Ludwig-Maximilians-Universität München

Statistical Consulting Project

E-Complaints of Citizens and Responsiveness Strategies of Political Candidates and Parties: Investigating Large-Scale Unstructured Text

Comparison of homepages of national and local political entities

Simon Wiegrebe, Patrick Schulze

Abstract

TBD

supervised by
Prof. Dr. Christian Heumann and Prof. Dr. Paul W. Thurner

June 29, 2020

Contents

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.

2 Theoretical Framework

2.1 Topic Modeling - Overview

Topic models seek to discover latent thematic clusters, called topics, within a collection of discrete data, usually text; therefore, topic modeling can be regarded as dimensionality reduction technique. Furthermore, since both the number and content of topics is unknown beforehand (and can never be truly verified), topic modeling is an instance of unsupervised learning. Information retrieval (IR) research generally proposes the reduction of text documents to vectors of real numbers, each number representing (modified) counts of "words" or "terms". An instance of this proposed methodology is the *tf-idf* scheme by Salton and McGill (1983), which for a collection of documents returns a term-by-document matrix where each row corresponds to a document in the corpus and the columns contain the respective *tf-idf* term count. Since only words in a vocabulary of fixed length V are considered, documents of unrestricted length are being reduced to vectors of a fixed length V. To further reduce dimensionality, the *latent semantic indexing* (LSI) by Deerwester et al. (1990) applied singular value decomposition (SVD) to the *tf-idf* document-term matrix. However, as Blei et al. (2003) put it, the idea should be to develop a generative probabilistic model of text, in order to estimate to which extent the LSI methodology can align data with the generative text model; yet, given such a model, "it is not clear why one should adopt the LSI methodology - one can attempt to proceed more directly, fitting the model to data using maximum likelihood or Bayesian methods" (p. 994). Picking up this shortcoming of LSI, Hofmann (1999) introduced the *probabilistic LSI* (pLSI) model. This generative data model allows for individual words to be sampled from a mixture model: they are drawn from a multinomial distribution, with latent random variables determining the mixture proportions, which in turn can be viewed as topics. However, the pLSI can only be regarded as partly probabilistic text model, since the mixing components themselves are fixed on a document level, thus lacking a probabilistic generating process.

In their *Latent Dirichlet Allocation* (LDA) model, Blei et al. (2003) included the generation of topic proportions into the generative probabilistic model, the resulting 3-level hierarchical Bayesian mixture model marking the starting point of modern topic modeling. In order to present the main idea of LDA, we first introduce some notation and terminology that we will use throughout the remainder of this paper.

- A *word* (also called *term*) is the smallest unit of discrete text data. Words are elements of a vocabulary of length V and can thus be indexed by $\{1, \dots, V\}$. Mathematically, the v-th word in the vocabulary can be represented as a vector of length V, whose v-th component equals one, with all other components equalling zero. We will sometimes refer to the v-th word of the vocabulary simply as v. Apart from document-level

covariates, words are the only random variables within the topic model that we actually observe, the rest being latent.

- A *document* $d \in \{1, \dots, D\}$ is a sequence of words of length N_d. For a given document d, we denote its words by $d = (w_{d,1}, \dots, w_{d,N_d})$. Consequently, the n-th word of document d is denoted by $w_{d,n}$.
- A *corpus* is a collection (or set) of D documents. Therefore, $d \in \{1, \dots, D\}$ means that our corpus contains D documents.
- A *topic* is a latent thematic cluster within a text corpus. The idea is that any collection of documents is made up of K such topics, where the number of topics K is an (unknown) hyperparameter which needs to be determined ex ante (see Section 4.1 for hyperparameter determination in our specific use case). We will refer to topics simply by the actual *topic index* (or *topic number*) $k \in \{1, \dots, K\}$.
- A *topic-word distribution* β is a probability distribution over words, i.e., over the vocabulary. This is what actually characterizes a topic. For a model containing K topics (and no topical content variable, see section 2.2 below), topic-word distributions do not vary across documents and uniquely characterize a topic: we denote the word distribution corresponding to the k-th topic by β_k and the matrix whose k-th column is topic β_k by $B := \beta_{1:K} = [\beta_1 | \dots | \beta_K]$. Each vector β_k thus has length V, while B is a $V \times K$-matrix. Therefore, k refers to the latent thematic cluster with topic index k in general, and β_k refers to the underlying word distribution in particular.
- A *topic assignment* $z_{d,n}$ is the assignment of the n-th word of document d to a specific topic $k \in \{1, \dots, K\}$ (i.e., to the corresponding word distribution β_k). Therefore, $z_{d,n}$ is simply a vector of length K whose k-th entry equals one and all other entries equal zero. This way, we can represent the word distribution corresponding to the n-th word in document d as $\beta_{d,n} := Bz_{d,n}$ (again, for a model without topical content variable).
- For a given document d, the corresponding *topic proportions*, denoted by θ_d, are the proportions of the document's terms assigned to each of the topics $k \in \{1, \dots, K\}$. Topic proportions vary across documents. Since for each document d the proportions of all K topics must add up to one ($\sum_{k=1}^{K} \theta_{d,k} = 1, \forall d \in \{1, \dots, D\}$), topic proportions represent probabilities.
- The *bag-of-word* assumption is an assumption used in all (probabilistic) text models referenced in this paper, including LSI and pLSI, and states that only words themselves (and their counts) carry meaning, while word order or grammar do not. Statistically, this is equivalent to assuming that words within a document are *exchangeable* (Aldous, 1985).

As mentioned above, LDA is the first generative probabilistic model of an entire text corpus. (Recall that pLSI is only probabilisitic for a fixed document.) Now, LDA can be

neatly described by the following 2-step procedure, given the hyperparameter (number of topics) K: For each document $d \in \{1, \ldots, D\}$:

1) Draw topic proportions $\theta_d \sim Dir_K(\alpha)$.
2) For each word $n \in \{1, \ldots, N_d\}$:
   a) Draw a topic assignment $z_{d,n} \sim Multinomial_K(\theta_d)$.
   b) Draw a word $w_{d,n} \sim Multinomial_V(\beta_{d,n})$.

Thus, topic proportions are drawn from a Dirichlet distribution with K-dimensional hyperparameter vector α, with all components $\alpha_k > 0$; this vector is estimated from the data. This means that for each document $d \in \{1, \ldots, D\}$, the corresponding topic proportions θ_d represent a K-dimensional vector which can take on values on the $(K-1)$-simplex: $\theta_{d,k} \geq 0, \sum_{k=1}^{K} \theta_{d,k} = 1$. Also note that the Dirichlet distribution is the conjugate prior of the multinomial distribution, which greatly facilitates estimation (see section 2.3 on variational inference below). Put simply, for each document LDA first generates topic proportions, which are then used as weights for topic assignment. Finally, each "word spot" - which now already has a topic assigned to it - is filled with a draw from the topic-specific word distribution. These topic-specific word distributions β_k need to be estimated from data. Note that LDA is a very simple, restrictive model in (at least) three ways:

(i) By using the Dirichlet distribution to generate topic proportions, potential correlations between topics cannot be captured due to the neutrality of the Dirichlet distribution.[1]. As a consequence, the occurrence of one topic within a document is not correlated with the occurrence of another topic (Blei et al., 2007). This is a restrictive simplification, as topics such as "sports" and "health" are much more likely to co-occur within a document than, say, "sports" and "war".

(ii) Second, while topic proportions vary stochastically across documents, they do so given a single, global hyperparameter vector α, essentially implying that topic proportions are generated based merely on word counts (occurrences and co-occurrences). This is another unrealistic and limiting simplification, since researchers usually possess further document-specific information indicative of the topics addressed within the individual documents.

(iii) Third, the topic-specific word distributions β_k are estimated identically for all documents, by construction. Similarly to the second restriction, this prevents researchers

[1]Due to the constraint $\sum_{k=1}^{K} \theta_k = 1$, there is clearly some degree of dependence between topic proportions. However, the dependence is minimal, as the Dirichlet distribution is characterized by complete neutrality: the components $\theta_1/(1-S_0), \theta_2/(1-S_1), \ldots, \theta_K(1-S_{K-1})$ are mutually independent, where $S_0 := 0$ and $S_k = \sum_{i=1}^{k} \theta_k, k \in \{1, \ldots, K\}$. Stated differently, for each component $\theta_k, k \in \{1, \ldots, K\}$, it holds that $\theta_k/(1-S_{k-1})$ is independent of the vector constructed by weighting all *remaining* components by their total proportion (James et al., 1980)

from using (document-level) information which might potentially influence the weighting of specific words within a topic.

Due to its simplicity and the resulting restrictions, the LDA has been used used as building block for more advanced (and usually more specified) generative topic models. One model that builds on LDA, addressing some of its shortcomings, is the *Correlated Topic Model* (CTM) by Blei et al. (2007). Specifically, the CTM addresses the first one of the above-mentioned restrictions: the inability to cope with inter-topic correlations. The model no longer uses a Dirichlet distribution to sample topic proportions; instead, a logistic normal distribution is employed, which can capture correlations between topics due to the incorporated covariance structure between its components (Atchison and Shen, 1980). The resulting generative process for the CTM can be stated as follows:

For each document $d \in \{1, \ldots, D\}$:

1) Draw unnormalized topic proportions $\eta_d \sim N_{K-1}(\mu, \Sigma)$, with $\eta_{d,k} := 0$ for model identifiability.
2) Normalize η_d by mapping it to the simplex: $\theta_{d,k} = \frac{exp(\eta_{d,k})}{\sum_{j=1}^{K} exp(\eta_{d,j})}, \forall k \in \{1, \ldots, K\}$.
3) For each word $n \in \{1, \ldots, N_d\}$:
    a) Draw a topic assignment $z_{d,n} \sim Multinomial_K(\theta_d)$.
    b) Draw a word $w_{d,n} \sim Multinomial_V(\beta_{d,n})$.

Steps 1 and 2 constitute the sampling from a logistic normal distribution: a K-dimensional vector η_d is drawn from a multivariate normal distribution and subsequently transformed to a vector of proportions (or probabilities) by applying the *softmax* function to each of its elements. The number of topics K is again a hyperparameters which must be determined ex ante. As in LDA, the parameters of the normal distribution in step 1, $\mu \in \mathbb{R}^{K-1}$ and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$, as well as the topic-specific word distributions β_k need to be estimated from the data. As mentioned above, this process now allows for inter-topic correlation. Yet this comes at a cost: unlike the Dirichlet distribution, the logistic normal distribution is no longer conjugate to the multinomial distribution. As explained in more detail in section 2.3 below, this renders standard variational inference algorithms inapplicable, since these rely on conjugacy and the implied closed-form solutions. However, using the Laplace variational inference developed by Wang and Blei (2013), which is a generic method for variational inference when dealing with nonconjugate models, solves the inference problem for the CTM.

As for the inability to integrate covariate information into the determination of topic proportions, Mimno et al. (2011) were the first to model topic proportions as a function of *observable* document-level metadata. Specifically, their *Dirichlet-Multinomial Regression* (DMR) model still samples topic proportions θ_d from a Dirichlet distribution (thus, not allowing for inter-topic correlations), yet unlike in LDA, the Dirichlet prior α_d is no longer

global but topic-specific. This topic prior α_d, in turn, is log-linear in the document-level features $\mathbf{x_d}$ and the (topic-specific) priors for the coefficients of these features, $\lambda_\mathbf{t}$, have a normal prior. With coefficients being updated through numerical optimization as part of the EM algorithm used for training, the DMR model thus actively uses document features to model topic proportions.

Finally, the third restictiveness of LDA, the inflexibility of the topic-word distributions β_k when document-level metadata is available, is addressed by Eisenstein et al. (2011) in their *Sparse Additive General* model (SAGE). The authors propose to start off with a background word distribution m containing log frequencies and to model additive deviations from this baseline for each class. The idea behind SAGE can be used to model differences in topic-word distributions according to the category of some document-level covariate.

Based on the foundational LDA as well as its extensions, Roberts et al. (2013) developed the *Structural Topic Model* (STM), which combines the improvements over the original LDA discussed in this section. Due to its flexibility regarding the incorporation of document-level information, we choose the STM for our specific use case, a text-based analysis of German political entities (TBD, depends on final title of paper). Therefore, we discuss the model in greater detail in section 2.2 below.

2.2 The Structural Topic Model

Overview

The STM addresses the three main shortcomings of the LDA, as discussed in the previous section. In this subsection, we explain the corresponding modifications with respect to LDA and present the generative process of the STM.

(i) To allow for correlation among topics, the STM uses a logistic normal distribution to sample topic proportions. In fact, if no document-level metadata is fed into the STM, it simply reduces to the CTM.

(ii) The STM allows for the incorporation and use of document-level metadata when determining topic proportions. Similar to the DMR, topic proportions $(\theta_1, \ldots, \theta_D)^T$ are assumed to depend on P document-level *topical prevalence variables* (such as the author's name, her political party or her popularity on Twitter), yet now by following a multivariate logistic normal distribution with mean vector $X_d\Gamma$, where $X \in \mathbb{R}^{D\times P}$ and $\Gamma = [\gamma_1|\ldots|\gamma_k]$, and covariance matrix Σ. This way, the model accounts for the fact that document-level covariates might influence how much (that is, which percentage of the total number of words) the corresponding documents attribute to the different topics.

(iii) Within the STM, document-level covariate information can also be used to fine-tune the topic-word distributions β_k, the methodology being similar to the one in the SAGE

model. In particular, the STM allows for specifying a single categorical document-level *topical content variable* Y with A levels: $Y_d \in \{1, \ldots, A\}, \forall d \in \{1, \ldots, D\}$ (Roberts et al., 2019). Consequently, each topic $k \in \{1, \ldots, K\}$ is now associated with a total of A topic-word distributions $\beta_{k,a}, a \in \{1, \ldots, A\}$ instead of a single one, β_k. For a given document d, this means the K topic-word distributions β_k are determined according to the level a assumed by Y_d and are identical across all documents with $Y_d = a$ (Roberts et al., 2016). This way, for a given document d, document-level metadata can not only impact the weighting of topics θ_d, but also the weighting over words for each topic β_k. Note that for a given topic k, the word distributions $\beta_{k,a}$ are similar to each other for all values of a; that is, the content variable Y is really an A-level refinement of β_k and does *not* affect the number of topics K.

