Stefan Müller, PhD
Postdoctoral Researcher
Chair of Policy Analysis
Department of Political Science
University of Zurich
https://muellerstefan.net

Seminar 615786

Quantitative Text Analysis

Last update: May 23, 2019

Latest version: https://muellerstefan.net/teaching/2019-spring-qta.pdf

Term: Spring term 2019
Time (Group 1): Mon., 12:15–13:45
Time (Group 2): Wed., 10:15–12:00
Room (Group 1): AFL-E-019 (Affolternstr. 56)
Room (Group 2): BIN-1-D.22 (Binzmühlestr. 14)

Lecturer: Stefan Müller
Office: AFL-H-349
Office hours: Tuesday, 16:00–17:00
E-Mail: mueller@ipz.uzh.ch
ECTS: 6.0

Course Content

In recent times the availability of textual data has increased massively, and there are multiple opportunities for analysing these data to answer social science research questions. This course introduces students of political science to the quantitative analysis of textual data. We cover a treatment of underlying theoretical assumptions, applications of these methods in the scholarly literature, and the respective implementations in the R statistical programming language.

Each session contains practical, hands-on exercises to apply the methods to real texts. Most of these methods can be reduced to a three-step process: first, identifying texts and units of texts for analysis; second, extract quantitatively measured features from these texts and converting them to a quantitative feature matrix; third, analyse this matrix with statistical methods, such as dictionary construction and application, scaling models, and topic models, to draw inferences about the texts. Students will learn how to apply these steps to various types of texts. There will be two homeworks which cover the theoretical assumptions as well as modelling and coding of text data. Moreover, students will use their own text corpus (or one of various text corpora provided for this course) to answer a substantive question from their personal research interests for a final project.

Details

- MA/PhD seminar
- Language: English

- Grading: 2 Homeworks (20% each); Research Paper (60%)

Learning Outcomes

At the completion of this course, students will be able to:

1. Understand fundamental issues in (quantitative) text analysis such as inter-coder agreement, reliability, validation, accuracy, and precision.
2. Convert texts into quantitative matrices of features, and then analyse those features using statistical methods.
3. Use human coding and annotations of texts to train supervised classifiers.
4. Apply these methods to a custom text corpus in order to tackle a substantive research question.

Introductory Readings

General Readings

The seminar does not build on a single text book, but relies mostly on papers and chapters of books. For a general overview of quantitative text analysis, natural language processing, and computational social science, the following books are recommended.

- Daniel Jurafsky and James H. Martin (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press.
- Matthew J. Salganik (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Technical Background

The following books and websites are helpful to refresh and extend the knowledge of R, RMarkdown, and the `quanteda` package. Websites such as Stack OverFlow, R bloggers, and the documentation of R packages will be useful for solving practical problems. The books below are published in print, but also legally available online.

R, RMarkdown, and `quanteda`

- Hadley Wickham and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly.
- Yihui Xie, J.J. Allaire, and Garrett Grolemund (2018). *R Markdown: The Definite Guide*. Boca Raton: CRC Press, Taylor & Francis Group.
- Kohei Watanabe and Stefan Müller (2019). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io.

Data Visualisation

- Kieran Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press.
- Claus O. Wilke (Forthcoming). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly.

Software and Packages

The applications of the course are based on the R statistical programming language. Participants should download and install the latest versions of R and RStudio. Students should also install the latest releases of the following R packages, which will be used throughout the course.

- Quantitative text analysis: `quanteda`
- Importing text data: `readtext`
- Topic models: `topicmodels` and `stm`
- Data wrangling and visualisation: `tidyverse` (esp. `dplyr`, `tidyr`, `lubridate`, and `ggplot2`)
- Creating documents and reports: `rmarkdown` and `knitr`
- Part-of-speech tagging and lemmatisation: `spacyr` (installation not mandatory)

Additionally, I strongly encourage students to get used to `git` and set up a GitHub account (recently, GitHub started to provide unlimited private repositories even in their free version). The free and open-source software GitHub Desktop allows to use git and GitHub without having to rely on the terminal. The following sites contain comprehensible introductions to `git` and GitHub:

- https://guides.github.com/activities/hello-world/
- https://help.github.com/desktop/guides/getting-started-with-github-desktop/
- https://happygitwithr.com

Syllabus Modification Rights

I reserve the right to reasonably alter the elements of the syllabus at any time by adjusting the reading list to keep pace with the course schedule. Moreover, I may change the content of specific sessions depending on the participants' prior knowledge and research interests.

