Stefan Müller, PhD
Assistant Professor and Ad Astra Fellow
School of Politics and International Relations
University College Dublin
Belfield, Dublin 4, Ireland
✉ stefan.mueller@ucd.ie
☞ https://muellerstefan.net

Level 4 Module; Spring Trimester 2023

Quantitative Text Analysis (POL42050)

Draft (Version: December 6, 2022)

Latest version at: https://muellerstefan.net/teaching/2023-spring-qta.pdf

Time: Tuesday, 09:00–11:00	Module coordinator: Stefan Müller, PhD
Location: QUI-113 (Quinn School of Business)	stefan.mueller@ucd.ie | https://muellerstefan.net
Credits: 10.0	Office: Newman Building, G312
Format: Lecture and computer labs	Office hours: Tuesday, 11:15–13:00 (sign up here)

Course Content

Automated text analysis has become very popular in political science over the past years. With the massive availability of text data on the web, political scientists increasingly recognize automated text analysis (or "text as data") as a promising approach for analyzing various kinds of social and political behaviour. This module introduces students of political science to the quantitative analysis of textual data. We discuss the underlying theoretical assumptions, substantive applications of these methods, and the respective implementations in the R statistical programming language.

Each session combines lectures with practical, hands-on exercises to apply the methods to political text, dealing with practical issues in each step of the research process. 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. The course will also introduce advanced methods, including word embeddings, speech transcription, machine translation, and computer vision.

Learning Outcomes

Upon successful completion of the 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 of texts to train supervised classifiers.
4. Apply these methods to their own text corpus to address a substantive research question.
5. Critically evaluate (social science) research that uses automated text analysis methods.

General Readings

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

- K. Benoit (2020). "Text as Data: An Overview". *Handbook of Research Methods in Political Science and International Relations*. Ed. by L. Curini and R. Franzese. Thousand Oaks: Sage: 461–497.
- K. Watanabe and S. Müller (2022). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io. https://tutorials.quanteda.io
- K. Benoit and S. Müller (Work in Progress). *Text Analysis Using R*. URL: https://quanteda.github.io/Text-Analysis-Using-R/.
- E. Hvitfeldt and J. Silge (2021). *Supervised Machine Learning For Text Analysis in R*. Boca Raton: CRC Press.
- J. Grimmer, M. E. Roberts, and B. M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. 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 helpful for solving practical problems. Most books listed in the syllabus are published in print, but also freely available online.

R and RMarkdown

- C. Ismay and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press.
- H. Wickham and G. Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly.

Data Visualisation

- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press.

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:

- Quantitative text analysis:

  - `quanteda`
  - `quanteda.textmodels`
  - `quanteda.textstats`
  - `quanteda.textplots`
- Importing text data: `readtext`
- Topic models: `stm`, keyATM
- Data wrangling and visualisation: `tidyverse` (esp. `dplyr`, `tidyr`, `lubridate`, and `ggplot2`)
- Creating documents and reports: `Quarto`
- Part-of-speech tagging and lemmatisation: `spacyr` (not mandatory to install this package)

Plagiarism

Although this should be obvious, plagiarism – copying someone else's text without acknowledgement or beyond 'fair use' quantities – is not allowed. Plagiarism is an issue we take very serious here in UCD. Please familiarize yourself with the definition of plagiarism on UCD's website[1] and make sure not to engage in it.

Late Submission Policy

If a student submits an assignment late, the following penalties will be applied:

- Coursework received at any time within two weeks of the due date will be graded, but a penalty will apply.
  - Coursework submitted at any time up to one week after the due date will have the grade awarded reduced by two grade points (for example, from $B-$ to C).
  - Coursework submitted more than one week but up to two weeks after the due date will have the grade reduced by four grade points (for example, from $B-$ to $D+$). Where a student finds they have missed a deadline for submission, they should be advised that they may use the remainder of the week to improve their submission without additional penalty.
- Coursework received more than two weeks after the due date will not be accepted. Regulations regarding extenuating circumstances apply.

