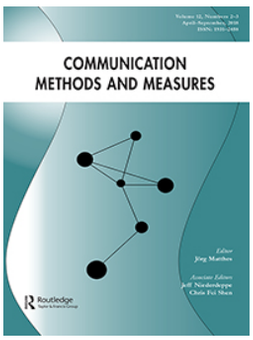


Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology

Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri & S. Adam

To cite this article: Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri & S. Adam (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

To link to this article: https://doi.org/10.1080/19312458.2018.1430754

COMMUNICATION METHODS AND MEASURES
2018, VOL. 12, NOS. 2–3, 93–118
https://doi.org/10.1080/19312458.2018.1430754

Routledge
Taylor & Francis Group

Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology

Daniel Maier[a], A. Waldherr[b], P. Miltner[a], G. Wiedemann[c], A. Niekler[c], A. Keinert[a], B. Pfetsch[a], G. Heyer[c], U. Reber[d], T. Häussler[d], H. Schmid-Petri[e], and S. Adam[d]

[a]Institute for Media and Communication Studies, Free University Berlin, Berlin, Germany; [b]Department of Communication, University of Münster, Münster, Germany; [c]Computer Science Institute, University of Leipzig, Leipzig, Germany; [d]Institute of Communication and Media Studies, University of Bern, Bern, Switzerland; [e]University of Passau, Passau, Germany

ABSTRACT

Latent Dirichlet allocation (LDA) topic models are increasingly being used in communication research. Yet, questions regarding reliability and validity of the approach have received little attention thus far. In applying LDA to textual data, researchers need to tackle at least four major challenges that affect these criteria: (a) appropriate pre-processing of the text collection; (b) adequate selection of model parameters, including the number of topics to be generated; (c) evaluation of the model's reliability; and (d) the process of validly interpreting the resulting topics. We review the research literature dealing with these questions and propose a methodology that approaches these challenges. Our overall goal is to make LDA topic modeling more accessible to communication researchers and to ensure compliance with disciplinary standards. Consequently, we develop a brief hands-on user guide for applying LDA topic modeling. We demonstrate the value of our approach with empirical data from an ongoing research project.

Introduction

Topic modeling with latent Dirichlet allocation (LDA) is a computational content-analysis technique that can be used to investigate the "hidden" thematic structure of a given collection of texts. The data-driven and computational nature of LDA makes it attractive for communication research because it allows for quickly and efficiently deriving the thematic structure of large amounts of text documents. It combines an inductive approach with quantitative measurements, making it particularly suitable for exploratory and descriptive analyses (Elgesem, Steskal, & Diakopoulos, 2015; Koltsova & Shcherbak, 2015).

Consequently, LDA topic models are increasingly being used in communication research. However, communication scholars have not yet developed good-practice guidance for the many challenges a user faces when applying LDA topic modeling. Important methodological decisions must be made that are rarely explained at length in application-focused studies. These decisions relate to at least four challenging questions: (a) How does one pre-process unstructured text data appropriately? (b) How does one select algorithm parameters appropriately, e.g., the number of topics to be generated? (c) How can one evaluate and, if necessary, improve reliability and interpretability of the model solution? (d) How can one validate the resulting topics?

These challenges particularly affect the approach's reliability and validity, both of which are core criteria for content analysis in communication research (Neuendorf, 2017), but they have, nevertheless, received little attention thus far. This article's aim is to provide a thorough review and

CONTACT Daniel Maier maier@zedat.fu-berlin.de Garystr. 55, 14195 Berlin, Germany

discussion of these challenges and to propose methods to ensure the validity and reliability of topic models. Such scrutiny is necessary to make LDA-based topic modeling more accessible and applicable for communication researchers.

This article is organized as follows. First, we briefly introduce the statistical background of LDA. Second, we review how the aforementioned questions are addressed in studies that have applied LDA in communication research. Third, drawing on knowledge from these studies and our experiences from an ongoing research project, we propose a good-practice approach that we apply to an empirical collection of 186,557 web documents. Our proposal comprises detailed explanations and novel solutions for the aforementioned questions, including a practical guide for users in communication research. In the concluding section, we briefly summarize how the core challenges of LDA topic modeling can be practically addressed by communication scholars in future research.

Statistical background of LDA topic modeling

LDA can be used to identify and describe latent thematic structures within collections of text documents (Blei, 2012). LDA is but one of several statistical algorithms that can be used for topic modeling; however, we are concentrating on LDA here as a general and widely used model. Blei, Ng, and Jordan (2003) introduced LDA as the first approach that allows for modeling of topic semantics entirely within the Bayesian statistical paradigm.

The application of LDA is based on three nested concepts: the text collection to be modelled is referred to as the *corpus*; one item within the corpus is a *document*, with words within a document called *terms*. Thus, documents are nested within the corpus, with terms nested within documents (see Figure 1, left side).

The aim of the LDA algorithm is to model a comprehensive representation of the corpus by inferring latent content variables, called *topics*. Regarding the level of analysis, topics are heuristically located on an intermediate level between the corpus and the documents and can be imagined as content-related categories, or clusters. A major advantage is that topics are inferred from a given collection without input from any prior knowledge. Since topics are hidden in the first place, no information about them is directly observable in the data. The LDA algorithm solves this problem by inferring topics from recurring patterns of word occurrence in documents.

In their seminal paper, Blei et al. (2003, p. 996) propose that documents can be "represented as random mixtures over latent topics, where each topic is characterized by a distribution over words." Speaking in statistical terms, the document collection (corpus) can equally be described as a distribution over the latent topics, in which each topic is a distribution over words. In linguistic theories, topics can be

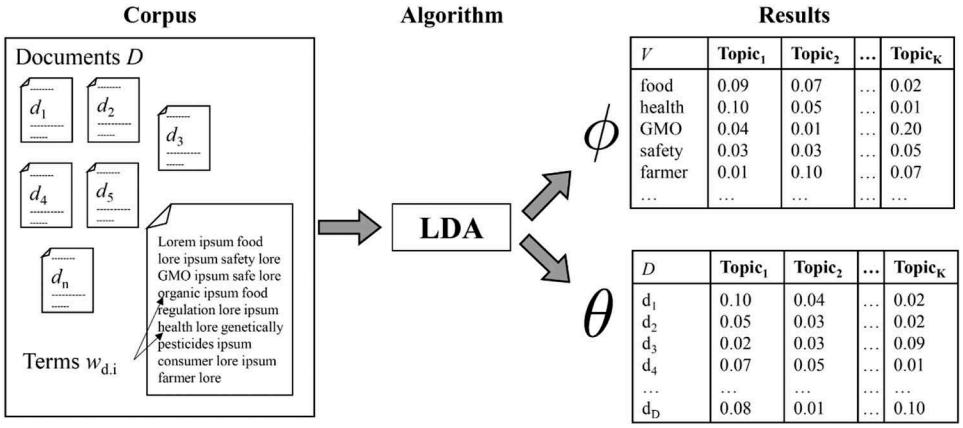


Figure 1. Application of LDA to a Corpus. *Note*. LDA = latent Dirichlet allocation.

seen as factors that consist of sets of words, and documents incorporate such factors with different weights (Lötscher, 1987). Topic models draw on the notion of distributional semantics (Turney & Pantel, 2010) and particularly make use of the so-called *bag of words* assumption, i.e., the ordering of words within each document is ignored. To grasp the thematic structure of a document, it is sufficient to describe its distribution of words (Grimmer & Stewart, 2013).

Although it appears fairly obvious what a *topic* is at first glance, there exists no clear-cut established definition of *topics* in communication research (Günther & Domahidi, 2017, p. 3057). Following Brown and Yule (1983, p. 73), Günther and Domahidi (2017, p. 3057) conclude that a "topic" can only vaguely be described as "what is being talked/written about". In the context of LDA topic modeling, the concept of a topic also takes on an intuitive and rather "abstract notion" of a topic (Blei et al., 2003, p. 995). However, what *topic* actually means in theoretical terms remains unclear. The meaning of a topic in an LDA topic model must be assessed empirically instead (Jacobi, Van Atteveldt, & Welbers, 2015, p. 91) and defined against the background of substantive theoretical concepts, such as "political issues" or "frames" (Maier, Waldherr, Miltner, Jähnichen, & Pfetsch, 2017).

LDA's core: the data-generating process

LDA relies on two matrices to define the latent topical structure: the word-topic assignment matrix ϕ and the document-topic assignment matrix θ (see Figure 1, right side). The word-topic assignment matrix ϕ has two dimensions, K and V, in which K is a numerical value defining the number of proposed topics in the model (which must be determined by the researcher), and V is the total number of words in the vocabulary of the corpus. Thus, any value of $\phi_{w,k}$ signifies the conditional probability with which the word $w = 1, \ldots, V$ is likely to occur in topic $k = 1, \ldots, K$. Analogously, θ has two dimensions, K and D, in which K, again, describes the number of proposed topics, and D is the number of documents in the corpus. Each value of $\theta_{d,k}$ discloses the conditional probability with which a topic k is likely to occur in a given document $d = 1, \ldots, D$ (see Figure 1, right side). In practice, the two resulting matrices are guiding the research process and enabling interpretation regarding content. For instance, from ϕ, researchers can identify the most salient, and thereby most characteristic, terms defining a topic, which facilitates the labeling and interpretation of topics. From θ, researchers can read the probability of the topics' appearance in specific documents; thus, documents may be coded for the presenceby Blei et al. (2003) of salient topics.

The computational core challenge is to estimate the two matrices, ϕ and θ. To master this challenge, Blei et al. (2003) designed a hypothetical statistical generative process within the Bayesian framework that tells us how documents are created and how words from unobserved topics find their way into certain places within a document.

Before we explicate this process, it is important to know that in Bayesian statistics, theoretically reasonable distributions are assigned to unknown variables, such as ϕ and θ. These distributions are called *prior distributions*, as they are assigned prior to data analysis and define their initial state. Here, two prior distributions are needed, one for ϕ and one for θ. LDA models use probability distributions from the Dirichlet family of distributions.[1] Each of the two Dirichlet priors is governed by the number of its dimensions K (the number of topics, which is equal for ϕ and θ) and an abstract (prior) parameter. As there are two prior distributions, there are also two prior parameters, which are sometimes also referred to as hyperparameters, i.e., α for θ and β for ϕ. In essence, α and β influence the shape and specificity of the word-topic and topic-document distributions. While the assignment of the prior parameters is included in the first two steps of the data-generating process, the remainder represents the stochastic core of the model.

What does the data-generating process by Blei et al. (2003) look like?

(1) We assume that each document, d, in a corpus can be described as a probability distribution over topics. This distribution, called θ_d (the topic distribution of document d), is drawn from a Dirichlet distribution with prior parameter α (which must be chosen by the researcher).