Expectations and Grading

- Students are expected to read all papers or chapters assigned under **Readings**. These readings serve as the basis for in-class discussions about the advantages, disadvantages, and applicability of the various approaches to social science questions. For each session, I also assign a variety of optional readings which are not mandatory, but I strongly encourage students to (at least) skim these reading. Both the required and the optional readings consist of technical readings and at least one practical application of the respective method. Note that the core readings for each week are highlighted with two stars (**) and usually appear on top of the reading list for the session.

- Students submit two **Homeworks**, each of which counts towards 20% of the final grade. The homeworks will be distributed via OLAT 14 days before the submission deadline as an RMarkdown file. Students fill in the answers and solutions in the same RMarkdown file, rename it to `hw_01/02_surname_firstname.Rmd`, knit it as an `html` file, and submit it via OLAT. Only knitted `html` files will be accepted. Homeworks will be sumitted via OLAT. The deadline for Homework 1 is **March 22, 2019 (8:00pm CET)**, the deadline for Homework 2 is **April 26, 2019 (8:00pm CET)**. More details on the homeworks will be provided in the first session(s) of the course.
- Students also submit a **Research Paper** which counts towards 60% of the final grade. The research paper is a written analysis consisting of 5,000–5,500 words (including bibliography, captions, and footnotes). Students are required to develop a research design to answer a question with textual data. Students are free to answer questions from all subfields of political science, but must justify their choice and the relevance of the question. Students registered for an MA degree in another social science discipline are encouraged to develop a research project answering a question from their subject. Students can use existing corpora, create their own text corpus, or access textual data that may be collected in spring at the Computational Social Science Hub (part of the Digital Democracy Lab). The research papers must be submitted via OLAT as a `pdf` document before **June 21, 2019 (8:00pm CET)**. In the 10th and 11th session, each student gives a short presentation, covering the research question, relevance, text corpus, and methodological approach. Alongside with the presentation, students will submit a 1,000 words research proposal to receive comments from peers and the lecturer. Each project will be discussed through *written* feedback by another seminar participant, and students will also receive written feedback from me. Detailed instructions on the research paper, the presentation, and the in-class discussion will be provided via OLAT.

Overview of deadlines

Date	Time	Assignment
Friday, March 22, 2019	8:00pm CET	Homework 1 (20%)
Friday, April 26, 2019	8:00pm CET	Homework 2 (20%)
Friday, June 14, 2019	8:00pm CET	Research Paper (60%)

Course Structure

Week 1: Organisation and Introduction (February 18)

- What are quantitative text analysis and natural language processing?
- What is the structure of the course and what are the expectations?
- *Application*: installing packages and setting up a Project in RStudio

Readings

- Justin Grimmer and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". *Political Analysis* 21 (3): 267–297.**
- Henry E. Brady (2019). "The Challenge of Big Data and Data Science". *Annual Review of Political Science* published ahead of print.

Optional

- Paul DiMaggio (2015). "Adapting Computational Text Analysis to Social Science (and Vice Versa)". *Big Data & Society* 2 (2): 1–5.
- David Lazer and Jason Radford (2017). "Data ex Machina: Introduction to Big Data". *Annual Review of Sociology* 43: 19–39.
- Julia Hirschberg and Christopher D. Manning (2015). "Advances in Natural Language Processing". *Science* 349 (6245): 261–266.
- Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy (2017). "Text as Data". NBER Working Paper Series, Working Paper 23276.
- Martijn Schoonvelde, Gijs Schumacher, and Bert N. Bakker (2019). "Friends with Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology". *Journal of Social and Political Psychology* 7 (1): 124–143.

Week 2: Assumptions and Workflow (February 25)

- What are the underlying assumptions of text-as-data approaches?
- *Application*: importing textual data, creating a text corpus, and adding document-level variables

Readings

- John Wilkerson and Andreu Casas (2017). "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges". *Annual Review of Political Science* 20: 529–544.**
- Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). "quanteda: An R Package for the Quantitative Analysis of Textual Data". *The Journal of Open Source Software* 3 (30): 774.