Office Hours

My office hours take place on Tuesday from 11:15–13:00, either in in person (Room G312, Newman Building) or online. Please sign up for a meeting at https://calendly.com/mueller-ucd/office-hours.

Questions and Problems

In this module, we will discuss concepts, methods, and software you might not have heard of before. I am aware that parts of this module could be challenging, and I will assist you as best as I can.

[1] https://libguides.ucd.ie/academicintegrity.

We will use Slack for in this module. Make sure to create a Slack account before the first seminar and join the Slack workspace. If you have a question that involve code or concepts, please share your question in #questions, #homework, or #research-paper.

If you struggle to solve problems relating to `R` or RStudio, please follow the steps outlined below before contacting your peers or me. It is very likely that at least one other person faced the same problem before or received the same error message.

1. Try to summarise the problem in your own words and then google this summary. If the problem relates to `R`, add `rstats` to your search query. For example: `how to import csv file in rstats`. I am almost certain that you find a solution to most of your questions.

2. If your `R` code returns an error, I would advise you to Google the text of the error message. For example, you google the error message "`Error: Can't subset columns that don't exist.`"

⟶ If steps 1–2 still do not solve your problem or question, please ask your question in the Slack channel devoted to this module. Your peers and I will help you.

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. If I make adjustments, I will email all seminar participants and upload the revised syllabus to Brightspace.

Expectations and Grading

- Students are expected to read all papers or chapters assigned under **Mandatory 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. I strongly encourage students to (at least) skim these readings. Both the required and the optional readings consist of technical readings and at least one practical application of the respective method.

- Students submit two **Homework assignments**, each of which counts towards 25% of the final grade. The assignments will be distributed as a Quarto file 14 days before the submission deadline. Students fill in the answers and solutions in the same Quarto file, rename it to `hw_01/02_surname_firstname.qmd`, render it as an `html` file, and submit it via Brightspace. Only rendered `html` files will be accepted! Homework 1 will be submitted at the end of Week 5. Homework 2 will be submitted at the end of Week 9. More details on the homework will be provided in the first session(s) of the course.

- Students also submit a short **Research Paper** of 3,000 words (excluding references and appendices). The research paper counts towards 50% of the final grade and must be submitted by 5 May 2023. In Homework 2, students will be required to briefly outline the research question they want to test in the research paper and describe which textual data they will use for testing this question.

  In the research paper, the students should succinctly but clearly write up the results of a small research project using quantitative text analysis methods. Students are free to collect their own data or use existing data. Creativity is encouraged. Students are free to answer questions from all subfields of political science but must justify their choice and the relevance of the question. This paper should contain the following elements:

1. Introduction and research question: introduction to the topic, research question, and relevance.
2. Expectations: a concise overview of the theoretical expectation(s) that will be tested in the results section.
3. Data and methods: description of the data sources as well as the methods employed.
4. Results: a discussion (with figures and tables) of the results of the analysis. This section forms the bulk of the paper.
5. Conclusion: a brief evaluation of the results and steps to push the research forward.

Overview of deadlines

Date	Assignment
End of Week 5	Homework 1 (25%)
End of Week 9	Homework 2 (25%)
Friday, 5 May 2023	Research Paper (50%)

Grading Criteria

In essence, markers assess four crucial elements in any answer:

- Analysis/understanding
- Extent and use of reading
- Organisation/structure
- Writing proficiency

The various grades/classifications listed below reflect the extent to which an answer displays essential features of each of these elements (and their relative weighting). At its simplest: the better the analysis, the wider the range of appropriate sources consulted, the greater the understanding of the materials read, the clearer the writing style, and the more structured the argument, the higher will be the mark.

The following provides an indicative outline of the criteria used by markers to award a particular grade/classification. If you are in any confusion about how to correctly approach referencing and bibliography issues, please refer to the following guidelines: APSA Committee on Publications (2018). *Style Manual for Political Science (Revised 2018 Version)*. URL: https://connect.apsanet.org/stylemanual/.