The generative process of the STM can be stated as follows (Roberts et al., 2016):

For each document $d \in \{1, \ldots, D\}$:

1) Draw unnormalized topic proportions $\eta_d \sim N_{K-1}(X_d\Gamma, \Sigma)$, with $\eta_{d,k}$ set to zero for model identifiability.
2) Normalize η_d by mapping it to the simplex: $\theta_{d,k} = \frac{exp(\eta_{d,k})}{\sum_{j=1}^{K} exp(\eta_{d,j})}, \forall k \in \{1, \ldots, K\}$.
3) For each word $n \in \{1, \ldots, N_d\}$:
    a) Draw a topic assignment $z_{d,n} \sim Multinomial_K(\theta_d)$.
    b) If no topical content variable has been specified, simply draw a word $w_{d,n} \sim Multinomial_V(\beta_{d,n})$. Otherwise, first determine the document-specific word distributions $\beta_{k,a}$ based on the level a taken on by Y_d, for all topics $k \in \{1, \ldots, K\}$: $B_a := [\beta_{1,a}| \ldots |\beta_{K,a}]$; next, analogously define $\beta_{d,n} := B_a z_{d,n}$; finally, draw a word $w_{d,n} \sim Multinomial_V(\beta_{d,n})$.

This means that unnormalized topic proportions are sampled from a normal distribution with mean $\Gamma = [\gamma_1| \ldots |\gamma_k]$ and covariance Σ. Γ is the vector of coefficients corresponding to the topical prevalence covariates contained in X, with prior distributions $\gamma_k \sim N_p(0, \sigma_k^2 I_p)$. The unnormalized topic proportions η_d are then "sent through the softmax function" to yield normalized topic proportions θ_d, which in turn are used as weights for the subsequent topic assignment $z_{d,n}$. Finally, each word is sampled from the corresponding multinomial word probability distribution (over the vocabulary of length V), which depends on topic assignment $z_{d,n}$ and, for models containing a topical content variable, on its level a. In line with SAGE methodology, the topic-word distributions are modelled as deviations in log-frequency from a baseline vocabulary. (See Roberts et al. (2016), p. 991 for further details.) K and σ_k^2 are hyperparameters. The graphical model representation in Figure 1 below visualizes the generative process described.

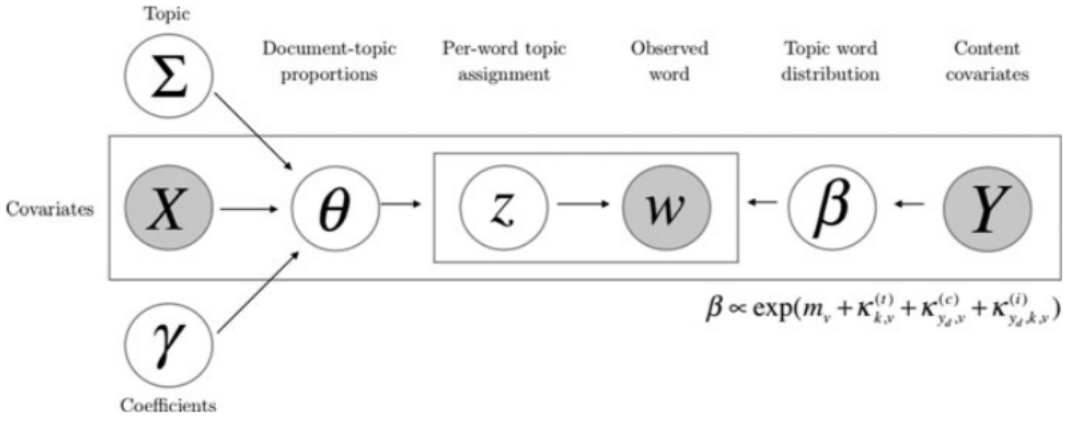

Figure 1: Graphical model representation of the STM (from Roberts et al. (2016), p. 990).

Scope

Topic models are unsupervised learning methods, since the true topics from which the text was generated are not known. Thus, topic models have been traditionally used as an exploratory tool providing a concise summary of topics, with the posterior ideally inducing a good decomposition of the corpus. Topic models have also been used for tasks such as collaborative filtering and classification (see e.g. (Blei et al., 2003). In particular, they can be used as dimensionality reduction method in semi-supervised learning methods. Such a process can in general be described as a two-stage approach, where in the first stage topic proportions and content are learned, and in the second stage a supervised method such as regression takes this learned representation as input.

The fundamental idea of the STM is to combine these two steps: Topics and their association with covariates are estimated jointly. For instance, the estimated effect of topical prevalence covariates X_d on topic proportions θ_d is reflected in the estimate of Γ. However, since the topic proportions are latent random variables, it is preferable to incorporate the uncertainty of θ_d, accesible through the estimated approximation of the posterior $p(\theta_d|\Gamma, \Sigma, X)$, when determinig the effect of covariates on topic proportions. This is achieved by what is called the "method of composition" in social sciences: By sampling from the approximate posterior for θ and subsequently regressing these topic proportions on X, it is possible to integrate out the topic proportions (since these are latent variables) and obtain an i.i.d. sample from the marginal posterior of the regression coefficients for the topical prevalence covariates (see Section 4.4).

A problem we see with this approach, however, is that the same covariates - and in general the same data - used to infer the topical structure are subsequently used to determine effects of the former on the latter (or vice versa). This problem has recently also been adressed by Egami et al. (2018). In our specific case, we find that the prevalence covariates do

not have much impact on the estimated topic proportions due to the regularizing priors for Γ (see Section 4.6). Thus, the regression coefficients (with topic proportions as the dependent variable) should not be largely affected by this problem of double usage. However, this begs the question why document-level covariates are being used to obtain the topical structure in the first place. In an empirical evaluation, Roberts et al. (2016) showed that the STM consistently outperforms other topic models, such as LDA, when comparing the respective heldout likelihoods in different settings. This indicates that the STM performs better at predicting the topical structure by incorporating covariates, regardless of their concrete specification.

Nevertheless, in each case it should be investigated whether the relationship of variables implied by the STM is valid. In line with Egami et al. (2018), we address this issue in section 4.7, where we split our data into a training and a test set. Similar topical structures on both datasets (as we find in our case) indicate that misspecification of topical prevalence or content variables is not a concern. However, since the topical prevalence covariates have almost no influence on the estimated topic proportions on the training set due to the regularizing priors (and likewise on the heldout likelihood that can be used for validation), it is practically impossible to validate a good prevalence specification.

Posterior Distribution

In this subsection, we briefly derive the posterior distribution of the STM (up to proportionality), as stated on p. 992 of Roberts et al. (2016). Recall that only words w, prevalence covariates X, and the content covariate Y are observable, while all other variables - unnormalized topic proportions η, topic assignments z, topic-word distribution deviations κ, prevalence coefficients Γ, and unnormalized topic proportion variance Σ - are latent.

$$
\begin{aligned}
p(\eta, z, \kappa, \Gamma, \Sigma | w, X, Y) &\propto \underbrace{p(w|\eta, z, \kappa, \Gamma, \Sigma, X, Y)}_{=p(w|z,\kappa,Y)} p(\eta, z, \kappa, \Gamma, \Sigma | X, Y) \\
&\propto p(w|z, \kappa, Y) p(z|\eta) p(\eta|\Gamma, \Sigma, X) \prod p(\kappa) \prod p(\Gamma) p(\Sigma) \\
&\propto \Big\{ \prod_{d=1}^{D} p(\eta_d|\Gamma, \Sigma, X_d) \Big(\prod_{n=1}^{N} p(w_n|\beta_{d,n}) p(z_{d,n}|\theta_d) \Big) \Big\} \prod p(\kappa) \prod p(\Gamma) p(\Sigma) \\
&\propto \Big\{ \prod_{d=1}^{D} \text{Normal}(\eta_d|X_d\Gamma, \Sigma) \Big(\prod_{n=1}^{N} \text{Multinomial}(z_{n,d}|\theta_d) \\
&\quad \times \text{Multinomial}(w_n|\beta_{d,n}) \Big) \Big\} \times \prod p(\kappa) \prod p(\Gamma) p(\Sigma),
\end{aligned}
$$

where $\beta_{d,n} \in \mathbb{R}^V$ is the topic-word distribution for word n in document d, which has been assigned to topic k through $z_{d,n}$. The topic-word distribution vectors $\beta_{k,a}$ have entries $\beta_{k,a,v} \propto \exp\Big(m_v + \kappa_{k,v}^{(t)} + \kappa_{a,v}^{(c)} + \kappa_{k,a,v}^{(i)}\Big)$, $v \in \{1, \ldots, V\}$, where $\kappa_{k,v}^{(t)}$, $\kappa_{a,v}^{(c)}$, and $\kappa_{k,a,v}^{(i)}$ are the

log-transformed rate deviations of word v for topic k, for content variable level a, and for the interaction of k and a, respectively.

2.3 Inference and Parameter Estimation

In this section, we describe how inference and parameter estimation for topic models, in particular for the STM, are performed. Inference is done using variational inference, and a variational Expectation-Maximization (EM) algorithm is used for empirical parameter estimation. As a detailed discussion of the underlying workings is outside the scope of this paper, we refer the reader to the appendix and the referenced papers.

Since the STM, as well as all models it builds on, are (hierarchical) Bayesian models, the central challenge we face is the exact determination of the posterior distribution. Recall that in the section above, we derived the posterior *up to proportionality*, neglecting the division by marginal distributions. The exact posterior distribution is intractable to compute due to the (high-dimensional) marginal distributions in the denominator, which is why exact inference is infeasible and variational inference is used instead. Generally, for a model with latent variables θ and z and observable data x, variational inference involves approximating the posterior $p(\theta, z|x)$ by postulating a simple distribution family for the (joint) distribution of latent model variables θ and z - $q(\theta, z)$ - and subsequently determining the member of this family which minimizes the "distance" to the true posterior distribution, measured using the Kullback-Leibler (KL) divergence (Wang and Blei, 2013). The approximations of variational inference bring a great amount of flexibility, but come at the cost of some bias, since the approximative distribution family usually does not contain the true posterior.

In the appendix, we show that minimizing KL divergence between true posterior p and the approximating variational distribution q is equivalent to maximizing a lower bound on $log(p(x))$, the log-likelihood of the observed data x. This lower bound is called *ELBO* and is defined as

$$ELBO := \mathbb{E}_q[\log\big(p(\theta, z, x)\big)] - \mathbb{E}_q[\log\big(q(\theta, z)\big)],$$

whose second component, $\mathbb{E}_q[\log\big(q(\theta, z)\big)$, is the entropy of the approximate distribution q. To be precise, maximizing *ELBO* (or minimizing KL divergence) refers to finding the governing parameter of the approximating distribution q which maximizes *ELBO*.

The optimality conditions resulting from maximizing *ELBO* lead to the *coordinate ascent algorithm* for variational inference (Wang and Blei, 2013), which converges towards a local optimum (Bishop, 2006). However, this algorithm only works for *conditionally conjugate* models, such as the LDA: all nodes in this model - in particular, the Dirichlet distribution for drawing topic proportions, the multinomial distribution for assigning topics, and the

multinomial for eventually picking words - are conditionally conjugate. The STM, however, as well as the CTM before it, are non-conjugate models due to the logistic normal distribution used to sample topic proportions, which is why algorithm updates are not feasible and the algorithm is not (direclty) applicable. As a remedy, Wang and Blei (2013) developed Laplace variational inference, which uses Laplace approximations within coordinate ascent algorithm updates and this way enables its use for the broader class of nonconjugate models, in particular for CTM and STM.

As stated above, the STM uses an Expectation-Maximization (EM) algorithm for empirical parameter estimation. In the E-step, the variational posterior distributions for topic proportions, $q(\theta_d)$, and for topic assignment, $q(z_{d,n})$, are updated using Laplace variational inference and coordinate ascent. In the M-step, the model parameters - specifically topical prevalence and content coefficients - are updated by maximizing $ELBO$ with respect to them (Roberts et al., 2016).

3 Data

3.1 Data Collection

The current political landscape of Germany consists of six parties: the right-wing *AfD*, the Greens (*Bündnis 90/Die Grünen*), the Christian Democrats (*CDU/CSU*), the Left Party (*Die Linke*), the liberal *FDP*, and the Social Democrats (*SPD*). These parties are represented in the German parliament (*Bundestag*) according to the votes obtained during the 2017 German federal election (*Bundestagswahl*), which took place on September 24, 2017. The legislative period amounts to 4 years, thus ending around September 2021. The parliament currently contains a total of 709 seats. The large majority of members of the German parliament (*Abgeordnete*) are assigned a single electoral district (*Wahlkreis*), the remaining ones do not have one.

In order to analyze German political entities based on text data, we constructed a broad database containing personal and Twitter data on an MP level as well as socioeconomic and election data on an electoral-district level, as detailed in the rest of this section. While parts of this database were used in the subsequent topic model analysis, it is also to be used in future text-based analyses regarding German politics. As a first step in constructing the database, we gathered personal information on all German MPs. Using Python's *Beautiful-Soup* web scraping tool as well as a selenium webdriver, we gathered data such as name, party, biographical information, electoral district, and social media accounts from the official parliament website (https://www.bundestag.de/abgeordnete) for all of the 709 members of the German parliament during its 19th election period, elected on September 24, 2017.[2] An additional source of personal MP-level information would be the MPs' personal homepages. However, after inspecting some of these personal homepages at random, we found that there is no systematic way to scrape them. Furthermore, hardly any of these websites contain any informative text data comparable to tweets or Facebook posts. As a consequence, we decided against further pursuing this potential source of information. Due to difficulties and recent restrictions when scraping Facebook data, caused in parts by the aforementioned data scandal, we also discarded Facebook as source of text data and focused solely on Twitter.

Since information on social media profiles was scarce and incomplete on the official parliament website, we additionally scraped official party homepages of all of the six political parties represented in the current parliament.[3] MPs who did not provide a Twitter account either on the official parliament website or on their party's official homepage were excluded. Using Python's *tweepy* library to access the official Twitter API, we scraped all tweets by

[2] As of March 30, 2020, the official parliament website contained information on 730 MPs. This is because MPs who resigned or passed away since the beginning of the election period are also listed on the website. These MPs were manually excluded from further analysis.

[3] The official homepage of the AfD party does not provide the Twitter profiles of their members, which is why for this party we had to manually gather the account names.

German MPs from September 24, 2017 through April 24, 2020, i.e., during a total of 31 months. The *tweepy* library offers a variety of additional features to be extracted apart from the mere tweet texts, such as the number of followers of an account, retweets, or how many times a tweet was like or retweeted. While we only use original tweets in the analysis presented in this paper, we included the most relevant additional Twitter features in our database, for use in future analyses. This initially yielded 342,542 tweets from a total of 470 members of parliament.[4]

To complement personal and Twitter data, we also gathered socioeconomic data such as GDP per capita and unemployment rate as well as 2017 election results on an electoral-district level for all of the 299 electoral districts from the official electoral website (https://www.bundeswahlleiter.de) After removing the only MP labeled as independent (*fraktionslos*) on the official electoral website as well as 19 MPs without a specific electoral district assigned to them (for matchability with socioeconomic data), the final dataset counted 450 MPs. Overall, 63% of all 709 MPs were thus included in the analysis. The corresponding total number of tweets amounted to 323,740. For those MPs without elecotoral district, electoral district-level socioeconomic variables could potentially be imputed by using state averages or values of nearby and/or similar districts. However, given that this only applies to 19 out of the remaining 450 MPs and since imputing covariates would introduce further uncertainty, we decided to exclude those MPs.

