Received: 14 September 2017 | Revised: 1 February 2019 | Accepted: 20 March 2019 | Published on: 2 May 2019

DOI: 10.1002/sam.11415

RESEARCH ARTICLE

WILEY

Assessing topic model relevance: Evaluation and informative priors

Angela Fan | **Finale Doshi-Velez** | **Luke Miratrix**

Department of Statistics, Harvard University, Cambridge, Massachusetts

Correspondence
Luke Miratrix, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138.
Email: lmiratrix@g.harvard.edu

Latent Dirichlet allocation (LDA) models trained without stopword removal often produce topics with high posterior probabilities on uninformative words, obscuring the underlying corpus content. Even when canonical stopwords are manually removed, uninformative words common in that corpus will still dominate the most probable words in a topic. In this work, we first show how the standard topic quality measures of coherence and pointwise mutual information act counter-intuitively in the presence of common but irrelevant words, making it difficult to even quantitatively identify situations in which topics may be dominated by stopwords. We propose an additional topic quality metric that targets the stopword problem, and show that it, unlike the standard measures, correctly correlates with human judgments of quality as defined by concentration of information-rich words. We also propose a simple-to-implement strategy for generating topics that are evaluated to be of much higher quality by both human assessment and our new metric. This approach, a collection of informative priors easily introduced into most LDA-style inference methods, automatically promotes terms with domain relevance and demotes domain-specific stop words. We demonstrate this approach's effectiveness in three very different domains: Department of Labor accident reports, online health forum posts, and NIPS abstracts. Overall we find that current practices thought to solve this problem do not do so adequately, and that our proposal offers a substantial improvement for those interested in interpreting their topics as objects in their own right.

KEYWORDS

informative priors, latent dirichlet allocation, topic modeling

1 | INTRODUCTION

Latent Dirichlet allocation (LDA) [4] is a popular model for modeling topics in large textual corpora as probability vectors over terms in the vocabulary. LDA posits that each document d is a mixture θ_d over K topics, each topic k is a mixture β_k over a common, set vocabulary of size V, and $w_{\mathrm{d,\,n}}$, the nth word in document d, is generated by first sampling a topic $z_{\mathrm{d,\,n}}$ from θ_{d} and then drawing a word from that topic:

$$\theta_d \sim Dirichlet(\alpha) \qquad \beta_k \sim Dirichlet(\eta)$$

$$z_{\mathrm{d,n}} \sim Mult(\theta_d) \qquad w_{d,n} \sim Mult(\beta_{z_{\mathrm{d,n}}}).$$

The α and η are hyperparameters to be selected by the user. Once such a model is fit, the K topics are then commonly interpreted by looking at the most probable words in their distributions β_k, $k = 1, \ldots K$. Unfortunately, *stopwords*—words with no contextual information—often dominate these lists of highest probability words. Stopword-dominated topics are uninterpretable as semantic themes, and even if canonical stopwords are removed, topics dominated by overly general and uninformative words still reduce the utility, reliability, and acceptance of statistical topic models by users outside of the machine learning community.

To improve topic quality, practitioners typically rely on heavy pre- and postprocessing, such as creating stopword lists and retraining the LDA models without those words. Broadly, stopwords can be divided into two categories: canonical ("the," "and") or domain-specific ("child," "son" in a corpus about children). Canonical stopwords can often be removed by referring to standard, publically available lists. Constructing lists of domain-specific stopwords, however, is a nontrivial task and risks introducing human bias if the model trainer builds these lists over repeated LDA runs. Such extensive processing is also a challenge for scientific reproducibility, as typically many preprocessing steps and deleted-word lists are not included in publications. Further, many proposed automated or technical methods to improve topic quality are complex and not easily integrated into existing software, particularly for the applied LDA community or as part of a larger and more complex graphical model.

In this work, we first expose a subtle but important concern regarding the evaluation of topic quality, as defined by concentration of information-rich words, when documents contain many irrelevant words: common metrics such as coherence [13] and pointwise mutual information (PMI) [16], actually *prefer* topics that place high probability on canonical stopwords. Furthermore, these standard topic quality metrics cannot compare LDA models trained across different vocabularies, as is the case when one is iteratively removing potential stopwords. Worse, we demonstrate that, across several data sets and stopword removal schemes, these metrics do not appropriately correlate with human evaluations of interpretability in the presence of stopwords.

