What's on the Agenda? Automated Text Classification for Political Text

Monday 17:15 18:45 weekly 28.09.2020 - 07.12.2020 SOWI-Zoomroom 6

https://uni-mannheim.zoom.us/j/5677743231?pwd=YzVDN0FqaWdYVlhNWXBwcVg1Zm1FQT09

Marius Sältzer

msaeltze@mail.uni-mannheim.de
A 339
Office Hours: TBD
+49(0)621-181-2078

Syllabus will be adapted over the course of the semester.

Course Description:
Politics is often about the importance of particular issues, from migration over climate change to COVID-19. To understand political processes from political communication to legislation, measuring what matters to political actors, newspaper or voters is of great importance. In recent times, the growing availability of text from social media to parliamentary procedures has increased the demand for automated content analysis. This methods course will introduce a number of basic clustering and machine learning approaches to automatically identify the topics of political text. It offers a hands-on introduction to big data with examples from newspapers, manifestos, speeches and social media data. Part of this course is an introduction to the statistical programming language R. It will be taught in the seminar and previous experience is not required, but the openness to learning a new language is key. Basic knowledge of statistics and respective software like STATA is required.

1. Data Science for Comparative Political Science
2. Get a first look at R
3. Find out what's out there in terms of data
4. Develop your own research project

Formalia:

- This is a methods course and I want you to get access to the tools to collect your own data.
- Your Grade depends on two Elements:
  - **Studienleistung:**
    - *Problem Sets*: You will hand in three problem sets (pass/fail) of which 2 have to be passed.
  - **Prüfungsleistung:**
    A *Term Paper* that is somehow connected to salience, issue emphasis or classifying text. It can be empirical but does not have to be.
- Software: R
  - Freeware that is used everywhere around the world
  - Easy to use on your home computer without license etc.
  - Introduction to programming and great job opportunities
  - Steep learning curve, but worth it!
- Helpful Textbooks:
  - Technical
    - concerning R
      - Andrie de Vries and Joris Meys. *R for dummies: Learn to: use R for data analysis and processing : write functions and scripts for repeatable analysis : create high-quality charts and graphics : perform statistical analysis and build models*. Wiley, Hoboken, NJ, 2. ed. edition, 2015. ISBN 9781119055839 (RFD)
      - Additional for the really curious: Hadley Wickham. *Advanced R*. Taylor and Francis, 2015 (A)
    - Text Analysis
      - Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, NJ, 2000. ISBN 978-0130950697
    - Substantive
      - Ian Budge, Hans-Dieter Klingeman, Andrea Volkens, Judith Bara, and Michael D. McDonald. *Mapping policy preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford University Press, Oxford, 2001. ISBN 0199244006

Part I: Theory:
The weekly coverage might change as it depends on the progress of the class. However, you must keep up with the reading assignments.

Week	Content
28.09.2020	• Introduction
05.10.2020 Salience and Agendas	Why is it important what issues we talk about in politics? • Peter Bachrach and Morton A. Baratz. Decisions and Nondecisions: An Analytical Framework. *American Political Science Review*, 57(3):632–642, 1963. ISSN 0003-0554 • Maxwell E. McCombs and Donald L. Shaw. The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, 36(2):176, 1972. ISSN 0033362X. doi: 10.1086/267990
12.10.2020 Issues and Politics	How do political parties engage with issues? Readings: • Stephen Ansolabehere, James M. Synder, and Jonathan Rodden. The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review*, 102(02):215–232, 2008. ISSN 0003-0554. doi: 10.1017/S0003055408080210 • Éric Bélanger and Bonnie M. Meguid. Issue salience, issue ownership, and issue-based vote choice. *Electoral Studies*, 27(3):477–491, 2008. ISSN 02613794. doi: 10.1016/j.electstud.2008.01.001
19.10.2020 Coding Political Text	How can we find out what parties emphasize? We engage with classification schemes that are used to analyze political text. • https://manifesto-project.wzb.eu/coding_schemes/mp_v5 • https://www.comparativeagendas.net/pages/master-codebook • Reading: – Heike Klüver and Jae-Jae Spoon. Who Responds? Voters, Parties and Issue Attention. *British Journal of Political Science*, 46(03):633–654, 2016. ISSN 0007-1234. doi: 10.1017/S0007123414000313 – Thomas M. Meyer and Markus Wagner. It Sounds Like They are Moving: Understanding and Modeling Emphasis-Based Policy Change. *Political Science Research and Methods*, 7(04):757–774, 2019. ISSN 2049-8470. doi: 10.1017/psrm.2017.30