The table below shows total monthly tweet frequencies for our period of analysis, September 24, 2017 through April 24, 2020. As can be seen, tweet frequencies - though fluctuating - show an increasing trend over time, peaking at almost 20,000 in March 2020. The decrease for April 2020 can partly be explained by the fact that only the first 24 days of the month were taken into account.

Next, data was grouped and tweets were concatenated on a per-user level (thus aggregating tweets across the entire 31 months) as well as on a per-user per-month level, yielding a user-level and a monthly dataset. This means that a document represents the concatenation of *all* of a single MP's tweets for the user-level dataset, while it represents a single MP's *monthly* tweets for the monthly dataset. This also means that MP-level metadata such as personal information and socioeconomic data (through the electoral-district matching) can be used as document-level covariates. For the monthly dataset, the temporal component (year and month) constitutes an additional covariate. Since it is reasonable to assume that the importance of topics varies over time and due to resulting documents being shorter and

[4] *tweepy* restricts the total number of tweets retrievable to 3,200. For those MPs who tweeted more than 3,200 tweets during our period of analysis, the most recent 3,200 tweets were taken into account. However, this only applied to two MPs.

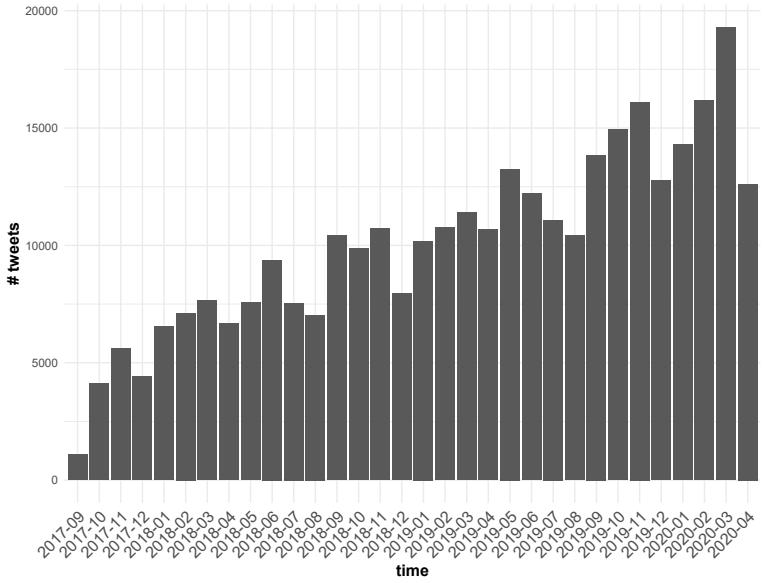

Figure 2: Monthly tweet volume by German MPs from September 24, 2017 through April 24, 2020.

more easily interpretable, we chose the monthly dataset for further analysis.[5] At this point, the data preparation was completed, marking the starting point of the preprocessing required for topic analysis, which is identical for both the user-level and the monthly dataset.

3.2 Data Preprocessing

We used the *quanteda* package within the R programming language for preprocessing. As a first step, we built a quanteda corpus from all documents. Next, we immediately transcribed German umlauts *ä/Ä*, *ö/Ö*, *ü/Ü* as well as German ligature *ß* as *ae/Ae*, *oe/Oe*, *ue/Ue*, and *ss*, respectively, and removed hyphens. Subsequently, we transformed the text data into a quanteda document-feature matrix (DFM), which essentially tokenizes texts, thereby converting all characters to lowercase. From the DFM, we removed an extensive list of German stopwords, using the stopwords-iso GitHub repository (https://github.com/stopwords-iso/stopwords-iso), as well as English stopwords included in the *quanteda* package. Moreover, hashtags, usernames, quantities and units (e.g., *10kg* or *14.15uhr*), interjections (e.g., *aaahhh* or *ufff*), terms containing non-alphanumerical characters, meaningless word stumps (e.g., *innen* from the German female plural declension, or *amp*, the remainder left after removing the ampersand sign, *&*) were removed. Terms with less than four characters and terms with a term frequency (overall number of occurrences) below five or with a document frequency

[5]For instance, as stated in section 4.2, one topic is about COVID-19, which is clearly a relatively recent topic. The monthly dataset allows for tracing the development of this topic's relevance over time: a flat curve until January 2020, followed by a sharp increase during the first months of 2020. The user-level dataset, on the other hand, would simply assign a low overall proportion to this topic.

(number of documents containing the word) below three were excluded. Finally, we manually removed overly frequent terms that would diminish the distinguishability of topics, such as *bundestag* or *polit* (see *semantic coherence* in section 4.1 for a technical explanation).

We also performed word-stemming, which means cutting off word endings to remove discrepancies arising purely from declensions or conjugations, being of particular importance for the German language. Due to the nature of the German language, the additional gains of lemmatization (which aims at identifying the base form of each word) would only be small as compared to the large increase in complexity, which is why we decided to use stemming only. Another issue when dealing with German language documents is represented by compound words, which are sometimes hyphenated, basically leading to a distinction where semantically there is none. We addressed this issue by removing hyphens in the very beginning of the preprocessing and converting all terms to lowercase, thus "gluing together" compound words; this way, terms like *Bundesregierung* and *Bundes-Regierung* are both transformed into *bundesregierung* (and, after stemming, into *bundesregier*). Finally, automatic segmentation techniques were not necesssary for the German language (Lucas et al., 2015). As a result of preprocessing, one empty MP-level document was dropped, so that a total of 10,998 (monthly) MP-level documents were eventually analyzed, each one associated with 90 covariates.

4 Results

4.1 Hyperparameter Search and Model Fitting

Throughout this section we use the *stm* package, which is implemented in the R programming language (Roberts et al., 2019). The most important hyperparameter choice when fitting an STM is the number of topics, K. While there is no *true* or *optimal* number of topics, we explore the hyperparameter space using the *searchK* function to get an understanding of the impact of K on model fit. We use four of the metrics that come with this function, *held-out likelihood*, *semantic coherence*, *exclusivity*, and *residuals*.

The *held-out likelihood* approach is based on document completion. The *searchK* function randomly holds out a proportion of some of the documents; both the number of documents from which a portion is held out and the respective held-out proportions can be specified by the user. This gives rise to a set of held-out words for which the likelihood is calculated, given the trained model. Thus, the higher this held-out likelihood, the more predictive power the model has on average. For more detailed information on held-out likelihood based on document completion and other types of held-out likelihoods, see Wallach et al. (2009).

Regarding the second metric, first introduced by Mimno et al. (2011), a model with K topics is *semantically coherent* whenever those words that characterize a specific topic k (i.e., the most frequent words within topic k) also do appear in the same documents. In order to formally define semantic coherence, let first $D(v)$ be the *document frequency* of word v (that is, the number of documents where v occurs at least once) and let $D(v, v')$ be the *co-document frequency* of words v and v' (that is, the number of documents where both v *and* v' occur at least once). Furthermore, consider the M most probable words in a given topic k. Then, semantic coherence for topic k, C_k, is defined as follows:

$$C_k = \sum_{i=2}^{M} \sum_{j=2}^{i-1} \log\left(\frac{D(v_i, v_j) + 1}{D(v_j)}\right).$$

That is, semantic coherence is the sum of (logarithmized) proportions of word co-occurrences to total word occurrences, the additive factor 1 in the numerator simply being a smoothness adjustment. It becomes apparent that by having some words that are very frequent across a couple of documents, we could achieve high semantic coherence without our topics being semantically coherent at all once we look beyond these common words (Mimno et al., 2011; Roberts et al., 2019). As a partial remedy, we previously excluded some of such overly frequent words (see section 3.2).

A natural "counter-metric" of semantic coherence is *exclusivity*, which basically tells us to which degree words within a given topic *only* occur in that topic. To formalize this, first

define the empirical frequency of word v, $v \in V$, within topic k as $\hat{\beta}_{k,v}$.[6] These empirical frequencies are then normalized across all topics $k \in \{1, ..., K\}$. This way, the normalized frequencies now represent the probability of observing topic k, conditional upon the word being v - that is, the exclusivity of word v regarding topic k. Formally, exclusivity of word v to topic k, $E_{k,v}$, is thus defined as:

$$E_{k,v} = \hat{\beta}_{k,v} / \sum_{j=1}^{K} \hat{\beta}_{j,v}.$$

Combining a word's frequency and exclusivity finally yields its Frequency-Exclusivity ($FREX$) score, explained in more detail in section 4.2 below and in Bischof and Airoldi (2012).

Finally, *residuals* is a metric based on residual dispersion. Recall that $z_{d,n}$ is drawn from a K-category multinomial distribution, which is a member of the exponential family. Therefore, its dispersion parameter is equal to one, according to theory. This way, an observed residual dispersion larger than one roughly indicates that the number of topics K was most likely chosen insufficiently small. See Taddy (2012) for a detailed derivation.

Another aspect to be taken into account when choosing K (or, to be precise, when choosing a search grid for searchK) is interpretability. While a large K certainly allows for a more fine-grained determination of topics, the resulting topics might be rather difficult to label. Furthermore, for large K we would obtain many topics which could be considered sub-topics of the topics we would obtain when using a smaller value for K. As a consequence, we select a search grid between 5 and 40, in steps of 5. Before fitting the model, we need to choose the document-level covariates we want to include. Since a topic model is explorative by definition, we simply include those covariates that seem to be most influential *a priori*: party and state (both categorical), date (as smooth effect), as well as percentage of immigrants, GDP per capita, unemployment rate, and the 2017 election results of the MP's respective party (the last four as smooth effects and on an electoral-district level). We choose degrees of freedom (df) = 5 for all smooth effects to avoid spurious wiggles due to overfitting.[7] No topical content variable is included at this stage.

The graph below shows the four metrics, as introduced above, for values of K between 5 and 40 (in steps of 5). Both 15 and 20 topics seem to be good trade-offs between the metrics used. As mentioned above, no true or optimal K exists. Taking into account the interpretability aspect, we opt for $K = 15$. For comparison, we also conducted the subsequent analysis for $K = 6$ and $K = 20$. In general, the topics generated are similar, but for $K = 6$ only around three of them are clear-cut, while for $K = 20$ some topics could easily be

[6] We use $\hat{\beta}_{k,v}$ for empirical frequencies (i.e., word counts) within topic k to distinguish them from the (normalized) word probabilities $\beta_{k,v}$.

[7] The graphical illustrations of the relationship between topic proportions and continuous covariates in sections 4.4 through 4.7 suggest that df = 5 is indeed sufficient.

grouped together. This further corroborates our choice that $K = 15$ indeed seems to be a good trade-off. Our model thus uses $K = 15$ as hyperparameter. For model fitting, we again need to choose document-level covariates. We initially select the same model specifications as in the hyperparameter search above (see sections 4.5 and 4.6 for modifications).

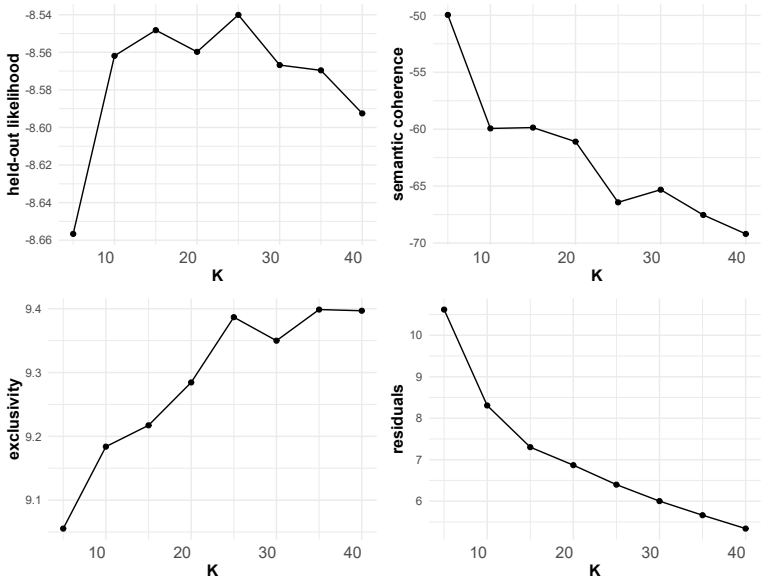

Figure 3: Model evaluation metrics for hyperparameter K (number of topics).

4.2 Labeling

As a first step after fitting the model, we would like to visually inspect the resulting topics, in particular their most representative words. However, representativeness of words for a given topic depends on the weighting metric used. The STM comes with four topic-word metrics - *highest probability*, *FREX*, *Lift*, and *Score* - which are discussed in the following.

Given a topic k, *highest probability* simply outputs those words in the topic-specific empirical word vector $\hat{\beta}_k$ with the highest corpus frequency, i.e, those with the highest absolute frequency across all documents within topic k. Using the same notation as in section 4.1 above, let $\hat{\beta}_{k,v}$ again be the empirical frequency of word v within topic k. Then the highest probability word within topic k is simply $\underset{v \in V}{\text{argmax}}\, \hat{\beta}_{k,v}$. This relatively simple measure only takes into account how often words occur in absolute terms, but not how specific those words are to the given topic. This is why we observe words like *wichtig*, *berlin*, or *frag* within the highest probability words for several topics. And since such words are very common, unspecific words, they are not particularly useful for distinguishing or labeling topics.

To also account for the degree to which a word *exclusively* belongs to a certain topic, we also consider the top words according to the *FREX* metric. It takes into account not only

how frequent but also how exclusive words are. Formally, the FREX score of word v with respect to topic k is calculated as follows:

$$FREX_{k,v} = (\frac{\omega}{ECDF(\hat{\beta}_{k,v}/\sum_{j=1}^{K}\hat{\beta}_{j,v})} + \frac{1-\omega}{ECDF(\hat{\beta}_{k,v})})^{-1} \quad = (\frac{\omega}{ECDF(E_{k,v})} + \frac{1-\omega}{ECDF(\hat{\beta}_{k,v})})^{-1},$$

where ω is the weight assigned to exclusivity (set to 0.7 by default in the STM), $E_{k,v}$ is the word's exclusivity as defined in section 4.1, and *ECDF* is the empirical CDF. Thus, for a given topic, $FREX_{k,v}$ is simply the harmonic mean of i) the rank of word v by frequency within topic k (frequency rank) and ii) the rank of topic k by the frequency of word v, across all topics $j \in \{1, ..., K\}$ (exclusivity rank). Further information on the estimation of *FREX* can be found in Roberts et al. (2019) and in Bischof and Airoldi (2012).