In sum, this work shows that (a) conventional approaches to the stopword problem for topic modeling are inadequate and (b) this fact is possibly obscured because common measures of topic quality can be deceptive if the vocabulary used in the modeling is not heavily and carefully curated as a preprocessing step. Not only do we demonstrate these two concerns, but we suggest a simple and easily implementable solution: use a heterogeneous collection of informative asymmetric priors on the topics to generate, in one model-fitting, different topics for the different words of interest. We show that this approach can both reduce the presence of stopwords and improve the domain relevance of topic models, assessed with human evaluation. We also provide an alternate evaluation metric, based on lift, that correlates well with human judgment of average word quality in topics to serve as a proxy when human evaluation is not feasible or practical. We suggest that lift be used in combination with other metrics to assess the many characteristics that contribute to high-quality topics.

2 | RELATED WORK

2.1 | Metrics for evaluating topic quality

The difficulties of measuring topic quality are well known. Traditional evaluation has used perplexity, but this has been shown to negatively correlate with human-measured topic interpretability using novel word and topic intrusion tasks [5]. Since then, several other methods have been proposed to automatically evaluate topic quality. For instance, Newman et al. [15] show that PMI correlates strongly with human assessments of semantic coherence. PMI measures the word association between pairs of topic words by using external data (often from English Wikipedia). Mimno et al. [13] propose the topic coherence metric that measures topic word co-occurrence across documents to detect low quality topics, and show it correlates with expert topic annotations.

However, previous work has noted that using single metrics to evaluate topic quality is problematic, as different metrics typically capture different facets of quality [18]. For example, one might measure whether topics represent coherent, readable ideas, or measure coverage of the range of topics actually present in the corpus, or, as in our case, measure concentration of substantive, domain-relevant words in the final topics. In this work, we focus on the problem of stopwords, which present a modeling obstacle as they dominate the word frequency and co-occurrence statistics of a corpus. In many LDA models, topics mainly represent these common words, which obscure relevant corpus content. Further, we find that in the presence of stopwords, LDA metrics meant to evaluate other aspects of topic quality perform counterintuitively (see Section 3).

2.2 | Methods for increasing topic relevance and reducing stopwords

To produce topics with more domain-relevant words, much work has focused on automatic stopword detection and removal. Popular techniques for automatic stopword detection include keyword expansion and other information retrieval approaches [6,10,19,23]. Several approaches identify stopwords based on term weighting schemes [14] or word occurrence distributions [2,27]. Makrehchi and Kamel [11] assume that every document has a type or label and only include the words that are most correlated with the document label while minimizing information loss. Lo et al. [10] begin with a set of pregenerated search engine queries and quantify word informativeness via the KL divergence between the query term distribution and the corpus background distribution. These approaches require parameter tuning to set various penalty cutoffs. Several require document-specific labels and/or query terms. In contrast, we propose a simple fix that can be easily applied within existing LDA software frameworks.

More broadly, there are many efforts to improve the semantic interpretability of topic models [1,3,12,28,29]. In particular, much work has improved topic quality via different priors: Wallach et al. [26] show the effectiveness of general asymmetric priors to improve topic quality, Newman et al. [16] use an informative prior capturing short range dependencies between words, and Andrzejewski et al. [1] use Dirichlet Forest priors to capture corpus structure. Other models modify LDA to incorporate corpus-wide data of word frequency and exclusivity [3,7], and focus on relative as well as absolute word frequencies in a topic.

These modifications, however, are much more effective if overwhelmingly high-frequency words are removed first, as the majority of these models are not targeted towards isolating stopwords to improve topic readability. This stopword removal, needed to produce coherent output, is often conducted as a preprocessing step [1,7,9,16,30]. Some models have some robustness toward the presence of stopwords, but perform noticeably better with canonical stopword deletion [25,30]. Further, while many methods have been proposed to identify stopwords and model only domain-relevant words, many LDA users still extensively use canonical stopword deletion, particularly for more complex graphical models, possibly because of the additional modeling burden of many of the methods above. Some work instead removes stopwords as a postprocessing step rather than preprocessing [21], but this still necessitates a curated list of words to delete.

3 | TRADITIONAL TOPIC QUALITY METRICS ARE NOT ROBUST TO STOPWORDS

We next show that two standard measures of topic quality—coherence and PMI—perform counter-intuitively in situations in which the corpus contains many common but irrelevant words. This situation is common in many real corpora, where there is standard vocabulary that is often repeated in the text but is generally uninformative. That said, our analysis does not invalidate the use of these measures in cases where the vocabulary has been carefully curated for relevance.

After a discussion of coherence and PMI, we introduce another metric, log lift, that alleviates these found concerns in the case of the stopword problem.