Optional

- Fabrizio Gilardi and Bruno Wueest (2018). *Text-as-Data Methods for Comparative Policy Analysis*. URL: https://www.fabriziogilardi.org/resources/papers/Gilardi-Wueest-TextAsData-Policy-Analysis.pdf.
- Burt L. Monroe, Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair (2016). "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science". *PS: Political Science & Politics* 48 (1): 71–74.

Week 3: Tokenisation and Document-Feature Matrix (March 4)

– What are tokens, types, and features? What is the difference between stemming and lemmatisation?

– *Application*: tokenising texts, and creating a document-feature matrix

Readings

- Kasper Welbers, Wouter Van Atteveldt, and Kenneth Benoit (2017). "Text Analysis in R". *Communication Methods and Measures* 11 (4): 245–265.**
- Matthew W. Denny and Arthur Spirling (2018). "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It". *Political Analysis* 26 (2): 168–189.**
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press: Chapter 2.
- Kohei Watanabe and Stefan Müller (2019). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io: Chapter 3.

Week 4: Dictionaries and Sentiment Analysis (March 11)

– What are automated dictionary approaches? How can we create, test, and refine dictionaries?

– *Application*: creating multiword expressions and applying dictionaries to tokens objects and document-feature matrices

Readings

- Michael Laver and John Garry (2000). "Estimating Policy Positions from Political Texts". *American Journal of Political Science* 44 (3): 619–634.**
- Matthijs Rooduijn and Teun Pauwels (2011). "Measuring Populism: Comparing Two Methods of Content Analysis". *West European Politics* 34 (6): 1272–1283.**
- Stuart N. Soroka (2012). "The Gatekeeping Function: Distributions of Information in Media and the Real World". *The Journal of Politics* 74 (2): 514–528.
- Stuart N. Soroka and Christopher Wlezien (2018). "Tracking the Coverage of Public Policy in Mass Media". *Policy Studies Journal* published ahead of print (doi: 10.1111/psj.12285).
- Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart N. Soroka (2019). "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches". *Legislative Studies Quarterly* 44 (1): 97–131.

Optional

- Robert A. Stine (2019). "Sentiment Analysis". *Annual Review of Statistics and Its Application* published ahead of print (doi: 10.1146/annurev-statistics-030718-105242).
- Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair (2018). "More than Bags of Words: Sentiment Analysis with Word Embeddings". *Communication Methods and Measures* 12 (2–3): 140–157.
- Yla R. Tausczik and James W. Pennebaker (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". *Journal of Language and Social Psychology* 29 (1): 24–54.
- Ashley Muddiman, Shannon C. McGregor, and Natalie Jomini Stroud (2018). "(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries". *Political Communication* published ahead of print (doi: 10.1080/10584609.2018.1517843).
- Christian Rauh (2018). "Validating a Sentiment Dictionary for German Political Language: A Workbench Note". *Journal of Information Technology & Politics* 15 (4): 319–343.

Week 5: Textual Statistics, Text Similarity and Reuse (March 18)

- How do texts differ in their 'readability' and complexity? What are measures to estimate the similarity and distance between texts?
- *Application*: creating n-grams; estimating complexity and similarities/distances of texts

Readings

- James P. Cross and Henrik Hermansson (2017). "Legislative Amendments and Informal Politics in the European Union: A Text Reuse Approach". *European Union Politics* 18 (4): 581–602.**
- Daniel Bischof and Roman Senninger (2018). "Simple Politics for the People? Complexity in Campaign Messages and Political Knowledge". *European Journal of Political Research* 57 (2): 473–495.**

- John Wilkerson, David Smith, and Nicholas Stramp (2015). "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach". *American Journal of Political Science* 59 (4): 943–956.
- Kenneth Benoit, Kevin Munger, and Arthur Spirling (2019). "Measuring and Explaining Political Sophistication Through Textual Complexity". *American Journal of Political Science* published ahead of print (doi: 10.1111/ajps.12423).

