Stefan Müller, PhD
Assistant Professor
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 2024

Quantitative Text Analysis (POL42050)

Version: April 5, 2024

Latest version at: https://muellerstefan.net/teaching/2024-spring-qta.pdf

---

Time: Monday, 10:00–11:50
Location: QUI-113 (Quinn School of Business)
Credits: 10.0
Format: Lecture and computer labs

Module coordinator: Stefan Müller, PhD
stefan.mueller@ucd.ie | https://muellerstefan.net
Office: Newman Building, G312
Office hours: Monday, 12:30–14:00 (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. The module will also introduce advanced methods, including word embeddings, speech transcription, machine translation, and computer vision. Furthermore, we will explore the Hugging Face Python library, a powerful resource for implementing transformer models and other state-of-the-art natural language processing techniques.

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. Through hands-on experience with word embeddings and transformer models, students will gain insights into the latest advancements in text analysis and their applications in political science research.

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.

Prerequisites

Prior familiarity with the statistical programming language R (or Python) is a prerequisite for this course due to its direct relevance to the content and assignments. Below are some reasons why prior experience with R or Python is crucial for students to follow the course and apply the methods effectively:

- Implementation of Text Analysis Methods: Text analysis is a central component of the course, and R is widely used for implementing text analysis techniques. R provides a comprehensive set of libraries and packages specifically designed for text processing, natural language processing (NLP), and sentiment analysis. Students with prior experience in R will be able to navigate and utilize these tools more efficiently, enabling them to implement text analysis methods covered in the course effectively.
- Course Content Alignment: The course content, lectures, and materials are designed with a focus on R-based implementation. The examples, code snippets, and demonstrations provided throughout the course will be predominantly in R. Some of the advanced methods are implemented in Python, but a good understanding of R will make it much easier to write and run code in Python. Without prior familiarity, students may struggle to comprehend and replicate these examples, hindering their understanding of the core concepts and methodologies.
- Homework Assignments and Research Papers: The assignments and research papers in this course will require students to apply the text analysis methods discussed in class to real-world data. Students without prior experience with R may find it challenging to write code to preprocess large text corpora, visualise results, and interpret the findings. Their lack of proficiency in R could impede their ability to complete assignments accurately and efficiently.

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 (2023). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io. https://tutorials.quanteda.io
- 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.
- W. Van Atteveldt, D. Trilling, and C. A. Calderón (2022). *Computational Analysis of Communication: A Practical Introduction to the Analysis of Texts, Networks, and Images with Code Examples in Python and R*. Hoboken: Wiley-Blackwell.

Technical Background

The following books and websites are helpful to refresh and extend the knowledge of R, Quarto, 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 Quarto

- 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, M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly.

Python for Social Scientists

- A. Turrell (2024). *Coding for Economists*. URL: https://aeturrell.github.io/coding-for-economists.

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)
- Applying and fine-tuning transformer-based machine learning models: `HuggingFace` & `transformers` (Python) and `flaiR` (R)

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 Mondays from 12:30–14:00, 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.

We will use Slack for in this module.[2] 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 #coding, #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] https://libguides.ucd.ie/academicintegrity.

[2] I have had very positive experiences with Slack in my modules. Müller (2023) discusses both the advantages and shortcomings of Slack for teaching and learning.

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.

Use of Generative Artificial Intelligence (AI) Tools

I encourage the use of generative AI tools when completing the assignments for this module but all work relying on AI-generated content must adhere to the highest academic standards. Users of this technology must be aware of what it can and more importantly what it cannot do well. It is crucial for you to exercise judgement when evaluating the quality and reliability of content generated through AI platforms. AI is not a panacea for all writing challenges; it will not automatically generate a flawless, logically coherent, and factually correct assignment. Instead, use AI as a tool to tackle specific issues such as brainstorming and idea formation, literature discovery, and text drafting issues. View your preferred AI platform(s) as useful but imperfect tools that can offer inspiration, new perspectives, and supplementary areas for research for your own work. In-depth research on your part remains essential to ensure coherent, factual, and scientifically informed perspectives in your assignment. Always cross-reference the information AI offers against other independent and reliable sources.

AI use must be in line with UCD's policies on academic integrity and adhere to the highest academic standards. See here for details: https://libguides.ucd.ie/academicintegrity.