3.1 | Coherence

The *Coherence* of topic t is defined as

$$\text{coherence}(t) := \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \log \frac{D(v_i^t, v_j^t) + 1}{D(v_i^t)},$$

where v_i^t is the ith most probable word in topic t and M represents the number of top topic words to evaluate. $D(x)$ represents the number of documents word x appears in, $D(x, y)$ represents the number of documents x and y coappear in, and the $+1$ ensures the *log* is defined [13]. Coherence is largest when $D(v_i^t, v_j^t) = D(v_i^t)$, which occurs when either (a) the words co-occur in a very small subset of documents and are absent elsewhere or (b) the words appear in all documents. The former case is unlikely, particularly as topic evaluation is conducted on the top M most probable topic words. The latter case is not: it occurs when the M words evaluated are common words appearing in every document, that is, are stopwords. For concreteness, in the Autism Spectrum Disorder (ASD) corpus, a corpus about children with autism, "Autism" and "child" only co-occur in 3% of documents, but "and" and "the" co-occur in 58% of documents. The stopword "the" appears in 93% of OSHA documents, 74% of ASD documents, and 99% of NIPS documents. When averaged across all topics, coherence is maximized when all top topic words are common and overlapping.

3.2 | PMI score

The *PMI Score* of topic t is the median of $\log p(v_i^t, v_j^t)/p(v_i^t)p(v_j^t)$ calculated for all pairs of the most probable words v_i^t, v_j^t within topic t, with $i, j \leq M$, where $p(x)$ is the probability of seeing word x in a random document, and $p(x, y)$ is the joint probability of seeing x and y appearing together in a random document. These frequencies are traditionally estimated from text outside of the corpus, for example, from a snapshot of Wikipedia. The PMI for a pair i, j of words is maximized if v_i^t and v_j^t co-occur. In practice, this is easily achieved with high frequency words that appear with high probability in all documents—stopwords. Particularly in real world, noisy corpora, domain words alone are relatively rare, so multiple domain-relevant words co-occuring strongly is incredibly rare. For example, on the ASD data set, the words "school" and "read," both fairly common words in English and topics of high concern for parents, only co-occur in 1.5% of documents. Variants of PMI, such as Normalized PMI [8] adjust the frequencies to reflect specialized or technical corpora but suffer similar drawbacks—they are still maximized if topics are full of the same, high frequency, co-occurring words.

3.3 | An alternate measure: The lift-score

In topic modeling, *lift* [24] is the ratio of a word's probability within a topic to its marginal corpus probability. The lift of word j in topic t is defined as

$$\text{lift}(j, t) := \frac{\beta_{tj}}{b_j}$$

where β_{tj} represents the probability mass of word j in topic t and $b_j = \frac{\sum_{d=1}^{D} n_{d_j}}{\sum_{d=1}^{D} n_d}$ is the empirical probability mass of word j in the entire corpus [24]. Previous work has used lift to sort top topic words for each topic. This use of lift reduces the appearance of globally frequent terms [22,24] as the b_j term accounts for the overall appearance of word j. In this work, we use the lift to generate an overall topic quality metric by averaging the log lift of the top M words of each topic, such that

$$\log\ \text{lift}(b) = \frac{1}{M}\sum_{j=1}^{M} \text{loglift}(v_j^t, t)$$

This will ideally achieve two ends: (a) if the top words in topics generally do not appear in other topics, we will tend to find the topics to be well-separated and distinct and (b) common words, such as stopwords, will tend to have comparatively lower lift and so stopword-laden topics will have lower scores. Given this intuition, we expect this metric to better target the stopword aspect of topic quality. Our experiments show that lift is robust to the presence of stopwords, and we suggest that it can be used in combination with other LDA metrics for holistic topic evaluation.

4 | EMPIRICAL EVALUATION OF METRICS AND FITTING METHODS

In this section, we demonstrate the inadequacies that we mathematically argued regarding traditional evaluation measures in Section 3 do indeed present themselves on three data sets with varying characteristics. The Autism Spectrum Disorder (ASD) corpus contains 656,972 posts from three online support communities for autism patients and their caretakers. Posts contain nonclinical medical vocabulary (eg, “potty going” instead of “toilet training”) and abbreviations (eg, “camhs” for “Child and Adolescent Mental Health Services”). The Occupational Safety and Health Administration (OSHA) corpus contains 49 558 entries from the Department of Labor Occupational Safety and Health database of casualties. Each entry describes a workplace accident. Unlike the ASD corpus, the OSHA posts are short and structured. The Neural Information Processing Systems (NIPS) corpus contains 403 abstracts from the Neural Information Processing Systems Conference 2015 accepted papers. These concisely written abstracts are of medium length with a highly technical vocabulary and comparatively few traditional stopwords.