(2) Thus, each topic can be defined as a probability distribution over the entire corpus vocabulary, i.e., all the different words that appear in the documents. More technically, for each topic k, we draw ϕ_k, a distribution over the V words of the vocabulary from a Dirichlet distribution with prior parameter β (which must be chosen by the researcher).
(3) Within each document ($d = 1, \ldots, D$) and for every word in that document ($i = 1, \ldots, N_d$), in which i is the index count for each word in document d and N_d is the total length of d, we sample:
    (a) a topic ($z_{d,i}$) from the respective topic distribution in the document (θ_d), and
    (b) a word ($w_{d,i}$) from the respective topic's word distribution ϕ_k, in which k is $z_{d,i}$, the topic we sampled in the previous step.

The core concept of the model implies a statistical creation of a document as a process of randomly drawing topics (3a), then randomly drawing words associated with these topics (3b). This process has a crucial function: It explicates the dependency relationship between the observed variables (words in documents $w_{d,i}$) and the unobserved variables (word-topic distribution ϕ and document-topic distribution θ), thereby paving the way for the application of statistical inference (Griffiths & Steyvers, 2004).

Although the inference procedures cannot be addressed here in detail, it is essential to understand that the statistical theory sketches a joint-probability distribution of the observed *and* latent variables altogether (see Blei, 2012, pp. 79–80). From this joint-probability distribution, defined by the generative process, the *conditional probability distribution of the latent variables* ϕ and θ can be estimated (see Blei, 2012, pp. 79–80) using variational inference (Blei et al., 2013) or Gibbs sampling (see Griffiths & Steyvers, 2004). Therefore, for application on an empirical corpus, the algorithm makes use of the generative process and inverts the aforementioned steps. LDA starts with a random initialization, i.e., it randomly assigns term probabilities to topics (i.e., the initial state of ϕ) and topic probabilities to documents (i.e., the initial state of θ). The algorithm then aims to maximize joint likelihood of the model by iteratively adapting values of the word-topic distribution matrix ϕ and document-topic distribution matrix θ.

Advantages, limitations, and challenges of applying LDA

In summary, LDA models draw on an abstract hypothetical probabilistic process that implies different assumptions. It has proved to be a powerful approach to quickly identify major thematic clusters in large text corpora and model topics as latent structures in a text corpus. Compared with simple co-occurrence analysis, a topic model can reveal a latent semantic connection between words, even if they never actually occurred in a document together. Compared with other topic-clustering methods, a further advantage of LDA topic modeling is its *mixed membership* approach (Grimmer & Stewart, 2013, p. 18), i.e., one document can contain several topics, which is a useful assumption.

Another condition is the aforementioned bag-of-words assumption. In the context of topic modeling, it proves useful and efficient to explore global and general topic clusters in document collections, which is a frequent task in communication research. However, by discarding word order, specific local context information on semantic relations between words is lost, which otherwise might help interpret deeper meanings and solve ambiguities (Lenci, 2008, p. 21). Therefore, some researchers developed variations on topic modeling that consider word order (Wallach, 2006). Another limitation is that LDA assumes topics are independent of each other; thus, correlations between topics or hierarchical structures in terms of meta-topics and sub-topics are not part of the analysis. For this purpose, Blei and Lafferty (2007) developed the Correlated Topic Model (CTM), which also models relations between topics.

When applying LDA, it is important to keep in mind that the model results are not deterministic. Instead, the results are affected by the researcher's choices about the input parameters and the built-

in stochastic processes. Reliability and validity cannot be taken for granted. In the remainder of this article, we highlight four challenges with LDA topic modeling and propose guidelines as to how to deal with them.

(1) Before a topic model can even be estimated for an empirical corpus, the text collection must be sanitized of undesirable components and further pre-processed. Cleaning and pre-processing affect the input vocabulary and the documents included in the modeling process. Until now, little is known about the impact of preprocessing on reliability, interpretability, and validity of topic models. However, recent studies (e.g., Denny & Spirling, 2017) suggest that preprocessing strongly affects all these criteria. We provide suggestions on how text data can be cleaned, which pre-processing steps are reasonable to include, and in which order these steps should be applied.
(2) Three model parameters must be selected (K, α, and β), which affect the dimensions and *a priori* defined distribution of the target variables, ϕ and θ. All three parameters (i.e., K, α, and β) are of substantial importance for the resulting topic model. Thus, the selection of appropriate prior parameters and the number of topics is crucial to retrieve models that adequately reflect the data and can be meaningfully interpreted. Thus far, there is no statistical standard procedure to guide this selection; thus, this remains one of the most complicated tasks in the application of LDA topic modeling. Our proposal suggests a two-step approach: In the first step, the prior parameters are calibrated along the mean intrinsic coherence of the model, i.e., a metric focused on the interpretability (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) to find appropriate candidate models with different numbers for the K proposed topics. In the second step, a qualitative investigation of these candidates follows, which aims to match the models' results with the theoretical concept under study.
(3) The random initialization of the model and the sequence of multiple random processes are integral parts of LDA. The fact that topical contexts are manifested by combining certain words throughout multiple documents will guide the inference mechanism to assign similar topics to documents containing similar word distributions. Inference, itself, is also governed by stochastic random processes to approach a maximum joint probability of the model based on the evidence in the data. Due to both random initialization and stochastic inference, the results from topic models are not entirely deterministic. This calls for reliability checks that indicate the robustness of the topic solutions. We provide an easy-to-calculate reliability metric (Niekler & Jähnichen, 2012) and show that random initialization is a weakness in the LDA architecture. It is clearly inferior to non-random initialization methods, which, as we demonstrate, can improve the reliability of an LDA topic model.
(4) Most importantly, topics are latent variables composed of word distributions. We agree with DiMaggio, Nag, and Blei (2013, p. 586), who write "[P]roducing an interpretable solution is the beginning, not the end, of an analysis." To draw adequate conclusions, the interpretation of the latent variables must be substantially validated. We advise researchers to use systematically structured combinations of existing metrics and in-depth investigation to boost the significance of the validation process.

The four challenges are not independent of each other. Having a clean text corpus and finding a parameter setting that generates interpretable topics are important prerequisites for valid interpretation. Just as well, reliability of the topic solution is an essential precondition for validity.

Literature review

In this section, we systematically review how communication-related research has responded to these challenges so far. We performed keyword searches in *EBSCO Communication Source* and *Web of*

Science (SSCI).[2] The search yielded 61 unique results, which two authors classified as focusing on communication research or other fields of study. Articles were considered further if they applied the LDA algorithm and set out to answer a question of communication research, or used mass-communication data (e.g., newspaper articles, public comments, tweets). Some studies have a substantive thematic research focus, while many others referred to methodological issues. Of the latter studies, only those that demonstrate the application of topic modeling with a sample corpus were included in our review, while general descriptions and discussions of the method were ruled out (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Günther & Quandt, 2016).

We completed our retrieval of relevant and recent studies by checking *Google Scholar* and also revisiting basic literature on topic modeling (e.g., Blei, 2012; Blei et al., 2003). The final collection of research articles contained 20 publications in communication research (listed in Appendix A), with 12 studies focusing on the method and only 8 studies dealing with thematic research questions. We reviewed all 20 studies for solutions regarding their approach to (a) preprocessing, (b) parameter selection, (c) reliability, and (d) validity.

Data cleaning and preprocessing of unstructured text data

All studies under review addressed the issue of data cleaning and preprocessing, but they differed in the level of detail used to describe the process. The process of cleaning text data is contingent on the research question and the type of data used. For instance, if a study's focus is on one language only, a language filter is used (e.g., Parra et al., 2016). In the case of web documents or tweets, boilerplate content, such as uniform resource locators (URLs) or hypertext markup language (HTML) markups, need to be removed prior to data analysis (e.g., Ghosh & Guha, 2013; Parra et al., 2016). Other studies consider the aggregation of distinct text elements necessary to obtain larger documents. These mergers are necessary, either because the text elements are too short for LDA to extract substantive topics, as in the case of tweets (Guo, Vargo, Pan, Ding, & Ishwar, 2016, pp. 9–10), or to facilitate analysis, e.g., when comparing topics on a monthly basis (Puschmann & Scheffler, 2016).

The standard procedures of language pre-processing include tokenization (breaking documents down into term components), discarding punctuation and capitalization of words, filtering out stop-words and highly frequent and infrequent terms (relative pruning), and stemming and/or lemmatizing. Stemming and lemmatizing are used to make inflected words comparable to each other. While stemming reduces each word to its stem by stripping "its derivational and inflectional suffixes" (Lovins, 1968, p. 22) (e.g., "contaminating" and "contamination" become "contamin"), lemmatizing converts them to their lemma form/lexeme (e.g., "contaminating" and "contamination" become "contaminate") (Manning & Schütze, 2003, p. 132). Recent work suggests that not only the pre-processing procedures as such, but also their ordering, significantly influence the results of subsequent (supervised and unsupervised) text-analysis techniques, including topic modeling (Denny & Spirling, 2017). These findings are reasonable because the various pre-processing steps depend on each other.

Choosing the number of topics and prior parameters

When specifying a topic model, several parameters, such as the number of topics, K, must be defined. With this parameter, the granularity of the topic model can be adjusted. Generally, the more topics we accept, the more specific and narrow the resulting topics are. However, accepting too many topics might result in similar entities that cannot be distinguished in a meaningful way (e.g., Grimmer, 2010, pp. 12–13). At the same time, too few topics might lead to very broad entities combining different aspects that should be separated (Evans, 2014, p. 2).

To determine an adequate number of topics, researchers usually run several candidate models with varying numbers of topics. Subsequently, the resulting models are compared for significant differences and interpretability (e.g., Biel & Gatica-Perez, 2014; Elgesem et al., 2015). Since the

objective is to find substantive topics, this approach also has been termed a *substantive search* (Bonilla & Grimmer, 2013, p. 656). Because the overall goal is to generate a topic solution that can be validly interpreted, some researchers also draw on further external and internal validation criteria (discussed below) to choose between different candidate models (Baum, 2012; Evans, 2014).

There are also different metrics used to inform the process of model selection. The most widely applied is the measure of *perplexity* (used by, e.g., Ghosh & Guha, 2013; Jacobi et al., 2015). The perplexity metric is a measure used to determine the statistical goodness of fit of a topic model (Blei et al., 2003). Generally, it estimates how well a model produced for the major part of the corpus predicts a held-out smaller portion of the documents.

Another strategy is to run a non-parametric topic model, such as a Hierarchical Dirichlet Process (HDP) topic model (see Teh, Jordan, Beal, & Blei, 2006) in which K does not need to be defined in advance. Instead, a statistically appropriate number of topics is estimated from the data (Bonilla & Grimmer, 2013). However, for such a model, other even more abstract parameters must be defined in advance, so that the problem about the model's granularity is not solved, but merely shifted to yet another parameter.