Optional

- Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N. Bakker (2019). "Liberals Lecture, Conservatives Communicate: Analyzing Complexity and Ideology in 381,609 Political Speeches". *PLoS One* 14 (2): e0208450.
- Todd Allee and Andrew Lugg (2016). "Who Wrote the Rules for the Trans-Pacific Partnership?". *Research and Politics* 3 (3): 1–9.
- Fridolin Linder, Bruce A. Desmarais, Matthew Burgess, and Eugenia Giraudy (2018). "Text as Policy: Measuring Policy Similarity through Bill Text Reuse". *Policy Studies Journal* published ahead of print (doi: 10.1111/psj.12257).

Week 6: Human Coding and Document Classification (March 25)

– How can we classify documents into known and pre-defined categories? What is crowd-sourced coding?

– *Application*: typical workflow of human coding using crowdsourcing; Naïve Bayes classification

Readings

- Kenneth Benoit, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov (2016). "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data". *American Political Science Review* 110 (2): 278–295.**
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press: Chapter 13 (Naïve Bayes).**
- Slava Mikhaylov, Michael Laver, and Kenneth Benoit (2012). "Coder Reliability and Misclassifcation in the Human Coding of Party Manifestos". *Political Analysis* 20 (1): 78–91.
- Kenneth Benoit, Michael Laver, and Slava Mikhaylov (2009). "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions". *American Journal of Political Science* 53 (2): 495–513.

Optional

- Daniel Jurafsky and James H. Martin (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition: Chapter 4 (Naïve Bayes).

- Gary King, Patricka Lam, and Margaret E. Roberts (2017). "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text". *American Journal of Political Science* 61 (4): 971–988.
- Daniel J. Hopkins and Gary King (2010). "A Method of Automated Nonparametric Content Analysis for Social Science". *American Journal of Political Science* 54 (1): 229–247.
- Kohei Watanabe (2018). "Newsmap: A Semi-supervised Approach to Geographical News Classification". *Digital Journalism* 6 (3): 294–309.
- Julio Cesar Amador Diaz Lopez, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo (2017). "Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data". *Statistics, Politics and Policy* 8 (1): 85–104.
- Andrew Peterson and Arthur Spirling (2018). "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems". *Political Analysis* 26 (1): 120–128.
- Matt W. Loftis and Peter B. Mortensen (2018). "Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents". *Policy Studies Journal* published ahead of print (doi: 10.1111/psj.12245).

Week 7: Supervised Scaling (April 1)

- What are the assumptions, advantages, and problems of supervised scaling?
- *Application*: Worscores

Readings

- Michael Laver, John Garry, and Kenneth Benoit (2003). "Extracting Policy Positions from Political Texts Using Words as Data". *American Political Science Review* 97 (2): 311–331.**
- Michael Laver (2014). "Measuring Policy Positions in Political Space". *Annual Review of Political Science* 17: 207–223.
- Alexander Baturo, Niheer Dasandi, and Slava Mikhaylov (2017). "Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus". *Research and Politics* 4 (2): 1–9.
- Alexander Herzog and Kenneth Benoit (2015). "The Most Unkindest Cuts: Speaker Selection and Expressed Goverment Dissent During Economic Crisis". *The Journal of Politics* 77 (4): 1157–1175.

Optional

- Will Lowe (2008). "Understanding Wordscores". *Political Analysis* 16 (4): 356–371.
- Lanny W. Martin and Georg Vanberg (2008). "A Robust Transformation Procedure for Interpreting Political Text". *Political Analysis* 16 (1): 93–100.
- Patrick O. Perry and Kenneth Benoit (2017). *Scaling Text with the Class Affinity Model*. arXiv PrePrint. URL: https://arxiv.org/abs/1710.08963v1.

Week 8: Unsupervised Scaling (April 15)

- What are differences between supervised and unsupervised scaling methods? How can we validate scaling models?
- *Application*: Wordfish

Readings

- Jonathan B. Slapin and Sven-Oliver Proksch (2008). "A Scaling Model for Estimating Time-Series Party Positions from Texts". *American Journal of Political Science* 52 (3): 705–722.**
- Will Lowe and Kenneth Benoit (2013). "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark". *Political Analysis* 21 (3): 298–313.
- Daniel Schwarz, Denise Traber, and Kenneth Benoit (2017). "Estimating Intra-Party Preferences: Comparing Speeches to Votes". *Political Science Research and Methods* 5 (2): 379–396.
- Heike Klüver (2009). "Measuring Interest Group Influence Using Quantitative Text Analysis". *European Union Politics* 10 (4): 535–549.

