chapter 10: computers. Second Edition. New York: Routledge.

Sommer, K., Wettstein, M., Wirth, W., & Matthes, J. (2014). *Automatisierung in der Inhaltsanalyse*. Köln: Halem.

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. Advance online publication. *Political Communication*.

Trilling, D., & Jonkman, J.G.F. (2018). Scaling up content analysis. *Communication Methods and Measures*, 12(2–3), 158–174.

Welbers, K., van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, 11(4), 245–265.

Wettstein, M. (2016). *Verfahren zur computerunterstützen Inhaltsanalyse in der Kommunikationswissenschaft*. Zürich: Zürich. Dissertation verfügbar via https://opac.nebis.ch/ediss/20162838.pdf

Wieling, M., Rawee, J., & van Noord, G. (2018). Reproducibility in computational linguistics: are we willing to share? *Computational Linguistics*, 44(4), 641-649.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: opportunities and challenges. *Annual Review of Political Science,* 20, 529–544.

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.

## Preprocessing & Feature Selection

Däubler, T., Benoit, K., Mikhaylov, S., & Laver, M. (2012). Natural sentences as valid units for coded political texts. *British Journal of Political Science*, 42(4), 937–951.

Denny, M.J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, why it misleads, and what to do about it. *Political Analysis*, 26(2), 168–189.

Monroe, B.L., Colaresi, M.P., & Quinn, K.M. (2008). Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4), 372–403.

Roberts, C.W. (2000). A conceptual framework for quantitative text analysis. *Quality & Quantity*, 34(3), 259–274.

Turney, P.D., & Pantel., P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.

## Document Similarity, Text Reuse, Clustering

Cross, J.P., & Hermansson, H. (2017). Legislative amendments and informal politics in the European Politics: A text reuse approach. *European Union Politics*, 18(4), 581–602.

Günther, E., Domahidi, E., & Quandt, T. (2017). Mediale Sichtbarkeit der WahlbewerberInnen und Themen der Bundestagswahl 2013. Eine automatisierte Analyse der Online-Berichterstattung. *Studies in Communication and Media*, 6(3), 262–299.

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. In *Proceedings of the National Academy of Sciences*, 108(7), 2643–2650.

Li, Y., Chung, S.M., & Holt, J.D. (2008). Text document clustering based on frequent word meaning sequences. *Data & Knowledge Engineering*, 64(1), 381–404.

Lindner,F., Desmarais, B., Burgess, M., & Giraudy, E. (2018). Text as policy: Measuring policy similarity through bill text reuse. Online first publication. *Policy Studies Journal.* doi:10.1111/psj.12257

Mozer, R., Miratrix, L., Kaufman, A. R., & Jason Anastasopoulos, L. (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*. Advance online publication.

Nicholls, T., & Bright, J. (2019). Understanding news story chains using information retrieval and network clustering techniques. *Communication Methods and Measures*, 13(1), 43–59.

Vogler, D., & Schäfer, M. S. (2020). Growing influence of university PR on science news coverage? A longitudinal automated content analysis of university media releases and newspaper coverage in Switzerland, 2003-2017. *International Journal of Communication*, 14, 3143-3164.

Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracking the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, 59(4), 943–956.

### Complexity / Lexical Diversity

Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2), 491–508.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2019). *Lexical Diversity and Language Development*. Palgrace Macmillan: London.

Spirling, A. (2016). Democratization and linguistic complexity: the effect of franchise extension on parliamentary discourse, 1832–1915. *The Journal of Politics*, 78(1), 120–136.

Tolochko, P., & Boomgarden, H.G. (2019). Determining political text complexity: conceptualizations, measurements, and application. *International Journal of Communication*, 13, 1784–1804.

