

Course Code: Advanced Methods: Quantitative Text Analysis, Final Version

Jens Wäckerle

Fall Semester, 2019

E-mail: jens.waeckerle@uni-koeln.de

Web: <https://sites.google.com/view/jens-waeckerle>

Office Hours: please send me an email if you want to meet individually.

Course Description

Quantitative Text Analysis has quickly become an important part of the quantitative methods political scientists use. It can help us to understand debates in parliaments, political manifestos, discourse online and debates in the media. Meanwhile, more and more data has become available: We have access to increasingly well organized datasets of parliamentary speech as well as video and newspaper archives. Additionally, translation services and automated transcription software break down boundaries of language and data format. This allows political scientists to gain insight into political processes with an unprecedented depth and scope.

This course is designed to teach students the tools required to work with text data. It will cover a sound introduction into handling text data and what promises and pitfalls lie in analyzing textual data in political science. We will then explore the most important descriptive and inferential methods that have dominated Quantitative Text Analysis in recent years.

Required Materials

- There is no single textbook or statistics book that is used throughout the course. For a general introduction (e.g. when deciding whether to take this course) on Quantitative Text Analysis (in German and English), I suggest:
 - Jonathan B. Slapin and Sven-Oliver Proksch. Words as Data: Content Analysis in Legislative Studies. In Shane Martin, Thomas Saalfeld, and Kaare W. Strøm, editors, *The Oxford Handbook of Legislative Studies*, pages 126–144. Oxford University Press, Oxford, 2014
 - Sven-Oliver Proksch. Computergestützte Textanalysen. In Claudius Wagemann, Achim Goerres, and Markus Siewert, editors, *Handbuch Methoden der Politikwissenschaft*, pages 1–19. Springer VS, Wiesbaden, 2018

- All materials (readings as described below and R code) will be uploaded to the Intranet or made available through the university library.

Prerequisites

You should have completed a basic introduction into quantitative analysis in the social sciences and have some knowledge in R. We will not focus too heavily on the statistical underpinnings of quantitative text analysis, so an advanced course in statistics is not required. We will spend a considerable amount of time on examples in R, so a basic knowledge of R is required. However, these examples will be based on built-in datasets in R using some well-documented functions in the package `quanteda`, so it should be easy to follow even if you have used R only sparingly in the past. We will also have enough time to go through code together and make sure everyone can follow along.

The course will use materials and examples from political science, but students from other disciplines in the social sciences should have little trouble following along.

Course Objectives

1. Understand how Quantitative Text Analysis works and how it can help us answer important questions in the social sciences.
2. Get familiar with state of the art methods in Quantitative Text Analysis, employing a widely used package in R, `quanteda`.
3. Understand both the promises and pitfalls of Quantitative Text Analysis and get a feeling for when and under what conditions Quantitative Text Analysis can help us answer political science questions.

Course Structure

Class Structure

The course will be taught on two weekends with courses on Friday (1.30 pm to 7 pm) and Saturday (10 am to 4 pm). We will alternate between lectures that introduce the methods and lab sessions in which students will work with the code that is supplied by the instructor. Each day is split into two sessions of 2.5 hours each and will feature both lecture and lab work.

Assessments

The examination will consist of a midterm research paper of 5-6 pages that is due four weeks after the last session of the course. You are free to choose any research question from political science and use any of the methods presented in the course. During the last session, we will have time to talk about your research questions and methods. You can work in groups of two people if you want to. In that case, the page limit will be 10-12 pages.

Schedule and Readings

Session 1: Introduction - Friday, September 20, 2019 - 1.30 pm to 4 pm

In the first session, we will talk about the role of Quantitative Text Analysis (QTA) in general. We will explore promises, limitations and pitfalls that come with QTA and see an overview of the most important contributions of the field. This session will answer the following questions:

- What is Quantitative Text Analysis and why should we use it?
- What contributions did QTA bring to political science in recent years?

Required Readings

- Jonathan B. Slapin and Sven-Oliver Proksch. Words as Data: Content Analysis in Legislative Studies. In Shane Martin, Thomas Saalfeld, and Kaare W. Strøm, editors, *The Oxford Handbook of Legislative Studies*, pages 126–144. Oxford University Press, Oxford, 2014

Additional Readings

- Sven Oliver Proksch and Jonathan B. Slapin. Institutional Foundations of Legislative Speech. *American Journal of Political Science*, 56(3):520–537, 2012

Session 2: Descriptives, Exploring Text Data and Text Complexity - Friday, September 20, 2019 - 4.30 pm to 7 pm

In the second session of the first day, we explore some basics of QTA: How to get data and explore descriptive statistics. We then go into some more technical aspects of dealing with text data: preprocessing and similarity measures. We also start exploring some data sets using the *quanteda* package in R. We then dig deeper into political text and also explore how we can use measures of text complexity to describe political speech and draw substantial conclusions about political actors.

- How can I get a first idea of what is going on in the text corpus I have?
- What is preprocessing and why is it important?
- What can the complexity of a text tell us?

Required Readings

- Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013
- Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N. Bakker. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PLoS ONE*, 14(2):1–15, 2019

Additional Readings

- Matthew Denny and Arthur Spirling. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, 26(2):168–189, 2018

- Arthur Spirling. Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915. *The Journal of Politics*, 78(1):120–136, 2015

Session 3: Dictionary-based approaches and sentiment - Saturday, September 21, 2019 - 10 am to 12.30 am

In this session, we tackle dictionaries and specifically sentiment analysis, an approach that has increased in importance in QTA recently. Dictionaries can be used to measure the prevalence of a pre-defined concept in a text, such as positive or negative sentiment or policy specific words. This session will answer the following questions:

- How can we measure the sentiment expressed in a speech?
- Can we use sentiment measures across languages?

Required Readings

- Sven Oliver Proksch, Will Lowe, Jens Wackerle, and Stuart Soroka. Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, 44(1):97–131, 2018

Additional Readings

- Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality and Quantity*, 51(6):2623–2646, 2017

Session 4: Text Classification and Machine Learning - Saturday, September 21, 2019 1.30 pm to 4 pm

In this session, we will explore how to classify text based on several machine learning algorithms. These methods are used to classify large quantities of text and/or let us draw wider conclusions about the circumstances that produced these texts (e.g. polarization). This session will answer the following questions:

- What is Machine Learning?
- How can we use classification in research?

Required Readings

- Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Par-net. A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, 66(6):1007–1031, 2016
- Takes a close look at the **Supplementary Appendix** of the Theocharis et al paper, which describes their strategy in more detail.

Additional Readings

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. 2013, chapters 2 and 6.
- Andrew Peterson and Arthur Spirling. Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, 26(1):120–128, 2018

Session 5: Supervised Scaling - Friday, September 27, 2019 - 1.30 pm to 4 pm

In this session, we will start with an important field in QTA: scaling political text. QTA provides political scientist with an important tool to infer positions of political actors that requires relatively little effort, money and time, compared to other approaches. This session will answer the following questions:

- How can we use QTA to infer political positions?
- How do these positions compare to other ways to estimate positions?

Required Readings

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

Additional Readings

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

Session 6: Unsupervised Scaling - Friday, September 27, 2019 - 4.30 pm to 7 pm

In the second session on day 3, we will use an unsupervised scaling method named Wordfish. Unsupervised scaling does require minimal assumptions about the nature of the political space and can recover political positions reliably. This session will answer the following questions:

- How can we scale political texts without requiring detailed assumptions about the nature of political conflict?
- How can we validate the positions we recover from unsupervised and supervised scaling?

Required Readings

- Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008
- Heike Klüver. Measuring interest group influence using quantitative text analysis. *European Union Politics*, 10(4):535–549, 2009

Additional Readings

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

Session 7: Social Media Data - Saturday, September 28, 2019 - 10 am to 12.30 am

In the first session of the final day, we will see how we can acquire, use and analyze social media data (e.g. from Twitter) with some of the tools described above. This session will answer the following questions:

- How can we acquire and analyze social media data to analyze political discourse?
- Which new research questions and hypotheses can we test with social media data?

Required Readings

- Pablo Barberá. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91, 2015

Additional Readings

- Armin Mertens, Franziska Pradel, Ayjeren Rozyjumayeva, and Jens Wäckerle. As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians. *Working paper*, 2019

Session 8: Recap and research papers - Saturday, September 28, 2019 - 10 am to 12.30 am

In the final session, we will recap the methods that we learned over the last two weeks. We will have time to work on the research projects and I will be available to answer any questions regarding what methods to use, how to get data, how to deal with common issues such as encodings and anything else that comes up.

References

- Pablo Barberá. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91, 2015.
- Matthew Denny and Arthur Spirling. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, 26(2):168–189, 2018.
- Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.
- Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality and Quantity*, 51(6):2623–2646, 2017.
- Alexander Herzog and Kenneth Benoit. The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent during Economic Crisis. *The Journal of Politics*, 77(4):1157–1175, 2015.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. 2013.
- Heike Klüver. Measuring interest group influence using quantitative text analysis. *European Union Politics*, 10(4):535–549, 2009.

- Benjamin E. Lauderdale and Alexander Herzog. Measuring Political Positions from Legislative Speech. *Political Analysis*, 24(3):374–394, 2016.
- Will Lowe. Understanding Wordscores. *Political Analysis*, 16(4):356–371, 2008.
- Armin Mertens, Franziska Pradel, Ayjeren Rozyjumayeva, and Jens Wäckerle. As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians. *Working paper*, 2019.
- Andrew Peterson and Arthur Spirling. Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, 26(1):120–128, 2018.
- Sven-Oliver Proksch. Computergestützte Textanalysen. In Claudius Wagemann, Achim Goerres, and Markus Siewert, editors, *Handbuch Methoden der Politikwissenschaft*, pages 1–19. Springer VS, Wiesbaden, 2018.
- Sven Oliver Proksch and Jonathan B. Slapin. Institutional Foundations of Legislative Speech. *American Journal of Political Science*, 56(3):520–537, 2012.
- Sven Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, 44(1):97–131, 2018.
- Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N. Bakker. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PLoS ONE*, 14(2):1–15, 2019.
- Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008.
- Jonathan B. Slapin and Sven-Oliver Proksch. Words as Data: Content Analysis in Legislative Studies. In Shane Martin, Thomas Saalfeld, and Kaare W. Strøm, editors, *The Oxford Handbook of Legislative Studies*, pages 126–144. Oxford University Press, Oxford, 2014.
- Arthur Spirling. Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915. *The Journal of Politics*, 78(1):120–136, 2015.
- Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. A Bad Workman Blames His Tweets: The Consequences of Citizens’ Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, 66(6):1007–1031, 2016.