Lift is another topic-word metric, where the frequency of word v within topic k, $\hat{\beta}_{k,v}$, is weighted by the inverse of v's relative frequency across the entire corpus, i.e., v's empirical corpus probability. Formally:

$$Lift_{k,v} = \hat{\beta}_{k,v}/(\omega_v/\sum_{v}\omega_v),$$

where ω_v denotes the word count of word v in the entire corpus. This way, Lift gives larger weight to those words that rarely appear in other topics. Further information on Lift can be found in Taddy (2012).

Finally, the *Score* metric for word v and topic k is formally defined as:

$$Score_{k,v} = \hat{\beta}_{k,v}(\log \hat{\beta}_{k,v} - 1/K \sum_{j}^{K} \log \hat{\beta}_{j,v}).$$

Thus, Score weights word v's frequency within topic k, $\beta_{k,v}$, by the difference between v's log frequency within topic k and the average of v's log frequencies across all K topics. This can roughly be interpreted as: $\beta_{k,v}$ is weighted by the proportion of v's log frequency within topic k to v's average logarithmic frequency across all topics. For further information on the Score metric, see the R package *lda* (Chang and Chang (2010)).

To get a broad overview of which words characterize each one of the topics, the output below shows the five top words according to each of the four topic-word evaluation metrics, for three selected topics (see appendix XXX for top words of all topics).

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 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 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

A key task of topic analysis is to actually ascribe a meaning to the topics identified, i.e., to label them. While this is clearly where human judgment should and does come into play, we attempt to conduct the labeling in a more strategic (and thus less subjective) manner, following a 3-step procedure. This procedure is exemplified by topic 1. First, we consider the *words* contained in the topic, for instance by simply inspecting the top words (see output above). For a better visualization, we use a word cloud. As shown below, for a given topic (i.e., conditional upon a specific topic being chosen), it shows highest probability words weighted by their frequency. Judging at first sight, topic 1 appears to be about right-wing nationalist issues, particularly immigration.

Figure 4: Word cloud for topic 1.

Second, to get a more thorough insight into the topic, we take a look at those documents that show the highest proportion for topic 1; that is, we consider the corresponding original,

unprocessed tweets. The most representative document for topic 1 has a topic proportion θ_1 equal to 98.86%. It contains tweets from MP Martin Hess, a member of the AfD party from Baden-Württemberg, during June 2018. That is, MP Martin Hess tweeted almost exclusively about topic 1 during June 2018. The monthly document starts with:

"Ehem. Verfassungsrichter bestätigt AfD-Forderung: Zurückweisung illegaler Migranten dringend geboten. Gegenwärtige Politik widerspricht dem Verstand und auch der Verfassung. Wir müssen zurück zu Recht & Ordnung, wie die #AfD seit fast 3 Jahren fordert!"

The second most representative document for topic 1, with an almost identical $\theta_1 = 98.37\%$, is from the same MP, this time from May 2018. The document begins with:

"Mio-Überweisungen u.a. an Kanzleien unter #BAMF-Außenstellenleiterin, die mit Anwälten bandenmäßig Asylbetrug begangen haben soll. Und die Frau ist noch frei und präsentiert sich als Gutmensch. #Staatsanwaltschaft muss hier handeln und Haftgründe prüfen."

The documents exclusively focus on immigration issues, confirming the first impression gained through top words and word cloud: topic 1 concerns right-wing nationalist issues, in particular immigration. As a third step in our labeling process we finally assign a label to the topic: in this case, "Right/Nationalist". We repeat this 3-step procedure (inspecting top words and word cloud, reading through top documents, assigning a 1- or 2-word label) for all remaining topics, arriving at the following manual labels:

Topic1	Right/Nationalist
Topic2	Miscellaneous 1
Topic3	Climate Economics
Topic4	Social/Housing
Topic5	Digital/Future
Topic6	Climate Protection
Topic7	Europe
Topic8	Corona
Topic9	Left/Anti-war
Topic10	Twitter/Politics 1
Topic11	Twitter/Politics 2
Topic12	Miscellaneous 2
Topic13	Twitter/Politics 3
Topic14	Right-wing Extremism
Topic15	Society/Solidarity

Table 1: List of topic labels.

4.3 Global-level Topic Analysis

Next, we identify two ways to calculate global topic proportions (for a given topic k): either as simple (unweighted) average of $\theta_{d,k}$ across all documents (i.e., as the average of MP-level proportions across all MPs): $\frac{1}{D}\sum_{d=1}^{D}\theta_{d,k}$; or by first weighting each $\theta_{d,k}$ by the number of words in the respective documents, N_d, and then averaging across documents. The table below shows all topics with their respective global proportions for both weighting methodologies. We observe that for most topics, weighted and unweighted proportions are rather similar, but there are exceptions. In particular, the topics concerned with everyday political tweets have much higher unweighted than weighted frequencies; this makes sense, however, since such "diplomatic" tweets tend to be shorter than those discussing specific content.

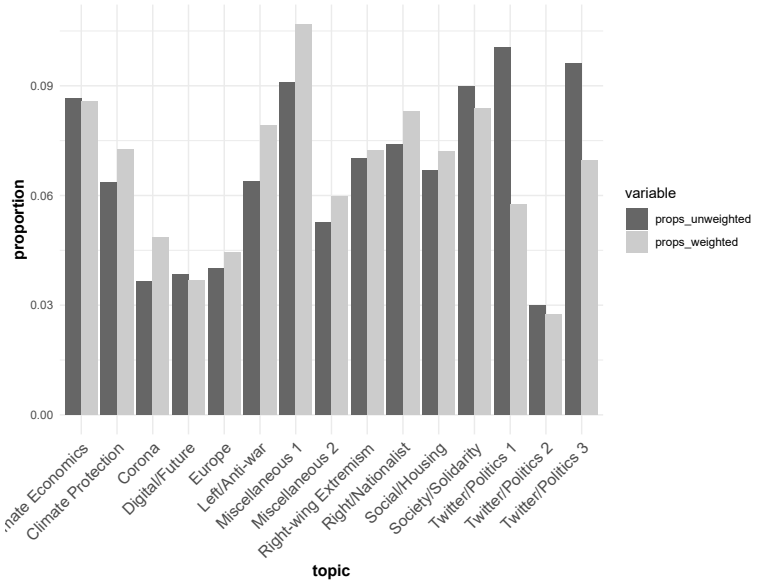

Figure 5: Weighted and unweighted global topic proportions.

While labeling tells us which words best represent each topic - and thus, what each topic truly represents - it does not yet tell us to which extent individual topics are related to each other. In the graph below, we visualize the similarity of two topics, Topic 3 (Climate Economics) and Topic 6 (Climate Protection), in terms of their vocabulary usage. As suggested by the topic labels already, there is a significant overlap in vocabulary usage.

More generally, we can evaluate the connectedness between different topics by means of a matrix of correlations between document-level topic proportions θ_d. This is visualized in Figure 7 (left panel). Most topics are negatively correlated with each other, which does not come as a surprise given the relatively low total number of topics, 15, and that topic proportions are "supplements": the higher one topic proportion, the lower the total of the others. Moreover, most topic correlations are rather weak in absolute size: the strongest negative correlation (-19.84%) is the one between topic 1 (Right/Nationalist) and topic 15

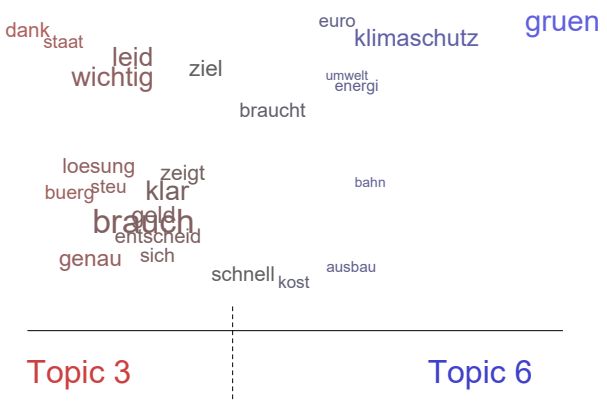

Figure 6: Comparison of vocabulary usage for two topics.

(Society/Solidarity), while the strongest positive correlation (11.79%) occurs between topic 10 (Twitter/Politics 1) and topic 13 (Twitter/Politics 3). We can also visualize these correlations using a network graph (Figure 7, right panel), where topics are connected by a dashed line whenever they are positively correlated. We observe three small clusters as well as some isolated topics, one of them being topic 8, Corona, which makes sense since this topic only entered the public sphere in early 2020, i.e., during the last months of our data collection period. In general, the relationships between the topics, as depicted below, are in line with their labeling.

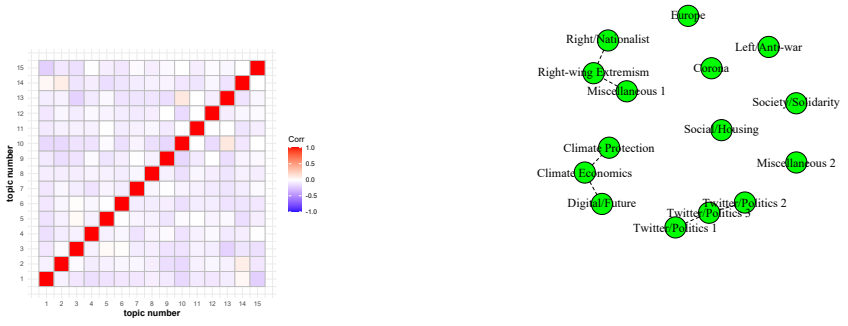

Figure 7: Global topic correlations as matrix (left) and graph (left).

5 Covariate-level Topic Analysis

We now proceed to analyze the relationship between metadata information (i.e., document-level covariates) and topic proportions. We specify topical prevalence as

$$\mu_{d,k} = x_d^T \gamma_k = \text{party}_{d,k} + \text{state}_{d,k} + f_k(\text{t}_d) + g_k(\text{struct}_d), \tag{5.1}$$

for all documents $d = 1, \ldots, D$, and for all topics $k = 1, \ldots, K$, where

$$g_k(\text{struct}_d) = g_k^{(1)}(\text{GDP}_d) + g_k^{(2)}(\text{unemployment}_d) + g_k^{(3)}(\text{immigrants}_d) + g_k^{(4)}(\text{votes}_d).$$

That is, the political party and federal state of the respective parliamentarian associated with a document are specified as simple categorial dummy effects, while date and electoral-district structural covariates (GDP per capita, unemployment rate, percentage of immigrants, and the 2017 vote share) are modeled as additive smooth functions.

Note that approximate inference implies replacing $\mu_{d,k}$ with $\lambda_{d,k}$, i.e., with the mean of the approximate Gaussian posterior $q(\eta_{d,k})$. The estimates of $\Gamma = [\gamma_1 | \ldots | \gamma_K]$ are updated in a Bayesian linear regression during each iteration of the EM algorithm in the M-step; for details see Roberts et al. (2013), p. 993.

While topical prevalence has an effect on the estimated topic proportions, the exact specification of topical prevalence is not a decisive factor. Both estimated topic proportions as well as heldout likelihood are in general only marginally affected by the concrete choice of the functional form. However, completely removing topical prevalence, in which case the model reduces to a CTM, does result in different topic proportions, as we show in section XXX. Since evaluation metrics such as heldout likelihood are mostly unaffected by the exact choice of topical prevalence and because the computational cost of fitting an stm is rather high, automatic model selection methods w.r.t. topical prevalence are not available. A reasonable specification of topical prevalence therefore relies on the domain knowledge of the researcher.

There exist different approaches to study the relationship between topic proportions and prevalence covariates. One possibility is to directly assess the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$, which are generated by the stm. Since the document-level topic proportions θ_d follow a logistic normal distribution (with mean μ_d and covariance matrix Σ), interpretation of the results can be difficult, since the logistic normal distribution is not very accessible. Nonetheless, we can visualize the relationship between a topic and a prevalence covariate, fixing other covariates at their median (for categorial variables the majority vote is used).

Alternatively, the estimated topic proportions can be used as the dependent variable of a new regression on prevalence covariates. However, in contrast to a standard regression setting, in this case the dependent variable has been estimated itself, before the regression is performed. Instead of simply using the maximum-a-posteriori (MAP) estimates of θ_d as

the dependent variable, having access to the posterior distribution of the topic proportions, we can take account for the uncertainty of the dependent variable. This can be achieved by employing a sampling procedure known as the method of composition in the social sciences; see Tanner (2012), p.52. This procedure is implemented in the *stm* package through its function *estimateEffect*.

In the following, we will first introduce the method of composition. We will discuss its implementation in the *stm* package and provide alternative regression approaches based on the method of composition. Subsequently, we will evaluate the relationship between prevalence covariates and topic proportions by directly assessing the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$, as outlined above, and compare the results of both approaches.

5.1 Method of Composition

Let $\theta_{(k)} := (\theta_{1,k}, \ldots, \theta_{D,k})^T \in [0,1]^D$ denote the proportions of the k-th topic for all D documents. As stated, we want to perform a regression of these topic proportions $\theta_{(k)}$ on a subset $\tilde{X} \in \mathbb{R}^{D \times \tilde{P}}$ of prevalence covariates X. The true topic proportions are unknown, but the stm produces an estimate of the approximate posterior of $\theta_{(k)}$. A naïve approach would be to regress the estimated mode of the approximate posterior distribution on $\tilde{X}$. However, this approach neglects much of the information contained in the distribution.

Instead, repeatedly sampling $\theta^*_{(k)}$ from the approximate posterior distribution, performing a regression for each sampled $\theta^*_{(k)}$ on $\tilde{X}$, and then sampling from the estimated distribution of icients, provides an i.i.d. sample from the marginal posterior distribution of regression coefficients.

Sampling $\theta^*_{(k)}$ is achieved by first sampling the unnormalized topic proportions η^* from the approximate posterior $q(\eta)$, applying the softmax $\theta^* = \text{softmax}(\eta^*)$ (element-wise, i.e., for each of the K-dimensional vectors of topic proportions), and lastly selecting the k-th column of θ^*. Precisely, $q(\eta) = \prod_d q(\eta_d)$ is a normal distribution, which emerges from the laplace approximation within the variational inference scheme; for details see Roberts et al. (2016), pp. 992-993. For clarity, we denote the approximate posterior of topic proportions as $q(\theta_{(k)}|X, W)$, in order to emphasize that the parameters of this distribution are learned from the observed data, i.e. prevalence covariates and words (note that we have no content variables included). Furthermore, let ξ denote the regression coefficients from a regression of $\theta_{(k)}$ on $\tilde{X}$, and let $q(\xi|\tilde{X}, \theta_{(k)})$ be the approximate posterior distribution of these coefficients, i.e. given design matrix $\tilde{X}$ and response $\theta_{(k)}$.

