4_5

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Results

1.1 Hyperparameter Search and Model Fitting

1.2 Labelling

1.3 Global-level Topic Analysis

1.4 Covariate-level Topic Analysis

1.5 Content Model

1.6 Train-test Split

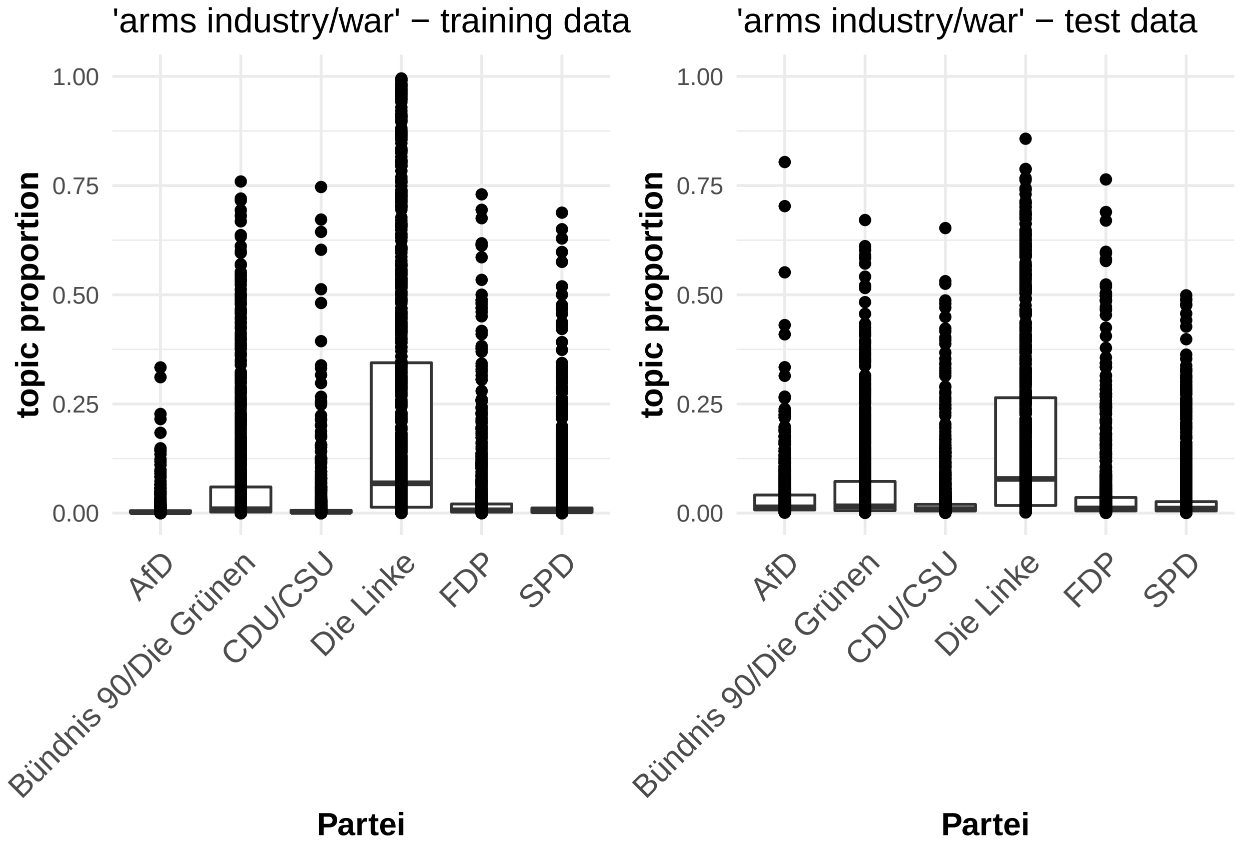
In section 4.4, we analyzed the relationship between metadata and topic proportions. From classical statistical modelling, we are used to interpret such relationships, oftentimes ascribing a causal interpretation to the corresponding coefficients; in our case, this would go along the lines of stating, for instance, that “a higher percentage of immigrants within an electoral district makes politicians prioritize issues that are not related to climate”, refering to Figure XXX.

Topic models, however, present a crucial difference as compared to classical statistical models: the target variable - θ - is latent and is 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 are used in both steps.

In their paper on causal inference for text data, Egami et al. (2018) introduce a train-test framework which avoids the aforementioned problem, since the model is fitted on training data and the fitted model is then used to predict topic proportions for the new, previously unseen test data, using the metadata corresponding to the test documents. This can be seen as predicting new values, given already estimated parameters and new covariate values, just like in classical statistical modelling.