Documenting AI Use (Mandatory)

Since generative AI is such a novel tool in an academic context, we do not yet fully understand what it is capable of, and these capabilities are evolving quickly. What was impossible today might well become trivial tomorrow (keeping in mind the academic standards mentioned above). In order to address this, you are expected to provide an account of the tools used and the way in which they were used in a mandatory appendix to your assignment. This appendix will be assessed as part of the assignment, with grade points awarded for effective communication of the methods used to generate content. For each instance where a generative AI tool is used, you need to provide:

1. An in-text citation or footnote. For example:
   - "Some AI-generated text (OpenAI 2024)"
   - "Some AI-generated text[3]"
2. A bibliographic reference to the tool used and the date of access.
3. An entry in the mandatory AI appendix detailing how the tool was used. For example:

Generative AI Tools and Apps

Below I have listed some AI tools that might help you drafting your policy brief:

[3]Text based on content generated by OpenAI's ChatGPT on 1 January 2024. See Appendix 1 for details.

Table 1: Example table demonstrating how generative AI was used to complete the research paper

AI Tool	Explanation	Prompt used
ChatGPT	Topic brainstorming	"Provide an overview of potential political science research questions that could be answered using quantitative text analysis methods."
Elicit	Literature search	"Compile a list of academic publications detailing advantages and shortcomings of topic models."

- Brainstorming
  - ChatGPT: https://chat.openai.com
  - Bing in creative mode (GPT4 for free): www.bing.com
- Literature Discovery
  - Elicit: https://elicit.org
  - Semantic Scholar: https://www.semanticscholar.org
  - Perplexity: https://www.perplexity.ai
  - SciSpace: https://typeset.io
- Structure and Drafting
  - Grammarly: https://www.grammarly.com
  - Quillbot: https://quillbot.com/
  - Jasper: https://www.jasper.ai/
  - Jenni: https://jenni.ai

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.

Dignity and Respect

UCD is committed to the promotion of an environment for work and study which upholds the dignity and respect of all members of the UCD community and which supports your right to study and/or work in an environment which is free of any form of bullying, harassment or sexual misconduct (including sexual harassment and sexual violence).

There are a number of supports in place if you are experiencing bullying, harassment or sexual misconduct and you are strongly encouraged to come forward to seek confidential support and guidance on the range of informal options and formal options for resolving issues as appropriate. Reports of bullying, harassment or sexual misconduct can also be made anonymously through UCD's Report and Support tool.

UCD is actively promoting a culture where bullying, harassment and sexual misconduct is not tolerated, where everyone is respected and feels valued, included and that they belong in UCD.

You can find more details on UCD's Dignity and Respect Website at: https://www.ucd.ie/equality/support/dignityrespect/.

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.

Table 2: Grade conversion scheme for homeworks

Homeworks	UCD Grade
97–100%	A+
94–96%	A
91–93%	A–
88-90%	B+
85–87%	B
83–84%	B
80–82%	C+
77–79%	C
74–76%	C–
71–73%	D+
68–70%	D
65–67%	E+
54–64%	E
44–53%	E–
0–32%	F

- 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 3 May 2024. 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 discussed in this module. 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, 3 May 2024	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 (22 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 Quarto (29 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 (2023). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io: chapter 1.

Week 3: Assumptions and Workflow (5 February [Public Holiday] – Session will be rescheduled

- 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.
- **ash23**.

Optional

- C. D. Manning, P. Raghavan, and H. Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press: chapter 2.

Week 4: Tokenisation and Document-Feature Matrix (12 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 (2023). *Quanteda Tutorials*. URL: https://tutorials.quanteda.io: chapter 2–3.

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 (19 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 (26 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.
- 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.
- J. Blumenau (2021). "The Effects of Female Leadership on Women's Voice in Political Debate". *British Journal of Political Science* 51 (2): 750–771.

Optional

- 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 (4 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?

– Difference between bag-of-words classifier and state-of-the art approaches (transformers)

– *Application*: supervised machine learning in `quanteda` and the `transformers` Python library

Mandatory Readings

- S. Müller (2022). "The Temporal Focus of Campaign Communication". *The Journal of Politics* 84 (1): 585–590.
- C. Hanretty (2023). *Replicating Mueller, 'The Temporal Focus of Campaign Communication'*. URL: http://chrishanretty.co.uk/posts/finetuning/.
- HuggingFace (2023). *Transformers: State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX*. V4.35.0/V4.35.1/V4.35.2. URL: https://huggingface.co/docs/transformers/: skim tutorials for a basic understanding of the transformers library.
- F. Gilardi, M. Alizadeh, and M. Kubli (2023). "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". *Proceedings of the National Academy of Sciences of the United States of America* 120 (3): e2305016120.

