How Document Sampling and Vocabulary Pruning Affect the Results of Topic Models

Daniel Maier, Andreas Niekler, Gregor Wiedemann, Daniela Stoltenberg

CCR 2 (2): 139–152
DOI: 10.5117/CCR2020.2.001.MAIE

Abstract

Topic modeling enables researchers to explore large document corpora. Large corpora, however, can be extremely costly to model in terms of time and computing resources. In order to circumvent this problem, two techniques have been suggested: (1) to model random document samples, and (2) to prune the vocabulary of the corpus. Although frequently applied, there has been no systematic inquiry into how the application of these techniques affects the respective models. Using three empirical corpora with different characteristics (news articles, websites, and Tweets), we systematically investigated how different sample sizes and pruning affect the resulting topic models in comparison to models of the full corpora. Our inquiry provides evidence that both techniques are viable tools that will likely not impair the resulting model. Sample-based topic models closely resemble corpus-based models if the sample size is large enough (> 10,000 documents). Moreover, extensive pruning does not compromise the quality of the resultant topics.

Keywords: latent Dirichlet allocation, model selection, preprocessing, text analysis, topic model

Introduction

Latent Dirichlet allocation (LDA) topic modeling has become a popular technique in communication research. It is used to identify hidden topical categories within large document collections (Blei, Ng, & Jordan,

2003). The LDA algorithm inductively searches for latent content variables (topics) inferred from recurring word patterns (Blei, 2012).

Although coding large corpora is one of the main advantages of LDA, the processing can be extremely costly in terms of time and computing resources. In order to accelerate the modeling process, two well-established techniques may be applied: (1) modeling only random document samples during learning, and (2) pruning the vocabulary of the corpus. Although both techniques are frequently applied, there has been no systematic study of how they affect the results of the topic model. The present study tries to address this research desideratum.

Topic modeling can be a lengthy and tedious process due to at least three reasons. First, LDA is an iteration-based algorithm, and the calculation of a single model, therefore, cannot be parallelized. Second, an increasing number of documents results in a linear increase in computing time and memory space (Niekler, 2018), which is why corpora of upwards of a hundred thousand documents might cause time issues. Finally, the literature suggests that multiple models with different parameter settings should be pretested and compared (Denny & Spirling, 2018) before a decision can be made regarding the model that fits best in terms of interpretability and the theoretical concept at hand (Bonilla & Grimmer, 2013; Maier et al., 2018). If the computational work for only one model is high, the estimation of multiple models can appear impractical, which, in turn, limits the usability of the method. Consequently, the application of LDA is less accessible, particularly for researchers with limited financial and computational resources.

Two techniques have been suggested in the literature to (partly) overcome this problem and accelerate the modeling process: modeling document samples and vocabulary pruning.

Modeling Document Samples

Maier et al. (2018) proposed using random document samples (in the learning phase) instead of fitting a model for the whole corpus. This appears plausible: if a document sample resembles the vocabulary distribution of its corpus, equivalent topics should be the result. Until now, linguistic research has only confirmed this precondition: sufficiently large samples resemble the vocabulary distribution of the complete corpus (Hanks, 2012). The question of whether topic models from document samples also resemble the topics of full corpus models still lacks empirical inquiry.

Vocabulary Pruning

Pruning means cutting the most frequently and infrequently used words of the vocabulary of a corpus. In theory, the most frequent and infrequent words do not contribute useful information to a topic model (Denny & Spirling, 2018). Due to their low conditional probability, infrequent words will never appear in a list of topic's top words. In contrast, very frequent words are not specific enough; such words will appear in the top words of every topic, and thus add no specific or exclusive information.[1]

As most words in the corpus vocabulary only occur once or twice (Manning & Schütze, 2003, p. 23-29), a relatively small fraction of words remains after pruning. A pruned document-term matrix, thus, has a considerably reduced dimensionality compared to an unpruned one. Pruning, therefore, leads to a considerable reduction of computational work, enhancing the algorithm's performance, and stabilizing the stochastic inference (Maier et al., 2018, p. 101).

Disciplinary standards have emerged regarding the question of just how much to prune. For infrequent words, most authors discussing the technique suggest removing words that occur in fewer than 0.5% (Grimmer, 2010; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010; Denny & Spirling, 2018) – or, at most, fewer than 1% (Hopkins & King, 2010; Grimmer & Stewart, 2013; Denny & Spirling, 2018) – of all documents. For frequent words, the majority of authors recommend removing those that occur in more than 99% of all documents (Hopkins & King, 2010; Grimmer & Stewart, 2013). Adhering to these relatively conservative pruning thresholds should ensure no valuable information is lost, while drastically reducing the vocabulary size of the corpus. However, although this technique is often used, there is no systematic study about how pruning affects the results of a topic model (Denny & Spirling, 2018, p. 172).

