Computational Analysis of Political Texts: Bridging Research Efforts Across Communities

Goran Glavaš and **Federico Nanni** and **Simone Paolo Ponzetto**
Data and Web Science Group, University of Mannheim
{goran,federico,simone}@informatik.uni-mannheim.de

1 Introduction

The development and adoption of natural language processing (NLP) methods by the political science community dates back to over twenty years ago. In the last decade the usage of computational methods for text analysis has drastically expanded in scope and has become the focus of many social science studies, allowing for a sustained growth of the text-as-data community (Grimmer and Stewart, 2013). Political scientists have in particular focused on exploiting available texts as a valuable (additional) data source for a number of analyses types and tasks, including inferring policy positions of actors from textual evidence (Laver et al., 2003; Slapin and Proksch, 2008; Lowe et al., 2011, *inter alia*), detecting topics (King and Lowe, 2003; Hopkins and King, 2010; Grimmer, 2010; Roberts et al., 2014), and analyzing stylistic aspects of texts, e.g., assessing the role of language ambiguity in framing the political agenda (Page, 1976; Campbell, 1983) or measuring the level of vagueness and concreteness in political statements (Baerg et al., 2018; Eichorst and Lin, 2018).

Just like in many other domains, much of the work on computational analysis of political texts has been enabled and facilitated by the development of dedicated resources and datasets such as, the topically coded electoral programmes (i.e., the Manifesto Corpus) (Merz et al., 2016) developed within the scope of the Comparative Manifesto Project (CMP) (Werner et al., 2014; Mikhaylov et al., 2012) or the topically coded legislative texts annotated for numerous countries within the scope of the Comparative Agenda Project (Baumgartner et al., 2006; Bevan, 2019).

While political scientists have dedicated a lot of effort to creating resources and using NLP methods to automatically process textual data, they have largely done so in isolation from the NLP community. For example, *political text scaling* – one of the central tasks in quantitative political science, where the goal is to quantify positions of politicians and/or parties on a scale based on the textual content they produce – has not received any attention by the NLP community until last year, whereas it has been at the core of political science research for almost two decades. At the same time, NLP researchers have addressed closely related tasks such as election prediction (O'Connor et al., 2010), ideology classification (Hirst et al., 2010), stance detection (Thomas et al., 2006), and agreement measurement (Gottipati et al., 2013), all rarely considered in the same format by the text-as-data political science community. In summary, these two communities have been largely agnostic of one another, resulting in NLP researchers not contributing to relevant research questions in political science and political scientists not employing cutting-edge NLP methodology for their tasks.

The main goal of this tutorial is to systematize and analyze the body of research work on computational analysis of political texts from both communities. We aim to provide a gentle, all-round introduction to methods and tasks related to computational analysis of political texts. Our vision is to bring the two research communities closer to each other and contribute to faster and more significant developments in this interdisciplinary research area. To that effect, this tutorial presents a continuation of our efforts which started with a very successful cross-community event organized in December 2017 (Nanni et al., 2018). In parallel with this tutorial at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), we will give a complementary tutorial at the 5th International Conference on Computational Social Science (IC^{2}S^2 2019).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 18–23
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

2 Tutorial Overview

This introductory tutorial aims to systematically organise and analyse the overall body of research in computational analysis of political texts. This body of work has been split between two largely disjoint research communities – researchers in natural language processing and researchers in political science – and the tutorial is designed bearing this in mind. We first explain the role that textual data plays in political analyses and then proceed to examine the concrete resources and tasks addressed by the text-as-data political science community. Continuing, we present the research efforts carried out by the NLP researchers. We close the tutorial by presenting text scaling, a challenging task that is at the center of the quantitative political science and has recently also attracted attention of NLP scholars. Accordingly, we divide the tutorial into the following four parts:

1. **Text as Data in Political Science.** We begin with an overview of the role that textual data has always played in political science research as a source for determining leader's positions (Winter and Stewart, 1977), campaign strategies (Petrocik, 1996), media attention (Semetko and Valkenburg, 2000), and crowd perception of the democratic process (Miller, 1990). We will further analyze the inherent difficulties in collecting political texts and political data in general and analyze crowdsourcing as an efficient and agile method for producing political data (Benoit et al., 2016).