The method of composition can now be described by repeating the following process m times:

1. Draw $\theta^*_{(k)} \sim q(\theta_{(k)}|X, W)$.

2. Draw $\xi^* \sim q(\xi|\tilde{X}, \theta_{(k)})$.

It then holds that $\xi_1^*, \ldots, \xi_m^*$ is an i.i.d. sample from the marginal posterior

$$q(\xi|X, W) := \int_{\theta_{(k)}} q(\xi|\tilde{X}, \theta_{(k)}) q(\theta_{(k)}|X, W) \mathrm{d}\theta_{(k)} = \int_{\theta_{(k)}} q(\xi, \theta_{(k)}|X, W) \mathrm{d}\theta_{(k)},$$

where $q(\xi, \theta_{(k)}|X, W) := q(\xi|\tilde{X}, \theta_{(k)}) q(\theta_{(k)}|X, W)$. Thus, by integrating over $\theta_{(k)}$, this approach allows incorporating information contained in the posterior distribution of $\theta_{(k)}$ when determining ξ.

5.1.1 Implementation in the *stm* package

The R package *stm* implements a simple OLS regression through its *estimateEffect* function. However, this approach ignores that the sampled topic proportions are restricted to $(0, 1)$. As expected, using this framework we frequently observe predicted proportions outside of $(0, 1)$. Moreover, credible intervals are non-informative, due to violated model assumptions.

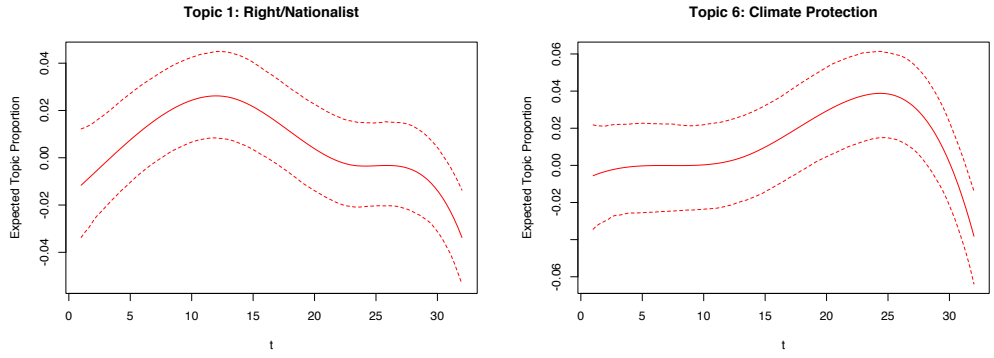

Figure 8: Estimated prevalence of topics 1 and 6 over time, generated using *estimateEffect* from the *stm* package

5.1.2 Alternative implementation

We can attempt to improve the approach employed within the *stm* package by replacing the OLS regression with a regression model that assumes a dependent variable in the interval $(0, 1)$. However, note that since topic proportions are modeled separately, regardless of the specific model implied, distributional assumptions about $\theta_{(k)}$ will be violated. This is due to the fact that the the distribution of a subvector - and thus particularly of a single component - of θ_d is not of a simple form, when θ_d follows a logistic normal distribution, see e.g. Atchison and Shen (1980).

As shown by Atchison and Shen (1980), a distribution that can be used to approximate a logistic normal distribution is the Dirichlet distribution. However, note that the Dirichlet

distribution assumes less interdependence among components than implied by the logistic normal distribution. In case of the Dirichlet distribution the univariate marginal distributions are beta. One possibility is thus to perform a separate beta regression for each topic proportion on $\tilde{X}$.

As an alternative approximation we can employ a quasibinomial generalized linear model (GLM). Topic proportions can be rescaled and discretized and topics comprehended as classes, such that each rescaled topic proportion can be interpreted as the "number of successes" for the respective class. To match the underlying logistic normal distribution more closely, the quasi-likelihood furthermore allows for a flexible variance specification.

Note that $q(\xi|\theta_{(k)}, \tilde{X})$ is asymptotically normal for both the beta regression, see Ferrari and Cribari-Neto (2004), p. 17, and the quasibinomial GLM, see e.g. Fahrmeir et al. (2007), p. 285. Furthermore, in both cases we use a logit-link.

5.1.3 Visualization

We now apply the method of composition, based on either a beta regression or a quasibinomial GLM, in order to visualize covariate effects. Here we only visualize the results obtained by the quasibinomial GLM; the results of the beta regression, which show similar trends, are found in the appendix. Setting the number of simulations to 100, we sample $\xi_1^*, \ldots, \xi_{100}^*$ from the marginal posterior distribution $q(\xi|X, W)$. As mentioned, when visualizing the impact of a particular covariate, all other covariates are held at their median (or majority vote, if categorial), in line with the methodology employed in the *stm* package. Let $\tilde{X}^*$ denote the subset of X where, apart from the variable of interest, each selected column consists of the median of the respective column of X. In order to plot the predicted effects, we then input $\tilde{X}^*\xi^*$ into the sigmoid function, which is the response function corresponding to a regression with logit-link, and calculate the predicted proportions.

We exemplarily illustrate the relationship between covariates and topic proportions for topic 4 ("Social/Housing") and topic 6 ("Climate Protection"). The linear predictor of our regressions takes the same form as in (5.1), i.e., we do not use a subset $\tilde{X}$, but the full set of prevalence covariates X, in order to estimate the effects, although we do not display each covariate included. For smooth effects, it is important to recall that their borders are inherently unstable, which is why one should refrain from (over-)interpreting them. For both continuous and categorical variables, black lines indicate the mean, and the shaded area represents 95% credible intervals.

For topic 4, "Social/Housing", we observe that most continuous variables have a small effect in absolute terms: the absolute variation in topic proportion across the covariate domains merely amounts to 4%, compared to 8% for topic 6. For most covariates the trend is rather ambiguous. Somewhat surprisingly, a very high unemployment rate is negatively

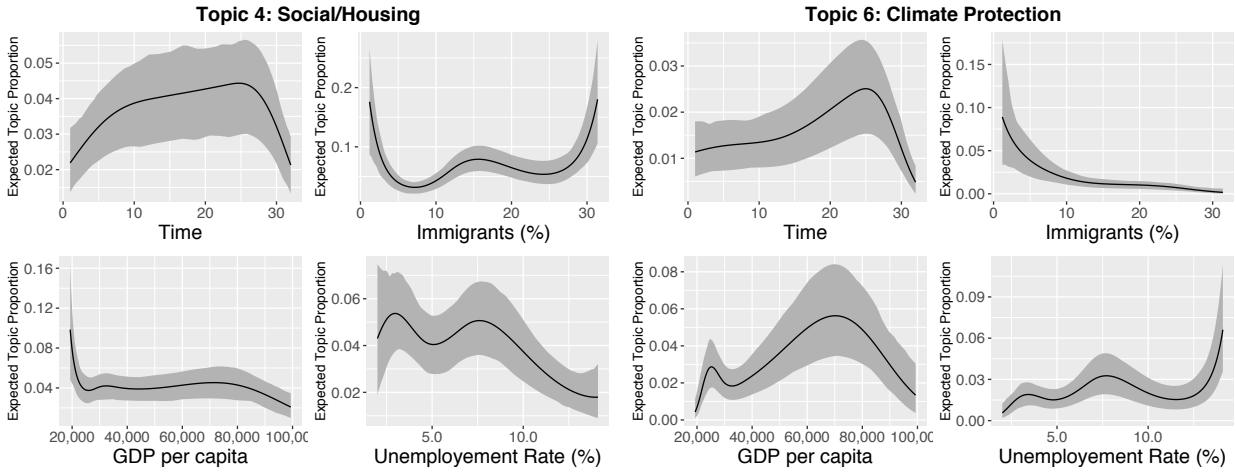

Figure 9: Mean and 95% credible intervals for smooth effects, obtained using a quasibinomial GLM.

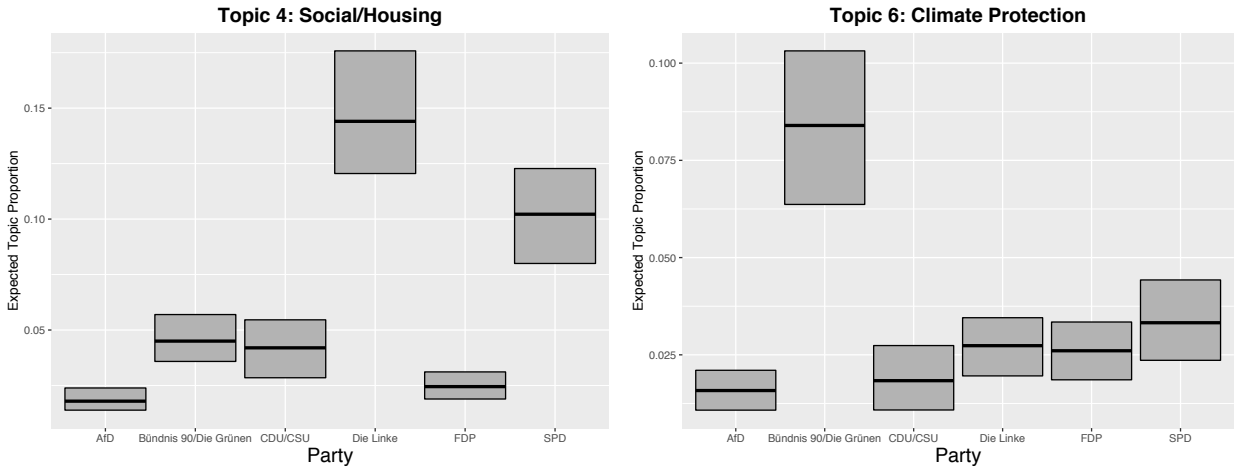

Figure 10: Mean and 95% credible intervals for different political parties, obtained using a quasibinomial GLM.

linked to topic 4.

The effect of the political party on the relevance assigned to the topic "Social/Housing" is very much in line with a priori expectations: the left party and social democrats have the highest topical prevalence (15% and 10%, respectively), and the nationalist party the lowest (2%).

For the smooth effects of topic 6, we observe its prevalence peaks in September 2019, corresponding to month t=25, decreasing afterwards. The absolute changes in topic proportions over time are rather small (around 3%). The percentage of immigrants within an electoral district shows a negative relation to topic 6. Furthermore, topic 6 tends to be discussed more frequently in mid-income electoral districts than in high- and low-income districts. Finally, the link to the unemployment rate is somewhat ambiguous, although generally rather positive.

Regarding the relationship between the political party and the prevalence of topic "Climate Protection", as to be expected, we find high topical prevalence for the green party. Similar to the smooth effects, total variation in topic proportions across parties amounts to approximately 8%.

Finally, the graph below shows a summary comparison of topical prevalence across all parties, for topics "Right/Nationalist", "Climate Protection" and "Social/Housing". The results are generally consistent with expectations. The proportions of topics "Climate Protection" and "Social/Housing" vary between 2% and 9% and between 2% and 15%, respectively. For topic 1, "Right/Nationalist", note how topical prevalence for the AfD party amounts to more than 40%, implying that more than 40% of the total content tweeted by AfD party members is about right-wing/nationalist issues, particularly immigration; for all other parties, topic 1 is rather marginal below 3%.

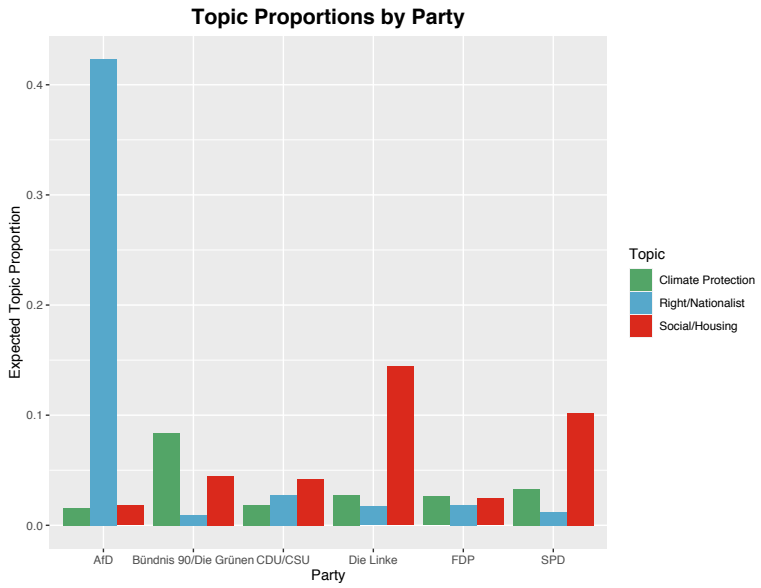

Figure 11: Topical prevalence by political party for topics 1, 4, and 6.

5.2 Direct assessment using $\hat{\Gamma}$ and $\hat{\Sigma}$

The *stm* being an extension to the correlated topic model (CTM), it is assumed that the topic proportions follow a logistic normal distribution, such that $\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma^T x_d^T, \Sigma)$. Within the CTM, the Dirichlet distribution of the LDA has been replaced with a logistic normal distribution, in order to allow for a joint dependence among topics. Therefore, as mentioned above, separately modeling topic proportions is a simplification; in particular credible intervals should be treated with caution.

In order to examine the relation of prevalence covariates and topic proportions considering the joint dependence among the latter, we can attempt to directly use the output produced

by the *stm*: inference of the *stm* involves finding the maximum-a-posteriori (MAP) estimate $\hat{\Gamma}$ and the maximum likelihood estimate $\hat{\Sigma}$.

If we are interested how a specific prevalence variable is related to topic proportions, similar to previous analyses, we can attempt to predict topic proportions based on a new design matrix X^*, where each column apart from the variable of interest corresponds to the median of the respective column of X. Ideally, in order to directly predict topic proportions, we would first draw a sample Γ^* from the posterior distribution of Γ, and subsequently sample the topic proportions θ_d^* from a logistic normal with mean parameters $((\Gamma^*)^T(x_d^*)^T, \hat{\Sigma})$, where $\hat{\Sigma}$ is the maximum likelihood estimation of Σ. The resulting topic proportions would then correspond to a sample of the posterior predictive distribution of topic proportions. Unfortunately, the output of the stm does not allow for the possibility to draw a sample from the posterior distribution of Γ, but only provides its MAP estimate $\hat{\Gamma}$.