Proper referencing is ESSENTIAL in a good assignment.

Grade Explanation for Research Paper

Grade: A (Excellent Performance)

A deep and systematic engagement with the assessment task, with consistently impressive demonstration of a comprehensive mastery of the subject matter, reflecting:

- A deep and broad knowledge and critical insight as well as extensive reading
- A critical and comprehensive appreciation of the relevant literature or theoretical, technical or professional framework
- An exceptional ability to organise, analyse and present arguments fluently and lucidly with a high level of critical analysis, amply supported by evidence, citation or quotation;
- A highly-developed capacity for original, creative and logical thinking
- An extensive and detailed knowledge of the subject matter
- A highly-developed ability to apply this knowledge to the task set
- Evidence of extensive background reading
- Clear, fluent, stimulating and original expression
- Excellent presentation (spelling, grammar, graphical) with minimal or no presentation errors
- Referencing style consistently executed in recognised style

Grade: B (Very Good Performance)

A thorough and well organised response to the assessment task, demonstrating:

- A thorough familiarity with the relevant literature or theoretical, technical or professional framework
- Well-developed capacity to analyse issues, organise material, present arguments clearly and cogently well supported by evidence, citation or quotation;
- Some original insights and capacity for creative and logical thinking
- A broad knowledge of the subject matter
- Considerable strength in applying that knowledge to the task set
- Evidence of substantial background reading
- Clear and fluent expression
- Quality presentation with few presentation errors
- Referencing style for the most part consistently executed in recognised style

Grade: C (Good Performance)

An intellectually competent and factually sound answer with, marked by:

- Evidence of a reasonable familiarity with the relevant literature or theoretical, technical or professional framework
- Good developed arguments, but more statements of ideas
- Arguments or statements adequately but not well supported by evidence, citation or quotation
- Some critical awareness and analytical qualities
- Some evidence of capacity for original and logical thinking
- Adequate but not complete knowledge of the subject matter
- Omission of some important subject matter or the appearance of several minor errors
- Capacity to apply knowledge appropriately to the task albeit with some errors
- Evidence of some background reading
- Clear expression with few areas of confusion
- Writing of sufficient quality to convey meaning but some lack of fluency and command of suitable vocabulary
- Good presentation with some presentation errors
- Referencing style executed in recognised style, but with some errors

Grade: D (Satisfactory Performance)

An acceptable level of intellectual engagement with the assessment task showing:

- Some familiarity with the relevant literature or theoretical, technical or professional framework
- Mostly statements of ideas, with limited development of argument
- Limited use of evidence, citation or quotation
- Limited critical awareness displayed
- Limited evidence of capacity for original and logical thinking
- Basic grasp of subject matter, but somewhat lacking in focus and structure
- Main points covered but insufficient detail
- Some effort to apply knowledge to the task but only a basic capacity or understanding displayed
- Little or no evidence of background reading
- Several minor errors or one major error
- Satisfactory presentation with an acceptable level of presentation errors
- Referencing style inconsistent

Grade: D– (Acceptable)

The minimum acceptable of intellectual engagement with the assessment task which:

- The minimum acceptable appreciation of the relevant literature or theoretical, technical or professional framework
- Ideas largely expressed as statements, with little or no developed or structured argument
- Minimum acceptable use of evidence, citation or quotation
- Little or no analysis or critical awareness displayed or is only partially successful
- Little or no demonstrated capacity for original and logical thinking
- Shows a basic grasp of subject matter but may be poorly focused or badly structured or contain irrelevant material
- Has one major error and some minor errors
- Demonstrates the capacity to complete only moderately difficult tasks related to the subject material
- No evidence of background reading
- Displays the minimum acceptable standard of presentation (spelling, grammar, graphical)
- Referencing inconsistent with major errors

Grade: E (Fail [marginal])