Optional

- Benjamin E. Lauderdale and Alexander Herzog (2016). "Measuring Political Positions from Legislative Speech". *Political Analysis* 24 (3): 374–394.
- Amy Catalinac (2018). "Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos". *American Political Science Review* 112 (1): 31–48.
- Zachary Greene and Matthias Haber (2016). "Leadership Competition and Disagreement at Party National Congresses". *British Journal of Political Science* 46 (3): 611–632.
- Nicole Baerg and Will Lowe (2018). "A Textual Taylor Rule: Estimating Central Bank Preferences Combining Topic and Scaling Methods". *Political Science Research and Methods* published ahead of print (doi: 10.1017/psrm.2018.31).
- Anna Storz and Julian Bernauer (2018). "Supply and Demand of Populism: A Quantitative Text Analysis of Cantonal SVP Manifestos". *Swiss Political Science Review* 24 (4): 525–544.

Week 9: Topic Models (April 29)

- How does unsupervised document classification work? What are the assumptions, advantages, and caveats of topic models?
- *Application*: Latent Dirichlet allocation (LDA) and structural topic models (STM)

Readings

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3: 993–1022.**

- David M. Blei (2012). "Probabilistic Topic Models". *Communications of the ACM* 55 (4): 77–84.**
- Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand (2014). "Structural Topic Models for Open-Ended Survey Responses". *American Journal of Political Science* 58 (4): 1064–1082.

Optional

- Gregory J. Martin and Joshua McCrain (2019). "Local News and National Politics". *American Political Science Review* 113 (2): 372–384.
- Justin Grimmer (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". *Political Analysis* 18 (1): 1–35.
- Derek Greene and James P. Cross (2017). "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach". *Political Analysis* 25 (1): 77–94.
- Constantine Boussalis and Travis G. Coan (2016). "Text-Mining the Signals of Climate Change Doubt". *Global Environmental Change* 36: 89–100.
- Amy Catalinac (2016). "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections". *The Journal of Politics* 78 (1): 1–18.
- Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers (2016). "Quantitative Analysis of Large Amount of Journalistic Texts Using Topic Modelling". *Digital Journalism* 4 (1): 89–106.

Week 10: Presentation of Projects [I] (May 6)

In this session, the first half of students will present their projects. The remaining projects will be presented in the following session. Detailed instructions on the presentations, the written outline of the research design, and how to discuss each other's proposal will be distributed through OLAT.

Week 11: Presentation of Projects [II] (May 13)

In this session, the second half of students will present their projects.

Week 12: Social Media and Multilingual Analysis (May 20)

- How can we analyse social media posts with text-as-data approaches? In what ways can we conduct multilingual analyses?
- *Application*: scraping Twitter data using an API; introducing platforms for machine translation

Readings: Social Media

- Tamar Mitts (2019). "From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West". *American Political Science Review* 113 (1): 173–194.

- Jürgen Pfeffer, Katja Mayer, and Fred Morstatter (2018). "Tampering with Twitter's Sample API". *EPJ Data Science* 7 (50): 1–21.

Readings: Machine Translation

- Erik De Vries, Martijn Schoonvelde, and Gijs Schumacher (2018). "No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications". *Political Analysis* 26 (4): 417–430.**
- James A. Evans and Pedro Aceves (2016). "Machine Translation: Mining Text for Social Theory". *Annual Review of Sociology* 42: 21–50.
- Christopher Lucas, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley (2015). "Computer-Assisted Text Analysis for Comparative Politics". *Political Analysis* 23 (2): 254–277.

Week 13: New Directions and Applications (May 27)

- What are future directions in natural language processing?
- *Application*: introducing assumptions of word2vec and deep learning approaches

Readings

- Arthur Spirling and Pedro L. Rodriguez. *Word Embeddings: What Works, What Doesn't, and Hot to Tell the Difference for Applied Research*. Unpublished Manuscript, New York University. URL: https://www.nyu.edu/projects/spirling/documents/embed.pdf.**
- Gary King, Jennifer Pan, and Margaret E. Roberts (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression". *American Political Science Review* 107 (2): 326–343.**
- Sven-Oliver Proksch, Christopher Wratil, and Jens Wäckerle (2019). "Testing the Validity of Automatic Speech Recognition for Political Text Analysis". *Political Analysis* published ahead of print (doi: 10.1017/pan.2018.62).
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep Learning". *Nature* 521: 436–444.
- Hannes Mueller and Christopher Rauh (2018). "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text". *American Political Science Review* 112 (2): 358–375.
- Jungseock Joo and Zachary C. Steinert-Threlkeld (2018). *Image as Data: Automated Visual Content Analysis for Political Science*. arXiv PrePrint. URL: https://arxiv.org/abs/1810.01544.