Optional

- C. D. Manning, P. Raghavan, and H. Schütze (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press: Chapter 13 (Naïve Bayes).
- S. Wankmüller (2022). "Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis". *Sociological Methods & Research* published ahead of print (doi: 10.1177/00491241221134527).
- 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.
- L. Birkenmaier, C. M. Lechner, and C. Wagner (2023). "The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation". *Communication Methods and Measures* published ahead of print (doi: 10.1080/19312458.2023.2285765).
- 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.
- L. Tunstall, L. von Werra, and T. Wolf (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Beijing: O'Reilly.

Week 8: Supervised, Unsupervised and Semi-Supervised Scaling (25 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, Wordfish, and Latent Semantic Scaling

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

- G. Le Mens and A. Gallego (2023). *Scaling Political Texts With ChatGPT*. arXiv:2311.16639. URL: https://doi.org/10.48550/arXiv.2311.16639.
- D. Zollinger (2024). "Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses". *American Journal of Political Science* 68 (1): 139–159.
- S. Müller and N. Fujimura (2024). "Campaign Communication and Legislative Leadership". *Political Science Research and Methods* published ahead of print (doi: 10.1017/psrm.2024.11).
- T. Gessler and S. Hunger (2022). "How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration". *Political Science Research and Methods* 10 (3): 524 –544.
- 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.
- S. Müller, S. Brazys, and A. Dukalskis (2022). "Discourse Wars and 'Mask Diplomacy': China's Global Image Management in Times of Crisis". AidData Working Paper 113.

Week 9: Retrieving, Loading and Wrangling Text Corpora (1 April [Public Holiday] – Session will be rescheduled)

– 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, M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly: ch. 5–7.
- P. C. Bauer and C. Landesvatter, eds. (2023). *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.
- M. J. Salganik (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press: ch. 6.

Week 10: Topic Models (8 April)

– How does unsupervised document classification work? What are the assumptions, advantages, and caveats of topic models?

– *Application*: Keyword-assisted topic models (keyATM) and structural topic models (STM)

Mandatory Readings

- T. Gessler (2022). "Topic Models". *Elgar Encyclopedia of Technology and Politics*. Ed. by A. Ceron. Cheltenham: Edward Elgar Publishing: 108–111.
- S. Eshima, K. Imai, and T. Sasaki (2023). "Keyword-Assisted Topic Models". *American Journal of Political Science* published ahead of print (doi: 10.1111/ajps.12779).

Optional

- 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.
- D. M. Blei (2012). "Probabilistic Topic Models". *Communications of the ACM* 55 (4): 77–84.
- 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.
- S. Müller, G. Kennedy, and T. Maher (2023). "Reactions to Experts in Deliberative Democracy: The 2016–2018 Irish Citizens' Assembly". *Irish Political Studies* 38 (4): 467–488.

Week 11: Moving Beyond Bag-of-Words: Word Embeddings (15 April)

Instructor: Yen-Chieh Liao, PhD

– Vector representation (one-hot vector), word and document embeddings

– Implementing word similarity with Word2Vec with different various pre-trained embedding models

– *Application*: apply word embeddings to political science research questions

Mandatory Readings

- S. Raaijmakers (2022). *Deep Learning for Natural Language Processing*. Shelter Island: Manning Publications: ch. 3.
- L. Rheault and C. Cochrance (2020). "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora". *Political Analysis* 28 (1): 112–133.

Optional

- S. Raaijmakers (2022). *Deep Learning for Natural Language Processing*. Shelter Island: Manning Publications: ch. 1.3; ch. 1.4.
- P. L. Rodriguez, A. Spirling, and B. M. Stewart (2023). "Embedding Regression: Models for Context-Specific Description and Inference". *American Political Science Review* 117 (4): 1255–1274.

Week 12: New Developments: Transformers (22 April)

Instructor: Yen-Chieh Liao, PhD

– Basic knowledge of Understanding the inner workings of Transformers

– Deriving word embeddings with BERT

– Comparing BERT and Word2Vec

– *Application*: Exploring the `transformers` Python library and `flaiR` R package

Mandatory Readings

- S. Raaijmakers (2022). *Deep Learning for Natural Language Processing*. Shelter Island: Manning Publications: ch. 9.
- M. Laurer, W. van Atteveldt, A. Casas, and K. Welber (2024). "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI". *Political Analysis* 32 (1): 84–100.