Nevertheless, in order to get an impression how the assumed generative process of topic proportions in the stm behaves, we can plug in the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$ into the logistic normal distribution and visualize sampled values from this distribution. Given a new observation x_d^*, we can sample θ_d^* from $\text{LogisticNormal}_{K-1}(\hat{\Gamma}^T(x_d^*)^T, \hat{\Sigma})$ by

1. Drawing $\eta_d^* \sim \mathcal{N}_{K-1}(\hat{\Gamma}^T(x_d^*)^T, \hat{\Sigma})$ and setting $\eta_{d,K}^* = 0$.
2. Mapping to the simplex, i.e., for all $k = 1, \ldots, K$: $\theta_{d,k}^* = \frac{\exp\left(\eta_{d,k}^*\right)}{\exp\left(\sum_{i=1}^K \eta_{d,i}^*\right)}$.
3. Setting $\theta_d^* := (\theta_{d,1}^*, \ldots \theta_{d,K}^*)^T$.

We have repeated the above steps 1000 times for each input value of a selected variable, while fixing other variables at their median, and obtained the empirical mean as well as 95% credible intervals. Plotting the results, we observe that while the mean shows a similar trend to our previous analyses, the obtained credible intervals are much broader.

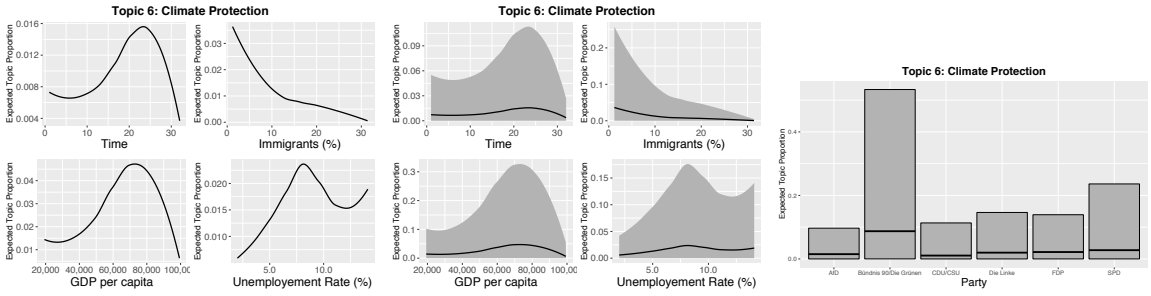

Figure 12: Smooth effects without credible intervals (left), smooth effects with credible intervals (mid), and effect of the political party (right).

The large fluctuations for a specific topic proportion can be ascribed to the fact that the unnormalized topic proportions are drawn from a $K-1$-dimensional *multivariate* normal

distribution, before the softmax is applied. Therefore, a single normalized proportion depends heavily on the sampled unnormalized proportions of the remaining topics. While the variance of a topic-specific unnormalized proportion is independent of the remaining unnormalized proportions and c.p. constant for an increasing number of topics, the application of the softmax function induces a large increase in the variance of a topic-specific normalized proportion.

We suspect that the magnitude of credible intervals in figure 12 provides a more realistic picture than in case of a separate modeling of topic proportions, since the usage of the logistic normal distribution of topic proportions is an implicit assumption made within the stm that there is a dependence among topics, as argued above. This ultimately produces a large variance of the univariate marginal distributions of topic proportions, as can be observed. While ideally we should sample Γ from its posterior distribution instead of plugging in its MAP estimate, our results nevertheless suggest that there is a discrepancy between the assumed distribution of topic proportions in the generative process of the stm, and the impression we gain of the distribution of topic proportions from a separate modeling of topics within the method of composition.

5.3 Topical Content

The STM provides an additional way to integrate covariate effects into the model, apart from prevalence variables that impact topic proportions across documents. To be specific, a categorical variable can be selected as topical content variable. While the prevalence variables influence the propensity of the 15 topics for each document, the content variable now allows for the word distributions for a given topic to vary across documents, according to the content variable level. Note that this is a completely new model, which is why one should not expect the resulting topics to be similar.

Formally, recall that the word distribution used to eventually pick a word is $\beta_{d,n} := \beta(z_{d,n}, Y_d) \in \mathbb{R}^V$, where $z_{d,n}$ is a (latent) indicator variable determining the word's topic assignation and Y_d is the document-level topical content variable with A levels. In the prevalence model, no (document-level) topical content variable is specified, implying $\beta_{d,n} = \beta(z_{d,n})$; since $z_{d,n}$ is a word-level variable, $\beta_{d,n}$ is constant across all documents for a given topic k. When specifying a content variable Y, however, $\beta_{d,n}$ now varies for each document, according to the level $a \in \{1, ..., A\}$ the content variable takes on for document d. That is, the total number of β-vectors, each one of length V, now increases from K to $K \times A$.

For our specific case, since the topical content variable needs to be categorical, we choose the variable *party*, being categorical by definition and as it is arguably the most significant factor in determining topic prevalence. In doing so, we implicitly posit that for a given topic, an MP's party additionally influences the vocabulary used when tweeting about that specific topic. For instance, this implies that an AfD party member tweets about immigration issues in a different linguistic manner than, say, a green MP. Since for the 2017 election period the German parliament contains members of 6 parties, $\mathbf{Y}$ is now a matrix with 10998 rows and 6 columns, yielding a total of 90 β-vectors.

After fitting the model, we proceed as for the prevelance model, that is, by inspecting top words and identifying topic labels. An additional difficulty, however, is that we do not have clear-cut top words per topic anymore; instead, we now have topic-level top words for each of the 15 topics, party-level top words for each of the 6 parties, as well as interaction top words for each of the 90 topic-party combinations. The table below presents topic labels for all 15 topics, identified by using the same 3-step procedure as for the prevalence model before. As can be seen, five topics are labeled as *miscellaneous*, reflecting the complexity caused by the large number of β-vectors.

The topical content model allows for vocabulary usage to differ across political parties, given a topic. In Figure 13 below, we visualize this effect for the Corona topic, contrasting the green party "Bündnis 90/Die Grünen" with the right-wing nationalist party "AfD". The result is very insightful: even for a topic as clear-cut and novel as COVID-19, stark differences in terms of vocabulary usage arise. In particular, the AfD uses language suitable to describe

Topic1	Right/Nationalist 1
Topic2	Miscellaneous 1
Topic3	Left/Humanitarian
Topic4	Housing
Topic5	Innovation
Topic6	Green/Energy
Topic7	Miscellaneous 2
Topic8	Corona
Topic9	Foreign Affairs
Topic10	Election
Topic11	Right/Nationalist 2
Topic12	Miscellaneous 3
Topic13	Miscellaneous 4
Topic14	Twitter/Politics
Topic15	Miscellaneous 5

Table 2: List of topic labels for STM with topical content variable (party).

immigration (*migration*, *grenz*) in order to discuss Corona, which very much reflects the unimodality of the party's political orientation (as can also be seen in Figure 21 at the end of section 4.4.1. The green party, on the other hand, seems to address the topic much more specifically, mentioning key words like *massnahm* or *kind*.

Figure 13: Differences in vocabulary usage across parties for the Corona topic.

While this type of visualization is indeed insightful, several concerns regarding the topical content model prevail: first of all, there is no natural candidate for the content variable, which

- for labeling and interpretational purposes - should ideally be binary. Our dataset contains very few categorical variables, none of them binary. Furthermore, there is no natural, non-arbitrary way to binarize any of the covariates; for instance, binarizing the variable party into conservative and liberal would misclassify at least one party. Therefore, our choice to use party as content variable is the result of a lack of alternatives, rather than being based on sound statistical or theoretical considerations. This, in turn, is reflected in the difficult labeling: recall that one third of all topics were eventually being labeled as miscellaneous. And while the previous illustration of inter-party differences in vocabulary usage is indeed insightful in terms of topic exploration and visualization, the aforementioned doubts lead us to discard the topical content variable for further analysis. In fact, in the next section we consider a model without any covariates in order to perform a clean 2-step procedure for covariate effect estimation.

5.4 2-step Approach: CTM

In sections 4.4 and 4.5, we analyzed the relationship between topic proportions and metadata, visualizing the effect of prevalence covariates and deciding against the further inclusion of a topical content variable. As briefly mentioned in section 2 already, a point of concern when using the STM is the double usage of covariates: they are used in the estimation of the topics themselves (and thus, in the estimation of the latent topic proportions) and subsequently they are again used to estimate their relationship with topic proportions. From classical statistical modeling, we are used to interpret such relationships, oftentimes ascribing a causal interpretation to the corresponding coefficients; in our case, this would go along the lines of stating, for instance, that "a higher percentage of immigrants within an electoral district makes politicians prioritize issues other than climate protection", referring to Figure XXX in section 4.4.1. Topic models, however, present a crucial difference as compared to classical statistical models: the target variable - θ - is latent and thus itself being estimated. For explorative or descriptive purposes, this does not pose a problem because there is only a single step: discovering topics in the text documents. Yet whenever in a second step, after estimating the model, we wish to conduct (causal) inference, we face an overfitting problem, since the *same* documents and covariates are used in both steps. In this section, we focus on the double usage of (prevalence) covariates, while section 4.7 deals with double usage of documents, i.e., words.

To avoid overfitting due to double usage of covariates, we fit an STM without including any covariates in the model estimation, thus reducing the model to a simple CTM. In a second, isolated step, we estimate the relationship between topic proportions and covariates. That is, we forgo the potential (small) gains of joint estimation of the STM in favor of a clear-cut 2-step procedure which avoids overfitting. As a first step, we fit the CTM analogously to the original STM (which includes topical prevalence variables), the only difference being that no document-level metadata is used in the estimation of the CTM. In line with the performance results in Roberts et al. (2016), we observe a slightly higher held-out likelihood for the STM (-8.5478) than for the CTM (-8.5492) when holding out a random 50% of the words from a randomly chosen 10% of the documents. Moreover, we notice that the topics themselves (in terms of their top words) are almost identical to those of the STM, which is why we use the same topic labeling as in section 4.4. As for differences in topic proportions between the two models on a document level, we consider the average topic proportion deviation per document, $\frac{1}{K}\sum_{k=1}^{K}|\theta_{d,k}(STM) - \theta_{d,k}(CTM)|$. The resulting average difference between topic proportions per topic, averaged across all documents, amounts to 1.61%; that is, for an average document, the absolute difference in the proportion of each topic is less than 2%, which is rather moderate. These differences in topic proportions between STM and CTM further cancel each other out across documents: when comparing global topic proportions

(i.e., topic proportions simply averaged across all documents), the results are very similar, with the average difference per topic only amounting to 0.23%. Altogether, topic proportions seem to be affected by the topical prevalence covariates only to a small degree on an individual document level, and this effect almost disappears entirely if we consider corpus-wide topic proportions.

In the second step, we consider the relationship between topic proportions and prevalence covariates for the CTM and compare the resulting relationships with those of the originally fitted STM (which contains prevalence covariates). For comparability, we use the same methodology as in section 4.4: applying the method of composition with a quasibinomial regression of individual topic proportions on covariates. The only difference is that prevalence covariates were not included in the model used to generate topic proportions. Consequently, sampling all (unnormalized) topic proportions jointly via the logistic normal distribution (as in Figure 4.XXX) is not applicable here, as no Γ-vector is being estimated at all. In the figures below, we visualize the CTM topic proportions of topics 4 (Social/Housing) and 6 (Climate Protection) in relation with continuous covariate values and across parties and compare the results to those of the STM (Figures XXX and XXX). As for the relationship between continuous covariates and topic proportions, the results for STM and CTM are very similar: for both topic 4 and topic 6, the trends across the respective covariate range are almost identical for the two models, while the scale differs slightly (with scale differences hardly exceeding 2%). Turning to the categorical variables, in particular party, the conclusion is very similar for topic 4: we observe minor scale differences and very similar patterns. For topic 6, the scale of the topic proportions is again slightly different compared to the STM, and now we also observe some (minor) difference in the relative positioning of the different parties.

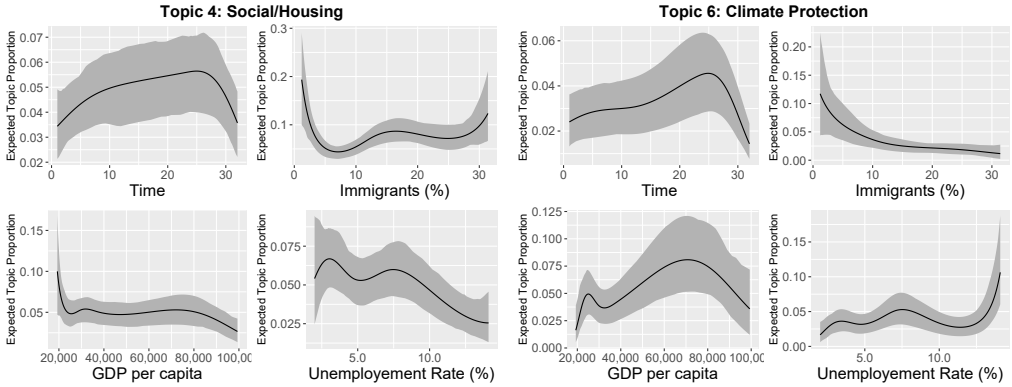

Figure 14: Mean and 95% credible intervals for smooth effects, obtained using a quasibinomial GLM (no covariates included in model estimation).

Topic proportions across parties for topics 3, 6, and 1 (Right/Nationalist) are further summarized in the plot below. Comparing the results to those of the STM for the additional

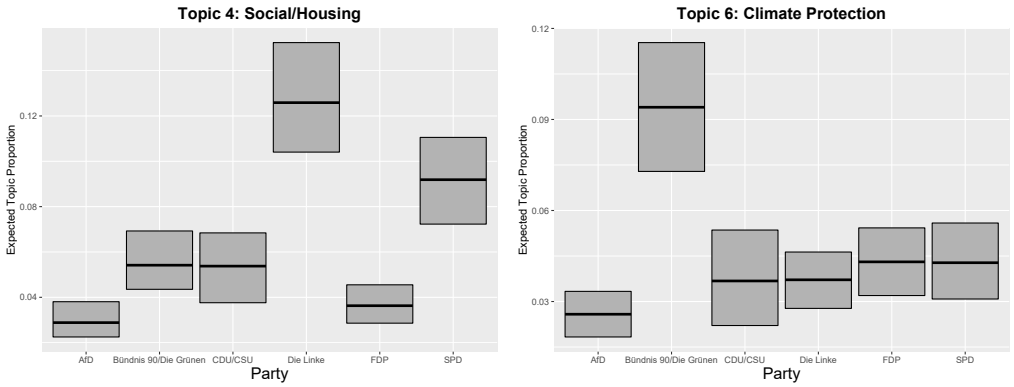

Figure 15: Mean and 95% credible intervals for different political parties, obtained using a quasibinomial GLM (no covariates included in model estimation).

topic 1, a rather large difference can be seen: the overall topic proportion for the AfD party is now almost 10% lower than in the STM (though still at almost 35%). Furthermore, for all topics and covariates, the comparison between STM and CTM does not change if we use Beta regression instead of quasibinomial regression within the method of composition, corroborating our results (see appendix XXX).