Optional

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781. URL: https://arxiv.org/abs/1301.3781.
- François Chollet and J.J. Allaire (2018). *Deep Learning with R*. Shelter Island: Manning.

- Matthew J. Salganik (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

References

Allee, Todd and Andrew Lugg (2016). "Who Wrote the Rules for the Trans-Pacific Partnership?". *Research and Politics* 3 (3): 1–9.

Baerg, Nicole and Will Lowe (2018). "A Textual Taylor Rule: Estimating Central Bank Preferences Combining Topic and Scaling Methods". *Political Science Research and Methods* published ahead of print (doi: 10.1017/psrm.2018.31).

Baturo, Alexander, Niheer Dasandi, and Slava Mikhaylov (2017). "Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus". *Research and Politics* 4 (2): 1–9.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov (2016). "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data". *American Political Science Review* 110 (2): 278–295.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov (2009). "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions". *American Journal of Political Science* 53 (2): 495–513.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling (2019). "Measuring and Explaining Political Sophistication Through Textual Complexity". *American Journal of Political Science* published ahead of print (doi: 10.1111/ajps.12423).

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). "quanteda: An R Package for the Quantitative Analysis of Textual Data". *The Journal of Open Source Software* 3 (30): 774.

Bischof, Daniel and Roman Senninger (2018). "Simple Politics for the People? Complexity in Campaign Messages and Political Knowledge". *European Journal of Political Research* 57 (2): 473–495.

Blei, David M. (2012). "Probabilistic Topic Models". *Communications of the ACM* 55 (4): 77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3: 993–1022.

Boussalis, Constantine and Travis G. Coan (2016). "Text-Mining the Signals of Climate Change Doubt". *Global Environmental Change* 36: 89–100.

Brady, Henry E. (2019). "The Challenge of Big Data and Data Science". *Annual Review of Political Science* published ahead of print.

Catalinac, Amy (2016). "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections". *The Journal of Politics* 78 (1): 1–18.

Catalinac, Amy (2018). "Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos". *American Political Science Review* 112 (1): 31–48.

Chollet, François and J.J. Allaire (2018). *Deep Learning with R*. Shelter Island: Manning.

Cross, James P. and Henrik Hermansson (2017). "Legislative Amendments and Informal Politics in the European Union: A Text Reuse Approach". *European Union Politics* 18 (4): 581–602.

De Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher (2018). "No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications". *Political Analysis* 26 (4): 417–430.

Denny, Matthew W. and Arthur Spirling (2018). "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It". *Political Analysis* 26 (2): 168–189.

Diaz Lopez, Julio Cesar Amador, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo (2017). "Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data". *Statistics, Politics and Policy* 8 (1): 85–104.

DiMaggio, Paul (2015). "Adapting Computational Text Analysis to Social Science (and Vice Versa)". *Big Data & Society* 2 (2): 1–5.

Evans, James A. and Pedro Aceves (2016). "Machine Translation: Mining Text for Social Theory". *Annual Review of Sociology* 42: 21–50.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy (2017). "Text as Data". NBER Working Paper Series, Working Paper 23276.

Gilardi, Fabrizio and Bruno Wueest (2018). *Text-as-Data Methods for Comparative Policy Analysis*. URL: https://www.fabriziogilardi.org/resources/papers/Gilardi-Wueest-TextAsData-Policy-Analysis.pdf.

Greene, Derek and James P. Cross (2017). "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach". *Political Analysis* 25 (1): 77–94.

Greene, Zachary and Matthias Haber (2016). "Leadership Competition and Disagreement at Party National Congresses". *British Journal of Political Science* 46 (3): 611–632.

Grimmer, Justin (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". *Political Analysis* 18 (1): 1–35.