In the present study, we set out to test the effect of both techniques on the resulting topic models systematically. More specifically, we (1) investigate how document sampling affects the resulting topics and (2) assess the role of relative pruning, i.e., stripping the corpus of both the most frequent and infrequent terms. To ensure results are generalizable, our study builds on three empirical corpora, which represent different usage contexts in communication research and exhibit different features in terms of content heterogeneity, vocabulary, and document size: one corpus of Twitter messages, one corpus of news articles, and one corpus of thematically focused web page content.

Study Design

The feasibility of (a) sampling documents or (b) pruning vocabulary critically depends on whether a topic model for (a) a random document subset or (b) the pruned vocabulary yields similar topics as a model for (a) the whole corpus or (b) the unpruned vocabulary. Thus, we posed the following research questions:

> RQ1: How large must a random document sample for a topic model, be for its topics to resemble the topics of full corpus models?
> RQ2: Does the relative pruning of the corpus vocabulary impair the model quality as compared to models of corpora with unpruned vocabularies?

In order to answer these questions, we investigated the effects of document sampling and pruning on three document corpora, each representing a typical use case in communication research: a website corpus, a news article corpus, and a Tweet corpus. The corpora differed considerably on several relevant features, such as the heterogeneity of vocabulary and content, writing styles, and the institutional context in which they originated (see Table 1). The diversity of the cases allowed us to test the generalizability of the findings.

All three corpora were cleaned and preprocessed (Denny & Spirling, 2018), following the suggestions of Maier et al. (2018, i.e., tokenization; lowercasing; removal of stop words, punctuation, and special characters; lemmatization). All corpora were prepared for two pruning modes that affect the vocabulary sizes of the corpora. While in the first mode, the vocabulary remained untouched (referred to as "unpruned"), the second mode was a relative pruning approach, where the top 1% most frequent and 0.5% most infrequent terms were stripped. Here, we followed the low pruning thresholds commonly recommended in the literature, which minimized the risk of removing valuable information while still drastically reducing the corpus vocabulary size (for our cases by between 92.2% for the website corpus and 95.8% for the news corpus, see Table 1).

Then, five full corpus models were calculated for each of the three corpora and two pruning modes (see the two bottom lines in Figure 1).[2] These were referred to as reference models. Additionally, for each corpus and each pruning mode, we calculated five models for five different sample size categories (1%, 5%, 10%, 20%, and 50%; see the two top lines in Figure 1).[3] Within a corpus and sample size category, we opted to calculate multiple models for the same sample of documents, rather than draw

different samples. The latter option would introduce an additional source of variance, as it would mean including different sets of documents in each model. Since our aim was to assess within-sample size variance due to the stochasticity of LDA, this would have confounded two sources of variance and ultimately obscured the answer to RQ1.

Subsequently, our analysis proceeded in three steps (as indicated by the arrows in Figure 1, right side).[4] Each step consisted of assessments of the similarity of word-topic matrices φ using a measure based on top word-comparisons proposed by Niekler and Jähnichen (2012). First, we calculated the reliability, that is, the similarity of models in the same sample size and corpus category (comparison *a* Figure 1). The reliability can be thought of as a baseline indicator for the stability of the topic modeling results. Second, we calculated the similarity of the sample model φ's to the reference model φ's (comparison *b* Figure 1 targeting RQ1). Third, we assessed the similarity between the φ's of the pruned and unpruned models (comparison *c* Figure 1 targeting RQ2). The next section provides details about the three corpora and the applied metrics.

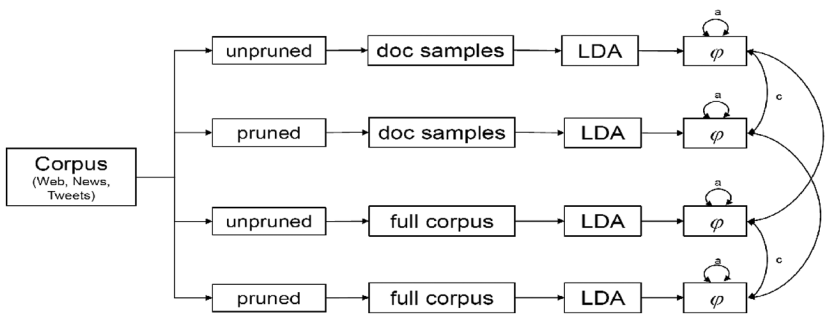

Figure 1. Study design. a = reliability; b = reproducibility sampled vs. full models; c = reproducibility pruned vs. unpruned models.

