

402150 VU Applied Methods – Enhanced Level

Introduction to Computational Text Analysis

Summer term 2026

version 2026-03

Instructor:	Ass.-Prof. Hauke Licht, Ph.D.	Dates:	Tuesdays March - June 2026
Affiliation:	University of Innsbruck	Times:	9:45 - 13:00
Email:	hauke.licht@uibk.ac.at	Place:	HS 1/SR 1 (<i>Theologie</i> building) Karl-Rahner-Platz 1, 6020 Innsbruck
Materials:	Code and data (Github), readings and slides (OLAT)		

Course short description Mastering computational text analysis (CTA) allows social scientists to gain quantitative insights into the communication behavior of a broad range of societal actors, ranging from political parties to non-governmental organizations. This course develops students' theoretical knowledge of and experience with applying CTA methods in the R programming language. Students will learn different methods for assigning scores or categories to text documents that require different amounts of prior data knowledge and gain practical experience with implementing them in R.

About the instructor Hauke Licht is an Assistant Professor of Computational Political Science at the Department of Political Science and the Digital Science Center, University of Innsbruck. He has received his PhD from the University of Zurich and previously worked as post-doctoral researcher at the University of Cologne. In his research, Hauke develops and applies deep learning-based computational text analysis methods to study political communication, electoral competition, and democratic representation (e.g. Licht et al., [2025a](#)) with a focus on multilingual analyses (Licht, [2023](#); Licht and Lind, [2023](#)) and text annotation (Licht et al., [2025b](#); Licht and Szczepanski, [2025](#)).

1 Course description	2
2 Sessions	7
3 Software and Packages	15
4 Policies	17

Course description

Vast and increasing amounts of text data are available in digitized formats online and in public archives. This presents exciting opportunities for social scientists to gain quantitative insights into the communication behavior of a broad range of societal actors, ranging from political parties to non-governmental organizations. Mastering **computational text analysis** (CTA) thus has become an essential skill for social scientists.

This course develops students' theoretical knowledge of and experience with applying CTA methods to analyze text data and quantify communication behavior. Building on a solid understanding of the goals and limitations of quantitative content analysis and its relation to CTA, students will learn different methods for analyzing texts computationally and how to implement them in the **R** programming language. In particular, **the course covers** how to

1. convert unstructured text data into quantitative representations ready for statistical analyses
2. produce quantitative measures for texts with supervise, semi-supervised, and unsupervised bag-of-words methods
3. apply deep learning methods like word embeddings and large language models (LLM)
4. critically evaluate application of CTA in political science and neighboring disciplines
5. align research questions, concepts, and measurement strategies in applied research

Courses goals

The **courses goals** are to

1. acquire a solid understanding of CTA methods
2. gain practical experience with applying CTA methods to real-world text data in R
3. develop the ability to critically assess the measurement quality and methodological choices in applied research

Learning objectives

Upon successful completion of the course, students will be able to:

1. Understand the fundamental principles, goals, and potential limitations of CTA methods.
2. Know how to describe, categorize, or score documents with bag-of-words methods.
3. Know the basics of texts scoring and classification with deep learning-based approaches.
4. Know how to critically evaluate the quality of measurements generated with CTA methods.
5. Implement these analysis in R.

Prerequisites

Study Program and Modules

Students in the Bachelor's Programme *Political Science* must have successfully completed *each* of the [compulsory modules](#) 1 through 3:

1. positive assessment in module 1 *Introduction to Political Science*, and
2. positive assessment in module 2 *Applied Methods of Social Science*, and
3. positive assessment in module 3 *Statistics*

Students in other programs may take the course as part of the Minor [Methods of Empirical Social Research](#) and should consult LFU online to determine whether they meet the specific prerequisites for registration in the VU.

Other prerequisites

- Mandatory participation in the first session on 10. March 2026, 9:45 – 13:00 CET. Failure to meet this requirement will lead to de-registration from the course and a VIS status “cancellation (unexcused)” (*Abbruch (unentschuldigt)*).
- Strong command of English. All course readings are in English, all discussions and presentations will be in English, and participants are expected to write their final paper in English.
- Participation in at least 12 of the 7x2 sessions (given seven double sessions), regardless of reason (this includes having a doctor's note). If you miss three or more sessions, you will receive a failing grade (NGD5).