A factually sound answer with a partially successful, but not entirely acceptable, attempt to:

- Integrate factual knowledge into a broader literature or theoretical, technical or professional framework develop arguments
- Support ideas or arguments with evidence, citation or quotation
- Engages with the subject matter or problem set, despite major deficiencies in structure, relevance or focus
- Has two major error and some minor errors
- Demonstrates the capacity to complete only part of, or the simpler elements of, the task
- An incomplete or rushed answer (e.g. the use of bullet points through part/all of answer)
- Little or no referencing style evident

Grade: F (Fail [unacceptable])

An unacceptable level of intellectual engagement with the assessment task, with:

- No appreciation of the relevant literature or theoretical, technical or professional framework
- No developed or structured argument
- No use of evidence, citation or quotation
- No analysis or critical awareness displayed or is only partially successful
- No demonstrated capacity for original and logical thinking
- A failure to address the question resulting in a largely irrelevant answer or material of marginal relevance predominating
- A display of some knowledge of material relative to the question posed, but with very serious omissions / errors and/or major inaccuracies included in answer
- Solutions offered to a very limited portion of the problem set
- An answer unacceptably incomplete (e.g. for lack of time)
- A random and undisciplined development, layout or presentation
- Unacceptable standards of presentation, such as grammar, spelling or graphical presentation
- Evidence of substantial plagiarism
- No referencing style evident

Grade: G (Fail [wholly unacceptable])

No intellectual engagement with the assessment task

- Complete failure to address the question resulting in an entirely irrelevant answer
- Little or no knowledge displayed relative to the question posed
- Little or no solution offered for the problem set
- Evidence of extensive plagiarism
- No referencing style evident

Grade: NG (No Grade)

No work was submitted by the student or student was absent from the assessment, or work submitted did not merit a grade.

Course Structure

Week 1: Introduction to Quantitative Text Analysis (24 January)

– What are quantitative text analysis and natural language processing?

– What is the structure of the module, and what are the expectations?

– *Application*: installing packages and setting up a project in RStudio

Readings

- K. Benoit (2020). "Text as Data: An Overview". *Handbook of Research Methods in Political Science and International Relations*. Ed. by L. Curini and R. Franzese. Thousand Oaks: Sage: 461–497.
- J. Grimmer and B. M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". *Political Analysis* 21 (3): 267–297.

Optional

- J. Wilkerson and A. Casas (2017). "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges". *Annual Review of Political Science* 20: 529–544.
- M. Gentzkow, B. T. Kelly, and M. Taddy (2019). "Text as Data". *Journal of Economic Literature* 57 (3): 535–574.

Week 2: R and RMarkdown (31 January)

– How to use R?

– An overview of important R functions

– How to structure the workflow for a quantitative research project?

Mandatory Readings

- C. Ismay and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: chapter 1.
- K. Watanabe and S. Müller (2022). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io: chapter 1.

Week 3: Assumptions and Workflow (7 February)

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

Mandatory Readings

- J. Grimmer, M. E. Roberts, and B. M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press: chapter 4.
- K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo (2018). "quanteda: An R Package for the Quantitative Analysis of Textual Data". *The Journal of Open Source Software* 3 (30): 774.

Optional

- C. D. Manning, P. Raghavan, and H. Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press: chapter 2.
- M. Schoonvelde, G. Schumacher, and B. 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 4: Tokenisation and Document-Feature Matrix (14 February)

- What are tokens, types, and features?
- What is the difference between stemming and lemmatisation?
- What information can we extract from a document-feature matrix?
- *Application*: tokenising texts, removing features, and creating a document-feature matrix

Mandatory Readings

- J. Grimmer, M. E. Roberts, and B. M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press: chapter 5.
- K. Watanabe and S. Müller (2022). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io: chapter 3.

Optional

- M. W. Denny and A. 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.
- K. Welbers, W. Van Atteveldt, and K. Benoit (2017). "Text Analysis in R". *Communication Methods and Measures* 11 (4): 245–265.