Data and Methods

In order to assess the robustness of our findings, we used three real-world corpora, a collection of Twitter messages, a website corpus and a corpus of news media articles. Together these corpora represent typical cases in communication research (see Table 1).

The Tweet corpus consisted of messages published on the microblogging platform Twitter. The corpus was compiled by identifying public user profiles located in the city of Berlin, Germany, and subsequently capturing their messages as well as those of the users with whom they interacted.

Consequently, the corpus was not thematically focused but could be topically diverse. As Tweets are rather short (maximum of 280 characters), raw Tweet collections were not considered suitable corpora to be modeled (Guo, Vargo, Pan, Ding, & Ishwar, 2016). We, therefore, followed recommendations to aggregate Tweets into larger text entities (Guo et al., 2016; Hong & Davison, 2010), in this case, by concatenating all Tweets from the same user into what we called user-documents. In order to guarantee sufficient lengths of user-documents, we limited our analysis to those with a minimum of 50 terms. Altogether, the corpus contained a total of 101,638 user-documents, spanning two natural weeks in the summer of 2018. For practical reasons, we limited our analysis to messages in English language.

The website corpus represented the textual content published by various social actors (e.g., civil society actors, bloggers, economic actors, media organizations) concerned with the topic of food safety. The data was retrieved monthly over two and a half years using the web crawling tool Issue Crawler. Starting from several seed websites of actors located in the United States that were deemed central to the issue, hyperlinks to other websites were followed with a crawling depth of two and a degree of separation of one (for more detailed information, see Waldherr, Maier, Miltner, & Günther, 2017). In order to retain only relevant, food safety-related content, a keyword filtering procedure was applied. The final corpus consists of 84,268 documents. Due to the keyword filtering, the website corpus was more thematically focused, and, thus, more homogeneous than the other two corpora. Although the corpus may be regarded as relatively homogenous content-wise, the document authors stemmed from heterogeneous institutional backgrounds with diverse interests and writing styles.

Finally, the news corpus contained articles published by the London-based newspaper The Guardian. The Guardian provides an open platform for researchers and developers that allows them to retrieve news articles covering a large period. The dataset was a full retrieval of all available articles between 2015 and 2017. There was no search string involved and, thus, the corpus had no thematic delimitation. A total of 238,031 articles were acquired for our experiment.

In order to carry out the comparisons of the word-topic matrices φ as outlined above (i.e., reliability; reproducibility sample vs. full models; reproducibility pruned vs. unpruned models), we used a measure developed by Niekler and Jähnichen (2012). For each topic of a model j, the probability values of the $n = 20$ top words were compared to the probability values of each of the $n = 20$ top words of the topics in another model k. Two topics were regarded as a matched pair if their top word probability cosine

Table 1. Corpus Characteristics

	Website Corpus	Tweet Corpus	News Corpus
Type of communication	Issue-focused communication	Social media communication	News coverage
Timespan covered	2012-2014	Two weeks in 2018	2015-2017
Content heterogeneity	Medium	High	Medium
N documents	84,268	101,638	234,031
Vocabulary size			
Unpruned	98,526	125,731	167,527
Pruned	7,651	6,635	7,029
Terms per document			
Mean (SD)	601 (752)	392 (315)	427 (499)
Median	375	310	343

distance was minimal and less than $t = 0.5$. The fraction of matched topics was defined as the share of reproduced topics. To obtain confidence intervals for the share of reproduced topics, all possible model combinations were compared.[5]

Results

This section is divided into three subsections. First, in order to assess the stability of sample models, we look at their reliability.[6] The guiding question is how similar the topics of models are within the same (sample size or reference model) category. Reliability checks are needed because a certain degree of reliability is a necessary precondition for the sample models to resemble the topics of full corpus (reference) models. Second, we compare the sample model topics with the reference model topics (reproducibility sample vs. reference models). Third, we assess the deviance of pruned from unpruned model topics to see if they differ significantly (reproducibility pruned vs. unpruned models).