Grimmer, Justin and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". *Political Analysis* 21 (3): 267–297.

Healy, Kieran (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press.

Herzog, Alexander and Kenneth Benoit (2015). "The Most Unkindest Cuts: Speaker Selection and Expressed Goverment Dissent During Economic Crisis". *The Journal of Politics* 77 (4): 1157–1175.

Hirschberg, Julia and Christopher D. Manning (2015). "Advances in Natural Language Processing". *Science* 349 (6245): 261–266.

Hopkins, Daniel J. and Gary King (2010). "A Method of Automated Nonparametric Content Analysis for Social Science". *American Journal of Political Science* 54 (1): 229–247.

Jacobi, Carina, Wouter Van Atteveldt, and Kasper Welbers (2016). "Quantitative Analysis of Large Amount of Journalistic Texts Using Topic Modelling". *Digital Journalism* 4 (1): 89–106.

Joo, Jungseock and Zachary C. Steinert-Threlkeld (2018). *Image as Data: Automated Visual Content Analysis for Political Science*. arXiv PrePrint. URL: https://arxiv.org/abs/1810.01544.

Jurafsky, Daniel and James H. Martin (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition.

King, Gary, Patricka Lam, and Margaret E. Roberts (2017). "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text". *American Journal of Political Science* 61 (4): 971–988.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression". *American Political Science Review* 107 (2): 326–343.

Klüver, Heike (2009). "Measuring Interest Group Influence Using Quantitative Text Analysis". *European Union Politics* 10 (4): 535–549.

Lauderdale, Benjamin E. and Alexander Herzog (2016). "Measuring Political Positions from Legislative Speech". *Political Analysis* 24 (3): 374–394.

Laver, Michael (2014). "Measuring Policy Positions in Political Space". *Annual Review of Political Science* 17: 207–223.

Laver, Michael and John Garry (2000). "Estimating Policy Positions from Political Texts". *American Journal of Political Science* 44 (3): 619–634.

Laver, Michael, John Garry, and Kenneth Benoit (2003). "Extracting Policy Positions from Political Texts Using Words as Data". *American Political Science Review* 97 (2): 311–331.

Lazer, David and Jason Radford (2017). "Data ex Machina: Introduction to Big Data". *Annual Review of Sociology* 43: 19–39.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep Learning". *Nature* 521: 436–444.

Linder, Fridolin, Bruce A. Desmarais, Matthew Burgess, and Eugenia Giraudy (2018). "Text as Policy: Measuring Policy Similarity through Bill Text Reuse". *Policy Studies Journal* published ahead of print (doi: 10.1111/psj.12257).

Loftis, Matt W. and Peter B. Mortensen (2018). "Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents". *Policy Studies Journal* published ahead of print (doi: 10.1111/psj.12245).

Lowe, Will (2008). "Understanding Wordscores". *Political Analysis* 16 (4): 356–371.

Lowe, Will and Kenneth Benoit (2013). "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark". *Political Analysis* 21 (3): 298–313.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley (2015). "Computer-Assisted Text Analysis for Comparative Politics". *Political Analysis* 23 (2): 254–277.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press.

Martin, Gregory J. and Joshua McCrain (2019). "Local News and National Politics". *American Political Science Review* 113 (2): 372–384.

Martin, Lanny W. and Georg Vanberg (2008). "A Robust Transformation Procedure for Interpreting Political Text". *Political Analysis* 16 (1): 93–100.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit (2012). "Coder Reliability and Misclassifcation in the Human Coding of Party Manifestos". *Political Analysis* 20 (1): 78–91.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781. URL: https://arxiv.org/abs/1301.3781.

Mitts, Tamar (2019). "From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West". *American Political Science Review* 113 (1): 173–194.

Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair (2016). "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science". *PS: Political Science & Politics* 48 (1): 71–74.

Muddiman, Ashley, Shannon C. McGregor, and Natalie Jomini Stroud (2018). "(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries". *Political Communication* published ahead of print (doi: 10.1080/10584609.2018.1517843).

Mueller, Hannes and Christopher Rauh (2018). "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text". *American Political Science Review* 112 (2): 358–375.

Perry, Patrick O. and Kenneth Benoit (2017). *Scaling Text with the Class Affinity Model*. arXiv PrePrint. URL: https://arxiv.org/abs/1710.08963v1.