All in all, the relationships between topical prevalence variables and topic proportions are very similar to those of the STM when instead using a clean 2-step estimation procedure where no covariate information is used in the model estimation. This indicates that the problem of double usage of covariate information in the STM, potentially leading to overfitting, is not overly severe. However, we wish to remind the reader that at this point, we have not yet accounted for the double usage of documents - even in this "clean" 2-step procedure, the estimation of topic proportions is based on the same documents which are associated with the covariate values used in the second step. Section 4.7 finally addresses this open issue.

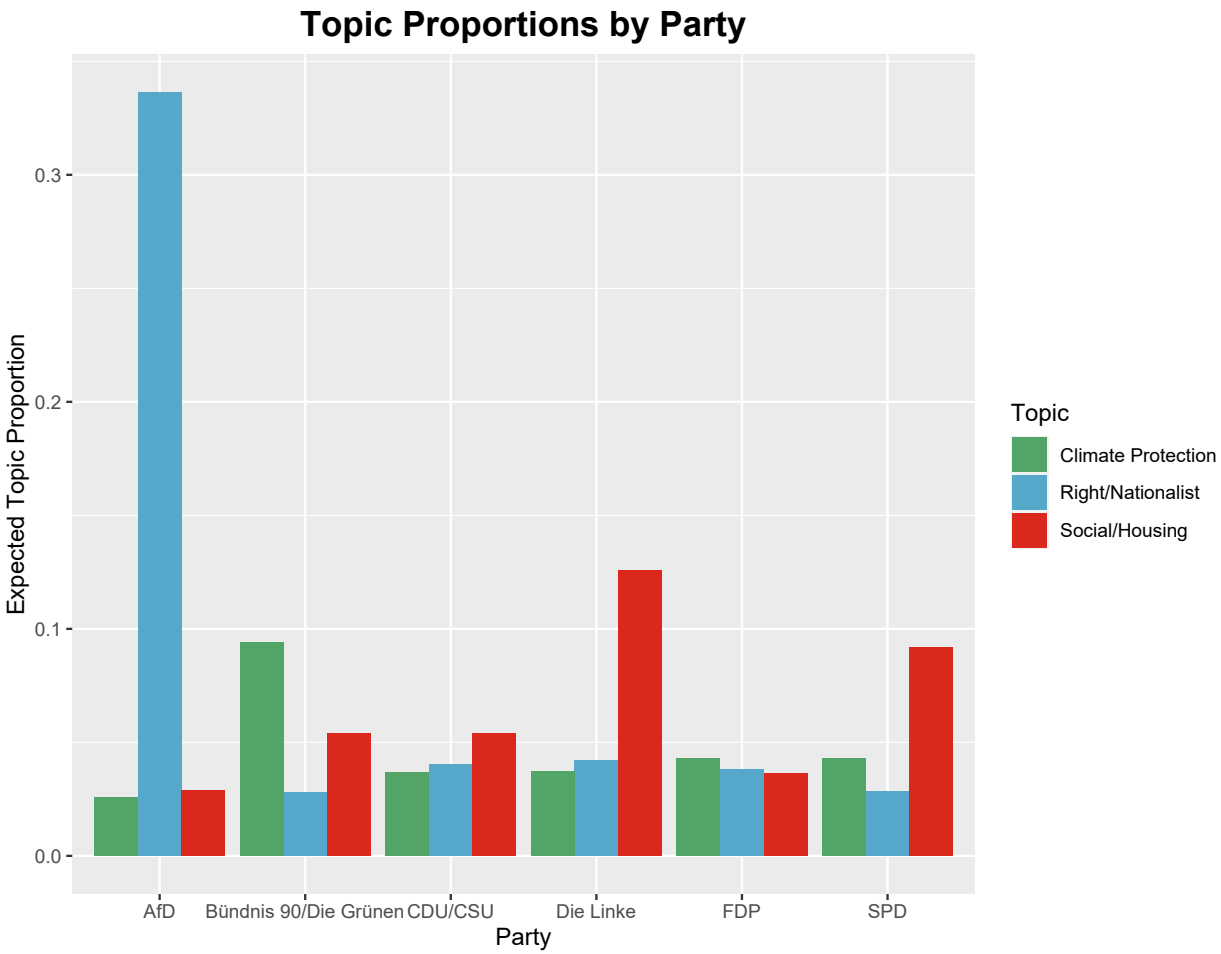

Figure 16: Mean and 95% credible intervals for different political parties, obtained using a quasibinomial GLM (no covariates included in model estimation).

6 Causal Inference: Train-test Split

In our analyses from section 4.4, we first estimated the latent topic proportions using the stm, and then assessed the relation between these document-level topic proportions and prevalence covariates. In particular, the documents that were used to obtain the topic proportions were the same that were subsequently used to quantify relationships between covariates and topic proportions. As Egami et al. (2018) argue, this double usage of data is a form of overfitting and hence inferences about covariate effects are biased. Additionally, since in the stm prevalence covariates affect estimated topic proportions, there is not only a mere double usage of data (i.e., in the sense that the same documents are used twice), but also a direct double usage of prevalence covariates, as the estimated latent topic proportions are regressed on the former.

Both problems can be addressed using the framework proposed by Egami et al. (2018). The general idea is to split the data $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$, and utilize the training set in order to determine a model to infer latent topic proportions from any text assumed to be generated by the same underlying process as the training set. Subsequently, this estimated model is applied on the test set, in order to assess the relation between test set topic proportions and test set prevalence covariates. In the following, we will explain the exact procedure for the stm (note that Egami et al. (2018) focus, for the most part, on the general framework, while the exact application within the stm is not discussed in depth) and evaluate the results when applied to our data.

6.1 Model Estimation on the Training Set

On the training set, we estimate components of the stm similarly to the estimation on the full data set. That is, we input documents, i.e., words and metadata from the training set, and obtain estimates $(\hat{\beta}_{\text{train}}, \hat{\Gamma}_{\text{train}}, \hat{\Sigma}_{\text{train}})$, where $\hat{\beta}_{\text{train}}$ is associated with the topic-word distribution, and $\hat{\Gamma}_{\text{train}}$ as well as $\hat{\Sigma}_{\text{train}}$ are the topical prevalence parameters.

6.2 Prediction of Topic Proportions on the Test Set

Prediction of the topic proportions on the test is not straightforward, since the topic proportions are latent and the stm is not built for the purpose of predicting these latent variables on a set of new, unseen data. The fundamental idea is to estimate the variational posterior of the latent variables, that is, the topic proportions θ_d, where $d \in \mathcal{D}_{\text{test}}$ (note that z_d is integrated out in the stm), conditioned on the model parameters $(\hat{\beta}_{\text{train}}, \hat{\Gamma}_{\text{train}}, \hat{\Sigma}_{\text{train}})$ from the training set, as well as the words W_{test} from the test set. This functionality is implemented in the *stm* package through the function *fitNewDocuments*, which per default outputs the MAP estimates of topic proportions θ_d, for all $d \in \mathcal{D}_{\text{test}}$. Note that estimating the vari-

ational posterior of the latent variables, conditioned on the parameters and the words, is precisely what occurs during each E-step of the EM Algorithm. Thus, the implementation of *fitNewDocuments* simply consists of one E-step with inputs $(\hat{\beta}_{\text{train}}, \hat{\Gamma}_{\text{train}}, \hat{\Sigma}_{\text{train}}, W_{\text{test}})$. It is, however, not obvious how to exactly input $\hat{\Gamma}_{\text{train}}$ and $\hat{\Sigma}_{\text{train}}$ into the E-step. Depending on the characteristics of the specific analysis conducted by the researcher, Egami et al. (2018) propose three different alternatives:

1. **Covariate-specific prior**: Before applying the E-step, $\hat{\Gamma}_{\text{train}}$ is used to obtain $\hat{\mu}_d := (\hat{\Gamma}_{\text{train}})^T(x_d)^T$, for each document $d \in \mathcal{D}_{\text{test}}$ in the test set. Each document is then updated performing the E-step with inputs $(\mu_d, \Sigma) = (\hat{\mu}_d, \hat{\Sigma}_{\text{train}})$ together with the respective document specific words as well as $\hat{\beta}_{\text{train}}$ (for the exact update machanism see pp. 992-993, Roberts et al. (2013)). The problem with this approach is, however, that for two documents from the test set containaing the exact same words, different topic proportions are predicted if the prevalence covariates differ. However, in such a case we would want the causal effect of the covariates on the topic proportions to be zero.

2. **Average prior**: The average prior circumvents the above described problem of the covariate-specific prior by simply using - for each document in the test set - the average $\overline{\mu}_{\text{train}} := \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{d \in \mathcal{D}_{\text{train}}} (\hat{\Gamma}_{\text{train}})^T(x_d)^T$ of all document-specific means from the training set. The covariance $\hat{\Sigma}_{\text{train}}$ is recalculated based on the new average $\overline{\mu}_{\text{train}}$ according to formula (11) on p. 993, Roberts et al. (2013). In this scenario, prevalence covariates from the test have no influence at all on the predicted topic proportions.

3. **No prior**: If no prior is used, then for each document $d \in \mathcal{D}_{\text{test}}$ in the test set the E-step is performed using $\mu_d = 0$ and replacing $\hat{\Sigma}_{\text{train}}$ with a diagonal covariance matrix with very large diagonals.

The covariate-specific prior cannot be used in our case due to the above described problem, that different topic proportions are predicted for identically worded test set documents, if their prevalence covariates differ. The option "no prior" can be useful if the metadata on the test set is believed to be linked differently to topics than is the case on the training set. In most cases the second option, "average prior", should provide the best trade-off, since in this case metadata from the training set is directly used to predict topic proportions, but the problem of the covariate-specific prior is solved. Note that hence in this case there is no double usage of covariates.

6.3 Results

We now depict the results obtained conducting a train-test split, where we split the data into two equally sized sets, for the options "average prior" and "no prior". Note that the

test data cannot consist of words which have not been seen in the training data. Therefore, all previously unseen words are removed from the test data. After removing the words, the test data contains 80.6% of the original words. Since we use only a subset of the full data, the estimated topics are slightly different than those obtained using the full data; however, most topics are similar. We assigned new labels to the topics, a complete list of which can be found in the accompanying R code of this paper.

In contrast to section 4.4., the focus of this section will be on quantifying causal effects between covariates and the amount a topic is discussed, since the train-test framework is most appropriate in order to conduct such analyses. As mentioned, the function *fitNewDocuments* outputs the MAP estimates of the variational posterior of topic proportions for the test set. In Figure 17 we depict these MAP estimates of topic proportions, together with the topic proportions obtained for the training data.

The UN Climate Action Summit 2019 was held on 23 September 2019. As can be observed, the topic associated with climate issues was discussed to a much larger extent during this time than a year earlier. While the MAP estimates for the different prior specifications on the test set are rather similar, the estimated effect is much larger training for the training data. If we compare the estimated topic proportions for a topic we labelled as 'Emancipation' for the two opposing parties 'AfD' and 'Bündnis 90/Die Grünen', we find similar results: the average difference of estimated topic proportions between both parties is larger for the training data. Also, note that the variation is higher on the training data compared to the test data in both cases.

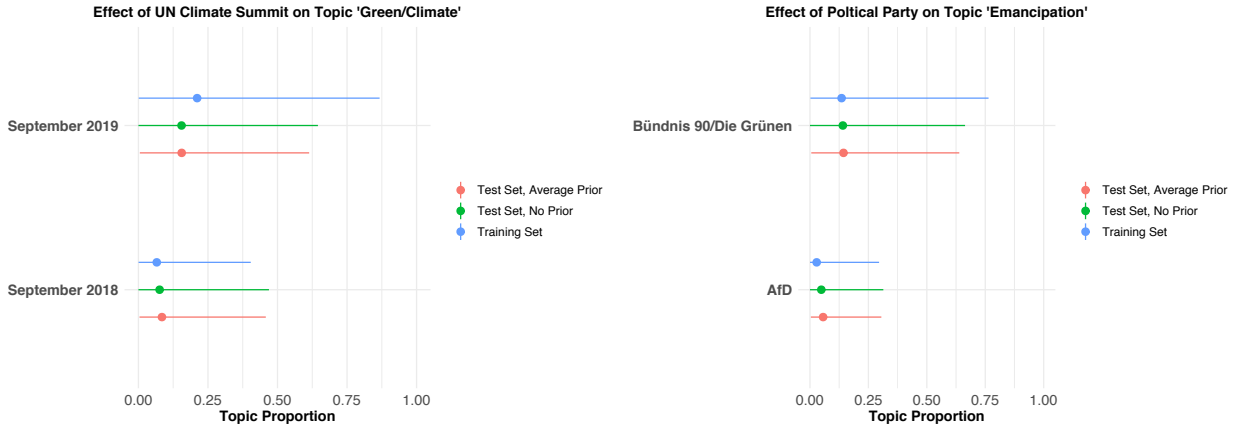

Figure 17: Maximum-a-posteriori (MAP) estimates of topic proportions on training and test data. Points display the mean, lines 2.5% and 97.5% credible intervals.

If we want to estimate the treatment effect, we can estimate the average difference of MAP estimates between both groups. Following Egami et al. (2018), we obtain an estimate

of the Average Treatment Effect (ATE) on a data set $\mathcal{D}$ as

$$\widehat{\text{ATE}} = \frac{1}{|\mathcal{D}_{\text{treatment}}|} \sum_{i \in \mathcal{D}_{\text{treatment}}} \hat{\theta}_i - \frac{1}{|\mathcal{D}_{\text{control}}|} \sum_{i \in \mathcal{D}_{\text{control}}} \hat{\theta}_i, \tag{6.1}$$

where $\hat{\theta}_i$ is the MAP estimate for the i-th document. Egami et al. (2018) show that, if additional conditions hold, the estimated $\widehat{\text{ATE}}$ on previously unseen test data $\mathcal{D}_{\text{test}}$ is an unbiased estimate of the ATE.

In Figure 18 we visualize the ATE estimated on training and on test data with different prior specifications (note that this is simply the difference of the means depicted in Figure 17). The results correspond to our classical idea of overfitting: since the characteristics of each parliamentarian associated with a document have been used to estimate the topic proportions in the first place, when evaluating the effects of these characteristics on topic proportions on the same data, the effect is optimistically biased.