### Keywords in Context, Collocations, Co-Occurrence

Arendt, F., & Karadas, N. (2017). Content analysis of mediated associations: an automated text-analytic approach. *Communication Methods and Measures*, 11(2), 106–120.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Holtzman, N.S., Schott, J.P., Jones, M.N., Balota, D.A., & Yarkoni, T. (2011). Exploring media bias with semantic analysis tools: validation of the contrast analysis of semantic similarities (CASS). *Behavior Research Methods*, 43(1), 193–200.

Ruigrok, N., & van Atteveldt, W. (2007). Global angling with a local angle: how U.S., British and Dutch newspapers frame global and local terrorist attacks. *The International Journal of Press/Politics*, 12(1), 68–90.

### Dictionary approaches

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2020). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behavior Research Methods. Advance online publication

Lind, F., Eberl, J.-M., Heidenreichm T., & Boomgarden, H.G. (2019). When the journey is as important as the goal: roadmap to multilingual dictionary construction. *International Journal of Communication*, 13, 4000–4020.

Loughran, T., & McDonals, B. (2011). When is a liability not a liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance,* 66(1), 35–65.

Muddiman, A., McGregor, S., Stroud, N.J. (2019). (Re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, 36(2), 214–226.

Pennebaker, J.W., & Chung, C.K. (2008). Computerized text-analysis of Al-Qaeda transcripts. In K. Krippendorff & M. Bock (Eds.), *A content analysis reader*. Thousand Oaks: Sage.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PLoS One*, 11(12), 1–18.

Rooduijn, M., & Pauwels, T. (2011). Measuring populism: comparing two methods of content analysis. *West European Politics*, 34(6), 1272–1283.

Soroka, S., Young, L., & Balmas, M. (2015). Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 108–121.

Tausczik, Y.R., & Pennebaker, J.W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology,* 29(1), 24–54.

**Sentiment analysis**

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2020). Automated Text Classification of News Articles: A Practical Guide. Advance online publication. *Political Analysis*.

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2019). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, 14(2), 83-104.

Dodds, P.S., & Danforth, C.M. (2010). Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4), 441–456.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27–38). Verfügbar via: https://arxiv.org/abs/1406.0032

González-Bailón, S., & Paltoglou, G. (2015). Aignals 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), 95–107.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: combining a dictionary approach with crowdsourcing. *Quality & Quantity*, 51(6), 2623–2646.

Rauh, C. (2018). Validating a sentiment dictionary for German political language – a workbench note. *Journal of Information, Technology & Politics*, 15(4), 319–343.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, S., & Sedlmair, M. (2018). More than bags of words: sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2–3), 140–157.

Stine, R.A. (2019). Sentiment analysis. *Annual Review of Statistics and Its Application*, 6(1), 287–308.

Su, L.Y-F., Cacciatore, M.A., Liang, X., Brossard, D., Scheufele, D., & Xenos, M.A. (2017). Analyzing public sentiments online: combining human- and computer-based content analysis. *Information, Communication & Society*, 20(3), 406–427.

Taboada, M. (2016). Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics*, 2(1), 325–347.

Young, L., & Soroka, S. (2012). Affective news: the automated coding of sentiment in political texts. *Political Communication*, 29(2), 205–231.

**Topic Modeling / Supervised machine learning approaches to topic classification**

van Atteveldt, W., Welbers, K., Jacobi, C., & Vliegenthart, R. (2014). LDA models topics… But what are 'topics'? Verfügbar via: http://vanatteveldt.com/wp-content/uploads/2014_vanatteveldt_glasgowbigdata_topics.pdf

Blei, D. M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3: 993–1022.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22*. Verfügbar via: http://www.cs.columbia.edu/~blei/papers/ChangBoyd-GraberWangGerrishBlei2009a.pdf

Airoldi, E. M., & Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, 111(516), 1381-1403.

Eshima, S., Imai, K., & Sasaki, T. (2020). Keyword assisted topic models. https://arxiv.org/abs/2004.05964

Evans, M.S. (2014). A computational approach to qualitative analysis in large textual datasets. *PLoS One*, 9(2), 1–10.