Week 5: Dictionaries and Sentiment Analysis (21 February)

- What are automated dictionary approaches? How can we create, validate, refine, and apply dictionaries?
- *Application*: creating multi-word expressions and applying dictionaries to tokens objects and document-feature matrices

Mandatory Readings

- S.-O. Proksch, W. Lowe, J. Wäckerle, and S. N. Soroka (2019). "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches". *Legislative Studies Quarterly* 44 (1): 97–131.
- S. Müller (2020). "Media Coverage of Campaign Promises Throughout the Electoral Cycle". *Political Communication* 37 (5): 696–718.

Optional

- A. Muddiman, S. C. McGregor, and N. J. Stroud (2019). "(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries". *Political Communication* 36 (2): 214–226.
- C. Rauh (2018). "Validating a Sentiment Dictionary for German Political Language: A Workbench Note". *Journal of Information Technology & Politics* 15 (4): 319–343.

Week 6: Describing and Comparing Texts (28 February)

- How do texts differ in their 'readability' and complexity? What are measures to estimate the similarity and distance between texts?
- How can we identify distinct features in texts?
- *Application*: estimating readability, similarity, and "keyness"

Mandatory Readings

- D. Bischof and R. Senninger (2018). "Simple Politics for the People? Complexity in Campaign Messages and Political Knowledge". *European Journal of Political Research* 57 (2): 473–495.
- J. P. Cross and H. Hermansson (2017). "Legislative Amendments and Informal Politics in the European Union: A Text Reuse Approach". *European Union Politics* 18 (4): 581–602.
- K. Benoit, K. Munger, and A. Spirling (2019). "Measuring and Explaining Political Sophistication Through Textual Complexity". *American Journal of Political Science* 63 (2): 491–508.

Optional

- J. Blumenau (2021). "The Effects of Female Leadership on Women's Voice in Political Debate". *British Journal of Political Science* 51 (2): 750–771.
- E. Hengel (2022). "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review". *The Economic Journal* 132 (648): 2951–2991.

Week 7: Human Coding and Document Classification (7 March)

- How can we classify documents into known and pre-defined categories? How do we create a training set? How do we assess the classification performance?
- *Application*: supervised machine learning using a Naïve Bayes classifier and a Support Vector Machine

Mandatory Readings

- C. D. Manning, P. Raghavan, and H. Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press: Chapter 13 (Naïve Bayes).
- S. Müller (2022). "The Temporal Focus of Campaign Communication". *The Journal of Politics* 84 (1): 585–590.
- A. Peterson and A. Spirling (2018). "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems". *Political Analysis* 26 (1): 120–128.

Optional

- D. Jurafsky and J. H. Martin (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition: Chapter 4 (Naïve Bayes).
- K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2016). "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data". *American Political Science Review* 110 (2): 278–295.
- S. Mikhaylov, M. Laver, and K. Benoit (2012). "Coder Reliability and Misclassification in the Human Coding of Party Manifestos". *Political Analysis* 20 (1): 78–91.

Week 8: Supervised, Unsupervised and Semi-Supervised Scaling (28 March)

- What are the assumptions, advantages, and problems of supervised and unsupervised scaling?
- How can we use supervised scaling to answer substantive questions?
- *Application*: Wordscores and Wordfish

Mandatory Readings

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

- J. B. Slapin and S.-O. Proksch (2008). "A Scaling Model for Estimating Time-Series Party Positions from Texts". *American Journal of Political Science* 52 (3): 705–722.
- K. Watanabe (2021). "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages". *Communication Methods and Measures* 14 (2): 81–102.