Course Requirements

The coursework comprises

1. active participation during our meetings,
2. reading papers assigned as required readings,
3. two take-home assignments,
4. a final term paper

The take-home assignment and final term paper will be graded. The **final grade** will be computed by applying the following formula:

$$\begin{aligned}\text{final grade} &= 10\% \text{ active participation} \\ &+ 25\% \text{ take-home assignment 1} \\ &+ 25\% \text{ take-home assignment 2} \\ &+ 40\% \text{ final term paper}\end{aligned}$$

Required readings There are **2-3 required readings assigned for each double session**. These readings are academic articles, book chapters, and/or R programming tutorials. Copies of the readings will be made available through **OLAT** unless they are freely accessible online. Participants are expected to read each reading listed as required *before* the respective session meeting. Participants should take notes and bring them to class.

Take-home midterm exam After double sessions 3 (14. April 2026) and 5 (5. May 2026), participants will work on **graded take-home assignments**. Participants will be required to implement and discuss their implementation of the methods learned in the given blocks' sessions. The first take-home assignment covering the second topical block will be released on 15. April 2026, 12:00 CEST, and participants must upload your solutions via OLAT by **23. April 2026, 23:59 CEST**. The second take-home assignment covering the third topical block will be released on 6. May 2026, 12:00 CEST, and participants must upload your solutions via OLAT by **18. May 2026, 23:59 CEST**. Participants' grades in the two take-home assignments \alert{each contribute 25% to their final grade}.

Final term paper After the course, you will write a **final term paper**. In the term paper, you will be expected to **apply a method** covered in the course to a dataset of your choice. The lecturer will provide guidance to identify and source suitable datasets if needed.

You will be expected to **report and present your application in a short paper of 2500–3000 words** and accompany your submission with your R code.

Your paper must discuss the following aspects:

1. dataset description
2. the phenomenon you aim to measure and analyze
3. your research question and your motivation for studying it
4. your chosen method and your implementation of it
5. your results
6. potential limitations

The paper will be due on is **22. July 2026, 23:59 CEST**, and you must upload your paper and code via OLAT.

Reasons for cancellation of your participation

- **You decide to drop the course:** This option is available until the second session on 17. March 2026. Simply inform me via email.
- **If you do not participate in the first session:** Should you not participate in the first meeting on 10. March 2026, you will be de-registered from the course (VIS/LFU online status “cancellation (unexcused)” (*Abbruch (unentschuldigt)*)).
- **If you fail to meet the course prerequisites:** Should you not fulfill the course prerequisites by the second meeting, you will be de-registered from the course (VIS/LFU online status “registration declined”/“Anmeldung abgewiesen”). ERASMUS students are an exception.
- **If you miss three or more sessions:** If you miss three or more sessions, you will receive a failing grade (NGD5)
- **If you violate the GenAI policy:** Repeated violations of the AI policy listed above will lead to failing the course (NGD5).

Meeting dates and times

We will meet for double sessions on seven Tuesdays in either HS1 and SR1 ([Theologie, Karl-Rahner-Platz 1](#)) from 09:45 - 13:00.

Table 1: Meeting dates and times and important deadlines

	Date	Time	Room
session 1	10. March 2026	09:45 - 13:00	HS I
session 2	17. March 2026	09:45 - 13:00	HS I
session 3	14. April 2026	09:45 - 13:00	SR I
take-home assignment 1, release due	15. April 2026	12:00 CET	
	23. April 2026	12:00 CET	
session 4	28. April 2026	09:45 - 13:00	SR I
session 5	5. May 2026	09:45 - 13:00	SR I
take-home assignment 2, release due	6. May 2026	12:00 CET	
	18. May 2026	23:59 CET	
session 6	2. June 2026	09:45 - 13:00	HS I
session 7	16. June 2026	09:45 - 13:00	SR I
final term paper, due	22. July 2026	23:59 CET	

Sessions

The course is organized into 4 thematic blocks:

1. **Introduction and basics** (session 1): In this first block of the course, we will discuss the basic concepts, principles, and assumptions of computational text analysis methods and why we use them political science and neighboring fields. We will make first steps in processing unstructured text data into quantitative representations with the `quanteda` R package.
2. **Distinguishing between groups and categories** (sessions 2 and 3): In this thematic block, we will address questions like what distinguishes the texts of different groups of authors or periods and how can we develop models that automatically classify texts into categories? In session 2 will learn metrics for measuring word usage differences between groups of documents. In session 3, we will focus on *supervised learning*. The block will be concluded by a *take-home assignment*.
3. **Topic modeling** (sessions 4 and 5): In this block, we will learn about methods for “discovering” topical patterns in large text corpora with algorithms that require no (or only very little) prior knowledge about relevant patterns in the documents under study. The block will be concluded by a *take-home assignment*.
4. **Deep learning approaches** (session 6): In this block, we move beyond bag-of-words analysis and dive into the world of deep learning-powered methods. We will learn about using generative Large Language Models (LLMs) as annotators.
5. **Concluding Block** (session 7): In the concluding session, we will discuss how to connect research ideas to text analysis methods. This will prepare students to develop their concepts for their final term papers.