In this section, we pursue the same strategy as Egami et al. (2018): we split the data into equally sized training and test sets, train the model on the training set and predict topic proportions for the documents in the test set, using the *fitNewDocuments* function. In doing so, it is important to set the priors for the prevalence covariates, μ and Σ, so that they are not document-specific functions of the covariates. Therefore, we choose an average global prior, which can roughly be seen as average about training-data priors (see Egami et al. (2018) for further details).

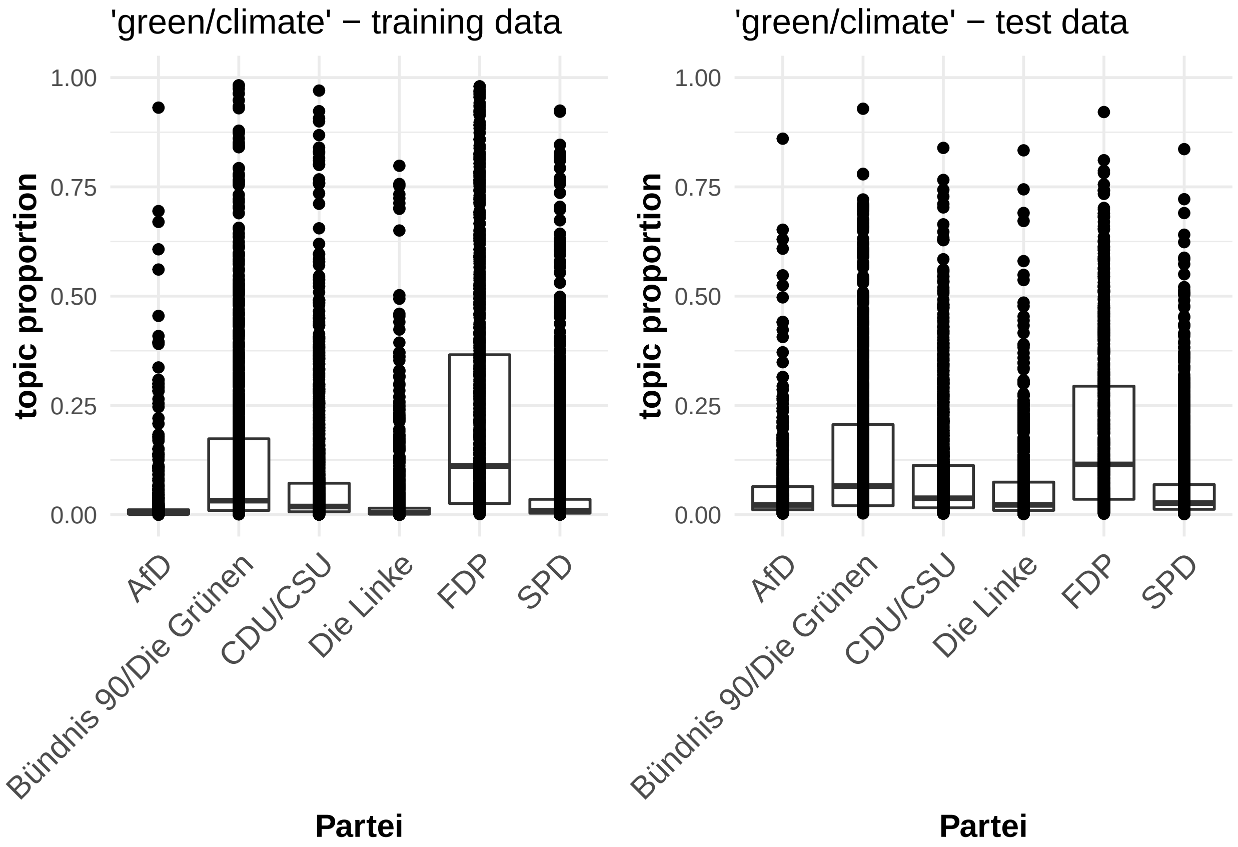
The graph below shows the difference in topic proportions between training and test data for the arms industry/war topic, across all parties. It is important to note that, as the method of composition does not apply here, the values plotted - for both training and test data - are the simple modes of the posterior distributions of θ_d. As can be seen, the topic proportions are very similar on the training and test set for all parties, even for the left party with an average topic proportion of almost 10%.