Peterson, Andrew and Arthur Spirling (2018). "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems". *Political Analysis* 26 (1): 120–128.

Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter (2018). "Tampering with Twitter's Sample API". *EPJ Data Science* 7 (50): 1–21.

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart N. Soroka (2019). "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches". *Legislative Studies Quarterly* 44 (1): 97–131.

Proksch, Sven-Oliver, Christopher Wratil, and Jens Wäckerle (2019). "Testing the Validity of Automatic Speech Recognition for Political Text Analysis". *Political Analysis* published ahead of print (doi: 10.1017/pan.2018.62).

Rauh, Christian (2018). "Validating a Sentiment Dictionary for German Political Language: A Workbench Note". *Journal of Information Technology & Politics* 15 (4): 319–343.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand (2014). "Structural Topic Models for Open-Ended Survey Responses". *American Journal of Political Science* 58 (4): 1064–1082.

Rooduijn, Matthijs and Teun Pauwels (2011). "Measuring Populism: Comparing Two Methods of Content Analysis". *West European Politics* 34 (6): 1272–1283.

Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair (2018). "More than Bags of Words: Sentiment Analysis with Word Embeddings". *Communication Methods and Measures* 12 (2–3): 140–157.

Salganik, Matthew J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Schoonvelde, Martijn, Anna Brosius, Gijs Schumacher, and Bert N. Bakker (2019). "Liberals Lecture, Conservatives Communicate: Analyzing Complexity and Ideology in 381,609 Political Speeches". *PLoS One* 14 (2): e0208450.

Schoonvelde, Martijn, Gijs Schumacher, and Bert N. Bakker (2019). "Friends with Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology". *Journal of Social and Political Psychology* 7 (1): 124–143.

Schwarz, Daniel, Denise Traber, and Kenneth Benoit (2017). "Estimating Intra-Party Preferences: Comparing Speeches to Votes". *Political Science Research and Methods* 5 (2): 379–396.

Slapin, Jonathan B. and Sven-Oliver Proksch (2008). "A Scaling Model for Estimating Time-Series Party Positions from Texts". *American Journal of Political Science* 52 (3): 705–722.

Soroka, Stuart N. (2012). "The Gatekeeping Function: Distributions of Information in Media and the Real World". *The Journal of Politics* 74 (2): 514–528.

Soroka, Stuart N. and Christopher Wlezien (2018). "Tracking the Coverage of Public Policy in Mass Media". *Policy Studies Journal* published ahead of print (doi: 10.1111/psj.12285).

Spirling, Arthur and Pedro L. Rodriguez. *Word Embeddings: What Works, What Doesn't, and Hot to Tell the Difference for Applied Research*. Unpublished Manuscript, New York University. URL: https://www.nyu.edu/projects/spirling/documents/embed.pdf.

Stine, Robert A. (2019). "Sentiment Analysis". *Annual Review of Statistics and Its Application* published ahead of print (doi: 10.1146/annurev-statistics-030718-105242).

Storz, Anna and Julian Bernauer (2018). "Supply and Demand of Populism: A Quantitative Text Analysis of Cantonal SVP Manifestos". *Swiss Political Science Review* 24 (4): 525–544.

Tausczik, Yla R. and James W. Pennebaker (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". *Journal of Language and Social Psychology* 29 (1): 24–54.

Watanabe, Kohei (2018). "Newsmap: A Semi-supervised Approach to Geographical News Classification". *Digital Journalism* 6 (3): 294–309.

Watanabe, Kohei and Stefan Müller (2019). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io.

Welbers, Kasper, Wouter Van Atteveldt, and Kenneth Benoit (2017). "Text Analysis in R". *Communication Methods and Measures* 11 (4): 245–265.

Wickham, Hadley and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly.

Wilke, Claus O. (Forthcoming). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly.

Wilkerson, John and Andreu Casas (2017). "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges". *Annual Review of Political Science* 20: 529–544.

Wilkerson, John, David Smith, and Nicholas Stramp (2015). "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach". *American Journal of Political Science* 59 (4): 943–956.

Xie, Yihui, J.J. Allaire, and Garrett Grolemund (2018). *R Markdown: The Definite Guide*. Boca Raton: CRC Press, Taylor & Francis Group.