Reliability – How Reliable Are Topic Models from Sampled Documents?
The trajectories of the reliability values indicate a clear pattern (Figure 2). Models in the small sample size categories (1% and 5%) are much less stable than models in larger size categories (≥ 10%). The full corpus category (sample size = 100%) typically achieves the highest reliability values and should be taken as the reference point for the other models. Due to the stochasticity of the LDA process, however, reliability will never approach a value of 1, not even in the full corpus models. The trajectories of both pruned and unpruned models (across all corpora) show a rapid onset of saturation.

Beyond sample sizes of 10%, only marginal reliability gains are achieved. This result indicates, first and foremost, that (large enough) samples do not impair the reliability of a topic model.

In almost all cases, pruned models perform slightly better than their unpruned counterparts. This is a clear indication of the superior stability of topic models calculated on pruned corpora. The massive reduction in the vocabulary has a stabilizing effect on reliability mainly because the number of extremely infrequent, noisy words has no chance to confuse the topic compositions. With the exception of the news corpus, the small differences between pruned and unpruned models remain more or less constant across sample sizes, i.e., pruning has a constant but mostly insignificant effect on reliability.

Comparing the three corpora, it is noticeable that the news corpus achieves higher reliability values throughout all sample size categories, most strikingly in the small size categories (1% and 5%). This finding has two potential causes: First, the news corpus is the largest, comprising more than 230,000 documents. Therefore, even small samples cover a large absolute number of cases (1% $\triangleq$ 2,340 documents, 5% $\triangleq$ 11,702 documents). We, therefore, see the first evidence that the absolute number of documents may be more critical than the sample size relative to the full corpus, with sample models of fewer than roughly 10,000 documents suffering in reliability. Second, the content heterogeneity of documents is restricted. The news corpus only contains editorially processed media texts. This feature translates into a relatively homogenous writing style within comparatively clearly delimited thematic sections. Such characteristics are beneficial for topic models because the topics' top words tend to be highly specific, and overlap of top words is less likely.

Reproducibility – Do Sampled Models Resemble Topics of Full Models?

Topics of sampled models resemble the topics of the reference models if the sample size approaches a threshold value of ≥ 10% of the full corpus size (Figure 3), but no less than approximately 10,000 documents total. This pattern holds across all three corpora. It is unsurprising that the looming saturation of the curves corresponds with the findings presented in the previous subsection (Reliability). The obvious similarity of the reliability and reproducibility graphs (Figures 2 and 3) suggest that only reliable sample models are capable of resembling full models. However, there is also a notable difference between the curves: while there is only a marginal reliability difference between pruned and unpruned sample

models, pruned sample models exhibit significantly better reproducibility than unpruned models. As with reliability, this finding can mainly be traced back to the restricted vocabulary in the pruned sample and reference models. The large share of infrequently used, noisy words is eliminated through pruning. Consequently, the top word distributions of topics are less likely to be disturbed by intruder terms. Topics of pruned sample and reference models are likely to be more similar than their unpruned counterparts.

Notably, as compared to the other two corpora, the news corpus stands out because the share of reproducible topics is generally higher. Again, this finding has the same two potential causes as laid out in the Reliability subsection: (1) Due to the larger size of the corpus in absolute terms, small samples comprise a large number of documents. (2) Because of the professional editorial processing of media texts, content heterogeneity is restricted, and the top terms of topics tend to be highly specific.

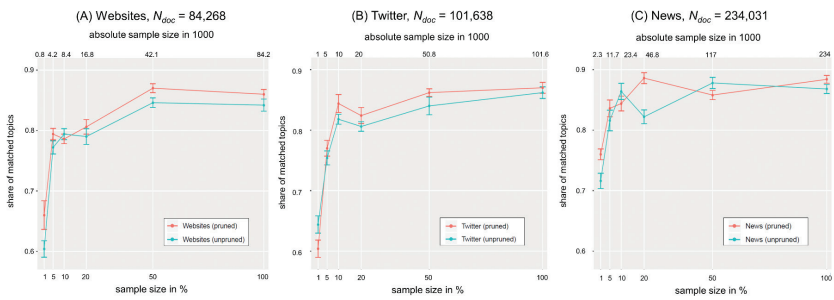

Figure 2. Reliability of topic models.

