Mirko Wegemann (he/him)
Comparative Politics
Scharnhorststraße 100
48151 Münster
m.wegemann@uni-muenster.de
Office hours: Tuesday, 15:00-16:00 (upon agreement)

Universität Münster

Summer term 2026

Text as Data

Logistics

The seminar takes place weekly on Wednesdays from 4:00 to 6:00 PM (c.t.). The first session is on April 15, 2026, and the last unit will take place on July 22, 2026. It is open to undergraduate students and will be held in English.

Course Outline

This method seminar provides an insight into quantitative text analysis, a type of content analysis that examines texts based on numerical similarities. Throughout the seminar, students will learn to (1) collect text data from publicly accessible websites, (2) prepare the raw material for various analyses, and (3) apply different techniques of quantitative text analysis. Since the advent of large language models (LLMs), the discipline has moved fast forward. The goal of the seminar is to give students both an understanding of so-called *bags-of-words* approaches, which are still suitable to understand what happens behind the black-box of deep learning networks, and an introduction into cutting-edge classification with *large language models*. The individual sessions will be very practice-oriented and give students the opportunity to realize their own project within the framework of the seminar. In doing so, they will develop their own research question, formulate theoretical expectations, access research data, and apply a suitable method of quantitative text analysis. By the end of the course, they will write a term paper in the form of a journalistic article or a policy brief.
A prerequisite for participation is prior experience with the statistical software R. Students are required to bring a laptop to all sessions.

Learning Outcomes

By the end of the seminar, students will have learned various methods to automatically download web content (web scraping). Additionally, students will be able to apply a range of procedures to quantitatively analyze the content of text documents. In this context, they will learn about the differences and respective advantages and disadvantages of methods that function with (supervised) and without (unsupervised) researcher input. They will learn how to transform raw data into a usable format and identify different topics within a text document. Furthermore, they will learn to apply this methodological knowledge to answer a substantial question in political science. Students will have acquired basic understanding of how neural networks and Large Language Models work. Through continuous practical application, students will deepen their knowledge of the statistical software `R` (and R Studio) and get a (very basic) introduction into Python. Although the seminar has a clear focus on learning a method, students will also gain a first insight into empirical application literature.

Requirements

The number of credits awarded depends on the degree program in which the students are enrolled. Approximately 30 working hours are planned per ECTS. The final grade consists of the following components (time effort measured at 4 ECTS).

- regular attendance and active participation in discussions — 18 hours
- **Seminarleistung**: application and presentation of a previously learned method
- **Prüfungsleistung**: Policy brief/newspaper report

} 102 hours

Method Assignment

As a mandatory seminar performance, students must apply at least one of the learned methods from week 5, 6, 9, 11, 12 or 13 to their own or other data. In doing so, they must develop a research question, but do not need to discuss it further.
Students must use an `R` Markdown or Quarto script, where critical steps in the analysis should be commented on. In the subsequent session (in the following week), the application should be presented in a brief impulse lecture. This lecture should also briefly address challenges and difficulties encountered during the analysis.

The seminar can only be passed if both the creation of a script and the presentation in the plenary session in the following week are completed.

Policy brief or newspaper report

Those who need a grade for their module must submit a policy brief **or** newspaper report by the end of the semester (no later than September 30, 2026). The paper should consist of approximately 4000 words ($\pm$ 10 percent). The bibliography and any potential appendices do not count towards the word limit.

Irrespectively of the format as a policy brief or newspaper report, the term paper should be based on the methods learned in the seminar. It must apply at least one text-as-data method. However, it is not a pure methods essay. Instead, the method(s) should be applied to a self-developed research question, one that appears relevent from a practitioner's or journalist's standpoint . The paper is mainly about raising a question of social relevance and finding an answer through text-as-data methods.

Accordingly, the seminar paper should begin with a puzzling question, motivate why its worth asking the question, identifying relevant literature and then proposing your own approach to test the question. Students should give a comprehensive explanation of why they chose a certain method, and how they implement it. Afterwards, they shall present results. Results should be visualized, so that policymakers or readers of a newspaper can easily comprehend what the paper finds. Finally, you should give a conclusion: what do the results you find imply for your research question? What are remaining limitations?