Part II: Programming in R:

Week	Content
26.10.2020 Introduction to R 1	We start the introduction into R by installing R and R Studio, learning about basic operations, objects and functions. Please install • R: https://cran.r-project.org/mirrors.html • R Studio: https://rstudio.com/products/rstudio/download/download • Suggested Reading: – Sessions 5 and 6 will cover RFD p.49-149
02.11.2020 Introduction to R 2	After we learned basic operations, we load data from the manifesto project! We learn how to import, display and manipulate data sets. We import the dataset of the Manifesto Project. • Suggested Reading: – Sessions 5 and 6 will cover RFD p.49-149 – https://manifestoproject.wzb.eu/information/documents/manifestoR
09.11.2020 Text as Data in R	After learning basic R, we directly enter the world of text as data. We learn about corpora, tokens, document-feature-matrices and word frequencies. • Software: https://tutorials.quanteda.io/basic-operations/ • Literature: Text as Data – Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(03):267–297, 2013. ISSN 1047-1987. doi: 10.1093/pan/mps028 – Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, NJ, 2000. ISBN 978-0130950697: **pages 2-28**

Part III: Text Classification:

Week	Content
16.11.2020 Supervised Learning	A brief introduction to Machine Learning to start Working with Labeled Data: Supervised Learning. • Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, NJ, 2000. ISBN 978-0130950697: **pages 56-74**
23.11.2020 Supervised Learning II	Learn evaluating your model by measuring precision and accuracy • Pablo Barberá, Andrue Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901, 2019. ISSN 0003-0554. doi: 10.1017/S0003055419000352
30.11.2020 Unsupervised Classification	We start with unsupervised machine learning, letting the computer choose the categories. We introduce so called topic models. • Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. ISSN 00925853. doi: 10.1111/ajps.12103 • David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55 (4):77–84, 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826
07.12.2020 Validating Topic Models	We look at ways to better intepreet and optimize topic models. • Jonathan Chang, Sean Gerrish, Wang. Chong, Jordan Boyd-graber, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in neural information processing systems*, pages 288–296, 2009 • Pablo Barberá, Andrue Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901, 2019. ISSN 0003-0554. doi: 10.1017/S0003055419000352

References

Stephen Ansolabehere, James M. Synder, and Jonathan Rodden. The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review*, 102(02):215–232, 2008. ISSN 0003-0554. doi: 10.1017/S0003055408080210.

Peter Bachrach and Morton A. Baratz. Decisions and Nondecisions: An Analytical Framework. *American Political Science Review*, 57(3):632–642, 1963. ISSN 0003-0554.

Pablo Barberá, Andrue Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901, 2019. ISSN 0003-0554. doi: 10.1017/S0003055419000352.

Éric Bélanger and Bonnie M. Meguid. Issue salience, issue ownership, and issue-based vote choice. *Electoral Studies*, 27(3):477–491, 2008. ISSN 02613794. doi: 10.1016/j.electstud.2008.01.001.

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826.

Ian Budge, Hans-Dieter Klingeman, Andrea Volkens, Judith Bara, and Michael D. McDonald. *Mapping policy preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford University Press, Oxford, 2001. ISBN 0199244006.

Jonathan Chang, Sean Gerrish, Wang. Chong, Jordan Boyd-graber, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in neural information processing systems*, pages 288–296, 2009.

Andrie de Vries and Joris Meys. *R for dummies: Learn to: use R for data analysis and processing : write functions and scripts for repeatable analysis : create high-quality charts and graphics : perform statistical analysis and build models*. Wiley, Hoboken, NJ, 2. ed. edition, 2015. ISBN 9781119055839.

Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(03):267–297, 2013. ISSN 1047-1987. doi: 10.1093/pan/mps028.

Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, NJ, 2000. ISBN 978-0130950697.

Heike Klüver and Jae-Jae Spoon. Who Responds? Voters, Parties and Issue Attention. *British Journal of Political Science*, 46(03):633–654, 2016. ISSN 0007-1234. doi: 10.1017/S0007123414000313.

Maxwell E. McCombs and Donald L. Shaw. The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, 36(2):176, 1972. ISSN 0033362X. doi: 10.1086/267990.

Thomas M. Meyer and Markus Wagner. It Sounds Like They are Moving: Understanding and Modeling Emphasis-Based Policy Change. *Political Science Research and Methods*, 7(04):757–774, 2019. ISSN 2049-8470. doi: 10.1017/psrm.2017.30.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. ISSN 00925853. doi: 10.1111/ajps.12103.

Hadley Wickham. *Advanced R*. Taylor and Francis, 2015.