Optional

- D. Zollinger (2022). "Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses". *American Journal of Political Science* published ahead of print (doi: 10.1111/ajps.12743).
- A. Baturo, N. Dasandi, and S. Mikhaylov (2017). "Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus". *Research & Politics* 4 (2): 1–9.
- T. Gessler and S. Hunger (2021). "How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration". *Political Science Research and Methods* published ahead of print (doi: 10.1017/psrm.2021.64).
- T. O'Grady (2019). "Careerists Versus Coal-Miners: Welfare Reforms and the Substantive Representation of Social Groups in the British Labour Party". *Comparative Political Studies* 52 (4): 544–578.
- A. Herzog and S. Mikhaylov (2020). "Intra-Cabinet Politics and Fiscal Governance in Times of Austerity". *Political Science Research and Methods* 8 (3): 409–424.

Week 9: Retrieving, Loading and Wrangling Text Corpora (4 April)

- What are typical text corpora you can use for your final research paper?
- What are APIs are how can we use them to retrieve data?
- How can we load various types of text copora and transform them into a quanteda corpus object?
- What are legal and ethical requirements and challenges when working with social media data?
- *Application*: Manifesto Corpus, UN General Debate Corpus, Guardian API, Twitter API

Mandatory Readings

- H. Wickham and G. Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly: ch. 11–12.
- P. C. Bauer and C. Landesvatter, eds. (2022). *APIs for Social Scientists: A Collaborative Review*: skim potentially relevant chapters.

Optional

- N. Merz, S. Regel, and J. Lewandowski (2016). "The Manifesto Corpus: A New Resource for Research on Political Parties and Quantitative Text Analysis". *Research & Politics* 3 (2): 1–8.
- Z. C. Steinert-Threlkeld (2018). *Twitter as Data*. Cambridge: Cambridge University Press.

- A. Baturo, N. Dasandi, and S. Mikhaylov (2017). "Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus". *Research & Politics* 4 (2): 1–9.
- M. J. Salganik (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press: ch. 6.

Week 10: Topic Models (11 April)

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

Mandatory Readings

- T. Gessler (2022). "Topic Models". *Elgar Encyclopedia of Technology and Politics*. Ed. by A. Ceron. Cheltenham: 108–111.
- D. M. Blei (2012). "Probabilistic Topic Models". *Communications of the ACM* 55 (4): 77–84.
- M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). "Structural Topic Models for Open-Ended Survey Responses". *American Journal of Political Science* 58 (4): 1064–1082.

Optional

- F. Gilardi, C. R. Shipan, and B. Wüest (2021). "Policy Diffusion: The Issue-Definition Stage". *American Journal of Political Science* 65 (1): 21–35.
- D. Greene and J. P. Cross (2017). "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach". *Political Analysis* 25 (1): 77–94.
- R. Parthasarathy, V. Rao, and N. Palaniswamy (2019). "Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies". *American Political Science Review* 113 (3): 623–640.
- A. Catalinac (2016). "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections". *The Journal of Politics* 78 (1): 1–18.

Week 11: New Developments in Data: Images, Speech Recognition, Machine Translation (18 April)

- How can we extract text from videos and audio files?
- How can we use computer vision techniques to address social science questions?
- How can we conduct multilingual text analysis?
- *Application*: an introduction to APIs and website for machine translation, speech transcription, and image recognition

Mandatory Readings

- S.-O. Proksch, C. Wratil, and J. Wäckerle (2019). "Testing the Validity of Automatic Speech Recognition for Political Text Analysis". *Political Analysis* 27 (3): 339–359.
- C. Boussalis, T. G. Coan, M. R. Holman, and S. Müller (2021). "Gender, Candidate Emotional Expression, and Voter Reactions During Televised Debates". *American Political Science Review* 115 (4): 1242–1257.
- E. De Vries, M. Schoonvelde, and G. 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.

Optional

- C. Schwemmer, C. Knight, E. D. Bello-Pardo, S. Oklobdzija, M. Schoonvelde, and J. W. Lockhart (2020). "Diagnosing Gender Bias in Image Recognition Systems". *Socius: Sociological Research for a Dynamic World* 6: 1–17.
- N. Düpont and M. Rachuj (2021). "The Ties That Bind: Text Similarities and Conditional Diffusion among Parties". *British Journal of Political Science* published ahead of print (doi: 10.1017/S0007123420000617).