The choice of the prior parameters α and β is rarely discussed in current studies. Ghosh and Guha (2013) apply default values that are set in the *R topicmodels* package by Grün and Hornik (2011). Biel and Gatica-Perez (2014) refer to standard values proposed by Blei et al. (2003). Evans (2014) uses an optimization procedure offered by the *MALLET* software package (McCallum, 2002) to iteratively optimize the Dirichlet parameter for each topic at regular intervals.

Reliability of topic solution

While reliability is usually not regarded as a major concern with computer-based content-analysis techniques, the random processes in the LDA algorithm make robustness in the sense of retest reliability of a topic model an important issue. However, few researchers ensure that the obtained topics are robust across multiple runs of the model, with the same parameter set (but different random seeds) (DiMaggio et al., 2013; Levy & Franklin, 2014). More researchers are examining whether the identified topics are reproducible across several runs of the topic model with different parameters, most often varying the number of topics (e.g., Levy & Franklin, 2014; Van Atteveldt, Welbers, Jacobi, & Vliegenthart, 2014). Biel and Gatica-Perez (2014) have checked whether they can replicate the model's topics with smaller samples of the dataset.

Topic interpretation and validity

The most straightforward approach of most studies regarding valid interpretation of the resulting topics is to review the words with the highest probabilities for each topic (top words) and try to find a label describing the substantive content of the topic. Often, researchers also read through a sample of documents featuring high proportions of the respective topic (e.g., Elgesem, Feinerer, & Steskal, 2016; Jacobi et al., 2015; Koltsova & Shcherbak, 2015). These strategies are applied to ensure *intra-topic semantic validity* of topics as the most crucial aspect of semantic validity (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010).

Additionally, some researchers use quantitative diagnostic metrics, such as topic coherence (e.g., Evans, 2014) or mutual information measures (e.g., DiMaggio et al., 2013). While (intrinsic) topic coherence measures how frequently the top words of a topic co-occur (Mimno et al., 2011), mutual information aims to identify which of the top words contributes the most significant information to a given topic (e.g., Grimmer, 2010). To ascertain whether topics are sufficiently distinct from each other (inter-topic validity) or to find patterns of semantics among topics, hierarchical clustering can be applied (e.g., Marshall, 2013; Puschmann & Scheffler, 2016).

In various studies, we also noticed strategies for external validation. External criteria can include expert evaluations (Levy & Franklin, 2014), manual codings, and code systems (e.g., Guo et al., 2016;

Jacobi et al., 2015). Some studies also checked whether the temporal patterns of topics corresponded with events that occurred in the study's time frame (e.g., Evans, 2014; Newman, Chemudugunta, Smyth, & Steyvers, 2006).

Summarizing our review, we agree with Koltsova and Koltcov (2013, p. 214) that "the evaluation of topic models is a new and still underdeveloped area of inquiry." While in the past few years, a range of strategies for testing the validity of topic models has been established, a standard methodology for ensuring the reliability of the topics has yet to be developed in communication research.

Toward a Valid and reliable methodology for LDA topic modeling

In this section, we propose our methodological approach to topic modeling with respect to cleaning and preprocessing, model selection, reliability, and valid interpretation of identified topics. We illustrate the soundness of our approach by using empirical data from an ongoing research project in which we investigate online communication of civil-society actors concerning the issue of food safety. The theory we drew on originates from political agenda-building research (Cobb & Elder, 1983). Hence, we are interested in exploring the spectrum of "political issues" discussed by civil-society organizations concerned about food safety on the Web. In political communication, the term "issue" is used to denote a contentious matter of dispute, with the potential of "groups taking opposing positions" (Miller & Riechert, 2001, p. 108).

Building and preprocessing the corpus

To identify websites on the Internet that are concerned with the issue of food safety, we collected hyperlink networks, i.e., websites connected by hyperlinks, on a monthly basis from June 2012 to November 2014 (30 months), starting with eight websites involving U.S.-based civil-society actors.[3] The networks were collected using the web-based software *Issue Crawler*.[4] Altogether, 575,849 webpage documents were identified in these networks, of which—for both technical and practical reasons—we downloaded only those pages that included (a combination of) issue-specific search terms (see Waldherr, Maier, Miltner, & Günther, 2017, p. 434), resulting in a collection of 344,456 webpages.

The web-crawling procedure resulted in a heterogeneously structured set of webpages. Since we were interested in analyzing substantive text only, the crawled webpages had to be further processed to remove so-called boilerplate content, such as navigation bars, page markups, ads, teasers, and other items regarded as irrelevant.

In the first step, we deleted the HTML-markups using the content-extraction library *Apache Tika*. Second, the text files were passed through the *openNLP* toolkit for sentence separation. The text of each page was separated into sentence candidates temporarily stored in separate lines. So far, candidates included navigation elements, teasers, or copyright information. We filtered out the boilerplate text and selected only valid sentences among all sentences on each page with a rule-based approach using *regular expressions* (see Manning & Schütze, 2003, p. 121).

The resulting main texts from each webpage were classified further using a language-detection algorithm to distinguish between documents written in English or German (the project languages), and other languages. Language detection was necessary for subsequent pre-processing steps. Since removal of boilerplate content from pages could reveal that an extracted document was not thematically relevant for our analysis, we filtered again for relevant content by only including those documents containing the (combination of) issue-specific key terms.These procedures resulted in a massive reduction of content. The final corpus included 186,557 documents stored in a database for further analysis.

In the final step, we ran a duplicate detection algorithm (Rajaraman & Ullman, 2011) on the filtered document set to identify near-duplicates in very large datasets efficiently. Documents were marked as duplicates if their similarity, defined by the Jaccard index on their word set, was above a

threshold of .95. For each duplicate, a reference to the first occurrence of that document was stored to allow for queries, including or excluding duplicates in the resulting set. Altogether, 87,692 documents were marked as being unique.

Generally speaking, we deem rigorous data cleaning to be necessary and suggest that text documents should be relieved of boilerplate content, such as ads, side bars, and links to related content. If boilerplate content either is not randomly distributed across all the documents in the corpus—which would be a naive assumption for most empirical corpora—or the documents are not cleaned extensively enough, the LDA algorithm could be distorted and uninterpretable, as messy topics could emerge.

Corpus cleaning is only the first step. Automated content-analysis procedures, such as topic modeling, need further specific preprocessing of textual data. "Preprocessing text strips out information, in addition to reducing complexity, but experience in this literature is that the trade-off is well worth it" (Hopkins & King, 2010, p. 223). As we pointed out in the literature review, many LDA studies have reported using a range of seemingly standard pre-processing rules. However, most studies fail to emphasize that these consecutively applied rules depend on each other, which implies that their ordering matters (see also Denny & Spirling, 2017). Although a single correct preprocessing chain cannot be defined, the literature provides reasons for proceeding in a specific order.

Thus, we suggest that after data cleaning, the documents should be divided into units, usually word units, called tokens. Hence, this step is called tokenization (Manning & Schütze, 2003, p. 124). After tokenization, all capital letters should be converted to lowercase, which should be applied for the purpose of term unification. After that, punctuation and special characters (e.g., periods, commas, exclamation points, ampersands, white-space, etc.) should be deleted. While punctuation may bear important semantic information for human readers of a text, it is usually regarded as undesirable and uninformative in automatic text analyses based on the bag-of-words approach (e.g., Scott & Matwin, 1999, p. 379). However, following Denny and Spirling (2017, p. 6), some special characters, such as the hashtag character, might be informative in specific contexts, e.g., modeling a corpus of tweets, and should be kept in such cases. The next step is to remove stop-words, which are usually functional words such as prepositions or articles. Their removal is reasonable because they appear frequently and are "insufficiently specific to represent document content" (Salton, 1991, p. 976). While lowercasing and removal of punctuation and special characters can be done in any order after tokenization, they should be done before the removal of stop-words to reduce the risk that stop-word dictionaries may be unable to detect stop-words in the corpus vocabulary. Unification procedures, such as lemmatization and stemming, should be used only after stop-word removal. As mentioned above, both techniques are used for the purpose of reducing inflected forms and "sometimes derivationally related forms of a word to a common base form" (Manning, Raghavan, & Schütze, 2009, p. 32). However, we prefer lemmatization over stemming because stemming "commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma" (Manning et al., 2009, p. 32). Thus, interpreting word stems correctly can be tough, or even impossible. For example, while the word *organized* is reduced to its stem, *organ*, its lemma is *organize*.

In the very last step, relative pruning should be applied. Due to language-distribution characteristics, we can expect a vast share of very infrequent words in the vocabulary of a collection. In fact, roughly half of the terms of the vocabulary occur only once (Zipf's Law, e.g., Manning & Schütze, 2003, pp. 23–29). Thus, relative pruning is recommended to strip very rare and extremely frequent word occurrences from the observed data. Moreover, relative pruning reduces the size of the corpus vocabulary, which will enhance the algorithm's performance remarkably (Denny & Spirling, 2017) and will stabilize LDA's stochastic inference. In our empirical study, relative pruning was applied, removing all terms that occurred in more than 99% or less than .5% of all documents (Denny & Spirling, 2017; Grimmer, 2010; Grimmer & Stewart, 2013).

If the unification of inflected words is not applied before relative pruning, chances are high that semantically similar terms such as *genetic* and *genetically* will be part of the vocabulary, i.e., if a user

complies with the suggested ordering, the corpus vocabulary will be reduced, while still maintaining a great diversity of substantively different words. In our empirical case, we followed the proposed ordering of the pre-processing steps.

Model selection: reliability issues and choosing appropriate parameters

Model selection is the process of determining a model's parameters, i.e., the number of topics, K, and the prior parameters, α and β. The objective of this process is to find the parameter configuration that leads to the most appropriate model available for the data and the research interest alike. Evaluating how well a model fits the data and whether it appropriately serves its purpose always should be guided by a study's research question and the theoretical concepts of interest. We note that communication researchers working with content data generally aim to gain knowledge about the content and its substantive meaning. A topic model provides information about both, but the quality of the information depends on how well human researchers can interpret the model with respect to theory. Thus, interpretability must be regarded as a necessary precondition for a model's validity. Hence, we argue that the interpretability of the modeled topics should be the prime criterion in the model-selection phase. However, a parameter configuration that leads to interpretable solutions is worthless if it cannot be replicated. From this perspective, interpretability and reliability are intertwined and directly related to a model's validity.

In this section, we first introduce two metrics, reliability and intrinsic coherence, which enable users to provide information about the quality of a topic model. To enhance both criteria right away, the topic-modeling literature puts forth techniques that have been discussed under the term *regularization*. We briefly discuss the findings of the regularization literature in the second part of the section and provide an easy-to-implement regularization technique to boosting the reliability of topic models. We confirm this approach by providing evidence from experiments we conducted. The final part of the section concentrates on selecting the most appropriate model using what we call *substantive search in coherence-optimized candidates*.