Günther, E., & Domahidi, E. (2017). What communication scholars write about: an analysis of 80 years of high-impact journals. *International Journal of Communication*, 11, 3051–3071.

Hase, V., Engelke, K., & Kieslich, K. (2020). The things we fear. Combining automated and manual content analysis to uncover themes, topics and threats in fear-related news. *Journalism Studies*. doi:10.1080/1461670X.2020.1753092

Hillard, D., Purpura, S., & Wilkerson, J. (2007). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4), 31–46.

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.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keiner, 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.

Mohr, J.W., & Bogdanov, P. (2013). Introduction – topic models: What they are and why they matter. *Poetrics*, 41(6), 545–569.

Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2), 358–375.

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.

Roberts, M.E., Stewart, B.M., & Tingley, D. (2016). Navigation the local modes of big data: the case of topic models. In R. Alvarez (Ed.), *Computational social science: discovery and prediction* (Analytical methods for social research, pp. 51–97). Cambridge: Cambridge University Press.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Lui, J., Gadarian, S.K., Albertson, B., & Rand, D.G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.

Scharkow, M. (2011). Zur Verknüpfung manueller und automatischer Inhaltsanalyse durch maschinelles Lernen. *Medien & Kommunikationswissenschaft*, 59(4), 545–562.

Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Quality & Quantity*. doi:10.1007/s11135-020-00976-w

Wallach, H., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1105–1112).

Walter, D., & Ophir, Y. (2019). News Frame Analysis: An Inductive Mixed-Method Computational Approach. *Communication Methods and Measures,* 13(4): 248–266.

Watanabe, K., & Zhou, Y. (2020). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. *Social Science Computer Review*. doi:10.1177/0894439320907027

## Mixed Methods

Burscher, B., Vliegenthart, R., & de Vreese, C.H. (2016). Frames beyond words: applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5), 530–545.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication,* 64(2), 317–332.

Dun, L., Soroka, S., & Wlezien, C. (2020). Dictionaries, Supervised Learning, and Media Coverage of Public Policy. Advance online publication. *Political Communication*.

Guo, L., Vargo, J.C., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, 93(2), 332–359.

Lemke, M., & Wiedemann, G. (2016). *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Springer: Wiesbaden.

Lewis, S.C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: a hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, 57(1), 34–52.

Nelson, L.K. (2017). Computational grounded theory: a methodological framework. *Sociological Methods & Research*. Online first publication. doi:10.1177/0049124117729703

Ophir, Y., Walter, D., & Marchant, E. R. (2020). A collaborative way of knowing: Bridging computational communication research and grounded theory ethnography. *Journal of Communication, 70*(3), 447–472.

Waldherr, A., Wehden, L.-O., Stoltenberg, D., Miltner, P., Ostner, S., & Pfetsch, B. (2019). Inductive Codebook Development for Content Analysis: Combining Automated and Manual Methods. *Forum Qualitative Sozialforschung*, 20(1).

Walter, D., & Ophir, Y. (2019). News frame analysis: an inductive mixed-methods computational approach. *Communication Methods and Measures*. Online first publication. doi:10.1080/19312458.2019.1639145

**Supervised Machine Learning**

Broersma, M., & Harbers, F. (2018). Exploring Machine Learning to Study the Long-Term Transformation of News. Digital Journalism, 6(9), 1150–1164.

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., & de Vreese, C.H. (2014). Teaching the computer to code frames in news: comparing two supervised maching learning approaches to frame analysis. *Communication Methods and Measures*, 8(3), 190–206.

Burscher, B., Vliegenhart, R., & de Vreese, C.H. (2015). Using supervised machine learning to code policy issues: can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science*, 659(1), 122–131.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

De Grove, F., Boghe, K., & de Marez, L. (2020). (What) Can Journalism Studies Learn from Supervised Machine Learning? Journalism Studies, 21(7), 912-927.