We will discuss the expectations in the seminar, especially in the last session. Additionally, all students should have a rough idea for their research project by the last session. It is recommended to discuss this idea with me during office hours.

Plagiarism

Plagiarism and all forms of ghostwriting are prohibited. Written works will be checked for plagiarism. During the seminar, we will see how important AI tools have become for document analysis. Therefore, the use of these tools is generally possible. However, students should ensure two things: (1) Always remain the *human in the loop*. You should lead AI, not the other way around. Your seminar assignments as well as your term paper should be based on your creative thinking. (2) You **must** be transparent about your usage of AI tools and declare which prompts you used. Failure to declare AI usage in the submission of any written work leads to failure of the course. In case of doubt, the final paper must be defended

orally. For coding, AI can be a helpful assistant **but** if you don't try and fail, you won't learn. So, consider doing assignments first by your own before asking for AI assistance.
It is expected that students will use scientific sources and cite them correctly. To simplify the workflow and prepare for future work (such as the bachelor's thesis), students are recommended to use citation software. An excellent tool for this purpose is the open-source software Zotero. If there are doubts about correct citation, please refer to the Guidelines for Academic Work, provided by the Institute of Political Science.

Inclusion

This seminar seeks to establish an inclusive learning environment, where students of all backgrounds can actively participate in the discussion. The instructor will use gender-sensitive language. Participants are invited to share their pronouns with the class.
Furthermore, an attempt will be made to establish a dynamic feedback culture, in which students can provide feedback on the seminar during the semester. Therefore, students are encouraged to regularly provide (anonymous) feedback via LamaPoll or by emailing me.

Readings and Timetable

Week	Date	Topic	Main Reading	Complementary Readings
1	15 April 2026	Introduction		
2	22 April 2026	Basics in R I		
3	29 April 2026	Basics in R II		
4	06 May 2026	Basics of Text Analysis	Krippendorff (2018) Grimmer und Stewart (2013)	Benoit (2020)
5	13 May 2026	Web-Scraping I	Munzert (2015): Chapter 1 & 2	
6	20 May 2026	Web-Scraping II	Munzert (2015): Section 9.1.9	
7	27 May 2026		*No Session (Whitsun Holidays)*	
8	03 June 2026	Data Preparation	Denny und Spirling (2017) Hvitfeldt und Silge (2022)	
9	10 June 2026	Topic Modeling	Roberts et al. (2019) Kangaslahti et al. (2026)	Blei et al. (2003)
10	17 June 2026		*No Session (EPSS conference)*	
11	24 June 2026	Supervised Text Analysis	James et al. (2021) Hvitfeldt und Silge (2022)	Stukal et al. (2017)
12	01 July 2026	Embeddings	P. Rodriguez und Spirling (2021)	P. L. Rodriguez et al. (2023)
13	08 July 2026	LLMs and Classification	Törnberg (2024) Mellon et al. (2024)	Alizadeh et al. (2025) Egami et al. (2024) Gao et al. (2025) Spirling (2023)
14	15 July 2026	Projects and Wrap-Up		
15	22 July 2026	No Session		

Textbooks

- **General workflow** Stoltz, D. S., & Taylor, M. A. (2024, März). *Mapping Texts: Computational Text Analysis for the Social Sciences* (1. Aufl.). Oxford University PressNew York. https://doi.org/10.1093/oso/9780197756874.001.0001
- Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (First edition). O'Reilly
- Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press
- **For literature scholars, but with a good insight into data preparation and some analysis formats:** Jockers, M. L., & Thalken, R. (2020). *Text Analysis with R: For Students of Literature* (2nd edition). Springer

Data

For your own projects (both the seminar and the exam performance) you can use various corpora. Some example corpora are listed here. Over the course of the seminar (especially in the scraping sessions) we will use additional data.