Block 1: Introduction and Basics

In this first block of the course, we will discuss why we use computational text analysis methods in political science and neighboring fields and their basic concepts, principles, and assumptions. We will make first steps in processing unstructured text data into quantitative representations with the `quanteda` R package.

Session 1 (10. March 2026): Why computational text analysis

Topics:

- course structure, topics, and requirements
- basic concepts (corpus, document, tokens, bag of words)
- how to create a *bag of words* with `quanteda`

- tokenization with `quantada`'s `tokens` function and tokenization choices ('stop word' removal, punctuation removal, lower casing, etc.)
- the `corpus` and `dfm` objects in `quantada`
- the document-feature matrix

Preparation

- install R, RStudio, and the required packages (see the installation guide on [Github](#))
- complete the required readings

Readings

required

`grimmer_text_2013`: Justin Grimmer and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3, pp. 267–297. DOI: [10.1093/pan/mps028](https://doi.org/10.1093/pan/mps028)

`watanabe_tokens_2025`: Kohei Watanabe and Stefan Müller (2025d). "Tokens". In: *quantada tutorials*. URL: <https://tutorials.quantada.io/basic-operations/tokens/>

`watanabe_corpus_2025`: Kohei Watanabe and Stefan Müller (2025a). "Corpus". In: *quantada tutorials*. URL: <https://tutorials.quantada.io/basic-operations/corpus/>

further readings Boumans and Trilling (2016), Wilkerson and Casas (2017), Atteveldt and Peng (2018), Gentzkow, Kelly, and Taddy (2019), Schoonvelde, Schumacher, and Bakker (2019), Bonikowski and Nelson (2022), Grimmer, Roberts, and Stewart (2022), Ash and Hansen (2023), Nielbo et al. (2024), and Bachl and Scharkow (2024)

Block 2: Introduction and Basics

In the second block of the course, we will continue learning about processing unstructured text data into quantitative representations with the `quantada` R package. We will implement our first quantitative text analyses using dictionaries. We will learn metrics for measuring word usage differences between groups of documents. And we will learn how to develop algorithms that automatically classify texts into categories.

The block will be concluded by a *take-home assignment*.

Session 2 (17. March 2026): Keyword analysis and dictionary-based measurement

Topics:

- the document-feature matrix
- dictionary analysis in *quanteda*
- prevalence and popular transformations
- sentiment analysis
- constructing dictionaries

Readings

required

watanabe_document-feature_2025: Kohei Watanabe and Stefan Müller (2025b).

“Document-Feature Matrix”. In: *quanteda tutorials*. URL:

<https://tutorials.quanteda.io/basic-operations/dfm/>

watanabe_lookup_2025: Kohei Watanabe and Stefan Müller (2025c). “Look up dictionary”. In: *quanteda tutorials*. URL:

https://tutorials.quanteda.io/basic-operations/tokens/tokens_lookup/

muddiman_reclaiming_2019: Ashley Muddiman, Shannon C. McGregor, and

Natalie Jomini Stroud (2019). “(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries”. In: *Political Communication* 36.2, pp. 214–226.

DOI: [10.1080/10584609.2018.1517843](https://doi.org/10.1080/10584609.2018.1517843)

further readings Gessler and Hunger (2021), Breyer (2023), and Proksch et al. (2019)

Session 3 (14. April 2026): Modeling word use differences and supervised classification

Topics:

- discovering distinctive patterns in word use between groups of documents
- supervised learning basics
- Naive Bayes
- evaluation metrics

Preparation

- work through the notebook on classification evaluation metrics that will be made available via on [Github](#)

Readings

required

monroe_fightin_2008: Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn (2008). “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict”. In: *Political Analysis* 16.4, pp. 372–403. doi: [10.1093/pan/mpn018](https://doi.org/10.1093/pan/mpn018)

kuhn_discriminant_2013: Max Kuhn and Kjell Johnson (2013). “Discriminant Analysis and Other Linear Classification Models”. In: *Applied Predictive Modeling*. Ed. by Max Kuhn and Kjell Johnson. Springer, pp. 275–328. doi: [10.1007/978-1-4614-6849-3_12](https://doi.org/10.1007/978-1-4614-6849-3_12). URL: https://doi.org/10.1007/978-1-4614-6849-3_12

further readings Park and Montgomery (2025), Silge and Hvitfeldt (2022), Peterson and Spirling (2018), and Barberá et al. (2021)

Block 3: Topic modeling

In this block, we will learn about methods for “discovering” topical patterns in text data with algorithms that require no (or only very little) prior knowledge about relevant patterns in the documents under study. Specifically, we will focus on topic modeling. The block will be concluded by a *take-home assignment*.