Hopkins, D.J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1), 229–247.

Kananovich, V. (2018). Framing the taxation-democratization link: an automated content analysis of cross-national newspaper data. *The International Journal of Press/Politics*, 23(2), 247–267.

Kelm, O., Gerl, K., & Meißner, F. (2020). Machine Learning. In I. Borucki, K. Kleinen-von Königslöw, S. Marschall, & T. Zerback (Eds.), Handbuch Politische Kommunikation. Springer: Wiesbaden.

Khan, A. Baharudin, B., Lee, L.H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, 1(1), 4–20.

King, G., Pan, J., & Roberts, M.E. (2013). How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107(2), 326–343.

Manning, C.D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval*. Cambridge: University Press.

Mirończuk, M.M, & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106, 36–54.

Opperhuizen, A.E., Schouten, K., & Klijn, E. H. (2019). Framing a conflict! How media report on earthquake risks caused by gas drilling. *Journalism Studies*, 20(5), 714–734.

Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: measuring polarization in Westminster systems. *Political Analysis*, 26(1), S. 120–128.

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2019). Using supervised machine learning in automated content analysis: an example using relational uncertainty. *Communication Measures and Methods*. Online first publication. doi:10.1080/19312458.2019.1650166

Vijayakumar, R., & Cheung, M. W.-L. (2019). Assessing replicability of machine learning results: An introduction to methods on predictive accuracy in social sciences. *Social Science Computer Review*. Advance online publication.

Scharkow, M. (2011). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, 47(2), 761–773.

Stoll, A. (2020). Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten. *Publizistik, 65*(2), 233–251.

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting Impoliteness and Incivility in Online Discussions: Classification Approaches for German User Comments. *Computational Communication Research, 2*(1), 109–134.

## More than Bag of Words: Methods with Focus on Semantic Structure / Word Embeddings

van Atteveldt, W. (2008). Semantic network analysis. Techniques for extracting, representing, and querying media content. Verfügbar via: http://vanatteveldt.com/wp-content/uploads/vanatteveldt_semanticnetworkanalysis.pdf

van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Analysis*, 16(4), 428–446.

Chang, C., & Masterson, M. (2020). Using word order in political text classification with long short-term memory models. *Political Analysis, 28*(3), 395–411.

Dhingera, B., Zhou, Z., Fitzpatrick, D., Muehl, M., & Cohen, W.W. (2016). Tweet2Vec: character-based distributed representations for social media. Verfügbar via: https://arxiv.org/abs/1605.03481

Fogel-Dror, Y., Shenhav, S.R., Sheafer, T., & van Atteveldt, W. (2019). Role-based associations of verbs, actions, and sentiments with entities in political discourse. *Communication Methods and Measures*, 13(2), 69–82.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. In *Proceedings of the National Academy of Science of the United States of America*, 115(16), E3635–E3644.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. Verfügbar via: https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. Verfügbar via: https://arxiv.org/abs/1310.4546

Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Spirling, A., & Rodriguez, P.L. (2019). Word embeddings. What works, what doesn't, and how to tell the difference for applied research. Verfügbar via: https://www.nyu.edu/projects/spirling/documents/embed.pdf

Tshitoyan, V., Dagdelen, J., Weston, L., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., & Jain, A. (2019). Unsupversied word embeddings capture latent knowledge from materials science literature. *Nature*, 571, 95–98.

## Tutorials

Automated content analysis with R und Automatisierte Inhaltsanalyse mit R

Lda and Learning R

Quanteda Tutorials

Teaching Materials for Computational Social Science

Text Analysis in R; see related github Account

Text mining for humanists and social scientists in R

Text Mining with R; Text mining with R: a tidy approach

Tidytext tutorials

Tutorials by Julia Silge

University of California: Computational Social Science

## R Packages for text analysis

Caret: Misc functions for training and plotting classification and regression models.