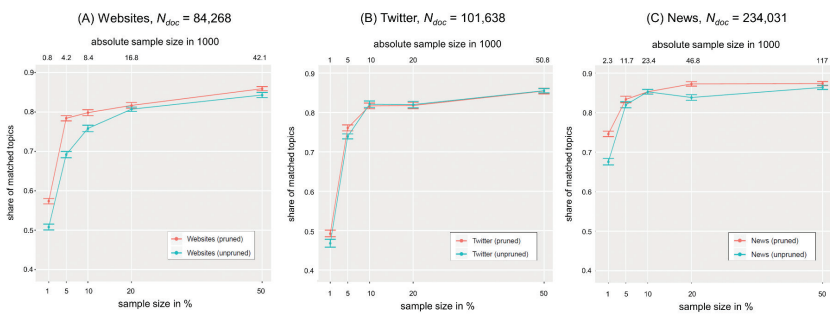

Figure 3. Reproducibility of reference models from sample models.

Pruning – How Different Are Pruned from Unpruned Models?

When comparing pruned and unpruned models, we find that the respective topics share great similarity. Figure 4 shows the share of matched topics between pruned and unpruned models of one kind (same corpus and same sample size category). The figure provides evidence for three aspects. First, there is again a clear saturation effect: beyond a certain sample size threshold, only marginal gains in the share of matched topics may be achieved. This means that pruned, small sample-size models are less similar to their unpruned counterparts. This makes sense as small sample-size models (pruned and unpruned) are less reliable than large sample-size models. Put differently, pruned or not, the topics of a topic model will not change considerably if the calculation is based on a large enough sample. This indicates that it is not necessary to take the vast majority of (infrequent or very frequent) terms into account. Extensive pruning boosts the performance of LDA as it massively reduces the dimensionality of the term-topic matrix (between 92.2% and 95.8% for our corpora, even with our conservative pruning approach).

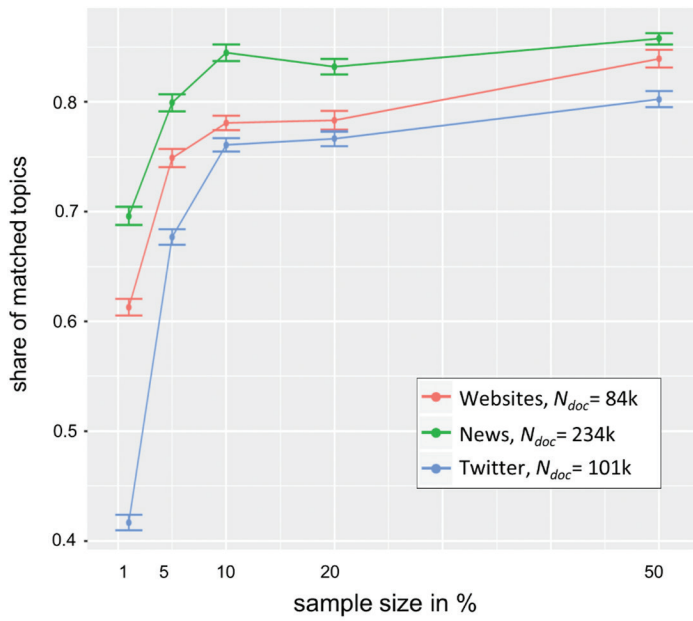

Figure 4. Reproducibility of unpruned models from pruned models.

Conclusion

In summary, we find that topic models may be calculated from sampled documents without impairing the resulting topics. It is important, however, to choose a sufficiently large sample size. If a sampled corpus contains too few documents, topic models will not be reliable, and unreliable topics will not resemble the topics of a full corpus topic model.

Our findings suggest that "large enough" should not be defined exclusively as a relative share of the overall number of documents, e.g., 10% of the documents. Instead, it depends on the absolute sample size. In our test cases, a sample of 10% was more than enough if the full corpus covered more than 230,000 documents (news corpus). For our smallest corpus, Websites, a 10% sample still yielded satisfactory results. If the corpus was slightly smaller (e.g., < 80,000 documents), a 10% sample could be borderline insufficient. To ensure the validity of document sampling, we suggest random samples of at least 10,000 documents or 10% to 20% of the corpus size if the total number of sampled documents is > 10,000 documents.

Our research also provides evidence that pruning does not impair the resulting topics. On the contrary, models based on pruned documents are more reliable than those based on unpruned documents. Pruned sample models resemble the topics of the full corpus models better. Moreover, topics from pruned sample models do not differ markedly from unpruned models.

Combining both approaches, or even using just one of them, drastically boosts the performance of the LDA algorithm and, therefore, makes topic modeling more feasible even with limited time and computational resources. Exemplary numbers for the Tweet corpus presented in Appendix A illustrate the size of this effect. Calculating models for the (unpruned) 10% sample of the full corpus takes about 12.5% of the time it takes to calculate a full corpus-model (29.9 vs. 238.6 minutes). Pruning the corpus vocabulary of the 10% sample model further decreases the required time to 18.8 minutes (62.9% of the time required for the unpruned model). Taken together, both recommendations, therefore, reduce the required computing time by 92.1% or, in absolute terms, 219.8 minutes (i.e., 3.7 hours).