Session 4 (28. April 2026): LDA topic modeling

Topics:

- the LDA topic model
- observed data and latent quantities
- LDA topic model estimation and interpretation in R

Preparation

- install the `topicmodels` R package

Readings

required

quinn_how_2010: Kevin M. Quinn et al. (2010). “How to Analyze Political Attention with Minimal Assumptions and Costs”. In: *American Journal of Political Science* 54.1, pp. 209–228. doi: [10.1111/j.1540-5907.2009.00427.x](https://doi.org/10.1111/j.1540-5907.2009.00427.x)

maier_applying_2018: Daniel Maier et al. (2018). “Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology”. In: *Communication Methods and Measures* 12.2, pp. 93–118. doi: [10.1080/19312458.2018.1430754](https://doi.org/10.1080/19312458.2018.1430754)

Denny and Spirling (2018)

further readings Blei (2012)

Session 5 (5. May 2026): Advanced topics in topic modeling

Topics:

- accounting for authorship and other metadata
- semi-supervising and ‘seeding’ topic models
- validating topic models

Preparation

- install the `stm` and `keyATM` R packages

Readings

required

roberts_structural_2014: Margaret E. Roberts et al. (2014). “Structural Topic Models for Open-Ended Survey Responses”. In: *American Journal of Political Science* 58.4, pp. 1064–1082. DOI: [10.1111/ajps.12103](https://doi.org/10.1111/ajps.12103)

eshima_keyword-assisted_2024: Shusei Eshima, Kosuke Imai, and Tomoya Sasaki (2024). “Keyword-Assisted Topic Models”. In: *American Journal of Political Science* 68.2, pp. 730–750. DOI: [10.1111/ajps.12779](https://doi.org/10.1111/ajps.12779)

further readings Greene and Cross (2017), Chen et al. (2023), and Bernhard-Harrer et al. (2025)

Block 4: Deep learning approaches

In this block, we move beyond bag-of-words analysis and dive into the world of deep learning-powered methods such as static word embeddings and generative large language models. We will learn about th

idea of representing texts in embedding spaces and how this enables using generative LLMs as annotators.

Session 9 (16.12.2025): Word embedding basics

Topics:

- from word frequency to ‘meaning’
- understanding the logic of word embedding
- the word2vec algorithm
- basic computations with word embeddings

Preparation

- complete the required readings
- work through short notebook on vector and matrix algebra basics and vector distance/similarity measures

Readings

required

voita_embeddings_2025: Lena Voita (n.d.[b]). *Word Embeddings*. NLP Course | For You. URL: https://lena-voita.github.io/nlp_course/word_embeddings.html

rodriguez_word_2021: Pedro L. Rodriguez and Arthur Spirling (2021). “Word Embeddings: What works, what doesn’t, and how to tell the difference for applied research”. In: *The Journal of Politics* 84.1, pp. 101–115. DOI: [10.1086/715162](https://doi.org/10.1086/715162)

further readings Gennaro and Ash (2022), Baroni, Dinu, and Kruszewski (2014), Caliskan, Bryson, and Narayanan (2017), Garg et al. (2018), Kozłowski, Taddy, and Evans (2019), Rodman (2020), and Rheault and Cochrane (2020)

Session 10 (13.01.2026): Large language model (LLM) basics

Topics:

- understanding generative language models
- learn about chat completion and prompting
- first steps in code automation with LLMs in `ellmer`

Preparation

- complete the required readings

Readings

required

voita_language_2025: Lena Voita (n.d.[a]). *Language Modeling*. NLP Course | For You. URL: https://lena-voita.github.io/nlp_course/language_modeling.html

weber_evaluation_2023: Maximilian Weber and Merle Reichardt (2023). *Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models*. DOI: [10.48550/arXiv.2401.00284](https://doi.org/10.48550/arXiv.2401.00284). arXiv: [2401.00284](https://arxiv.org/abs/2401.00284) [cs]

further readings Jurafsky and Martin (2025)

Session 11 (20.01.2026): Measurement with LLMs

Topics:

- using LLMs for text annotation tasks
- best practices in LLM-based annotation

Preparation

- complete the required readings
- short reading on the concept of a ‘policy pledge’ (used in practical in-class exercises)

Readings

required

cova_primer_2024: Joshua Cova and Luuk Schmitz (2024). *A primer for the use of classifier and generative large language models in social science research*. DOI: [10.31219/osf.io/r3qng](https://doi.org/10.31219/osf.io/r3qng)

ornstein_how_2025: Joseph T. Ornstein, Elise N. Blasingame, and Jake S. Truscott (2025). “How to train your stochastic parrot: large language models for political texts”. In: *Political Science Research and Methods* 13.2, pp. 264–281. DOI: [10.1017/psrm.2024.64](https://doi.org/10.1017/psrm.2024.64)

further readings Mens and Gallego (2024) and Licht et al. (2025b)

Block 5: Concluding block

In this concluding block, wrap up and look ahead to advanced topics and future directions in quantitative text analysis.

Session 12 (27.01.2026): Concluding session

Topics:

- wrap up and look ahead
- research thinking: linking research questions, concepts, measurement strategies, and data sources

Preparation

- prepare to discuss your idea for the final term paper with the lecturer

Software and Packages

We will use the **R** programming language in this course through the integrated development environment (IDE) **RStudio**.

Detailed guidelines for installing the relevant software are available on the accompanying Github page: [Github](#)

Important: Students must install and set up R and RStudio *before* our first meeting. If you have any question related to installing the software, please post your question to the course **Forum** in OLAT.

R packages

We will use the following packages in this module:

- Dependency management etc.:
 - [renv](#)
 - [reprex](#)
- Importing (text) data:
 - [readr](#)
 - [readtext](#)
- Text processing:
 - [stringr](#)
 - [stringi](#)
 - [quanteda](#)
 - [quanteda.textmodels](#)
 - [quanteda.textstats](#)
 - [quanteda.textplots](#)
 - [quanteda.corpora](#)
- Specific CTA methods:
 - [topicmodels](#)
 - [stm](#)
 - [keyATM](#)
- Using LLMs
 - [ollamar](#)
 - [ellmer](#)

- General data wrangling and visualisation:
 - `dplyr`
 - `tidyr`
 - `purrr`
 - `ggplot2`
 - `lubridate`

In case of questions or problems

You will likely encounter concepts, methods, and software you might not have heard of or used before. So it is natural that parts of this course might be challenging. The lecturer – and hopefully also your peers – will assist you to overcome such challenges.

We will use the OLAT forum in this course to document and discuss any questions and problems you may encounter in a course. Hence, you will be asked to post to the forum if you have a question or face a problem related to the course content. (Other, personal issues that affect your course work should be directly communicated to the lecturer via email).

Before you post to the forum

If you struggle to solve problems relating to R or RStudio, please try the following things below before posting to the forum:

- Summarize the problem in your own words and then google this summary. If the problem relates to R, add the word “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.
- If your R code returns an error, Google the text of the error message. For example, you google the error message `rstats Error: "Can't subset columns that don't exist."`

It is very likely that at least one other person faced the same problem before or received the same error message and that there is an accepted solution in a forum like StackOverflow or so.

If this does not solve your problem, consider discussing the problem with ChatGPT or another conversational AI or code assistant. If you choose to do so, verify the correctness of the provided solution based on a minimum working example (MWE). If you cannot come up with an MWE, compare and discuss the provided solution with the code of your peers. If this does not help you and your peers to determine the correctness of the solution, post the problem and your solution attempt(s) to the forum.

Posting to the forum

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

Policies

Late submission

If a student submits an assignment late (i.e., after the deadline communicated in this syllabus), the following penalties will be applied:

- Submissions received at any time within four days after the submission deadline will be graded, but the following penalties will apply:
 - $\leq$ two days late $\Rightarrow$ 20% off the total points scored
 - $\leq$ four (but $>$ two) days late $\Rightarrow$ 40% off the total points scored
- Submissions received more than four days after the submission deadline date will not be accepted and counted as a *fail* in the given assignment.

Regulations regarding extenuating circumstances apply.

Sharing solutions

You may discuss your code and with each other.

But *do not*:

- sharing solutions to assignments with peers. Reusing or recycling another students' course work in an assignment submission is considered plagiarism (see below).
- post your solutions to assignments in the OLAT forum. We will use the OLAT forum only to document and discuss any questions and problems you may encounter in this course, not to share full solutions.

Plagiarism

Plagiarism is not allowed in this course and will be penalized. *Plagiarism* means presenting someone else's words, ideas, or work as your own without proper acknowledgment or citation. "someone else" could be a peer, another researcher, another human, or a generative AI ("Artificial Intelligence") like a coding assistant.

Depending on severity, plagiarism in assignment submissions will be penalized with 25% off the score you achieved in this assignment or lead to a failing grade for the assignment.

Why? Plagiarism undermines trust credibility and the trust in academic work and institutions by creating doubts about the source(s) of originality and claims thereof. It is therefore generally considered an ethical violation in academic, professional, and creative contexts.

AI policy

Generative Artificial Intelligence (genAI) are software tools that create new content based on user input and pre-trained machine learning models. The use of genAI in this course is regulated by the AI policy of the University of Innsbruck's Computer Science Department: <https://iis.uibk.ac.at/courses/aipolicy>

Most importantly, the following general rules apply:

- You are *responsible* for your course work (e.g., a term paper), regardless of the information sources or tools you used to produce it.
- You must *declare* any use of GenAI for any specific part of your course work.
- You must *document* how you verified the GenAI output you use in your course work.
- *Violations* result in zero points for the course work in question. Repeated violations result in a failing grade ("NGD5") for the course.

References

- Ash, Elliott and Stephen Hansen (2023). “Text Algorithms in Economics”. In: *Annual Review of Economics* 15 (Volume 15, 2023), pp. 659–688. doi: [10.1146/annurev-economics-082222-074352](https://doi.org/10.1146/annurev-economics-082222-074352).
- Atteveldt, Wouter van and Tai-Quan Peng (2018). “When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science”. In: *Communication Methods and Measures* 12.2, pp. 81–92. doi: [10.1080/19312458.2018.1458084](https://doi.org/10.1080/19312458.2018.1458084).
- Bachl, Marko and Michael Scharkow (2024). *Computational text analysis*.
- Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler (2021). “Automated Text Classification of News Articles: A Practical Guide”. In: *Political Analysis* 29.1, pp. 19–42. doi: [10.1017/pan.2020.8](https://doi.org/10.1017/pan.2020.8).
- Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). “Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. doi: [10.3115/v1/P14-1023](https://doi.org/10.3115/v1/P14-1023).
- Bernhard-Harrer, Jana, Randa Ashour, Jakob-Moritz Eberl, Petro Tolochko, and Hajo Boomgaarden (2025). “Beyond standardization: a comprehensive review of topic modeling validation methods for computational social science research”. In: *Political Science Research and Methods*, pp. 1–19. doi: [10.1017/psrm.2025.10008](https://doi.org/10.1017/psrm.2025.10008).
- Blei, David M. (2012). “Probabilistic topic models”. In: *Communications of the ACM* 55.4, pp. 77–84. doi: [10.1145/2133806.2133826](https://doi.org/10.1145/2133806.2133826).
- Bonikowski, Bart and Laura K. Nelson (2022). “From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research”. In: *Sociological Methods & Research* 51.4, pp. 1469–1483. doi: [10.1177/00491241221123088](https://doi.org/10.1177/00491241221123088).
- Boumans, Jelle W. and Damian Trilling (2016). “Taking Stock of the Toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars”. In: *Digital Journalism* 4.1, pp. 8–23. doi: [10.1080/21670811.2015.1096598](https://doi.org/10.1080/21670811.2015.1096598).
- Breyer, Magdalena (2023). “Populist positions in party competition: Do parties strategically vary their degree of populism in reaction to vote and office loss?” In: *Party Politics* 29.4, pp. 672–684. doi: [10.1177/13540688221097082](https://doi.org/10.1177/13540688221097082).
- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017). “Semantics derived automatically from language corpora contain human-like biases”. In: *Science* 356.6334, pp. 183–186. doi: [10.1126/science.aal4230](https://doi.org/10.1126/science.aal4230).
- Chen, Yingying, Zhao Peng, Sei-Hill Kim, and Chang Won Choi (2023). “What We Can Do and Cannot Do with Topic Modeling: A Systematic Review”. In: *Communication Methods and Measures* 17.2, pp. 111–130. doi: [10.1080/19312458.2023.2167965](https://doi.org/10.1080/19312458.2023.2167965).
- Cova, Joshua and Luuk Schmitz (2024). *A primer for the use of classifier and generative large language models in social science research*.
- Denny, Matthew J. and Arthur Spirling (2018). “Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It”. In: *Political Analysis* 26.2, pp. 168–189. doi: [10.1017/pan.2017.44](https://doi.org/10.1017/pan.2017.44).
- Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki (2024). “Keyword-Assisted Topic Models”. In: *American Journal of Political Science* 68.2, pp. 730–750. doi: [10.1111/ajps.12779](https://doi.org/10.1111/ajps.12779).
- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644. doi: [10.1073/pnas.1720347115](https://doi.org/10.1073/pnas.1720347115).
- Gennaro, Gloria and Elliott Ash (2022). “Emotion and Reason in Political Language”. In: *The Economic Journal* 132.643, pp. 1037–1059. doi: [10.1093/ej/ueab104](https://doi.org/10.1093/ej/ueab104).

- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). “Text as Data”. In: *Journal of Economic Literature* 57.3, pp. 535–574. doi: [10.1257/jel.20181020](https://doi.org/10.1257/jel.20181020).
- Gessler, Theresa and Sophia Hunger (2021). “How the refugee crisis and radical right parties shape party competition on immigration”. In: *Political Science Research and Methods*, pp. 1–21. doi: [10.1017/psrm.2021.64](https://doi.org/10.1017/psrm.2021.64).
- Greene, Derek and James P. Cross (2017). “Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach”. In: *Political Analysis* 25.1, pp. 77–94. doi: [10.1017/pan.2016.7](https://doi.org/10.1017/pan.2016.7).
- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- Grimmer, Justin and Brandon M. Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. In: *Political Analysis* 21.3, pp. 267–297. doi: [10.1093/pan/mps028](https://doi.org/10.1093/pan/mps028).
- Jurafsky, Daniel and James H Martin (2025). “Large Language Models”. In: *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. 3rd edition. URL: <https://web.stanford.edu/~jurafsky/slp3/9.pdf>.
- Kozlowski, Austin C., Matt Taddy, and James A. Evans (2019). “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings”. In: *American Sociological Review* 84.5, pp. 905–949. doi: [10.1177/0003122419877135](https://doi.org/10.1177/0003122419877135).
- Kuhn, Max and Kjell Johnson (2013). “Discriminant Analysis and Other Linear Classification Models”. In: *Applied Predictive Modeling*. Ed. by Max Kuhn and Kjell Johnson. Springer, pp. 275–328. doi: [10.1007/978-1-4614-6849-3_12](https://doi.org/10.1007/978-1-4614-6849-3_12). URL: https://doi.org/10.1007/978-1-4614-6849-3_12.
- Licht, Hauke (2023). “Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings”. In: *Political Analysis* 31.3, pp. 366–379. doi: [10.1017/pan.2022.29](https://doi.org/10.1017/pan.2022.29).
- Licht, Hauke, Tarik Abou-Chadi, Pablo Barberá, and Whitney Hua (2025a). “Measuring and Understanding Parties’ Anti-elite Strategies”. In: *The Journal of Politics*. doi: [10.1086/730711](https://doi.org/10.1086/730711).
- Licht, Hauke and Fabienne Lind (2023). “Going cross-lingual: A guide to multilingual text analysis”. In: *Computational Communication Research* 5.2, p. 1. doi: [10.5117/CCR2023.2.2.LICH](https://doi.org/10.5117/CCR2023.2.2.LICH).
- Licht, Hauke, Rupak Sarkar, Patrick Y. Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle (2025b). *Measuring Scalar Constructs in Social Science with LLMs*. arXiv: [2509.03116\[cs\]](https://arxiv.org/abs/2509.03116).
- Licht, Hauke and Ronja Sczepanski (2025). “Detecting Group Mentions in Political Rhetoric A Supervised Learning Approach”. In: *British Journal of Political Science* 55, e119. doi: [10.1017/S0007123424000954](https://doi.org/10.1017/S0007123424000954).
- Maier, Daniel et al. (2018). “Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology”. In: *Communication Methods and Measures* 12.2, pp. 93–118. doi: [10.1080/19312458.2018.1430754](https://doi.org/10.1080/19312458.2018.1430754).
- Mens, Gaël Le and Aina Gallego (2024). *Scaling Political Texts with Large Language Models: Asking a Chatbot Might Be All You Need*. arXiv: [2311.16639\[cs\]](https://arxiv.org/abs/2311.16639).
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2008). “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict”. In: *Political Analysis* 16.4, pp. 372–403. doi: [10.1093/pan/mpn018](https://doi.org/10.1093/pan/mpn018).
- Muddiman, Ashley, Shannon C. McGregor, and Natalie Jomini Stroud (2019). “(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries”. In: *Political Communication* 36.2, pp. 214–226. doi: [10.1080/10584609.2018.1517843](https://doi.org/10.1080/10584609.2018.1517843).
- Nielbo, Kristoffer L., Folger Karsdorp, Melvin Wevers, Alie Lassche, Rebekah B. Baglini, Mike Kestemont, and Nina Tahmasebi (2024). “Quantitative text analysis”. In: *Nature Reviews Methods Primers* 4.1, pp. 1–16. doi: [10.1038/s43586-024-00302-w](https://doi.org/10.1038/s43586-024-00302-w).

- Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott (2025). “How to train your stochastic parrot: large language models for political texts”. In: *Political Science Research and Methods* 13.2, pp. 264–281. doi: [10.1017/psrm.2024.64](https://doi.org/10.1017/psrm.2024.64).
- Park, Ju Yeon and Jacob M. Montgomery (2025). “Toward a framework for creating trustworthy measures with supervised machine learning for text”. In: *Political Science Research and Methods*, pp. 1–17. doi: [10.1017/psrm.2025.10042](https://doi.org/10.1017/psrm.2025.10042).
- Peterson, Andrew and Arthur Spirling (2018). “Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems”. In: *Political Analysis* 26.1, pp. 120–128. doi: [10.1017/pan.2017.39](https://doi.org/10.1017/pan.2017.39).
- Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka (2019). “Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches”. In: *Legislative Studies Quarterly* 44.1, pp. 97–131. doi: [10.1111/lsq.12218](https://doi.org/10.1111/lsq.12218).
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev (2010). “How to Analyze Political Attention with Minimal Assumptions and Costs”. In: *American Journal of Political Science* 54.1, pp. 209–228. doi: [10.1111/j.1540-5907.2009.00427.x](https://doi.org/10.1111/j.1540-5907.2009.00427.x).
- Rheault, Ludovic and Christopher Cochrane (2020). “Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora”. In: *Political Analysis* 28.1, pp. 112–133. doi: [10.1017/pan.2019.26](https://doi.org/10.1017/pan.2019.26).
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand (2014). “Structural Topic Models for Open-Ended Survey Responses”. In: *American Journal of Political Science* 58.4, pp. 1064–1082. doi: [10.1111/ajps.12103](https://doi.org/10.1111/ajps.12103).
- Rodman, Emma (2020). “A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors”. In: *Political Analysis* 28.1, pp. 87–111. doi: [10.1017/pan.2019.23](https://doi.org/10.1017/pan.2019.23).
- Rodriguez, Pedro L. and Arthur Spirling (2021). “Word Embeddings: What works, what doesn’t, and how to tell the difference for applied research”. In: *The Journal of Politics* 84.1, pp. 101–115. doi: [10.1086/715162](https://doi.org/10.1086/715162).
- Schoonvelde, Martijn, Gijs Schumacher, and Bert N. Bakker (2019). “Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology”. In: *Journal of Social and Political Psychology* 7.1, pp. 124–143. doi: [10.5964/jpspp.v7i1.964](https://doi.org/10.5964/jpspp.v7i1.964).
- Silge, Julia and Emil Hvitfeldt (2022). “Classification”. In: *Supervised Machine Learning for Text Analysis in R*. URL: <https://smltar.com/mlclassification>.
- Voita, Lena (n.d.[a]). *Language Modeling*. NLP Course | For You. URL: https://lena-voita.github.io/nlp_course/language_modeling.html.
- (n.d.[b]). *Word Embeddings*. NLP Course | For You. URL: https://lena-voita.github.io/nlp_course/word_embeddings.html.
- Watanabe, Kohei and Stefan Müller (2025a). “Corpus”. In: *quanteda tutorials*. URL: <https://tutorials.quanteda.io/basic-operations/corpus/>.
- (2025b). “Document-Feature Matrix”. In: *quanteda tutorials*. URL: <https://tutorials.quanteda.io/basic-operations/dfm/>.
- (2025c). “Look up dictionary”. In: *quanteda tutorials*. URL: https://tutorials.quanteda.io/basic-operations/tokens/tokens_lookup/.
- (2025d). “Tokens”. In: *quanteda tutorials*. URL: <https://tutorials.quanteda.io/basic-operations/tokens/>.
- Weber, Maximilian and Merle Reichardt (2023). *Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models*. arXiv: 2401.00284[cs].

Wilkerson, John and Andreu Casas (2017). “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges”. In: *Annual Review of Political Science* 20.1, pp. 529–544. doi: [10.1146/annurev-polisci-052615-025542](https://doi.org/10.1146/annurev-polisci-052615-025542).