Corpus: Text corpus data analysis, with full support for international text (Unicode). Functions for reading data from newline-delimited 'JSON' files, for normalizing and tokenizing text, for searching for term occurrences, and for computing term occurrence frequencies, including n-grams.

keyATM: Fits keyword assisted topic models (keyATM) using collapsed Gibbs samplers. The keyATM combines the latent dirichlet allocation (LDA) models with a small number of keywords selected by researchers in order to improve the interpretability and topic classification of the LDA. The keyATM can also incorporate covariates and directly model time trends.

KoRpus: A set of tools to analyze texts. Includes, amongst others, functions for automatic language detection, hyphenation, several indices of lexical diversity (e.g., type token ratio, HD-D/vocd-D, MTLD) and readability (e.g., Flesch, SMOG, LIX, Dale-Chall). Basic import functions for language corpora are also provided, to enable frequency analyses (supports Celex and Leipzig Corpora Collection file formats) and measures like tf-idf.

Newsmap: Semi-supervised model for geographical document classification (Watanabe 2018). This package currently contains seed dictionaries in English, German, French, Spanish, Japanese, Russian and Chinese (Simplified and Traditional).

oolong: Intended to create standard human-in-the-loop validity tests for typical automated content analysis such as topic modeling and dictionary-based methods.

perspective: The 'Perspective' API uses machine learning models to score the perceived impact a comment might have on a conversation (i.e. TOXICITY, INFLAMMATORY, etc.). 'peRspective' provides access to the API and returns tidy data frames with results of the specified machine learning model(s).

Quanteda: A fast, flexible, and comprehensive framework for quantitative text analysis in R. Provides functionality for corpus management, creating and manipulating tokens and ngrams, exploring keywords in context, forming and manipulating sparse matrices of documents by features and feature co-occurrences, analyzing keywords, computing feature similarities and distances, applying content dictionaries, applying supervised and unsupervised machine learning, visually representing text and text analyses, and more.

Readtext: Functions for importing and handling text files and formatted text files with additional meta-data, such including '.csv', '.tab', '.json', '.xml', '.html', '.pdf', '.doc', '.docx', '.rtf', '.xls', '.xlsx', and others.

Rainette: An R implementation of the Reinert text clustering method.

Rnewsflow: A collection of tools for measuring the similarity of text messages and tracing the flow of messages over time and across media.

rsyntax: Various functions for querying and reshaping dependency trees, as for instance created with the 'spacyr' or 'udpipe' packages. This enables the automatic extraction of useful semantic relations from texts, such as quotes (who said what) and clauses (who did what).

Sentimentr: Calculate text polarity sentiment at the sentence level and optionally aggregate by rows or grouping variable(s).

Spacyr: An R wrapper to the 'Python' 'spaCy' 'NLP' library.

Stm: The Structural Topic Model (STM) allows researchers to estimate topic models with document-level covariates. The package also includes tools for model selection, visualization, and estimation of topic-covariate regressions.

Stringr: A consistent, simple and easy to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with "NA"'s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another.

Textdata: Provides a framework to download, parse, and store text datasets on the disk and load them when needed. Includes various sentiment lexicons and labeled text data sets for classification and analysis.

Tidytext: Text mining for word processing and sentiment analysis using 'dplyr', 'ggplot2', and other tidy tools. Also: Tidymodels: The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.

Tm: A framework for text mining applications within R.

Topicmodels: Provides an interface to the C code for Latent Dirichlet Allocation (LDA) models and Correlated Topics Models (CTM) by David M. Blei and co-authors and the C++ code for fitting LDA models using Gibbs sampling by Xuan-Hieu Phan and co-authors.

Tosca: A framework for statistical analysis in content analysis. In addition to a pipeline for preprocessing text corpora and linking to the latent Dirichlet allocation from the 'lda' package, plots are offered for the descriptive analysis of text corpora and topic models. In addition, an implementation of Chang's intruder words and intruder topics is provided.

Vader: A lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains