Text-mining international politics

Block Seminar
Institute of Political Studies, Charles University, Prague
May 7-10, 2019

Dr. Christian Rauh
WZB Berlin Social Science Center

christian.rauh@wzb.eu
www.christian-rauh.eu

I. Overview

Politics takes place in and through texts. Speeches, debates, position papers, press releases, traditional and social media coverage, as well as the resulting laws, agreements, or resolutions can tell us much on the priorities, preferences, and power of political actors. This holds especially for politics beyond the nation state. For studying international politics across long time periods or broad actor sets, political text is often the only consistently available information source that we have. The challenge, however, lies in extracting systematic information from largely unstructured texts in a reliable and systematic fashion. This is a promise of automated content analyses: various algorithms offer means to reveal relevant patterns in the vast amount of political language that we can nowadays access in digital formats.

The block seminar thus introduces the strengths but also the limitations of various approaches to treat text as data. Based on my own work with and on these tools, students will learn about the basic intuitions behind the most prominent text analysis methods in recent political science research. We will work along concrete examples by discussing extant and possible applications of these methods to EU and international politics. These examples are also useful to highlight the pragmatic issues involved in collecting, analysing, and visualizing large-scale digital text corpora.

The course targets advanced BA as well as MA students in the political sciences or related disciplines who wish to broaden their empirical toolkit. Prior knowledge in content analysis or quantitative methods is not required but may be an asset. The seminar pursues three related teaching goals:

- Enable students to read and to assess studies using automated text analyses
- Allow informed methodological choices and provide pragmatic tips/resources for conducting own text analyses
- Create awareness for the more general promises and pitfalls of analysing human language with automated algorithms

Successful participants will be awarded with 4 ECTS. The evaluation of student performance will be based on three criteria:

- Thorough reading of the obligatory literature marked with (O) in the syllabus below
- Regular and active participation in the individual six sessions of the seminar
- A short research-design paper (4,000-5,000 words) that discusses whether and which automated text analyses might be suited to address a freely chosen question on the EU, international politics, or related fields (more details during the seminar)

Ideally, this course convinces you that conducting text analyses can be fun and insightful at the same time. In any case, I am very much looking forward to work with you!

II: Course organization and literature

(O): marks literature that each participant should prepare *before* each session
(R): marks recommended literature going more into depth or presenting exemplary applications

All course materials will be available at latest 3 weeks beforehand in the University system (SIS).

Day 1: May 7 2019

Session 1: Fundamentals: Qualitative, quantitative, and automated content analysis

Time: *15:30 – 17:00*
Room: Celetna street 20, C216

This session sets the methodological cornerstones for content analysis in the social sciences more generally. Our discussion focusses on ubiquitous human biases in text interpretation and the resulting challenges of reliable and valid measurement of political variables from textual data. On this basis, we then contrast well-established content analysis approaches (qualitative interpretation and human coding) to automated text analyses.

(O) Krippendorff, Klaus (2004) *Content Analysis: An Introduction to Its Methodology*. London: Sage Publications: Chapters 1-2 (pp. 3-43)

(R) Neuendorf, Kimberly (2001) *The Content Analysis Guidebook*. SAGE Publications: Chapters 1-2

(R) Krippendorff, Klaus (2004) Content Analysis: An Introduction to Its Methodology. London: 2nd Edition. London: Sage Publications: Chapter 11 "Reliability" (pp. 211-23 & 241-50) and Chapter 13 "Validity" (pp. 313-38).

Session2: Automated text analysis: Corpus construction and discovery

Time: *17:00-18:00*
Room: Celetna street 20, C216

This session initially focusses on the practical requirements of conducting a text analysis: From where and how to collect political texts? How to store them and how to turn them into data? Then we will delve into initial exploratory analyses of large text corpora. Rather simple visualizations already produce interesting insights once a relevant corpus has been constructed. Amongst others, speeches on climate change in the United Nations General Assembly will provide an example here.

(O) Grimmer, Justin, and Brandon Stewart (2013) 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts', Political Analysis 21(3): 267-297.

(R) Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis (2015) *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Wiley.

(R) Lowe, Will (2003) 'The Statistics of Text: New Methods for Content Analysis', Paper presented as the *Midwest Political Science Association Conference* April 2003.

Day 2: May 9 2019

Session 3: **Categorize political text I: Dictionary-based approaches**

Time: *14:00 – 15:30*
Room: Celetna street 20, C216

This session looks into dictionary-based approaches that employ often very encompassing and flexible key word lists to identify issues of interest to the researcher in political text corpora. This also covers dictionary-based sentiment analyses which allow us to see how positively or how negatively political actors frame certain issues in the language they use. Showcase examples in this session cover political speeches on the European Union as well as media coverage of the WTO. While dictionary-based methods are highly intuitive, a number of methodological and pragmatic pitfalls need to be taken into account

(O) Rauh, Christian and De Wilde, Pieter (2018) 'The Opposition Deficit in EU Accountability: Evidence from over 20 years of plenary debate in four member states'. *European Journal of Political Research* 57(1): 194–216.

(O) Rauh, Christian (2018) 'Validating a sentiment dictionary for German political language', *Journal of Information Technology & Politics* 15(4): 319-343.

(R) Young, Lori, and Stuart Soroka (2012) 'Affective News: The Automated Coding of Sentiment in Political Texts', *Political Communication* 29(2): 205-231.

(R) Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka (2018) 'Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches', *Legislative Studies Quarterly*(Online first).

(R) Rauh, Christian, Bart Joachim Bes, and Martijn Schoonvelde (2018) 'Undermining, defusing, or defending European integration? Assessing public communication of European executives in times of EU politicization', *European Journal of Political Research,* forthcoming.

(R) Rauh, Christian, and Sebastian Bödeker (2013) The international trade regime in the public sphere, 1986-2012. In *European Workshop on International Studies (EWIS)*. Tartu, Estonia.

Session 4: **Categorize political text II: Topic models and machine learning**

Time: *15:30-17:00*
Room: Celetna street 20, C216

This session focusses on more advanced approaches that analyse term frequencies statistically to classify texts. First, we will look at topic models. Topic models – a tool originally develop to optimize search engines - are an unsupervised method in that promise to just learn the distribution of topics/issues/themes from the texts themselves without much researcher intervention. Exemplary applications model UN debates or preferential trade agreements. On the other hand, we will look into the basic intuition behind machine learning. Here, algorithms are trained and tested on human interpretations before they ‘amplify’ these interpretations to larger amounts of text.

(O) Blei, David (2012) 'Probabilistic topic models', Commun. ACM 55(4): 77-84.

(O) Quinn Albaugh, Stuart Soroka, Jeroen Joly, Peter Loewen, Julie Sevenans, and Stefaan Walgrave (2014) ‘Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding’, *Unpublished Manuscript*.

(R) Hopkins, Daniel, and Gary King (2010) 'A Method of Automated Nonparametric Content Analysis for Social Science', *American Journal of Political Science* 54(1): 229-247.

(R) Bagozzi, Benjamin E. (2015) 'The multifaceted nature of global climate change negotiations', *The Review of International Organizations* 10(4): 439–464.

(R) Greene, Derek, and James P. Cross (2017) 'Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach', *Political Analysis* 25(1): 77-94.

(R) Genovese, Federica (2015) 'Politics ex cathedra: Religious authority and the Pope in modern international relations', *Research & Politics* 2(4).

(R) Lewis Kraus, Gideon (2016) ‘The Great A.I. Awakening: How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself‘, *The New York Times Magazine*.

Day 3: May 10 2019

Session 5: Positioning political text: Scaling approaches

Time: *11:00-12:30*
Room: Celetna street 20, C216

In this session, we discuss text analyses that aim at placing speakers or document authors on latent dimensions of political conflict. Initially, we consider *Wordscores*, a supervised algorithm that places speeches or documents in between reference texts the researcher supplies. Then, we will look into *Wordfish*, an unsupervised algorithm that aims to uncover a single conflict dimensions solely from the observed word distributions. While these tools have been originally developed to study party positions, our examples on lobbying success in the EU or on states’ expressed positions on climate change in UN debates showcase their potential for studying international politics.

(O) Laver, Michael, Kenneth Benoit, and John Garry (2003) 'Extracting Policy Positions from Political Texts Using Words as Data', *The American Political Science Review* 97(2): 311-331.

(O) Slapin, Jonathan, and Sven-Oliver Proksch (2008) 'A Scaling Model for Estimating Time-Series Party Positions from Texts', *American Journal of Political Science* 52(3): 705-722.

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

(R) Klüver, Heike (2009) 'Measuring Interest Group Influence Using Quantitative Text Analysis', *European Union Politics* 10(4): 535-549.

(R) Genovese, Federica (2014) 'States' interests at international climate negotiations: new measures of bargaining positions', *Environmental Politics* 23(4): 610-631.

Session 6: Advanced methods and conclusions

Time: *12:30-14:00*
Room: Celetna street 20, C216

This final session will briefly present text analysis methods that go beyond mere frequency models. Text similarity approaches – known amongst other things from plagiarism detection - take word order into account and allow, for example, to study the evolution of international treaties across time and countries. Grammatical parsing facilitate extracting relations between different actors expressed in text and have been used to predict violent political conflict in the Middle East, for example. Word vector models - a key approach in contemporary artificial intelligence - exploit term co-occurrences in large text corpora. This allows studying the changing meaning of terms over time, making these tools relevant for understanding norm evolution in international relations, for example. We will then finally wrap up, by comparing the promises and pitfalls of the different methods to the methodological cornerstones developed in Session 1.

(O) Allee, Todd, and Andrew Lugg (2016) 'Who wrote the rules for the Trans-Pacific Partnership?', *Research & Politics* 3(3).

(O) Schrodt, Philip A., and David Van Brackle (2012) 'Automated Coding of Political Event Data'. In: V. Subrahmanian (ed.) *Handbook of Computational Approaches to Counterterrorism*. New York: Springer.

(R) Alschner, Wolfgang, and Dmitriy Skougarevskiy (2016) 'Mapping the Universe of International Investment Agreements', *Journal of International Economic Law* 19(3): 561-588.

(R) Cross, James P, and Henrik Hermansson (2017) 'Legislative amendments and informal politics in the European Union: A text reuse approach', *European Union Politics* 18(4): 581-602.

(R) Gurciullo, Stefano, and Slava Mikhaylov (2017) Detecting Policy Preferences and Dynamics in the UN General Debate with Neural Word Embeddings: https://arxiv.org/abs/1707.03490

(R) Bruchansky, Christophe (2017) Political Footprints: Political Discourse Analysis using Pre-Trained Word Vectors. https://arxiv.org/abs/1705.06353

Exemplary large-scale collections of political text

ParlSpeech corpus: 3.9 million parliamentary speeches (including CZ/PSP 1993-2016) http://bit.ly/ParlSpeech

EUspeech Corpus: 17,184 speeches of national, EU, and IMF executives during the 2007-2015 period: https://doi.org/10.7910/DVN/GKABNU

UNGD corpus: All speeches by national representatives in the United Nations General Assembly 1970-2017: https://doi.org/10.7910/DVN/0TJX8Y

Manifesto Corpus: Full texts of partisan election manifestos from more than 50 countries https://manifesto-project.wzb.eu/information/documents/corpus

US SOTU Corpus: Full texts of all State-of-the-Union by US presidents 1790-2018 https://www.kaggle.com/rtatman/state-of-the-union-corpus-1989-2017/data

Useful text analysis resources

JFreq and Yoshikoder by Will Lowe:	http://conjugateprior.org/software/
R packages for diff. text analysis tasks:	https://cran.r-project.org/web/views/NaturalLanguageProcessing.html
Quanteda tutorials:	https://tutorials.quanteda.io/
Text mining in R along ‘tidy’ principles:	https://www.tidytextmining.com/
Test your Regular Expressions:	https://regexr.com/
Search text file batches (with Regex):	https://www.digitalvolcano.co.uk/textcrawler.html