Measuring reliability and interpretability of topic models

Reliability of a topic model can be measured in different ways. We implemented an approach following the intuition of comparing two models, i and j, for their similarities. For each topic from model i, the probability values of the N topics' top words were compared with the probabilities of each of the N topics' top words from all topics in model j. Two topics, one from each model, were counted as a matched pair if the cosine similarity of their top-word probabilities was at a maximum *and* above a defined threshold ($t = .7$). The proportion of topic matches from models i and j over all K topics was defined as a reliability score (Niekler, 2016). Reliability between more than two models can be computed as an average between all model pairs.

Regarding a model's quality in terms of interpretability, multiple metrics are available. The most frequently used statistical measures are *held-out likelihood* or *perplexity* (Blei et al., 2003). For their application, a model needs to be computed on one (major) part of a collection, e.g., 90% of all documents, then applied to the (smaller) 10% of collection documents not included in the modeling process. The model's goodness of fit (likelihood) is estimated by how well the model predicts the held-out smaller portion of the documents. Higher likelihood corresponds to a lower perplexity measure.

A method of systematic manual evaluation has been proposed by Chang, Boyd-Graber, Gerrish, Wang, and Blei (2009). For a tested topic, they used the list of the top N terms of a fitted model and inserted a random term with high probability from another topic of that model. If human subjects (users) can identify this false intruder, the topic may be considered coherent. Surprisingly, the study demonstrated in a large user study that the widely used evaluation metrics based on perplexity do not correspond well with human results of intrusion detection, and in some cases, they are even negatively correlated. Also, LDA variants—such as correlated topic models (CTM) (Blei & Lafferty, 2006), which reportedly achieve a higher model likelihood—turned out to be less coherent.

In response to these findings, topic-coherence measures were proposed based on the assumption that the more frequently top words of a single topic co-occur in documents, the more coherent the topic. Studies have shown that coherence measured with respect to data that is external (Newman, Lau, Grieser, & Baldwin, 2010) or internal to the corpus (Mimno et al., 2011) correlates with human judgment on topic interpretability. The latter is also referred to as intrinsic coherence.[5]

For both interpretability and reliability, different *regularization* techniques have been tested. In this regard, regularization of topic models describes a process that helps mitigate ill-posed mathematical problems and guides them toward a more favorable solution.

Enhancing interpretability and reliability with regularization techniques

The seminal model proposed by Blei et al. (2003) is based on the idea that the clustering effect of the algorithm works well, even if the initial assignments for θ and ϕ are set completely at random. Although the generative model consists of successive random processes, in theory, many allocation iterations will lead to similar models because the allocations depend on distributions dominated by the data. However, experiments conducted by Lancichinetti et al. (2015), and Roberts, Stewart, and Tingley (2016) point to serious issues of topic models regarding reliability.

While interpretability of topic models has been extensively studied, reliability has been a much less discussed issue thus far. Hence, we distinguish between approaches that raise the interpretability of a model and approaches that aim at higher topic reliability among repeated inferences on the same data.

For the issue of interpretability, two branches of research can be identified. The first branch develops regularization techniques that alter the inference scheme of the original LDA model (Newman, Bonilla, & Buntine, 2011; Sokolov & Bogolubsky, 2015). The second branch of regularization techniques solely alters the initialization of the model to guide the inference process toward a desired local optimum. For instance, word co-occurrence statistics are used in conjunction with clustering techniques to assign words to semantic clusters for initializing the model (e.g., Newman et al., 2011; Sokolov & Bogolubsky, 2015). Without exception, all these studies demonstrate a positive effect from regularization strategy on topic interpretability.

Regarding reliability, only a few studies are available that propose improving strategies. Reliability problems can emerge from two model settings: (a) random initialization of the two result matrices and (b) successive random processes. For the latter, Koltcov, Nikolenko, Koltsova, Filippov, and Bodrunova (2016) introduce a slight variation to the LDA Gibbs sampler as originally proposed by Griffiths and Steyvers (2004). When drawing a topic for a word, they force the neighboring words into the same topic. This results not only in better coherence, but also in higher reliability. Unfortunately, there is no publicly available implementation for this approach. Alternatively, Lancichinetti et al. (2015) extract K semantic term clusters based on word co-occurrence statistics to initialize the LDA model's K topics. They show that this procedure leads to perfect reproducibility of the topic model when running the inference process for one iteration after initialization.

In acknowledgement of this research, we aim for a solution that raises both interpretability and reliability. Moreover, we prefer to rely on freely available, well-established implementations of the original LDA model. Therefore, we opt for a regularization strategy that is compatible and relatively easy to implement, namely an initialization strategy in which semantically pre-clustered terms are provided as an input to the inference algorithm. In accordance with Roberts et al. (2016), who expect "advances in areas such as optimal initialization strategies," we decided to refine the Lancichinetti et al. (2015) idea. The major drawbacks of their approach are that they use an artificial corpus and run Gibbs sampling for only one single iteration after initialization. Although this leads to perfect reliability, the effect on interpretability remained untested. We assume that not running multiple iterations of sampling has a severe negative influence on topic quality in real-world applications. Therefore, we conducted an experiment in which we evaluate the effects of different initialization strategies across a varying number of inference iterations with respect to reliability and coherence as measures for model quality. To examine whether our findings generalize across corpora and topic resolutions, we ran the

test for three different corpora (food-safety-related content from Germany, the U.K., and the U.S., which is the focal corpus of the empirical study), with different topic numbers K.[6]

As a baseline strategy, we tested the standard random initialization of LDA. As a second strategy, we fixed the random initialization with a specific seed value, but afterward, we reset the random-number generator. We ran this experiment to test the influence of random sampling during the inference algorithm, independent of initialization. As our own third strategy, we proposed a modification of clustered initialization from Lancichinetti et al. (2015). We also initialized the topics based on term-co-occurrence networks. In contrast to the original approach, which was tested on two highly artificial text collections, we observed that their proposed combination of significance measure (Poisson) and clustering algorithm (Infomap) does not perform well on real-world data to identify coherent semantic clusters. Thus, we selected alternatives to achieve a better pre-clustering of terms. For determining co-occurrence significance, we relied on Dunning's Log-Likelihood Ratio Test (LL) (Bordag, 2008). Subsequent semantic community detection is performed by applying the Partitioning-Around-Medoids (PAM) algorithm (see Kaufman & Rousseeuw, 1990).

Each experiment was repeated $n = 10$ times. Figure 2 displays the average reliability of the experiments over the progress of Gibbs sampling iterations. Confidence intervals for reliability are provided on the basis of $\frac{n*(n-1)}{2} = 45$ possible pairs for comparing models i and j. The results indicate that our cluster-initialization strategy significantly improves the reliability of the inference for all three corpora and leads to levels of reproducibility above 85% for the German and U.K. corpus, and above 75% for the U.S. corpus. The seeded initialization also outperforms the random standard initialization, but does not reach the performance of an initialization by semantic network clustering. From this result, we conclude that the stability of the inference algorithm itself actually can be quite high once it starts from the same position. We further conclude that providing semantic clusters of terms as a starting position leads to even more stable results in the inference process, thereby indicating why it is the preferred strategy to improve reliability.

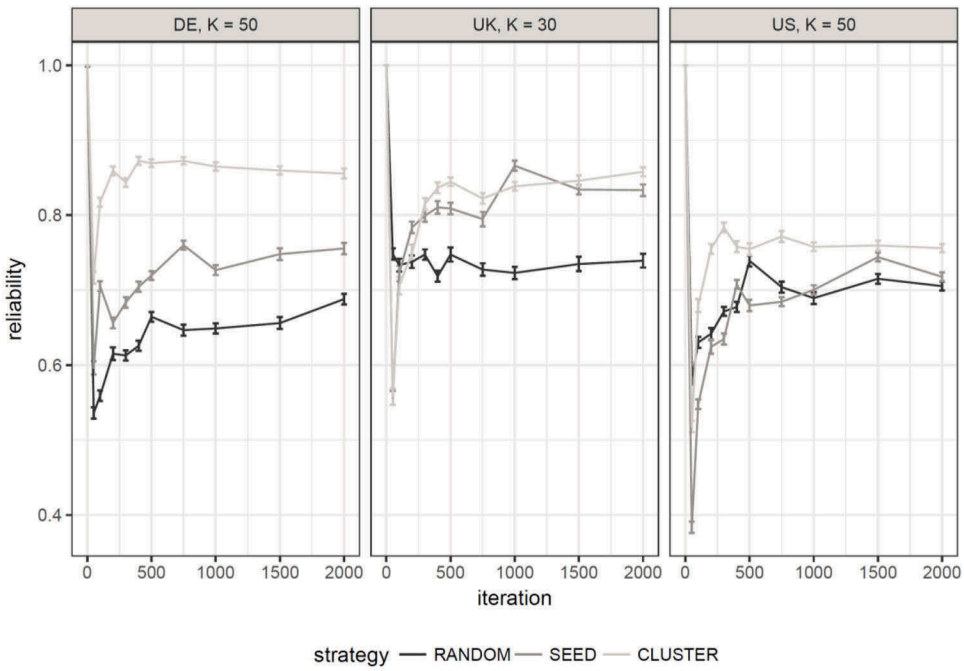


Figure 2. Reliability of topic models for three corpora (DE = Germany; UK = United Kingdom; US = United States) according to different initialization techniques (random = default random initialization; seed = fixed seed initialization; and cluster = semantic co-occurrence network initialization) and varying number of inference iterations; K = number of topics.

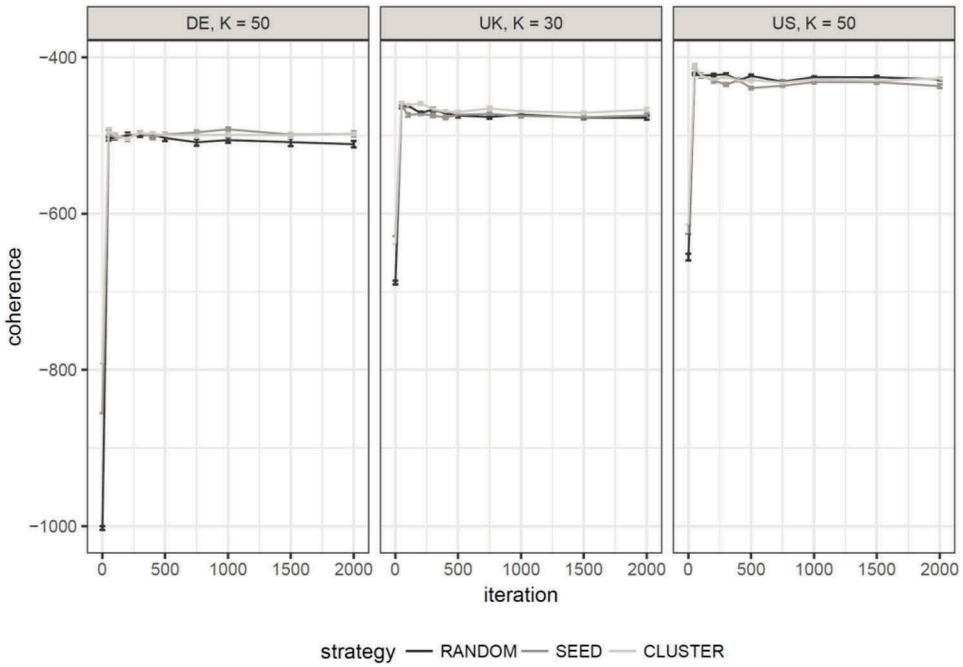


Figure 3. Mean coherence of topic models for three corpora (DE = Germany, UK = United Kingdom, US = United States) according to different initialization techniques (random = default random initialization; seed = fixed seed initialization; and cluster = semantic co-occurrence network initialization) and varying number of inference iterations; *K* = number of topics.

