1 Introduction

The rise in popularity of social media has changed various aspects of private, public, and professional life over the last two decades. From a data-analytical point of view, this has led to an unprecedented increase in the supply of publicly available unstructured (text) data, ready to be analyzed. In fact, unstructured data makes up the lion's share of what is called *big data* (Gandomi and Haider, 2015). At the same time, advances in the field of machine learning, particularly in *Natural Language Processing* (NLP), have created numerous new opportunities for the analysis of such large-scale unstructured texts.

A field which has been particularly impacted by the use of social media (and the information extracted from it) is politics. At least since the 2016 Brexit vote and US presidential election, politicians have come to recognize not only that social media presence is ever more important, but also how strong a message their social media behavior can transmit. Among social media networks, Twitter is of particular importance, since it allows for direct communication between politicians and voters - and even more so after the Facebook-Cambridge Analytica data breach in 2018. As a consequence, there has been increasing academic interest in text-based (intra- and inter-)party politics (e.g., Ceron, 2017; Daniel et al., 2019; Grimmer, 2010; Quinlan et al., 2018). Moreover, unstructured text and the insights generated from it can subsequently be used as input for a broad variety of tasks, ranging from election forecasts (e.g., Burnap et al., 2016; Jungherr, 2016; Tumasjan et al., 2010) to prediction of stock market movements (e.g., Nisar and Yeung, 2018).

The key challenge in analyzing large amounts of unstructured text is to reduce dimensionality and classify pieces of text: either into previously determined categories (for instance, sentiments), which corresponds to a supervised learning problem; or by trying to discover latent thematic clusters that govern the content of the documents, which is now an instance of unsupervised learning (since the number and labeling of clusters is to be determined). In this paper, we pursue the second strategy, usually referred to as *topic modeling*, and apply it to German politics. In particular, we construct a dataset where the text documents consist of Twitter messages by German Members of Parliament (MPs) and which furthermore contains a plenitude of personal MP-level data as well as socioeconomic data on an electoral-district level. Subsequently, we fit a topic model to the data to discover latent topics and analyze their relationship with document-level metadata. Due to the difficulties regarding causal inference within (latent variable-based) topic models, the analysis presented in this paper is mostly explorative/descriptive with a focus on statistical and methodological soundness instead of specific (politological) hypothesis testing.

[Short summary of key findings (TBD)]

The remainder of this paper is organized as follows. Section 2 provides the theoretical foundation of topic modeling, in particular the "component models" of the *Structural Topic Model* which we use for the major part of our analysis, as well as a brief discussion on inference and parameter estimation. Section 3 describes the data collection process, the data itself, and the data preprocessing necessary for topic modeling. Section 4 presents the results, including a discussion of inference strategies and an alternative modeling procedure. Finally, section 5 concludes.

References

Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233, 2016.

Andrea Ceron. Intra-party politics in 140 characters. *Party politics*, 23(1):7–17, 2017.

William T Daniel, Lukas Obholzer, and Steffen Hurka. Static and dynamic incentives for twitter usage in the european parliament. *Party Politics*, 25(6):771–781, 2019.

Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, 35(2):137–144, 2015.

Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, 13(1):72–91, 2016.

Tahir M Nisar and Man Yeung. Twitter as a tool for forecasting stock market movements: A short-window event study. *The journal of finance and data science*, 4(2):101–119, 2018.

Stephen Quinlan, Tobias Gummer, Joss Roßmann, and Christof Wolf. 'show me the money and the party!'–variation in facebook and twitter adoption by politicians. *Information, communication & society*, 21(8):1031–1049, 2018.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.