1 Results

1.1 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 $\hat{\beta}_{k,v}$.[1] 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.[2] 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

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

[2] 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).

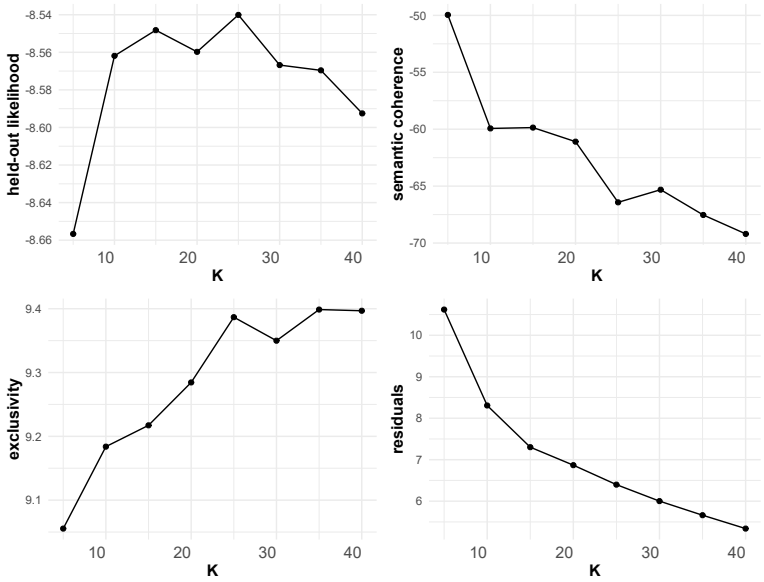

Figure 1: Model evaluation metrics for hyperparameter K (number of topics).

1.2 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 2: 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.

1.3 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.

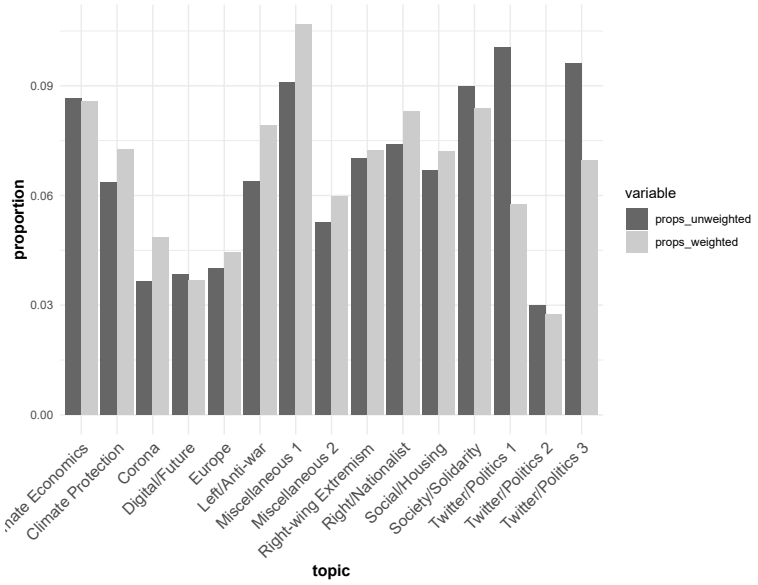

Figure 3: Weighted and unweighted global topic proportions.

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 5 (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

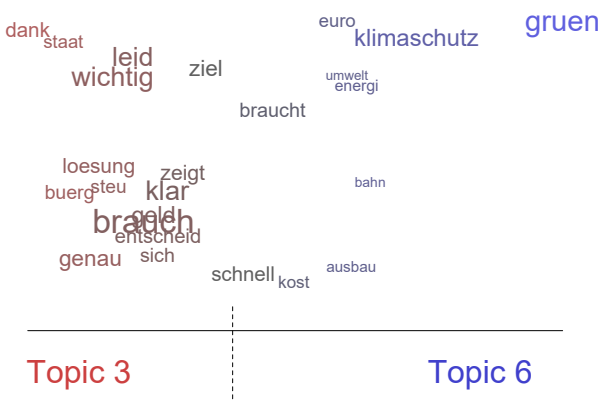

Figure 4: Comparison of vocabulary usage for two topics.

(Society/Solidarity), while the strongest positive correlation (11.79%) is the one shown before, between topic 3 (Twitter/Politics 1) and topic 6 (Twitter/Politics 3). We can also visualize these correlations using a network graph (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 it 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.

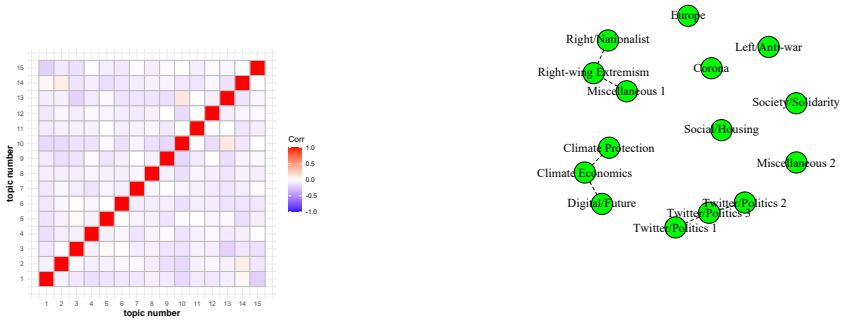

Figure 5: Global topic correlations as matrix (left) and graph (left).

References

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.

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

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.

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.

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

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.