Week 12: New Developments in Modeling: Word Embeddings (25 April)

- What are word embeddings, and how do they improve classic bag-of-words approaches?
- *Application*: using pre-trained and locally fit word embeddings; using word embeddings in a regression framework

Mandatory Readings

- E. Hvitfeldt and J. Silge (2021). *Supervised Machine Learning For Text Analysis in R*. Boca Raton: CRC Press: chapter 5.
- P. L. Rodriguez and A. Spirling (2022). "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research". *The Journal of Politics* 84 (1): 101–115.
- C. Baden, C. Pipal, M. Schoonvelde, and M. Van der Velden (2022). "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda". *Communication Methods and Measures* 16 (1): 1–18.

Optional

- P. L. Rodriguez, A. Spirling, and B. M. Stewart (Forthcoming). "Embedding Regression: Models for Context-Specific Description and Inference". *American Political Science Review*.
- M. Osnabrügge, S. B. Hobolt, and T. Rodon (2021). "Playing to the Gallery: Emotive Rhetoric in Parliaments". *American Political Science Review* 115 (3): 885–899.
- L. Hargrave (2022). "A Double Standard? Gender Bias in Voters' Perceptions of Political Arguments". *British Journal of Political Science* published ahead of print (doi: 10.1017/S0007123422000515).

References

APSA Committee on Publications (2018). *Style Manual for Political Science (Revised 2018 Version)*. URL: https://connect.apsanet.org/stylemanual/.

Baden, C., C. Pipal, M. Schoonvelde, and M. Van der Velden (2022). "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda". *Communication Methods and Measures* 16 (1): 1–18.

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

Bauer, P. C. and C. Landesvatter, eds. (2022). *APIs for Social Scientists: A Collaborative Review*.

Benoit, K. (2020). "Text as Data: An Overview". *Handbook of Research Methods in Political Science and International Relations*. Ed. by L. Curini and R. Franzese. Thousand Oaks: Sage: 461–497.

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2016). "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data". *American Political Science Review* 110 (2): 278–295.

Benoit, K. and S. Müller (Work in Progress). *Text Analysis Using R*. URL: https://quanteda.github.io/Text-Analysis-Using-R/.

Benoit, K., K. Munger, and A. Spirling (2019). "Measuring and Explaining Political Sophistication Through Textual Complexity". *American Journal of Political Science* 63 (2): 491–508.

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo (2018). "quanteda: An R Package for the Quantitative Analysis of Textual Data". *The Journal of Open Source Software* 3 (30): 774.

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

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

Blumenau, J. (2021). "The Effects of Female Leadership on Women's Voice in Political Debate". *British Journal of Political Science* 51 (2): 750–771.

Boussalis, C., T. G. Coan, M. R. Holman, and S. Müller (2021). "Gender, Candidate Emotional Expression, and Voter Reactions During Televised Debates". *American Political Science Review* 115 (4): 1242–1257.

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

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

De Vries, E., M. Schoonvelde, and G. 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, M. W. and A. 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.

Düpont, N. and M. Rachuj (2021). "The Ties That Bind: Text Similarities and Conditional Diffusion among Parties". *British Journal of Political Science* published ahead of print (doi: 10.1017/S0007123420000617).

Gentzkow, M., B. T. Kelly, and M. Taddy (2019). "Text as Data". *Journal of Economic Literature* 57 (3): 535–574.

Gessler, T. (2022). "Topic Models". *Elgar Encyclopedia of Technology and Politics*. Ed. by A. Ceron. Cheltenham: 108–111.

Gessler, T. and S. Hunger (2021). "How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration". *Political Science Research and Methods* published ahead of print (doi: 10.1017/psrm.2021.64).