2. **Resources and Tasks.** We then present computational research tasks based on textual data, which are relevant for the political science community (Grimmer and Stewart, 2013). We examine the type of applications and discuss the complex challenges currently faced, especially concerning cross-lingual and topic-based studies. We will analyze in detail the corpora developed within the scope of two major annotation projects: Comparative Manifesto Project (Werner et al., 2014; Mikhaylov et al., 2012) and Comparative Agendas Project (Baumgartner et al., 2006; Bevan, 2019). We will also describe other datasets, annotated corpora, gold standards, and benchmarks that are already promptly available (Bakker et al., 2015; Merz et al., 2016; Schumacher et al., 2016; Van Aggelen et al., 2017; Döring and Regel, 2019).

3. **Topical Analysis of Political Texts.** Next, we focus on a large body of work of topical analysis of political texts, covering unsupervised topic induction, including dictionary-based, topic-modelling and text segmentation approaches (Quinn et al., 2006, 2010; Grimmer, 2010; Albaugh et al., 2013; Glavaš et al., 2016; Menini et al., 2017), as well as supervised topic classification studies (Hillard et al., 2008; Collingwood and Wilkerson, 2012; Karan et al., 2016). We will also cover more recent work on cross-lingual topic classification in political texts (Glavaš et al., 2017a; Subramanian et al., 2018). We will further emphasize topic classification models that exploit large manually anotated corpora from CMP (Zirn et al., 2016; Subramanian et al., 2017) and CAP (Karan et al., 2016; Albaugh et al., 2013) projects, which we cover in the previous part.

4. **Political Text Scaling.** Finally, we present a detailed overview of the task of political text scaling, which has the goal of inferring policy position of actors from textual evidence. After introducing the text scaling task, we will present in detail the traditional scaling models that operate on lexical text representations such as Wordscores (Laver et al., 2003) and WordFish (Slapin and Proksch, 2008; Lowe et al., 2011) as well as a more recent scaling approach that exploits latent semantic text representations (Glavaš et al., 2017b; Nanni et al., 2019). Furthermore, we will discuss the task of scaling multilingual text collections, presenting potential approaches and inherent issues. We conclude the tutorial with a short discussion of key challenges and foreseeable future developments in computational analysis of political texts.

3 Tutorial Outline

Part I: Text-as-Data in Political Science (*30 min*)

- Quick introduction to quantitative methods in political science
- Reliability and suitability of textual data for political analyses
- Constructing corpora of political texts
- Crowdsourcing political data: advantages and potential pitfalls

Part II: Resources and Tasks (*30 minutes*)

- Overview of computational analysis of political texts in the political science community
- International annotation projects: Comparative Manifesto Project (CMP) and Comparative Agendas Project (CAP)
- Other large collection of political texts (EuroParl, UK Hansard Corpus, etc.) and associated tasks

Part III: Topical Analysis of Political Texts (*60 minutes*)

- Dictionary-based approaches to classification of political text
- Unsupervised topical analysis of political texts with topic models
- Models for supervised topic classification of political texts
- Hierarchical and fine-grained topic classification
- Cross-lingual topic classification

Part IV: Political Text Scaling and Conclusion (*60 minutes*)

- Lexical models for political text scaling: Wordscores and WordFish
- Text scaling using latent semantic text representations
- Policy dimensions in scaling: pitfalls and artefacts
- Cross-lingual scaling
- Conslusion: short discussion of key challenges and presumed future developments

4 Tutorial Breadth

In our previous work, we contributed to the research efforts on topic classification (Nanni et al., 2016; Zirn et al., 2016; Glavaš et al., 2017a), semantic scaling of political texts (Glavaš et al., 2017b) as well as (dis-)agreement detection in party manifestos (Menini et al., 2017). However, the key objective of this tutorial is to provide a comprehensive overview of recent and current research on computational analysis of political texts, both in NLP and political science communities. We estimate that at most one quarter of the tutorial will be dedicated to covering our own work.