'arms industry/war' − training data

'arms industry/war' − test data

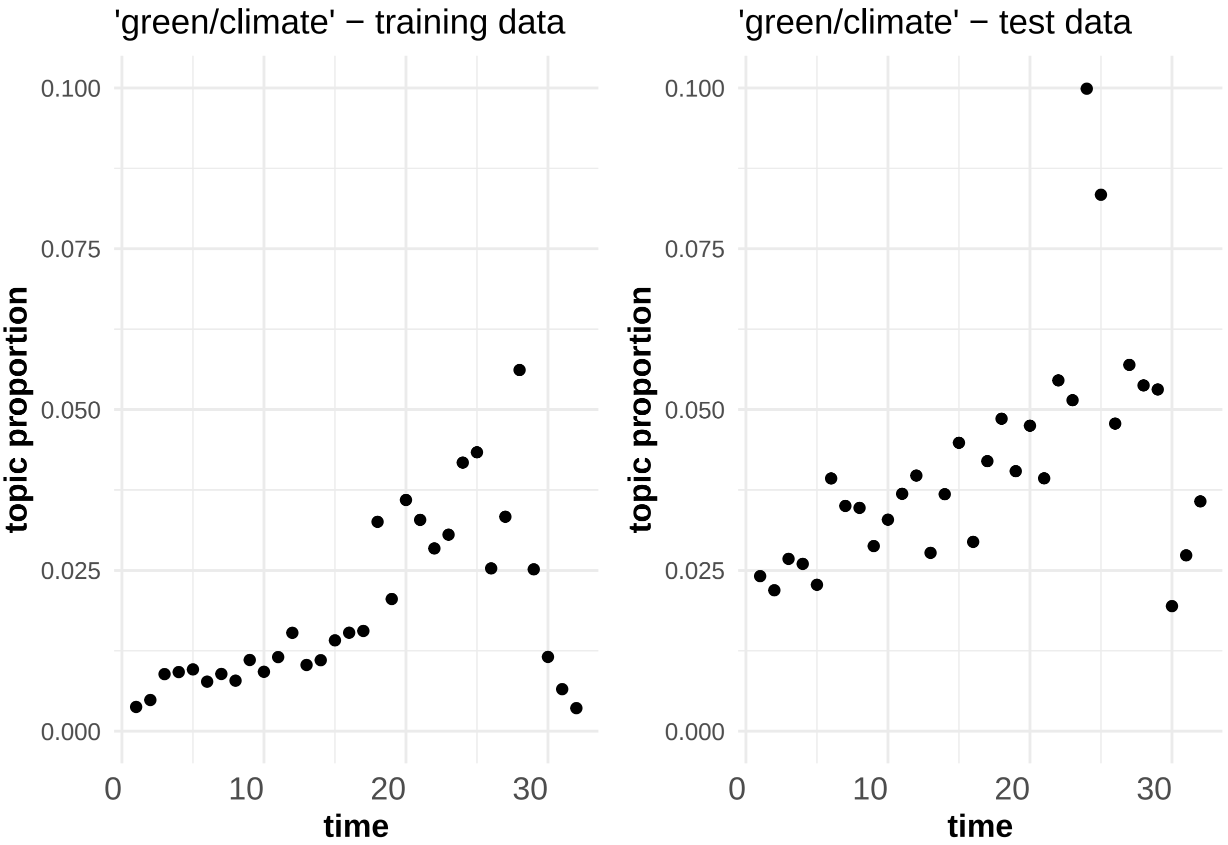
We compare training and test data proportions for another topic, "green/climate", again across all parties. As shown in the graph below, the distribution of topic proportions when fitting previously unseen documents to the estimated model, is again very similar for all parties. Moreover, the results are very much in line with those of the "green" topic in Figure XXX (section 4.4), obtained by fitting the model on the entire dataset and by applying the method of composition. (Recall that we are simply plotting the posterior modes here.)