Figure 3 displays the average topic-model coherences of the ten repeated runs of our experiment, including their confidence intervals. Compared with the reliability check, the results are rather mixed. Although the cluster initialization usually performs very well, differences between all the strategies are not very pronounced. The most important finding from this part of the experiment is that topic coherence is drastically lowered if sampling runs for only one iteration. Although it guarantees perfect reliability, the results of such an early stopped process cannot be used in a practical scenario. We conclude that to further improve interpretability, the process also needs to run for some time until the topic composition stabilizes. We recommend at least 1,000 iterations. Running only one iteration, as proposed in Lancichinetti et al. (2015), trades reliability for interpretability and appears to be a bad choice in practical scenarios.

Selecting the model: substantive search in coherence-optimized candidate models

In general, finding the *optimal* parameter set is not an easy task in an unsupervised, data-driven scenario. There is no gold standard to evaluate model results, as in a supervised scenario, and the best solution cannot be ensured by a single criterion independent of the research interest. The literature on natural language processing (NLP) provides various methods and evaluation metrics for topic models that can be utilized to find the optimal parameters. But it is still highly likely that solutions optimized along single metrics do not comply with the analytical requirements in communication research, such as the desired topic granularity necessary to obtain meaningful results. For this reason, we suggest avoiding the use of only one numerical optimization procedure for parameter selection, and instead combine different measures with intersubjective qualitative human judgment.

Like the procedure described by Marshall (2013, p. 709), we applied a systematic approach for the choice of the number of topics, *K*, and the prior parameters, α and β. Instead of using default suggestions, which often do not yield optimal results, we systematically varied different combinations for *K* (30, 50, and 70) and α (.01, .05, .1, .2, .5, 1). As the combinatorial set expands with the

number of parameters included, we fixed the value of β at $1/K$, the default value as proposed by the widely used topic model library *gensim* (Řehůřek & Sojka, 2010). The prior for the topic-document matrix α was found to be of greater importance for the quality of the topic model (e.g., Wallach, Mimno, & McCallum, 2009), which was the reason to fix β and let α vary. The model was run with 1,000 iterations. We calculated 6 different models (i.e., all possible combinations of α) for each of the three values in K (resulting in 18 models, see Appendix B) and chose the single best model for each K regarding the mean intrinsic topic coherence for further investigation. We refer to these three models as our candidate models.

Instead of using the whole corpus for the model creation, in this phase, we took a random sample of 10,000 non-duplicate documents (out of 87,692 unique documents) to calculate these models. Whether a document sample is representative "depends on the extent to which it includes the range of linguistic distributions in the population" (Biber, 1993, p. 243). Thus, for topic-modeling purposes, a valid sample must catch the variety of word co-occurrence structures in the document population. Random sampling can be regarded a valid procedure for topic modeling of very large document collections. Due to the characteristic distribution of language data, we can expect a huge share of very infrequent words in the vocabulary of a collection. This is also the reason why the pruning of infrequent vocabulary is a recommended and valid pre-processing step. In other words, applying relative pruning to the full corpus yields a very similar vocabulary, as would applying relative pruning to a random sample of 10% of the corpus. In both cases, document content is reduced to a very similar vocabulary. Thus, it is reasonable to expect that co-occurrence structures of these terms in a large-enough random sample would be very similar to those in the entire corpus.

Still, the size of the sample must be big enough to draw valid conclusions about which parameter configurations yielded solid models. Scholars from corpus linguistics (e.g., Hanks, 2012) argue that sample size is the most important criterion to consider in covering the thematic diversity of the corpus. As a rule of thumb for domain-specific corpora, we recommend using at least a two-digit fraction (10% minimum) of the overall corpus size. In our empirical case, we drew a random sample of 10,000 documents, or 11.4% (10,000/87,692) of the document total. However, it is important to note that it cannot be guaranteed that this technique will work well for corpora containing significantly smaller sized and/or more heterogeneous documents. In our view, the validity of this technique crucially depends on whether the sample size is big enough to capture the heterogeneity of the corpus vocabulary. In this regard, future research needs to figure out valid guidelines for sampling strategies and sample sizes.

The 10,000 sampled documents are used only for purposes of model creation and selection. Inference is conducted for the complete corpus. The separation of model creation and inference enables us to directly use the model that we created on the basis of the random sample and successively infer the topic composition of the remaining documents.

A group of four researchers discussed the three best topic models in terms of their mean coherence metric, one for each value in K. For the collaborative investigation of the three models, the LDA visualization software *LDAvis* was used (Sievert & Shirley, 2014). The question that was guiding the qualitative investigation of the group was: *Which topic model most suitably represents the contentious matters of dispute, i.e., the "issues," of the food-safety discourse in civil society on the Web?* The discussion and interpretation were based on the models' ϕ matrices, i.e., word-topic distributions, and also considered varying orders of the top words using Sievert and Shirley (2014) relevance metric (explained in the next section). The group discussion led to a consensus within the research group. The model with $K = 50$ offered the most reasonable topic solution to interpret the theoretical concept of "political issues," which was the focus of our research. While setting $K = 70$ led to too many topics that could easily be traced back to arguments put forth by single websites, minor events, or remaining boilerplate, $K = 30$ obfuscated and blurred issues that would otherwise be treated separately by the research group. We decided in favor of the model with the parameters $K = 50$, $\alpha = .5$, and $\beta = 1/K = .02$. This solution deserved further investigation in validity checks.

Topic validity and labeling

We regard interpretability as a necessary, but not a sufficient prerequisite for validity. With some exceptions (e.g., DiMaggio et al., 2013), interpretation, validity, and successive labeling of topics become blended and blurred in application-focused studies. We want to gain awareness that good interpretability of a topic's top-word list is not equivalent to its validity. Referring to Neuendorf (2017, p. 122), "validity is the extent to which a measuring procedure represents the intended – and only the intended – concept." To uncover whether the modeled topics represent the concept under study, such as the issue-concept, we developed a three-step procedure. First, we summarized the most important quantitative information from the model. Second, all topic models created for non-artificial text corpora will contain a fraction of uninterpretable topics, which cannot be valid by definition and thus must be excluded. The third step is an in-depth investigation that includes a close reading of documents and the labeling of the topics.

Summarizing topics

To summarize the topics, we used several auxiliary metrics to better understand the semantics of the topics' word distributions. Specifically, we used the following four metrics.

(1) *Rank-1*: The Rank-1 metric (see Evans, 2014) counts how many times a topic is the most prevalent in a document. Thus, the metric can help identify so-called background topics, which usually contribute much to the whole model, but their word distribution is not very specific. In the case of a high topic share in the entire collection being accompanied by a low Rank-1 value, we can make a reasonable guess that a topic occurs in many documents, but rarely can be found as the dominant topic of a document. The empirical example presented below contains several background topics, such as *economy, politics*, and *health care*, all of which constitute the setting in which the food-safety debate among civil-society actors takes place.

(2) *Coherence*: This metric, developed by Mimno et al. (2011, p. 264), already was used for model-selection purposes. However, applied to single topics, it also helps guide intuition and may help identify true topics in which a researcher might not see a coherent concept at first glance.

(3) *Relevance*: The word distributions within any topic of the model are based on the word probabilities conditioned on topics. However, provided that a given word, e.g., *food*, occurs frequently in many documents, it is likely to have a high conditional probability in many topics and thereby occurs frequently as a top-word. In this case, such a word does not contribute much to the specific semantics of a given topic. Sievert and Shirley (2014, pp. 66–67) developed the so-called *relevance metric*, which is used to reorder the top words of a topic by considering their overall corpus frequency. The researchers can decide how much weight should be ascribed to corpus frequencies of words by manipulating the weighting parameter λ, which can have values ranging from 0 to 1. For $\lambda = 1$, the ordering of the top words is equal to the ordering of the standard conditional word probabilities. For λ close to zero, the most specific words of the topic will lead the list of top words. In their case study, Sievert and Shirley (2014, p. 67) found the best interpretability of topics using a λ-value close to .6, which we adopted for our own case.

(4) Sources and *concentration*: In our empirical dataset, we selected sources of topics by asking which websites were promoting certain topics and how much a topic was concentrated in the potential sources. Therefore, we assessed the average source distribution of topics by computing the Hirschman-Herfindahl Index (HHI) as a *concentration* measure. The HHI ranges from 1/*number of sources* to 1. An HHI = 1 signifies maximum concentration, i.e., the topic is pronounced by only one source. A very low HHI value, conversely, indicates that a topic can be found in many sources.

For the interpretation of our topics, we summarized the aforementioned metrics on a single overview sheet, one for each topic in the model (see Appendix C for an example topic).

Exclusion of topics

After summarizing the topics in this manner, two researchers reviewed all the topic sheets independently from each other. By relying on both the metrics and their expert knowledge about food safety, they (independently) judged whether the topics should still be included for further investigation or not. More specifically, topics whose top-word lists were hard to interpret and which came with low values in Rank-1 and coherence while showing low prevalence and high concentration were excluded. If one author had judged that a topic deserved in-depth investigation, the topic was kept. In the case that both authors came to the conclusion that a topic should be discarded it was discarded. In other words, we kept a topic if there was at least one indication that it contained a meaningful, coherent concept.

Another peculiarity of topic models are boilerplate topics. Although we extensively cleaned the corpus (see the *Building and preprocessing the corpus* section), boilerplate content still showed up in some topics. Boilerplate topics are common phenomena in topic models (Mimno & Blei, 2011). They have no substantive meaning, but their emergence sharpens other meaningful topics "by segregating boilerplate terms in a distinct location" (DiMaggio et al., 2013, p. 586). Most often, the identified boilerplate topics coincide with the most unreliable and least-salient topics (see also Mimno et al., 2011).

After discussing the results of the separate investigations we made a consensual decision using the aforementioned criteria. The authors decided that 13 topics should be removed because they showed no indication of being either meaningful or coherent. The remaining 37 topics were subject to the final validation and labeling step.