5 Presenters

Goran Glavaš is an Assistant Professor for Statistical Natural Language Processing at the Data and Web Science group, University of Mannheim. He obtained his Ph.D. at the Text Analysis and Knowledge Engineering Lab (TakeLab), University of Zagreb. His research efforts and interests are in the areas of statistical natural language processing (NLP) and information retrieval (IR), with focus on lexical and computational semantics, multi-lingual and cross-lingual NLP and IR, information extraction, and NLP applications for social sciences. He has (co-)authored over 60 publications in the areas of NLP and IR, publishing at top-tier NLP and IR venues (ACL, EMNLP, NAACL, EACL, SIGIR, ECIR). He is a co-organizer of the TextGraphs workshop series on graph-based NLP. He is a research associate at the Collaborative Research Center SFB 884 "Political Economy of Reforms" where he participates in two projects.

Federico Nanni is a Post-Doctoral researcher in Political Text Analyisis at the Collaborative Research Center SFB 884 "Political Economy of Reforms" and at the Data and Web Science Group of the University of Mannheim. He obtained his Ph.D. in History of Technology from the University of Bologna. The focus of his research is on adopting (and adapting) Natural Language Processing methods for supporting studies in Computational Social Sciences and Digital Humanities. Currently, he works on developing new methods for cross-lingual topic detection and scaling in political texts. He actively works as a researcher on two projects of the Collaborative Research Center SFB 884 – Project C4: "Measuring a common space and the dynamics of reform positions: Non-standard tools, non-standard actors" and Project B6: "Nonparametric and nonlinear panel data and time series analysis".

Simone Paolo Ponzetto is Professor of Information Systems at the University of Mannheim and member of the Data and Web Science Group, where he leads the NLP and IR group. Simone obtained his Ph.D. from the Institute for Natural Lan-

guage Processing, University of Stuttgart and has spent almost 15 years of service in the ACL community, enthusiastically contributing as reviewer, area chair and tutorial presenter at various *ACL events. His main research interests lie in the areas of knowledge acquisition, text understanding, and the application of NLP methods for research in the digital humanities and computational social sciences. Simone is currently a principal investigator of the Collaborative Research Center SFB 884 "Political Economy of Reforms" where he is a co-PI on two projects (Project C4: "Measuring a common space and the dynamics of reform positions: Non-standard tools, non-standard actors'; and Project B6: "Nonparametric and nonlinear panel data and time series analysis").

6 Target audience / prerequisites

This tutorial is designed for students and researchers in Computer Science and Natural Language Processing. We assume only a basic, graduate-level understanding of NLP problems and machine learning techniques for NLP, as commonly possessed by the typical ACL event attendee. No prior knowledge of computational social science or political science is assumed.

Prerequisites

- *Math:* Basic knowledge of linear algebra, graph theory, and numeric optimization.
- *Linguistics:* None.
- *Machine Learning:* The tutorial will not go into the basics of underlying machine learning models. Knowledge of basic (supervised) machine learning concepts is required.

7 Recommended reading list

1. Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The Promise and Pitfalls of Automatic Content Analysis Methods for political texts.Political Analysis, 21(3): 267–297.
2. Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting Policy Positions from Political Texts Using Words as Data. American Political Science Review, 97(02): 311–331.
3. Jonathan B. Slapin 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.

8 Other Information

Tutorial type: Introductory.

Tutorial materials: All tutorial materials and other information related to the tutorial are available at: https://poltexttutorial.wordpress.com

Acknowledgments

The work on this tutorial is supported by the German Science Foundation (DFG) within the scope of activities of the Collaborative Research Center SFB 884: "Political Economy of Reforms".

References

Quinn Albaugh, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. The automated coding of policy agendas: A dictionary-based approach. In *Proc. of CAP Conf*.

Nicole Baerg, Dominik Duell, and Will Lowe. 2018. Central bank communication as public opinion: Experimental evidence. Work in Progress.

Ryan Bakker, Catherine De Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2015. Measuring party positions in europe: The Chapel Hill expert survey trend file, 1999–2010. *Party Politics* 21(1).

Frank R Baumgartner, Christoffer Green-Pedersen, and Bryan D Jones. 2006. Comparative studies of policy agendas. *Journal of European Public Policy* 13(7).

Kenneth Benoit, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review* 110(2).

Shaun Bevan. 2019. Gone fishing: The creation of the comparative agendas project master codebook. In Frank Baumgartner, Christian Breunig, and Emiliano Grossman, editors, *Comparative Policy Agendas: Theory, Tools, Data*, Oxford University Press.