Gilardi, F., C. R. Shipan, and B. Wüest (2021). "Policy Diffusion: The Issue-Definition Stage". *American Journal of Political Science* 65 (1): 21–35.

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

Grimmer, J., M. E. Roberts, and B. M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.

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

Hargrave, L. (2022). "A Double Standard? Gender Bias in Voters' Perceptions of Political Arguments". *British Journal of Political Science* published ahead of print (doi: 10.1017/S0007123422000515).

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

Hengel, E. (2022). "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review". *The Economic Journal* 132 (648): 2951–2991.

Herzog, A. and S. Mikhaylov (2020). "Intra-Cabinet Politics and Fiscal Governance in Times of Austerity". *Political Science Research and Methods* 8 (3): 409–424.

Hvitfeldt, E. and J. Silge (2021). *Supervised Machine Learning For Text Analysis in R*. Boca Raton: CRC Press.

Ismay, C. and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press.

Jurafsky, D. and J. H. Martin (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition.

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

Manning, C. D., P. Raghavan, and H. Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press.

Merz, N., S. Regel, and J. Lewandowski (2016). "The Manifesto Corpus: A New Resource for Research on Political Parties and Quantitative Text Analysis". *Research & Politics* 3 (2): 1–8.

Mikhaylov, S., M. Laver, and K. Benoit (2012). "Coder Reliability and Misclassification in the Human Coding of Party Manifestos". *Political Analysis* 20 (1): 78–91.

Muddiman, A., S. C. McGregor, and N. J. Stroud (2019). "(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries". *Political Communication* 36 (2): 214–226.

Müller, S. (2020). "Media Coverage of Campaign Promises Throughout the Electoral Cycle". *Political Communication* 37 (5): 696–718.

Müller, S. (2022). "The Temporal Focus of Campaign Communication". *The Journal of Politics* 84 (1): 585–590.

O'Grady, T. (2019). "Careerists Versus Coal-Miners: Welfare Reforms and the Substantive Representation of Social Groups in the British Labour Party". *Comparative Political Studies* 52 (4): 544–578.

Osnabrügge, M., S. B. Hobolt, and T. Rodon (2021). "Playing to the Gallery: Emotive Rhetoric in Parliaments". *American Political Science Review* 115 (3): 885–899.

Parthasarathy, R., V. Rao, and N. Palaniswamy (2019). "Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies". *American Political Science Review* 113 (3): 623–640.

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

Proksch, S.-O., W. Lowe, J. Wäckerle, and S. N. Soroka (2019). "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches". *Legislative Studies Quarterly* 44 (1): 97–131.

Proksch, S.-O., C. Wratil, and J. Wäckerle (2019). "Testing the Validity of Automatic Speech Recognition for Political Text Analysis". *Political Analysis* 27 (3): 339–359.

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

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). "Structural Topic Models for Open-Ended Survey Responses". *American Journal of Political Science* 58 (4): 1064–1082.

Rodriguez, P. L. and A. Spirling (2022). "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research". *The Journal of Politics* 84 (1): 101–115.

Rodriguez, P. L., A. Spirling, and B. M. Stewart (Forthcoming). "Embedding Regression: Models for Context-Specific Description and Inference". *American Political Science Review*.

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

Schoonvelde, M., G. Schumacher, and B. 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.

Schwemmer, C., C. Knight, E. D. Bello-Pardo, S. Oklobdzija, M. Schoonvelde, and J. W. Lockhart (2020). "Diagnosing Gender Bias in Image Recognition Systems". *Socius: Sociological Research for a Dynamic World* 6: 1–17.

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

Steinert-Threlkeld, Z. C. (2018). *Twitter as Data*. Cambridge: Cambridge University Press.

Watanabe, K. (2021). "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages". *Communication Methods and Measures* 14 (2): 81–102.

Watanabe, K. and S. Müller (2022). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io.

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

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

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

Zollinger, D. (2022). "Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses". *American Journal of Political Science* published ahead of print (doi: 10.1111/ajps.12743).