In-depth validation of topics and topic labeling

We investigated two criteria for topic validity explained by Quinn et al. (2010), i.e., *intra-topic semantic validity* and *inter-topic semantic validity*.

To evaluate intra-topic semantic validity, we reviewed the document-topic distributions from θ for the remaining topics. Ten randomly sampled documents were read, all containing relatively large proportions of the respective topic ($\theta_{d,k} > .5$).[7] For the sampled topics, brief summary descriptions of their content were written, and suggestions about the topic labels were proposed. Subsequently, the researchers deliberately decided in a discussion (a) whether a topic was semantically coherent and, thus, a valid topic in theoretical terms and (b) what label should be given to the topic. For our empirical case, the guiding question regarding (a) was*: Do the topics depict a contentious matter of discourse in the food-safety debate?* Regarding (b) we asked: *Which aspects of the sampled documents describe the issue most comprehensively?* Thus, the label is the product of determining what catches the notion of the underlying concept, in our case the "issues," most concisely.

In this phase of in-depth investigation, eight of the remaining 37 topics were further discarded because they either did not reveal a coherent semantic meaning or solely represented contents from a single website unconnected to aspects of the debate about food safety. Thus, 29 validated, manually labeled "issues" in the food-safety debate remained.

In a second step, we also investigated inter-topic semantic validity, i.e., the relationship between topics by using hierarchical cluster analysis (see Puschmann & Scheffler, 2016). More specifically, the top 30 words of the validated "issues" (from the ϕ_k matrix) were clustered using the cosine-similarity measure and the "complete" clustering method, as implemented in the "hclust" function in *R*. The resulting dendrogram served as an auxiliary guideline for grouping topics that are similar, according to their top words, into higher-order categories. However, clustering results need to be complemented with the results of the in-depth investigation. Relying on the clustering alone could lead to false conclusions because two topics might be distinct according to their top words, although they are semantically related.

Presentation and interpretation of the selected topic model

The valid topics of our empirical case are described in Table 1. For a more comprehensive presentation, we grouped the topics into six categories. The topics in the categories *Agriculture* and *Consumption and Protection* define core issues of food safety. The agricultural topics are especially concerned with economization trends, such as the use and consequences of genetically modified food and the overuse of antibiotics in industrial livestock farming. Consumer-protection topics deal with negative effects of contaminated food. Contamination can be caused by toxic chemicals (e.g., in packaging), as well as pathogenic bacteria such as salmonella, causing food-borne infections.

Another important topical aspect of food safety is visible in the category *Science and Technology*, in which topics deal with new knowledge and innovative means for making food production more efficient and safe. The *Environment* category demonstrates a dual capacity. On one hand, environmental damage can endanger food and water safety, e.g., when chemicals utilized for fracking natural gas out of the soil contaminate drinking water. On the other hand, food-production practices also can have negative consequences for the environment, e.g., the impact of the use of pesticides on bee populations. Another less-political, but still very important component of the food-safety debate concerns the category *Personal Health and Wellbeing*. Topics within this category include diets, which supposedly keep people healthy. Additionally, from the identified *Background Topics* category, it clearly can be induced that food safety in our empirical text corpus is a political and contentious issue, touching economic, legal, and health care issues alike.

Table 1. Validated topic model for the online text corpus about food safety in the U.S.

K	Label	Share % *M* (SD)	HHI *M* (SD)	Top-5 Words
Agriculture				
25	GM Food	3.94 (0.90)	0.04 (0.01)	food, label, genetically, monsanto, gmo
9	Organic Farming	2.58 (0.37)	0.02 (0.00)	organic, food, farm, farmer, agriculture
20	Livestock	2.55 (0.18)	0.03 (0.00)	meat, food, animal, beef, milk
10	Antibiotics	2.21 (0.46)	0.10 (0.02)	antibiotic, animal, health, drug, human
Consumption and Protection				
22	Foodborne Diseases	4.06 (1.34)	0.06 (0.02)	food, outbreak, salmonella, illness, report
8	FS Regulation	3.48 (0.40)	0.04 (0.01)	food, fda, safety, product, consumer
7	Contaminated Food	2.77 (0.63)	0.04 (0.01)	safety, recall, produce, fda, outbreak
29	Food Consumption	2.26 (0.14)	0.03 (0.01)	product, company, consumer, store, sell
27	Restaurant Inspection	2.14 (0.98)	0.09 (0.04)	food, restaurant, safety, health, inspection
16	Tap Water	1.53 (1.03)	0.22 (0.23)	water, food, public, protect, watch
39	BPA-packaging	1.50 (0.83)	0.15 (0.11)	chemical, bpa, safe, toxic, health
Science and Technology				
6	Health Reports	3.48 (0.25)	0.02 (0.00)	health, report, public, risk, datum
19	Chemicals	2.28 (0.28)	0.02 (0.00)	study, chemical, level, health, human
37	GM Technology	1.84 (0.12)	0.02 (0.00)	research, test, science, article, study
Environment				
44	Bees and Pesticides	3.14 (1.90)	0.41 (0.28)	bee, pesticide, epa, food, center
43	Environment	1.41 (0.28)	0.05 (0.02)	read, fish, salmon, environment, specie
50	Fracking	1.37 (0.30)	0.04 (0.02)	energy, gas, oil, water, environmental
31	Climate Change	1.34 (0.22)	0.03 (0.01)	climate, change, report, world, warm
Personal Health and Wellbeing				
21	(Un)healthy Diet	2.32 (0.44)	0.04 (0.01)	food, fat, sugar, diet, health
35	Health and Nutrition	2.31 (0.24)	0.01 (0.01)	program, community, work, education, child
38	Recipes	2.26 (0.41)	0.03 (0.01)	cook, eat, meat, make, recipe
1	School Food	2.00 (0.52)	0.17 (0.08)	food, school, pew, safety, project
12	Dietary Therapy/Prevention	1.42 (0.18)	0.03 (0.01)	cancer, disease, woman, blood, child
42	Medical Information	1.29 (0.39)	0.07 (0.08)	doctor, medicine, take, day, skin
Background Topics				
14	Politics	2.65 (0.28)	0.03 (0.01)	bill, state, obama, law, house
11	Economy	2.50 (0.29)	0.02 (0.01)	company, market, country, million, u.s.
24	Law and Order	2.20 (0.34)	0.02 (0.00)	report, year, police, official, court
2	Infectious Diseases	2.03 (0.62)	0.06 (0.02)	health, coli, pet, animal, case
48	Health Care	1.07 (0.46)	0.13 (0.11)	drug, health, care, medical, patient

Note. HHI = Hirschman-Herfindahl-Index; GM = genetically modified; BPA = Bisphenol A; FS = food safety; K = index of the topic.

In our view, a comprehensive presentation of a topic model also should encompass some of the most important measures, such as the salience of a topic and a fraction of the top-words (see Table 1). Top-word presentation is important to give readers insight into topics.

Conclusion: a good practice guide for communication researchers

The goal of this article is to make LDA-based topic modeling more accessible and applicable for communication researchers. Therefore, it focused on four challenging methodological questions: (a) appropriate pre-processing of unstructured text collections; (b) selection of a parameter set that ensures interpretability of the topic model; (c) evaluating and improving the reliability of a topic model, while at the same time keeping interpretability high; and (d) validation of resulting topics. The following paragraphs briefly recap our recommendations for communication scholars who want to apply LDA-based topic modeling in their research.

Pre-processing

LDA does not just work for "nice" and "easy" data. As our technically challenging case exemplifies, elaborate data cleaning is necessary, especially for unstructured text collections. Additionally, researchers may not only rely on a seemingly standard procedure for successively applied pre-processing steps. Instead, it is important to consider the specifics of the text corpus, including theoretical implications, as well as the proper ordering of pre-processing steps. For instance, the removal of some special characters, such as hashtag-symbols, might be reasonable for the analysis of newspaper article-collections, but not for tweet collections. Regarding proper ordering, we suggest proceeding in the following order: 1. tokenization; 2. transforming all characters to lowercase; 3. removing punctuation and special characters; 4. Removing stop-words; 5. term unification (lemmatizing or stemming); and 6. relative pruning. We prefer lemmatizing over stemming, because a word's lemma is usually easier to interpret than its stem.

Model selection

Also, the proposed model-selection process can be costly and time-consuming, but it will yield more reliable topic models with enhanced interpretability. We propose three considerations.

First, our approach suggests a two-step procedure for model selection that aims to optimize the human interpretation of topic models. In our view, interpretability should be the prime criterion in selecting candidate models. Communication researchers working with content data aim to gain knowledge about content characteristics and the substantive meaning of the text collection. Thus, the success of LDA applications for both objectives depends on how well the resulting model can be interpreted by human researchers. Therefore, we suggest first calculating candidate models with varying granularity levels (i.e., different values for K) and different combinations of prior parameters α and β. Then, choose one model for each K, in which the parameter configuration yields the best results regarding the intrinsic coherence metric. The chosen candidate models need to be further investigated in the second step with a substantive search in coherence-optimized candidate models. The purpose of the substantive search should be to select one of the candidates that matches the granularity level with the theoretical concept under study, such as *political issues* or *frames*. Substantive searches also may include qualitative techniques, such as group discussions, to ensure intersubjectivity. Software tools, such as *LDAvis* (Sievert & Shirley, 2014), proved to be extremely helpful to accomplish this task.

Second, if the size of a corpus is very extensive (e.g., $n > 50{,}000$ documents), large-enough samples (e.g., > 10% of the documents) can be used instead of the whole corpus to calculate the candidate models. It is clearly an intricate process to test various combinations of parameter settings, but using a significantly smaller random subset of the corpus turned out to be a viable approach for mastering this challenge. Using random samples will boost the algorithm's performance and enable researchers to test various parameter settings much faster. The separation of model creation and inference enabled us to

directly use the model that we created on the basis of the random sample and successively infer the topic composition of the remaining documents. However, the validity of the sampling technique crucially depends on whether the sample size is big enough to capture the heterogeneity of the corpus vocabulary. Thus, we cannot guarantee that a sample of roughly 10% of the documents will work equally well for more heterogeneous corpora, and corpora containing significantly smaller sized documents (e.g., a corpus of tweets). Future research needs to address the question of valid guidelines regardless of corpus characteristics.

Third, a well-fitted model with meaningful interpretation is worthless if the results cannot be reproduced. To tackle this issue, we advanced the regularization technique of Lancichinetti et al. (2015) using a semantic-network initialization approach. The literature, as well as our experiments which included multiple corpora, provided evidence that available regularization techniques, such as ours, significantly enhances the reliability of topic models. However, because reliability cannot be guaranteed for topic models generally, we believe that reliability reporting for LDA models should become a disciplinary standard in communication research. We suggest using the metric proposed by Niekler (2016) for this purpose.