James E Campbell. 1983. Ambiguity in the issue positions of presidential candidates: A causal analysis. *American Journal of Political Science* 27(2).

Loren Collingwood and John Wilkerson. 2012. Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics* 9(3).

Holger Döring and Sven Regel. 2019. Party facts: A database of political parties worldwide. *Party Politics* 25(2).

Jason Eichorst and Nick Lin. 2018. Resist to commit: Concrete campaign statements and the need to clarify a partisan reputation. *The Journal of Politics* 81(1).

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proc. of *SEM*.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017a. Cross-lingual classification of topics in political texts. In *Proc. of NLP+CSS*.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017b. Unsupervised Cross-Lingual Scaling of Political Texts. In *Proc. of EACL*.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from debatepedia. In *Proc. of EMNLP*.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* 18(1).

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3).

Dustin Hillard, Stephen Purpura, and John Wilkerson. 2008. Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics* 4(4).

Graeme Hirst, Yaroslav Riabinin, and Jory Graham. 2010. Party status as a confound in the automatic classification of political speech by ideology. In *Proc. of JADT*.

Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1).

Mladen Karan, Jan Šnajder, Daniela Sirinic, and Goran Glavaš. 2016. Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *Proc. of SIGHUM*.

Gary King and Will Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization* 57(3).

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97(02).

Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling Policy Preferences from Coded Political Texts. *Legislative Studies Quarterly* 36(1).

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in us electoral manifestos. In *Proc. of EMNLP*.

Nicolas Merz, Sven Regel, and Jirka Lewandowski. 2016. The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics* 3(2).

Slava Mikhaylov, Michael Laver, and Kenneth R. Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis* 20(01).

William Lockley Miller. 1990. *How Voters Change: the 1987 British election campaign in perspective*. Oxford University Press.

Federico Nanni, Goran Glavaš, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2019. Political text scaling meets computational semantics. Work in Progress.

Federico Nanni, Goran Glavaš, Simone Paolo Ponzetto, Sara Tonelli, Nicolò Conti, Ahmet Aker, Alessio Palmero Aprosio, Arnim Bleier, Benedetta Carlotti, Theresa Gessler, Tim Henrichsen, Dirk Hovy, Christian Kahmann, Mladen Karan, Akitaka Matsuo, Stefano Menini, Dong Nguyen, Andreas Niekler, Lisa Posch, Federico Vegetti, Zeerak Waseem, Tanya Whyte, and Nikoleta Yordanova. 2018. Findings from the hackathon on understanding euroscepticism through the lens of textual data. In *Proc. of ParlaCLARIN*.

Federico Nanni, Cäcilia Zirn, Goran Glavaš, Jason Eichorst, and Simone Paolo Ponzetto. 2016. TopFish: topic-based analysis of political position in US electoral campaigns. In *Proc. of PolText*.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of ICWSM*.

Benjamin Page. 1976. The theory of political ambiguity. *American Political Science Review* 70(3).

John Petrocik. 1996. Issue ownership in presidential elections, with a 1980 case study. *American journal of political science* 40(3).

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2006. An automated method of topic-coding legislative speech over time with application to the 105th-108th us senate. In *Proc of. APSA*.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1).

Margaret E. Roberts, 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. *American Journal of Political Science* 58(4).

Gijs Schumacher, Martijn Schoonvelde, Denise Traber, Tanushree Dahiya, and Erik De Vries. 2016. Eu-speech: a new dataset of eu elite speeches. In *Proc. of PolText*.

Holli A Semetko and Patti M Valkenburg. 2000. Framing european politics: A content analysis of press and television news. *Journal of communication* 50(2).

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3).

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. In *Proc. of NAACL*.

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin, and Julian Brooke. 2017. Joint sentence-document model for manifesto text analysis. In *Proc. of ALTA*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proc. of EMNLP*.

Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the european parliament as linked open data. *Semantic Web* 8(2).

Annika Werner, Onawa Lacewell, and Andrea Volkens. 2014. *Manifesto Coding Instructions: 5th fully revised edition*. Manifesto Project.

David Winter and Abigail Stewart. 1977. Content analysis as a technique for assessing political leaders. In *A psychological examination of political leaders*, New York: Free Press.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos. In *Proc. of PolText*.