- Baumann, M., & Gross, M. (2016). Where Is My Party? Introducing New Data Sets on Ideological Cohesion and Ambiguity of Party Positions in Media Coverage
- Jankin, S., Baturo, A., & Dasandi, N. (2024, August). United Nations General Debate Corpus 1946-2023. https://doi.org/10.7910/DVN/0TJX8Y
- Lehmann, P., Franzmann, S., Al-Gaddooa, D., Burst, T., Ivanusch, C., Regel, S., Riethmüller, F., Volkens, A., Weßels, B., Zehnter, L., Wissenschaftszentrum Berlin Für Sozialforschung (WZB) & Institut Für Demokratieforschung Göttingen (IfDem). (2024). Manifesto Project Dataset. https://doi.org/10.25522/MANIFESTO.MPDS.2024A
- Rauh, C., & Schwalbach, J. (2020, März). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. https://doi.org/10.7910/DVN/L4OAKN
- Open Knowledge Foundation Deutschland e.V. (2024). kleineAnfragen
- Various newspaper corpora, accessible via CLARIN

- Various legal corpora

Week 1: Introduction [15.04.2026]

No preparation is required, but participants should bring their laptops.

- Atwell, E. (1999). Computers break the language barrier. *The guardian*
- Tapper, J. (2023). Authors shocked to find AI ripoffs of their books being sold on Amazon. *The guardian*

Week 2: Basics in R I [22.04.2026]

In the next two weeks, we will do a brief recap of R as a statistical software. Since I am at a conference, you will be asked to complete a problem set by your own at home. We will discuss solutions and more details in the upcoming week.

- **Introduction to R:**
  Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, Juni). *R for Data Science*. O'Reilly Media, Inc.
- **Shorter introduction in German:**
  Ellis, A., & Mayer, B. (2024). Einführung in R
- For a **quick guide to essential *dplyr* functions** YouTube tutorials are helpful, e.g. the following playlist.

Week 3: Basics in R II [29.04.2026]

In this week we talk about your problem sets and deal with data manipulation (*dplyr*), regular expressions (*stringr*) and data analysis in R.

- Literature see previous week
- for regular expressions, Chapter 15:
  Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, Juni). *R for Data Science*. O'Reilly Media, Inc.

Week 4: Basics of Text Analysis [06.05.2026]

In this session we deal with quantitative text analysis as a specialized form of content analysis. We will learn, among other things, in which contexts quantitative analyses of text are appropriate.

- **Chapter 2 on the conceptual clarification of content analysis and potential difficulties of quantitative text analysis:**
  Krippendorff, K. (2018). *Content Analysis: An Introduction to its Methodology* (Fourth Edition). SAGE

- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Optional: Benoit, K. (2020). Text as Data: An Overview. In L. Curini & R. Franzese (Hrsg.), *The SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

Week 5: Automated Data Collection 1 – Web Scraping of Static Pages [13.05.2026]

This week starts the first of three blocks dealing with automated collection of web data. In this session we will work a lot with statistical websites that are based on HTML.

- **Chapters 1 and 2**:
  Munzert, S. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley

- To access specific objects on webpages we need a **basic understanding of HTML structures**. Accordingly, students should go through the following tutorial in advance

Week 6: Automated Data Collection 2 – Web Scraping of Dynamic Pages and APIs [20.05.2026]

The second part of the block on automated data acquisition deals with more complex webpages that are based on JavaScript and cannot be accessed solely via HTML. Additionally we will deal with the functioning of APIs. APIs are official interfaces that allow direct download of data. We will get to know various APIs, including the API of the Manifesto Project and the NYT.

- **Section 9.1.9:**
  Munzert, S. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley
- **Background on the package we will use:**
  Harrison, J., & Ju Yeong, K. (2022, September). RSelenium: R Bindings for 'Selenium WebDriver'

Week 7: No session (Whitsun Holidays) [27.05.2026]

There will be no session this week.

Week 8: Data Preparation and Preparation for Analyses [03.06.2026]

Bags-of-words approaches require well-founded decisions regarding data preparation. We discuss the implications of choices like stopword removal, pruning, stemming, etc., and learn how to implement them in R.

- Denny, M., & Spirling, A. (2017, September). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. https://doi.org/10.2139/ssrn.2849145
- Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press – especially Chapters 2–4
- Overview of regular expressions in R
- Playground for regular expressions in R

Week 9: Topic Modeling [10.06.2026]

In this session we apply a quantitative text analysis for the first time. Using topic models we try to classify texts into different categories.

- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of statistical software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02

- **Empirical case study:**
  Kangaslahti, S., Ebanks, D., Kossaifi, J., Liu, A., Alvarez, R. M., & Anandkumar, A. (2026). Analyzing Political Text at Scale with Online Tensor LDA. *Political analysis*, *34*(1), 53–77. https://doi.org/10.1017/pan.2025.10024

Optional: Mathematical background to Latent Dirichlet Allocation:
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. mach. learn. res.*, *3*(null), 993–1022

Week 10: Scaling [17.06.2026]

While in the previous week we followed an unsupervised approach (i.e., an analysis without our input as researchers), this week we will apply a semi-supervised method, i.e., we will feed some input into the model to scale texts along a latent dimension.

- Watanabe, K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication methods and measures*, *15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976

- **Empirical case study:**
  Zollinger, D. (2024). Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses. *American journal of political science*, *68*(1), 139–159. https://doi.org/10.1111/ajps.12743

Week 11: Supervised Text Analysis and Basics of Machine Learning [24.06.2026]

In this week we learn some basics for supervised analysis of text. In this context we also deal with the basics of machine learning.

- **Introduction to Machine Learning, Chapter 2:**
  James, G., Witten, D., Hastie, T., & Tibshirani. (2021). *An Introduction to Statistical Learning*
- **Chapter 7:**
  Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press
- **Empirical case study:**
  Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting Bots on Russian Political Twitter. *Big data*, *5*(4), 310–324. https://doi.org/10.1089/big.2017.0038

Week 12: Embeddings [01.07.2026]

So far we have dealt with bag-of-words approaches that ignore the context of a word in a sentence. Word embeddings incorporate this (partially). We will learn them here and apply them ourselves.

- **Introduction to Word Embeddings:**
  Rodriguez, P., & Spirling, A. (2021). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The journal of politics*. https://doi.org/10.1086/715162
- **Empirical case study:**
  Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding Regression: Models for Context-Specific Description and Inference. *American political science review*, *117*(4), 1255–1274. https://doi.org/10.1017/S0003055422001228

Week 13: LLMs and Classification [08.07.2026]

In recent years there have been dramatic advances in computer-assisted text analysis. Deep neural networks enable us, using embeddings and various layers, among other things, to sort texts into categories. We will learn the fundamentals of these structures and build our own deep neural network in R.
Tutorial on deep learning
Further reading (especially on transformer models)

- On LLMs
  - Törnberg, P. (2024). Best Practices for Text Annotation with Large Language Models. *Sociologica*, *18*(2), 67–85. https://doi.org/10.6092/ISSN.1971-8853/19461
  - Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & politics*, *11*(1), 20531680241231468. https://doi.org/10.1177/20531680241231468
  - Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2025). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of computational social science*, *8*(1), 17. https://doi.org/10.1007/s42001-024-00345-9
  - Egami, N., Hinck, M., Stewart, B. M., & Wei, H. (2024). Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses. *Preprint from november*, *17*, 2024
  - Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the national academy of sciences*, *122*(24), e2501660122. https://doi.org/10.1073/pnas.2501660122
  - Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4
- In case you want to know more:
  - **The basic idea behind transformer models:**
    Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,

ukasz Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*

- Wankmüller, S. (2022). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological methods & research*, 1–77. https://doi.org/10.1177/00491241221134527
- **Podcast on the basic idea of AI:**
  Klein, E. (n. d.). A.I. Could Solve Some of Humanity's Hardest Problems. It Already Has.

Week 14: Wrap-Up Session [15.07.2026]

In this session we will recap which methods we have learned during the past semester. The unit also offers students the opportunity to clarify open questions and start brainstorming for the term paper. Students should prepare a short 5 minute Elevator Pitch of the planned news story / policy brief.

Week 15: No Session (Exam week) [22.07.2026]

This week will be no session due to exams in other subjects.