Optional

- S. Müller and S.-O. Proksch (2023). "Nostalgia in European Party Politics: A Text-Based Measurement Approach". *British Journal of Political Science* published ahead of print.
- L. Tunstall, L. von Werra, and T. Wolf (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Beijing: O'Reilly.
- C.-h. Chan (2023). "grafzahl: Fine-tuning Transformers for Text Data from Within R". *Computational Communication Research* 5 (1): 76–84.

References

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

Bauer, P. C. and C. Landesvatter, eds. (2023). *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., 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.

Birkenmaier, L., C. M. Lechner, and C. Wagner (2023). "The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation". *Communication Methods and Measures* published ahead of print (doi: 10.1080/19312458.2023.2285765).

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.

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

Chan, C.-h. (2023). "grafzahl: Fine-tuning Transformers for Text Data from Within R". *Computational Communication Research* 5 (1): 76–84.

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.

Eshima, S., K. Imai, and T. Sasaki (2023). "Keyword-Assisted Topic Models". *American Journal of Political Science* published ahead of print (doi: 10.1111/ajps.12779).

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: Edward Elgar Publishing: 108–111.

Gessler, T. and S. Hunger (2022). "How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration". *Political Science Research and Methods* 10 (3): 524 –544.

Gilardi, F., M. Alizadeh, and M. Kubli (2023). "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". *Proceedings of the National Academy of Sciences of the United States of America* 120 (3): e2305016120.

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.

Hanretty, C. (2023). *Replicating Mueller, 'The Temporal Focus of Campaign Communication'*. URL: http://chrishanretty.co.uk/posts/finetuning/.

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.

HuggingFace (2023). *Transformers: State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX*. V4.35.0/V4.35.1/V4.35.2. URL: https://huggingface.co/docs/transformers/.

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.

Laurer, M., W. van Atteveldt, A. Casas, and K. Welber (2024). "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI". *Political Analysis* 32 (1): 84–100.

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.

Le Mens, G. and A. Gallego (2023). *Scaling Political Texts With ChatGPT*. arXiv:2311.16639. URL: https://doi.org/10.48550/arXiv.2311.16639.

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.

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.

Müller, S. (2023). "How Slack Facilitates Communication and Collaboration in Seminars and Project-Based Courses". *Journal of Educational Technology Systems* 51 (3): 303–316.

Müller, S., S. Brazys, and A. Dukalskis (2022). "Discourse Wars and 'Mask Diplomacy': China's Global Image Management in Times of Crisis". AidData Working Paper 113.

Müller, S. and N. Fujimura (2024). "Campaign Communication and Legislative Leadership". *Political Science Research and Methods* published ahead of print (doi: 10.1017/psrm.2024.11).

Müller, S., G. Kennedy, and T. Maher (2023). "Reactions to Experts in Deliberative Democracy: The 2016–2018 Irish Citizens' Assembly". *Irish Political Studies* 38 (4): 467–488.

Müller, S. and S.-O. Proksch (2023). "Nostalgia in European Party Politics: A Text-Based Measurement Approach". *British Journal of Political Science* published ahead of print.

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.

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.

Raaijmakers, S. (2022). *Deep Learning for Natural Language Processing*. Shelter Island: Manning Publications.

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

Rheault, L. and C. Cochrance (2020). "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora". *Political Analysis* 28 (1): 112–133.

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., A. Spirling, and B. M. Stewart (2023). "Embedding Regression: Models for Context-Specific Description and Inference". *American Political Science Review* 117 (4): 1255–1274.

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

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.

Tunstall, L., L. von Werra, and T. Wolf (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Beijing: O'Reilly.

Turrell, A. (2024). *Coding for Economists*. URL: https://aeturrell.github.io/coding-for-economists.

Van Atteveldt, W., D. Trilling, and C. A. Calderón (2022). *Computational Analysis of Communication: A Practical Introduction to the Analysis of Texts, Networks, and Images with Code Examples in Python and R*. Hoboken: Wiley-Blackwell.

Wankmüller, S. (2022). "Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis". *Sociological Methods & Research* published ahead of print (doi: 10.1177/00491241221134527).

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 (2023). *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., M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. 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. (2024). "Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses". *American Journal of Political Science* 68 (1): 139–159.