Validation

The sequential validation procedure approximates validity from different angles. The available metrics, which have different interpretations, are not treated as objective indicators for how well the model works or how good a topic is. Instead, our approach focuses on inter-individual interpretability using the metrics as a basis. Each step in the process involves deliberation among several researchers. Two criteria of validity were checked: intra-topic and inter-topic semantic validity (Quinn et al., 2010). Our case study teaches us that intra-topic semantic validity cannot be derived merely from a topic's word distribution. Several easy-to-calculate metrics definitely should be considered to sharpen the understanding of whether or not a topic refers to a coherent semantic concept. The most time-consuming, but indispensable, step is the manual check of documents with a high probability of containing a specific topic. This practice allows us to compare and check whether the notion that we sketch from the ϕ distribution matches the interpretation of several information-rich text documents. Labeling topics on the basis of broader context knowledge seems only fair.

Final thoughts

We emphasize that we do not propose a whole new method for topic modeling. Instead, we develop an approach to dealing with the methodological decisions one has to make for applying LDA topic modeling reliably and validly in communication research. With the exception of the regularization-technique which we demonstrated to work significantly better for multiple corpora, we used only a single corpus as a showcase for our explications. However, we deem our approach generalizable to other cases because every single component of our approach is either based on substantial existent studies and/or based on a theoretical rationale.

All in all, LDA topic modeling has proven to be a most promising method for communication research. At the same time, it does not work well with non-deliberate, arbitrary choices in model selection and validation. Our study proposes methods and measures to approximate and improve validity and reliability when using LDA. After all, we aim to provide a "good practice" example, bringing LDA into the spotlight as a method that advances innovation in communication research.

Notes

1. The Dirichlet distribution is a continuous multivariate probability distribution which is frequently used in Bayesian statistics.
2. EBSCO communication source (search in title OR abstract OR keywords; apply related words): "topic model", "topic modeling", "topic modelling", "latent Dirichlet allocation". Web of Science (only communication-related

categories: Sociology, Political Science, Psychology, Linguistics, Language Linguistics, Telecommunications, Communication, Social Science Interdisciplinary; search in Title, Abstract, Author Keywords, Keywords Plus): "topic model*", "latent Dirichlet allocation". The searches were run on 10.05.2016.

3. The websites were identified using a combination of a literature review, expert evaluations and Google searches; the starting URLs for the network collection were: http://www.centerforfoodsafety.org/, http://www.cspinet.org/foodsafety/, http://www.foodandwaterwatch.org/food/, http://www.organicconsumers.org/foodsafety.cfm, http://notinmyfood.org/newsroom, http://barfblog.foodsafety.ksu.edu/barfblog (until May 2013);http://barfblog.com (from June 2013), http://www.greenpeace.org/international/en/campaigns/agriculture/, http://www.pewhealth.org/topics/food-safety-327507.
4. For the gathering of the networks, we used the snowball procedure, with a crawling depth of 2 and a degree of separation of 1 (for detailed information see Waldherr et al. (2017, p. 432); for further, general information on the tool, please visit http://www.govcom.org/Issuecrawler_instructions.htm).
5. A topic's intrinsic coherence C of a topic t over the topic's M top-words ($V^{(t)} = \left(v_1^t, \ldots, v_M^t\right)$ is defined by Mimno et al. (2011, p. 265) as $C\left(t, V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=2}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right)+1}{D\left(v_l^{(t)}\right)}$, where $D\left(v_l^{(t)}\right)$ is the document frequency of word $v_l^{(t)}$ in the corpus and $D\left(v_m^{(t)}, v_l^{(t)}\right)$ is the co-document frequency of the words $v_m^{(t)}$ and $v_l^{(t)}$.
6. For the U.K. corpus number of topics was set to $K = 30$, $K = 50$ for both the U.S. and Germany; we set $\alpha = .5$ and $\beta = .02$ for all models in this experiment. The data as well as the scripts of our experiments can be retrieved from: https://github.com/tm4ss/lda-reliability.
7. If no or not enough documents were available for $\theta_{d,k} > .5$, we set the threshold to $\theta_{d,k} > .3$.

Acknowledgement

The first author claims single authorship for subsection *Topic validity and labeling* and section *Presentation and interpretation of the selected topic model* including *Table 1* and *Appendix C*.

Funding

This publication was created in the context of the Research Unit "Political Communication in the Online World" (1381), subproject 7, which was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The subproject was also funded by the Swiss National Science Foundation (SNF).

References

Baum, D. (2012). Recognising speakers from the topics they talk about. *Speech Communication*, *54*(10), 1132–1142.
Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.
Biel, J.-I., & Gatica-Perez, D. (2014). Mining crowdsourced first impressions in online social video. *Ieee Transactions on Multimedia*, *16*(7), 2062–2074.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.
Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic podels*. Paper presented at the International Conference on Machine Learning, Pittsburgh, PA.
Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3* (4/5), 993–1022.
Bonilla, T., & Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*, *41*, 650–669.
Bordag, S. (2008). A comparison of co-occurrence and similarity measures as simulations of context. *Proceedings of the 9th international conference on computational linguistics and intelligent text processing*, 52–63. doi: 10.1007/978–3–540–78135–6_5
Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge, UK: Cambridge University Press.
Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). *Reading tea leaves: How humans interpret topic models*. Paper presented at the Neural Information Processing System 2009.
Cobb, R. W., & Elder, C. D. (1983). *Participation in American politics: The dynamics of agenda-building*. Baltimore, MD: Johns Hopkins University Press.
Denny, M. J., & Spirling, A. (2017). *Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it*. New York University. Retrieved from http://www.nyu.edu/projects/spirling/documents/preprocessing.pdf
DiMaggio, P., Nag, M., & Blei, D. M. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*, 570–606.

Elgesem, D., Feinerer, I., & Steskal, L. (2016). Bloggers' responses to the Snowden affair: Combining automated and manual methods in the analysis of news blogging. *Computer Supported Cooperative Work (CSCW)*, *25*, 167–191.
Elgesem, D., Steskal, L., & Diakopoulos, N. (2015). Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication*, *9*(2), 169–188.
Evans, M. S. (2014). A computational approach to qualitative analysis in large textual datasets. *PLoS One*, *9*(2), 1–10.
Ghosh, D. D., & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90–102.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101* (1), 5228–5235.
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114* (2), 211–244.
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*, 1–31. doi:10.1093/pan/mps028
Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40* (13), 1–30.
Günther, E., & Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. *International Journal of Communication*, *11*, 3051–3071.
Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism*, *4*(1), 75–88.
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*, 1–28. doi:10.1177/1077699016639231
Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398–436.
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.
Jacobi, C., Van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 1–18. doi:10.1080/21670811.2015.1093271
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: Wiley.
Koltcov, S., Nikolenko, S. I., Koltsova, O., Filippov, V., & Bodrunova, S. (2016). Stable topic modeling with local density regularization. In F. Bagnoli, A. Satsiou, I. Stavrakakis, P. Nesi, G. Pacini, Y. Welp, & D. DiFranzo (Eds.), *Internet science: Third international conference, INSCI 2016, Florence, Italy, September 12–14, 2016, Proceedings* (pp. 176–188). Cham, Switzerland: Springer International Publishing.
Koltsova, O., & Koltcov, S. (2013). Mapping the public agenda with topic modeling: The case of the Russian LiveJournal. *Policy & Internet*, *5*(2), 207–227.
Koltsova, O., & Shcherbak, A. (2015). 'LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012. *New Media & Society*, *17*(10), 1715–1732.
Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review*, *5*(1). doi:10.1103/PhysRevX.5.011007
Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista Di Linguistica*, *20*(1), 1–31.
Levy, K. E. C., & Franklin, M. (2014). Driving regulation: Using topic models to examine political contention in the U. S. trucking industry. *Social Science Computer Review*, *32*(2), 182–194.
Lötscher, A. (1987). *Text und Thema. Studien zur thematischen Konstituenz von Texten. [Text and topic. Studies concerning thematical constituency of texts]*. Berlin, Germany: De Gruyter.
Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11* (1–2), 22–31.
Maier, D., Waldherr, A., Miltner, P., Jähnichen, P., & Pfetsch, B. (2017). Exploring issues in a networked public sphere: Combining hyperlink network analysis and topic modeling. *Social Science Computer Review*, Advance online publication. doi:10.1177/0894439317690337
Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval*. Cambridge, UK: Cambridge University Press.
Manning, C. D., & Schütze, H. (2003). *Foundations of statistical natural language processing (6. print with corr.)*. Cambridge, MA: MIT Press.
Marshall, E. A. (2013). Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics*, *41*(6), 701–724.
McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. Retrieved from http://mallet.cs.umass.edu
Miller, M. M., & Riechert, B. P. (2001). The spiral of opportunity and frame resonance: Mapping the issue cycle in news and public discourse. In S. D. Reese, O. H. Gandy Jr., & A. E. Grant (Eds.), *Framing public life: Perspectives on media and our understanding of the social world* (pp. 107–121). Mahwah, NJ: Lawrence Erlbaum Associates.
Mimno, D., & Blei, D. M. (2011). Bayesian checking for topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 227–237.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

Neuendorf, K. A. (2017). *The content analysis guidebook* (2nd ed.). Los Angeles, CA: Sage.

Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improving topic coherence with regularized topic models. *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 496–504. Retrieved from http://dl.acm.org/citation.cfm?id=2986459.2986515

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In S. Mehrotra, D. D. Zeng, H. Chen, B. Thuraisingham, & F.-Y. Wang (Eds.), *Intelligence and security informatics* (Vol. 3975, pp. 93–104). Berlin, Germany: Springer.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL*, 100–108.

Niekler, A. (2016). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen*. [Automated approaches for the analysis of topics in news sources]. (PhD dissertation). University of Leipzig, Leipzig, Germany. Retrieved from http://www.qucosa.de/fileadmin/data/qucosa/documents/19509/main.pdf

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

Parra, D., Trattner, C., Gómez, D., Hurtado, M., Wen, X. D., & Lin, Y.-R. (2016). Twitter in academic events: A study of temporal usage, communication, sentimental and topical patterns in 16 computer science conferences. *Computer Communications*, *73*, 301–314.

Puschmann, C., & Scheffler, T. (2016) Topic modeling for media and communication research: A short primer. *HIIG Discussion Paper Series* (No. 2016-05): Alexander von Humboldt Institut für Internet und Gesellschaft.

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.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. New York, NY: Cambridge University Press.

Rauchfleisch, A. (2017). The public sphere as an essentially contested concept: A co-citation analysis of the last 20 years of public sphere research. *Communication and the Public*, *2*(1), 3–18.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez (Ed.), *Analytical methods for social research. Computational social science: Discovery and prediction* (pp. 51–97). New York, NY: Cambridge University Press. doi:10.1017/CBO9781316257340.004

Salton, G. (1991). Developments in automatic text retrieval. *Science*, *253*(5023), 974–980.