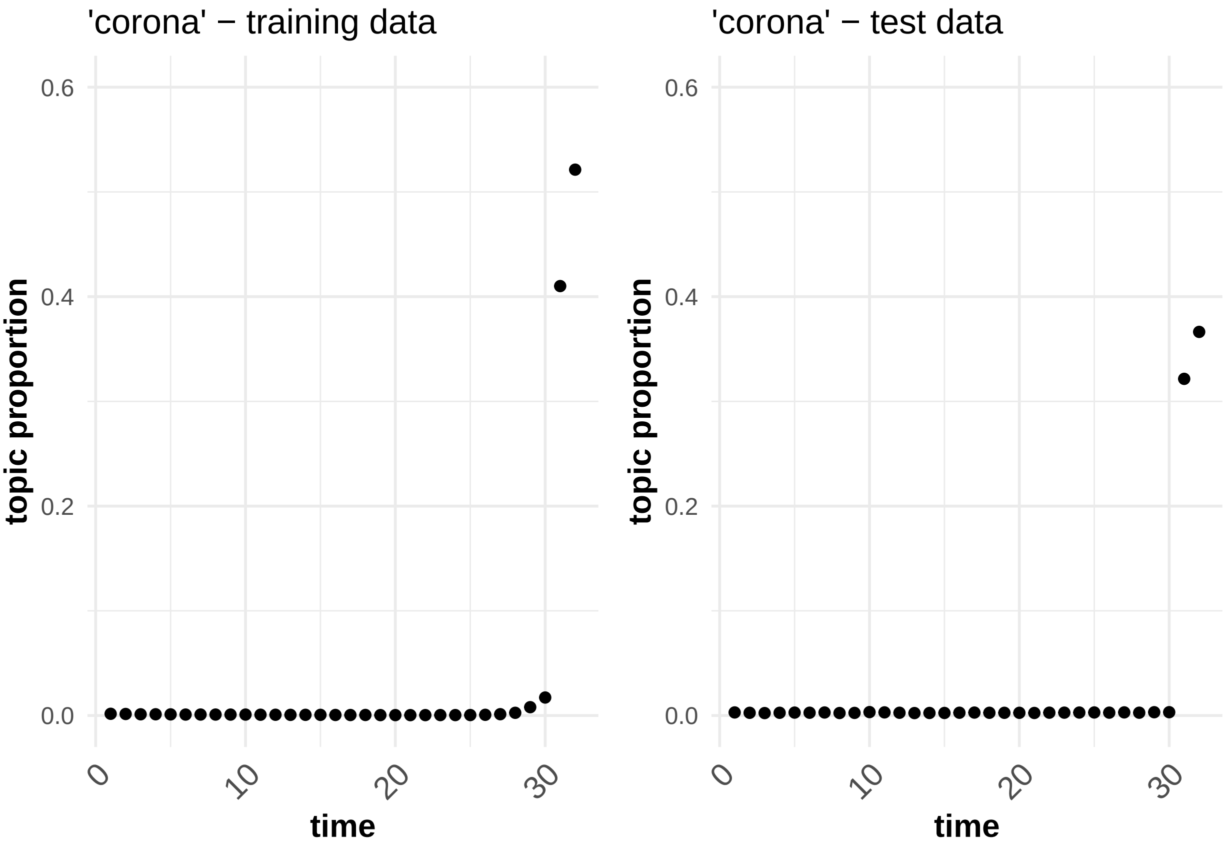
Next, one might also want to know how well the model predicts the evolution of topic revelance over time for a new set of documents. To do so, we simply calculate the monthly median across all document-level topic proportions, $\theta_{k,d}$, for each month from September 2017 through April 2020.

(Footnote: We chose the median because topic proportions tend to be heavily right-skewed and therefore, the average would not properly represent what holds for the “majority” of documents. The right-skewedness stems from the fact that there will always be some “specialized” MPs who tweet almost exclusively about a single topic, thus causing the respective topic proportions for some documents to be close to 100%. By choosing the median, our results are are also much more in line with those reported in section 4.4)

We visualize the comparison between training and test data over time for the green/climate topic. As demonstrated in the graph below, the trend is similar across the two datasets: the relevance of the topic increases steadily until September 2019, falls slightly for the last quarter of 2019 and then plunges in 2020. Note, again, that we observed the same trend in Figure 4.4.



As a final example, we show the development of the corona topic over time. As can be seen in the graph below, the topic is practically non-existent until early 2020, but its relevance increases sharply in March and April 2020. This is also a key explanatory factor of the plunge in relevance of the green/climate topic, as discussed above.



The main idea of the train-test split is to gauge the *predictive power* of the model: big differences between topic proportions in the training data and those in the test data would indicate that the estimated topics (and their proportions) are not very meaningful in summarizing texts, except the ones the model was trained with; this, in turn, would imply overfitting. Our results show that topic proportions for new documents, both across parties and over time, are distributed very similarly to those on training data. Intuitively, this means that the topics “make sense”.

This section demonstrated how the double usage of text documents - both in estimating the model and in making inference - and the resulting overfitting can be avoided by using a split-sample framework as in Egami et al. (2018). Since the STM distinguishes itself from the baseline LDA by incorporating covariate information into the modelling process, parameters for covariates and the target variable are estimated jointly. However, since the values of the target variable (i.e., the topic proportions) are estimated based on covariate values, we are still facing the problem of double usage of the covariates: they are used in estimating the model (which, in turn, estimates the latent target variable) and, subsequently, their coefficients are to be interpreted. Clearly, we cannot use the train-test framework to address this problem; indeed, it would only be exacerbated by *generating* topic proportions for the test documents, using (previously fitted) parameters and document metadata as covariates, and subsequently making inference about the effect of those covariates on the previously generated topic proportions. We address this issue in the next section.

To be addressed: * metadata for test data is entirely meaningless, does not affect topic proportions at all * manipulating covariate values neither

1.7 Two-step Approach

To avoid overfitting due to double usage of covariates, we decide to fit a simple CTM without including any covariates in the model estimation, and to estimate the relationship between topic proportions and covariates in a second, isolated step. That is, we forgo the potential (though limited) gains of joint estimation of the

STM in favor of a clear-cut two-step procedure which avoids overfitting.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. “How to Make Causal Inferences Using Texts.” *arXiv Preprint arXiv:1802.02163*.