The results presented in this paper will be beneficial to future topic modeling research. Topic modeling may be accelerated immensely if it is based on random document samples, and the vocabulary is pruned. We believe that sampling and pruning may help researchers to accelerate the otherwise lengthy and costly process of model selection (i.e., calculating

many topic models with differing parametrization and choosing the one that fits best). Additionally, our study provides evidence that a quick corpus exploration (using sampling and pruning) may now be done with confidence that the insights gained have value.

Although we used three text corpora with different characteristics for our inquiry, we cannot exhaustively rule out that diverging results might be observed for corpora created from other kinds of text sources. Corpora of political speeches or political manifestos, for example, usually feature single-issue documents, high term count per document, and long periods covered. For such corpora, samples of less than 10,000 documents may be sufficient, and pruning might also play out somewhat differently. Future applications should, thus, test the effect of sampling and pruning on their topic models for the case under study.

Funding Note

Work on this publication was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 290045248 – SFB1265.

Notes

1 Lack of exclusiveness and insufficient salience of words can also be countered with the application of top word-ordering strategies, such as the FREX ordering (Bischof & Airoldi, 2012). For an overview, see Sievert and Shirley (2014, p. 64-65).

2 LDA models are probabilistic by definition. Thus, their results do not feature deterministic robustness, i.e., two models for the same corpus with the same parametrization will yield slightly different results.

3 All models were calculated for the same parameter sets, i.e., $K = 50$ topics, $\alpha = 0.5$ and $\beta = 0.02$.

4 The word-topic matrix φ provides the conditional probabilities of the words by a given topic. The 20 most likely words of each topic are used for the topic interpretation.

5 For detailed information, visit https://github.com/danielmaier-fub/sampling_pruning_lda.

6 Calculating the reliability is not only a way to assess how well a model can be reproduced if calculated multiple times; in comparison with other corpora, it is also a way to assess which corpus characteristics impact the stability of the LDA process. Because the document samples are fixed (i.e., each model within one sample size category (and corpus) is calculated based on the exact same documents), one would expect the reliability to remain relatively constant across sample size categories. However, if the reliability

metric shows increasing trajectories across sample size categories, this is an indication that more documents in absolute numbers are required to stabilize the modeling process.

References

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning*, 201-208.

Blei, D., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3), 993-1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Bonilla, T., & Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*, *41*(6), 650-669. doi: 10.1016/j.poetic.2013.06.003

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168-189. doi: 10.1017/pan.2017.44

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1-35.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332-359. doi:10.1177/1077699016639231

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398-436.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the first ACM workshop on social media analytics*, 80-88.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229-247.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93-118. doi: 10.1080/19312458.2018.1430754

Manning, C. D., & Schütze, H. (2003). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Niekler, A. (2018). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen*. Köln: Herbert von Halem Verlag.

Niekler, A., & Jähnichen, P. (2012). Matching results of latent Dirichlet allocation for text. *Proceedings of 11th International Conference on Cognitive Modeling*, 317-322.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209-228.

Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings from the *Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, MD.

Waldherr, A., Maier, D., Miltner, P., & Günther, E. (2017). Big data, big noise: The challenge of finding issue networks on the web. *Social Science Computer Review*, *35*(4), 427-443.

About the authors

Daniel Maier, Freie Universität Berlin, Institute for Media and Communication Studies, Germany.
Correspondence address: Freie Universität Berlin, Institute for Media and Communication Studies, Garystr. 55, D-14195 Berlin (daniel.maier@fu-berlin.de)

Andreas Niekler, Universität Leipzig, Department for Computer Science, Germany

Gregor Wiedemann, Universität Hamburg, Department for Computer Science, Germany

Daniela Stoltenberg, Freie Universität Berlin, Institute for Media and Communication Studies, Germany.

Appendix A

Topic model processing time (in minutes) by sample size and pruning mode

Sample Size	1%	5%	10%	20%	50%	100%
Pruned	2.5	9.8	18.8	38.0	91.6	181.8
Unpruned	4.6	17.9	29.9	53.1	120.9	238.6

Note. Processing time in minutes for the Tweet corpus. Computer configuration CPU: Intel ® Core ™ i7-7820X (3.60GHz), Memory: 128 GB.