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 *Structural Topic Model* (STM) 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. Altogether, thus, the contribution of this paper is threefold: first, a broad and widely applicable dataset for future research, particularly in political science; second, a topic analysis of German parliamentarians' Twitter communication; and third, critical discussion of existing and development of new tools for (causal) inference within a topic modeling context.

We find that for most model specifications the majority of topics carry meaning, which can be regarded as a form of retrospective model validation. The fact that these topics are converted from mere word clusters into actually meaningful thematic clusters through manual labeling underlines the importance of human judgment in statistical topic modeling - this is in line with Chang et al. (2009), who show that solely focusing on quantitative metrics such as held-out likelihood does not guarantee meaningfulness of the latent space. As for document-level metadata, we discover some relevant associations between topic proportions and document features; particularly for the political parties, these relationships are in line with expectation. For continuous covariates such as unemployment and GDP the high degree of uncertainty induced by the underlying generative process of the STM renders relationships insignificant - though the observed tendencies are consistent across all modeling methodologies. The inclusion of a covariate to further model topical content (beyond its effect on topical prevalence) is found to reduce the meaningfulness of the latent space; furthermore, no natural candidate for the topical content variable exists in our case. Finally, we find that double usage of (prevalence) covariate information does not pose a problem, while double usage of document-level associations induces a substantial degree of overfitting.

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 STM which we use for the major part of our analysis, as well as a brief discussion of inference and parameter estimation. Section 3 describes the data collection process, the data itself, and the data preprocessing necessary for topic modeling. Section 4 discusses model selection, labeling as well as global characteristics of the latent space. In section 5, we include document-level metadata into the analysis, presenting the corresponding theory and results. Section 6 deals with alternative modeling approaches and inference strategies. Finally, section 7 concludes.

Topic 1 Top Words: **Highest Prob:** buerg, link, merkel, frau, sich **FREX:** altpartei, islam, linksextremist, asylbewerb, linksextrem **Lift:** eitan, 22jaehrig, abdelsamad, abgehalftert, afdforder **Score:** altpartei, linksextremist, frauenkongress, islamist, boehring
Topic 2 Top Words: **Highest Prob:** frag, einfach, find, genau, halt **FREX:** geles, tweet, sorry, quatsch, lustig **Lift:** baseball, demjen, duitsland, garn, haeh **Score:** schmunzel, tweet, fuerstenberg, sorry, geles
Topic 3 Top Words: **Highest Prob:** brauch, wichtig, leid, dank, klar

FREX: emissionshandel, soli, marktwirtschaft, feedback, co2steu **Lift:** aequivalenz, altersvorsorgeprodukt, bildungsqualitaet, co2limit, co2meng **Score:** emissionshandel, co2limit, basisrent, euet, technologieoff
Topic 4 Top Words: **Highest Prob:** sozial, miet, kind, arbeit, brauch **FREX:** mindestlohn, miet, wohnungsbau, mieterinn, loehn **Lift:** auseinanderfaellt, baugipfel, bestandsmiet, billigflieg, binnennachfrag **Score:** miet, mieterinn, mietendeckel, grundsicher, bezahlbar
Topic 5 Top Words: **Highest Prob:** digital, jung, duesseldorf, bildung, christian **FREX:** fdpbundestagsabgeordnet, duesseldorf, rimkus, intelligenz, startups **Lift:** boeing, dettenheim, duesseldorfbilk, eheim, elektrokleinstfahrzeug **Score:** fdpbundestagsabgeordnet, rimkus, digital, duesseldorf, uranfabr
Topic 6 Top Words: **Highest Prob:** gruen, klimaschutz, brauch, klar, euro **FREX:** fossil, erneuerbar, kohleausstieg, verkehrsminist, verkehrsw **Lift:** abgasbetrug, abgebaggert, abschalteinricht, abschaltet, ammoniak **Score:** erneuerbar, fossil, zdebel, verkehrsminist, klimaschutz
Topic 7 Top Words: **Highest Prob:** europaeisch, wichtig, europa, international, thank **FREX:** foreign, policy, clos, clear, important **Lift:** alam, bucerius, bulgaria, doping, judgment **Score:** need, important, great, foreign, today
Topic 8 Top Words: **Highest Prob:** kris, wichtig, brauch, kind, hilf **FREX:** corona, coronakris, virus, pandemi, coronavirus **Lift:** covid19, schutzmask, 600milliardenfond, abiturpruef, abstandhalt **Score:** corona, coronakris, pandemi, coronavirus, virus
Topic 9 Top Words: **Highest Prob:** krieg, link, europaeisch, regier, international **FREX:** milita, voelkerrechtswidr, aufruest, waffenexport, libysch **Lift:** katalan, abho, airbas, antimilitarist, aufklaerungsdat **Score:** voelkerrechtswidr, libysch, milita, iran, voelkerrecht
Topic 10 Top Words: **Highest Prob:** herzlich, glueckwunsch, wichtig, freu, gespraech **FREX:** gmuend, achim, backnang, sommertour, schwaebisch **Lift:** 24stundendien, abschlussfoto, absolventinn, abstandskriteri, afrikastrategi

Score: backnang, gmuend, achim, bentheim, sauerla
Topic 11 Top Words: **Highest Prob:** pfleg, versorg, wichtig, chemnitz, patient **FREX:** mention, neuwied, automatically, unfollowed, checked **Lift:** mention, unfollowed, alicia, alois.karl, altenkirch **Score:** mention, unfollowed, reach, automatically, windhag
Topic 12 Top Words: **Highest Prob:** frau, gruen, frag, antrag, debatt **FREX:** bielefeld, innenausschuss, streichung, selbstbestimm, bundesinnenminist **Lift:** abstammungsrecht, altruist, atrium, bundesgeschaeftsstell, cannabispolit **Score:** bielefeld, innenausschuss, u.spd, lobbyistengab, amri
Topic 13 Top Words: **Highest Prob:** berlin, schoen, dank, freu, woch **FREX:** buongiorno, moin, frank, kiel, leipzig **Lift:** altlandsberg, anrath, bergenenkheim, blindenleitsyst, bueromitarbeit **Score:** buongiorno, moin, schoen, neers, berlin
Topic 14 Top Words: **Highest Prob:** partei, link, demokrat, klar, wahl **FREX:** thuering, hoeck, faschist, neuwahl, kemmerich **Lift:** epost, gezittert, oktoberrevolution, parteischaed, uebergangsmp **Score:** faschist, kemmerich, thuering, ramelow, hoeck
Topic 15 Top Words: **Highest Prob:** dank, glueckwunsch, herzlich, gemeinsam, europa **FREX:** zusammenhalt, antisemitismus, lasst, hass, vielfalt **Lift:** 40jahr, afdtyp, dierk, fruendt, mutmacherinn **Score:** dank, hass, zusammenhalt, binding, antisemitismus

Table 1: Top words for all topics.

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.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

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.