Scott, S., & Matwin, S. (1999). Featuring engineering for text classification. *Proceedings of the ICML-99*, 379–388. Bled, Slovenia.

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

Sokolov, E., & Bogolubsky, L. (2015). Topic models regularization and initialization for regression problems. *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, 21–27. doi:10.1145/2809936.2809940

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic approaches to semantic representation. In T. K. Landauer, McNamara, S. Dennis, & W. Knitsch (Eds.), *Handbook of latent semantic analysis*, (pp. 424–440). Mahwah, NJ: Lawrence Erlbaum.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Tsur, O., Calacci, D., & Lazer, D. (2015). *A frame of mind: Using statistical models for detection of framing and agenda setting campaigns*. Paper presented at the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Van Atteveldt, W., Welbers, K., Jacobi, C., & Vliegenthart, R. (2014). *LDA models topics… But what are 'topics'?* Retrieved from http://vanatteveldt.com/wp-content/uploads/2014_vanatteveldt_glasgowbigdata_topics.pdf

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.

Wallach, H. M. (2006). *Topic modeling: Beyond bag-of-words*. Paper presented at the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 1973–1981). New York, NY: Curran Associates.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). *Comparing Twitter and traditional media using topic models*. Paper presented at the 33rd European Conference on IR Research, Dublin, Ireland.

Appendix A

Systematic review of studies in communication research, which uses LDA topic modeling

Reference	Type of Data	Preprocessing	Parameter Selection	Interpretability & Validity	Reliability
Studies with methodological focus					
Baum (2012)	Political speeches	Stemming Removing stop words *No specific sequence*	K (chosen after validation)	Review top words Review top documents Manual labeling External validation	—
Biel and Gatica-Perez (2014)	YouTube videos and comments	Removing punctuation and repeated letters Stemming *No specific sequence*	K (qualitative exploration), prior parameters (standard values)	Review top words Manual labeling Validation of topics via word intrusion tasks and topic intrusion tasks	Split sample test
DiMaggio et al. ((2013)	Newspaper articles	Removing stop words *No specific sequence*	K (qualitative exploration)	Review top words Review top documents Categorizing topics Statistical validation with mutual information (MI) criterion Internal validation via hand coding of sample texts External validation of topics with news events	Replication with variations of corpus, seeds and parameters
Evans (2014)	Newspaper articles	—	K (chosen after validation), prior parameters (optimization)	Review top words Manual labeling Quantitative metrics (topic coherence, etc.) External validation through qualitative domain knowledge	—
Ghosh and Guha (2013)	Tweets	1. Removing URLs and HTML entities 2. Removing punctuation and conversion to lowercase 3. Removing stop words 4. Stemming 5. Tokenization	K (quantitative metrics: perplexity); prior parameters (standard values)	Review top words Manual labeling External validation with political events	—
Guo et al. (2016)	Tweets	Stemming Removing punctuation, stop words, etc. *No specific sequence*	K (trial and error)	Review top words Manual labeling Comparison with manual coding	—
Jacobi et al. (2015)	News articles	1. Lemmatizing 2. Part of speech-tagging; Removing frequent and infrequent words; Removing terms with numbers/non-alphanumeric letters	K (qualitative exploration and quantitative metrics: perplexity)	Review top words Review top documents Review of co-occurrence of top words (topic coherence) Manual labeling Comparison with manual coding	—
Newman et al. (2006)	News articles	1. Tokenization; Removing stop words 2. Removing infrequent terms	K (no explanation)	Review top words and entities Manual labeling External validation of topics with news events	—
Puschmann and Scheffler (2016)	Newspaper articles	1. Removing numbers and punctuation, conversion in lower case 2. Removing stop words 3. Removing infrequent terms	K (quantitative metrics: perplexity and Euclidean distance)	Review top words Quantitative metrics (Euclidean distance) Manual evaluation Inter-topic semantic validation	—

(*Continued*)

(Continued).

Reference	Type of Data	Preprocessing	Parameter Selection	Interpretability & Validity	Reliability
Tsur, Calacci, and Lazer (2015)	Press releases and statements	—	K (qualitative exploration)	Review top words Manual labeling External validation by domain experts	—
Van Atteveldt et al. (2014)	News articles	Lemmatizing Removing frequent and infrequent words *No specific sequence*	K (high resolution)	Review top words Quantitative metrics (topic prevalence) Comparison with manual coding	Replication with different parameters
Zhao et al. (2011)	Tweets and newspaper articles	1. Removing stop words 2. Removing frequent and infrequent words 3. Removing tweets with less than three words/users with less than eight tweets	K (qualitative exploration)	Review top words Semi-automated topic categorization Manual labeling Manual judgement of interpretability	—
Studies with thematic research focus					
Bonilla and Grimmer (2013)	Newspaper articles and transcripts of newscasts	Stemming Removing punctuation and stop words *No specific sequence*	K (application of non-parametric topic model, qualitative exploration)	Review documents (random sample) Manual labeling Automated labeling (using mutual information)	Replication with varying number of topics
Elgesem et al. (2016)	Blog posts	—	K (qualitative exploration)	Review top words Review top documents Manual labeling	—
Elgesem et al. (2015)	Blog posts	—	K (qualitative exploration)	Review top words Review documents Manual labeling Quantitative metrics (mutual information, etc.)	—
Koltsova and Koltcov (2013)	Blog posts	Removing HTML tags, punctuation, etc. Lemmatization *No specific sequence*	K (quantitative metrics: perplexity)	Review top words Review top documents Manual labeling	—
Koltsova and Shcherbak (2015)	Blog posts	—	K (no explanation)	Review documents Manual labeling and evaluation	—
Levy and Franklin (2014)	Public comments	1. Stemming 2. Removing stop words 3. Removing terms with only single letters or numbers 4. Removing infrequent words	K (qualitative exploration)	Review top words External validation with expert evaluation	Replication with variations of corpus, seeds and parameters
Parra et al. (2016)	Tweets	Language filtering Removing stop words, special characters, URLs, words with less than three characters *No specific sequence*	K (qualitative exploration)	—	—
Rauchfleisch (2017)	Research articles	Removing stop words Removing numbers, replacing hyphens with space characters, conversion in lowercase Stemming *No specific sequence*	K (no explanation); parameters set according to Steyvers and Griffiths (2007)	Review top words Manual classification External validation	—

Note. K = number of topics. The ordering of pre-processing steps is numbered if the ordering was explicitly mentioned in the source.

Appendix B

Choice of candidate models from topic models with varying parameter sets

Nr.	*K*	*α*	*β*	Likelihood	Mean Coherence
1	30	0.01	0.033	−67464644.07	−399.49
2	30	0.05	0.033	−66324953.70	−399.60
3	30	0.10	0.033	−65740704.30	−401.30
4	30	0.20	0.033	−64822303.40	−393.80
5	30	0.50	0.033	−63435029.60	−396.90
6	30	1.00	0.033	−62317020.40	−393.30
7	50	0.01	0.020	−64835932.63	−423.18
8	50	0.05	0.020	−63182677.27	−421.18
9	50	0.10	0.020	−62079259.12	−421.67
10	50	0.20	0.020	−61058300.26	−427.59
11	50	0.50	0.020	−59290870.33	−404.24
12	50	1.00	0.020	−57956143.48	−405.92
13	70	0.01	0.014	−63164036.11	−438.95
14	70	0.05	0.014	−60895636.63	−426.21
15	70	0.10	0.014	−59579663.81	−422.07
16	70	0.20	0.014	−58399926.46	−423.43
17	70	0.50	0.014	−56628160.74	−404.67
18	70	1.00	0.014	−54896346.70	−411.50

Note. K = number of topics.

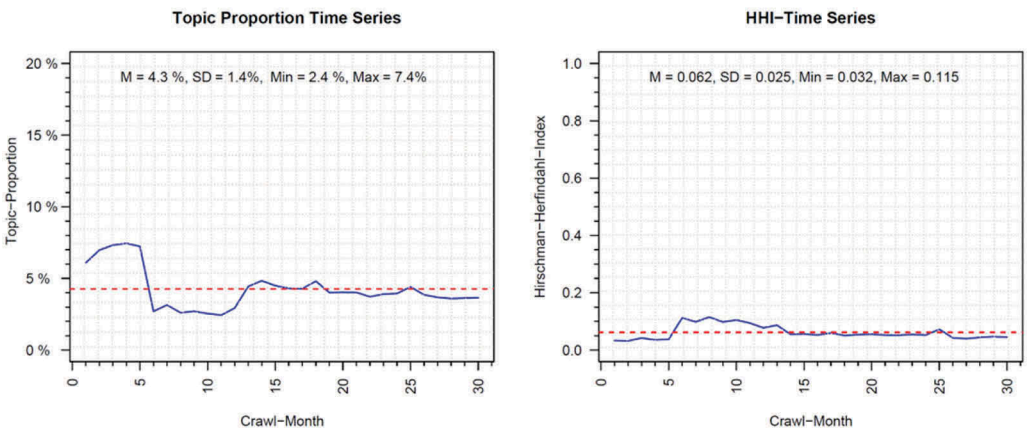
Appendix C

Summary statistics for the interpretation of a topic

Topic 22 — foodborne diseases

	Websites Top Websites	Share in %	Topwords $\lambda = 1$	$\lambda = 0.6$
1	fda.gov	14.40	food	outbreak
2	barfblog.com	11.40	outbreak	salmonella
3	cdc.gov	8.90	salmonella	illness
4	cspinet.org	4.20	illness	ill
5	notinmyfood.org	2.40	report	food
6	barfblog.foodsafety.ksu.edu	2.30	people	case
7	pewhealth.org	2.20	case	foodborne
8	foodsafety.gov	1.90	state	report
9	usatoday.com	1.60	eat	investigation
10	nytimes.com	1.50	ill	sick
11	organicconsumers.org	1.30	disease	contaminate
12	inspection.gc.ca	1.20	bacterium	people
13	bt.cdc.gov	1.10	contaminate	campylobacter
14	centerforfoodsafety.org	1.00	raw	bacterium
15	foodsafetytalk.com	0.80	foodborne	egg
16	eurosurveillance.org	0.80	infection	infection
17	foodsafetynews.com	0.80	investigation	raw
18	fsis.usda.gov	0.70	sick	hospitalize
19	oregonlive.com	0.70	chicken	strain
20	atwork.avma.org	0.60	egg	diarrhea

Rank1-Metric: Rank **26 out of 50**.
Coherence-Metric: Rank **18 von 50**.



Note. The figure depicts a divided table and two time-series plots. The left side of the table shows the average most prevalent sources of the topic while the right side maps out the top-words according to two different relevance values ($\lambda = 1$ and $\lambda = .6$). Below the table the ranks of the Rank-1 and the coherence metrics are given. The left time series shows the salience of the topic over time, while the right plot gives a sense of how concentrated the topic was over the course of investigation.