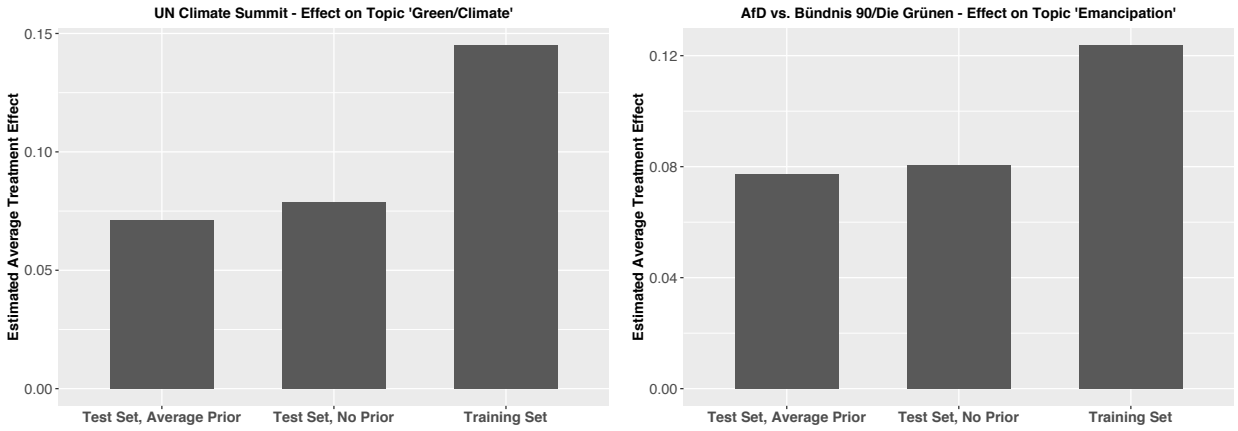

Figure 18: Estimated Average Treatment Effects (ATE) using training and test data.

Before we finish this section, we want to remind the reader that there are several general concerns when conducting a causal inference study. For instance, if the treatment group is not a random subsample of the population, but instead assignment to this group is related to the causal effect of interest, the resulting estimator of the treatment effect will suffer a selection bias.

7 Conclusion

TBD

8 Appendix 1

In line with Wang and Blei (2013), consider a generic topic model with latent variables θ and z as well as observed data x:

$$p(\theta, z, x) = p(x|z)p(z|\theta)p(\theta).$$

The exact posterior distribution

$$p(\theta, z|x) = \frac{p(\theta, z, x)}{\int p(\theta, z, x)dzd\theta}$$

is usually intractable due to the high-dimensional integral, which is why the distribution needs to be approximated.

As stated in section 2.3, in variational inference a simple distribution family $q(\theta, z)$ is posited and subsequently, we determine the member of this family - that is, the variational parameter(s) - that minimizes the KL divergence. Note that, for computational purposes, we compute KL divergence of the true posterior p from the approximating posterior q, $KL(q||p)$, whereas intuitively one would seek to minimize $KL(p||q)$.

The most popular variational inference technique is mean-field variational inference (also: mean-field variational Bayes), where we posit full factorizability of $q(\theta, z)$: $q(\theta, z) = q(\theta)q(z)$. That is, θ and z are assumed to be independent with their own distributions and variational parameters ϕ (which we suppress for improved readability). Since θ and z are actually dependent, this approximate distribution family $q(\theta, z)$ does not contain the true posterior $p(\theta, z|x)$.

Let us now write out the KL divergence of p from q:

$$\begin{aligned} KL(q||p) &= \mathbb{E}_q[\log \frac{q(\theta, z)}{p(\theta, z|x)}] \\ &= \mathbb{E}_q[\log\big(q(\theta, z)\big)] - \mathbb{E}_q[\log\big(p(\theta, z|x)\big)] \\ &= \mathbb{E}_q[\log\big(q(\theta, z)\big)] - \mathbb{E}_q[\log\big(p(\theta, z, x)\big)] + \log\big(p(x)\big) \end{aligned}$$

Since $KL(q||p) \geq 0$ (which can be easily shown using Jensen's inequality), it follows that:

$$\log\big(p(x)\big) \geq \mathbb{E}_q[\log\big(p(\theta, z, x)\big)] - \mathbb{E}_q[\log\big(q(\theta, z)\big)].$$

The left-hand side of the above inequality is the marginal log likelihood of observed data x and is also called evidence (of the observed data). Note that the evidence is not computable - otherwise we would not need to resort to variational inference in the first place. The right-hand side thus presents a lower bound on the evidence and we define the *Evidence Lower*

BOund (ELBO) as:

$$ELBO := \mathbb{E}_q[\log\big(p(\theta, z, x)\big)] - \mathbb{E}_q[\log\big(q(\theta, z)\big)],$$

where the second component of the ELBO, $\mathbb{E}_q[\log\big(q(\theta, z)\big)$, is the entropy of the approximate distribution q. Equivalently, we could say that the evidence constitutes an upper bound for the ELBO. This means that we actively maximize the ELBO (which is therefore also called *variational objective*), which in turn is equivalent to minimizing the KL divergence of the true posterior $p(\theta, z|x)$ from the approximate distribution $q(\theta, z)$. Therefore, the approximation $q(\theta, z)$ - or, more precisely, the variational parameters ϕ of $q(\theta)$ and $q(z)$ - that maximizes the ELBO simultaneously minimizes KL divergence (Blei et al., 2003; Wang and Blei, 2013). Wang and Blei (2013) show that for the chosen factorization of the joint distribution $p(\theta, z, x)$, and using the optimality conditions as derived in Bishop (2006), we obtain the following solutions when setting $\frac{\partial ELBO}{\partial q} \stackrel{!}{=} 0$:

$$q^*(\theta) \propto exp\{\mathbb{E}_{q(z)}[\log\big(p(z|\theta)\big)p(\theta)]\},$$
$$q^*(z) \propto exp\{\mathbb{E}_{q(\theta)}[\log\big(p(x|z)\big)p(z|\theta)]\}.$$

The coordinate ascent algorithm iteratively updates one of these two expressions while holding the other one constant, but requires closed-form updates to do so. This requirement is fulfilled as long as all model nodes are conditionally conjugate, i.e., as long as for each node in the model "its conditional distribution given its Markov blanket (i.e., the set of random variables that it is dependent on in the posterior) is in the same family as its conditional distribution given its parents (i.e., its factor in the joint distribution)" (Wang and Blei (2013), p. 1008). The authors consequently define a class of models where some nodes are not conditionally conjugate, the so-called *nonconjugate models*; for this class, using Laplace approximations, the variational family is shown to be $q(\theta, z) = q(\theta|\mu, \Sigma)q(z|\phi)$; that is, $q(\theta)$ is now Gaussian with variational parameters μ and Σ.

The STM in particular constitutes a nonconjugate model, since $p(\theta)$ is logistic normal and thus not conjugate with respect to the multinomial distribution $p(z|\theta)$. Consequently, no closed-form update is available for $q(\eta)$. Using mean-field variational inference, the approximate posterior family is $\prod_{d=1}^{D} q(\eta_d)q(z_d)$, where $q(\eta_d)$ is Gaussian and $q(z)$ is binomial (Roberts et al., 2016). Given the posterior, inference now consists in finding the particular member of the posterior distribution family that maximizes the approximate ELBO. (Due to the subsequent Laplace approximation, ELBO does not constitute a true lower bound on the evidence and the updates do not maximize ELBO directly, which is why Roberts et al. (2013) use the term *approximate* ELBO. See Wang and Blei (2013) for further discussion.) Applying Laplace variational inference, we approximate $q(\eta_d)$ using a (quadratic) Taylor expression

around the maximum-a-posteriori (MAP) estimate $\hat{\eta}_d$, which yields a Gaussian variational posterior $q(\eta_d)$, centered around $\hat{\eta}_d$, and allows for a closed-form solution of $q(z_d)$. Iteratively updating $q(\eta_d)$ and $q(z_d)$ thus constitutes the E-step of the EM algorithm.

The M-step consists in maximizing the approximate ELBO with respect to model parameters. Prevalence parameters Γ and Σ are updated through linear regression and maximum likelihood estimation (MLE), respectively. The updates for topic-word distributions β_k (or $\beta_{k,a}$ if a content covariate is specified) are obtained through multinomial logistic regression. Further details are provided in Roberts et al. (2013) and in the appendix of Roberts et al. (2013). Moreover, the appendix of Blei et al. (2003) provides a detailed description of variatonal inference and empirical parameter estimation for the (conditionally conjugate) LDA model.

9 Appendix 2

9.1 Plots of section 4.4

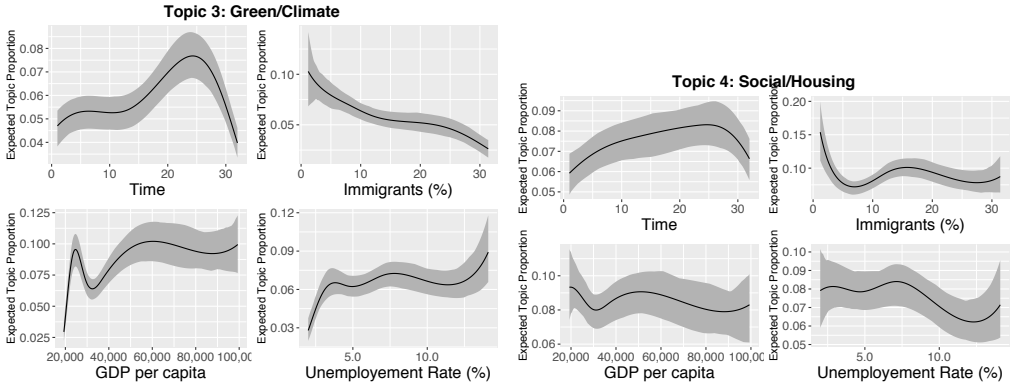

Figure 19: bla

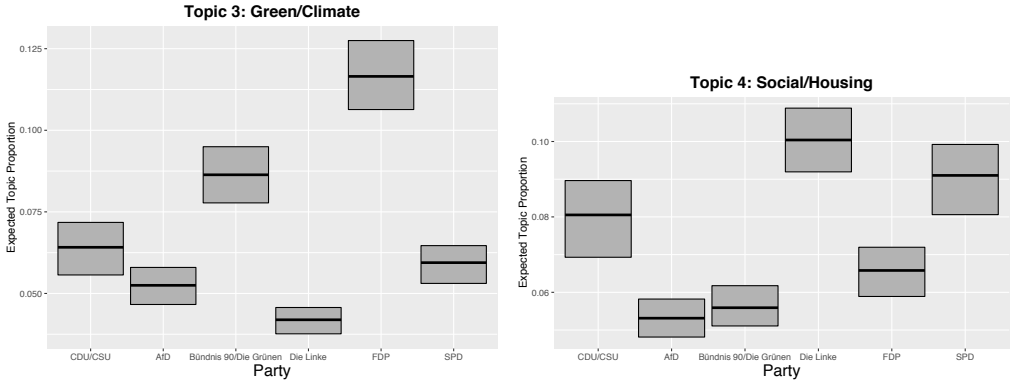

Figure 20: blabla

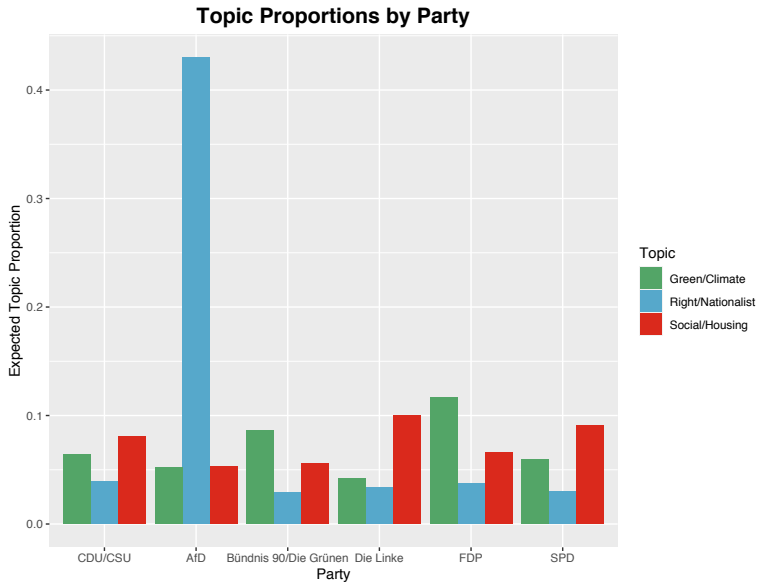

Figure 21: Topical prevalence by political party for topics 1, 2, and 3.

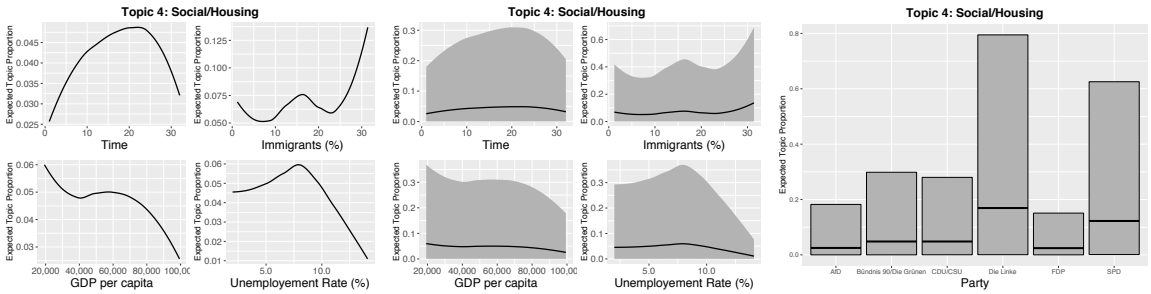

Figure 22: bla

References

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

J Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

Jonathan Bischof and Edoardo M Airoldi. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 201–208, 2012.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

David M Blei, John D Lafferty, et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

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 and Maintainer Jonathan Chang. Package 'lda', 2010.

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.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. 2011.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer, 2007.

Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

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.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

Ian R James, James E Mosimann, et al. A new characterization of the dirichlet distribution through neutrality. *The Annals of Statistics*, 8(1):183–189, 1980.

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

Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2):254–277, 2015.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

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.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, pages 1–20. Harrahs and Harveys, Lake Tahoe, 2013.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40, 2019. doi: 10.18637/jss.v091.i02.

Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

Matt Taddy. On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pages 1184–1193, 2012.

Martin A Tanner. *Tools for statistical inference*. Springer, 2012.

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.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.