Overall, for each corpus, we generate several collections of topics using several commonly used topic modeling methods from the literature for handling stopwords, as well as several variants of our new proposal designed specifically to accommodate and detect stopwords in a natural manner. We first evaluate the collections for richness of substantive words as marked by experts and stopword rates. We then see which quality metrics are appropriately associated with these metrics. We finally see which underlying approach for topic generation is most successful.

4.1 | Topic modeling methods

We use several different baselines and several versions of our proposed approach. We describe these methods in the following sections. The baselines we selected are extensively used by applied LDA users as well as the research community.

4.1.1 | Basic modeling approaches

We first consider three basic approaches to topic-modeling: (a) *No Deletion Baseline* — standard LDA without stopword removal, (b) *Stopword Deletion Baseline* — LDA deleting the 127 canonical stopwords from the Stanford Natural Language Toolkit, a common preprocessing step and a standard canonical stopword list, and (c) *TF-IDF Deletion Baseline* —LDA deleting words with TF-IDF scores in the lowest 5%, similar to the stopword removal work in Lo et al. [10] and Ming et al. [14]. These, particularly canonical stopword deletion and TF-IDF-based deletion, correspond to the approach many applied practitioners take.

We also have *Hyperparameter Opt Baseline*, which is LDA with hyperparameter optimization as part of the LDA training. Here researchers fit a series of models with different α and η, maximizing model fit and selecting the best fitting model, as measured by likelihood, as their final one. We use two versions of this baseline, one with the full vocabulary and one with canonical stopwords first deleted (the latter is most analogous to current state of practice). This general approach is commonly thought to solve the stopword problem. We will see, however, that our suggested priors can produce even more interpretable topics compared to optimizing these prior parameters with this approach.

4.1.2 | An alternate approach: Informative priors

We will see that the standard methods of practice outlined in Section 4.1.1 can fail, and fail quite badly by having topic lists dominated by words deemed unuseful. As an alternative we propose using an informative prior on η, which is just as simple to implement as stopword removal yet, as we will see in Section 4.4, yields more interpretable topics. The approach combines two ideas: (a) encourage the formation of different types of topics in the fitting process, in particular stopword topics and domain topics, by using different Dirichlet prior concentrations η_t for different t to model the corpus as a mixture of different types of topics that differently accommodate stopwords and domain-relevant words, and (b) for domain topics have an asymmetric prior η_t that penalizes likely stopwords and promotes domain-specific words.

Importantly, since we only need to change the prior concentrations η_t to implement this approach, this approach can easily be used to augment more complex LDA extensions. This concept of domain-relevant and stopword topics is similar to Paul and Dredze [17], which proposed that there are two separate distributions that generate corpus words—a background distribution that produces common words, and a foreground distribution that generates topical words.

To understand the core idea of our informative prior approach, recall that the posterior distribution of a topic is essentially a mix of the empirical distribution of words thought to be members of the topic and the prior distribution assigned to that topic. Typically this prior, a V-dimensional Dirichlet distribution, is symmetric with, say, all elements (weights) being 1, corresponding to a single pseudo-count for each possible word. This prior regularizes the topics, pulling the posterior toward the prior mean. If we increase the pseudo-counts proportionally, we regularize more. If we use asymmetric priors, then we will regularize toward the new prior mean defined by the normalized vector of weights. We exploit both having different levels of regularization and having asymmetry in our proposed alternate priors we discuss next.

Stopword topics (η_0)

The LDA model models all of the words in the document. Thus, for good model fit, it is important that high-frequency words, even those with little information-bearing content, be explained somehow—we cannot just relegate them to low probabilities in all our topics. Thus, of the K topics, we let I of them be stopword topics: $\beta_k \sim Dirichlet(\eta_o)$, where η_0 is uninformative (1, 1, …, 1). This prior only mildly regularizes the word probabilities, allowing these explicit stopword topics to give high-frequency, but uninteresting, words a place to go.

Choices for domain topics (η_1): Term weighting

One intuitive prior-based penalization is to set the Dirchlet weights η_1 for the domain topic priors to be the inverse of the corpus unigram frequency, which gives high frequency words low prior probabilities of occurring as a draw from a domain topic. This *Word Frequency Prior* is our most naïve approach. Even so, the overall model can achieve reasonable perplexities because frequent corpus words can still be explained in the stopword topics. This sequestering of high-frequency words to specific topics allows the domain topics to more accurately reflect the nuances of the corpus.

However, while penalizing words based on their frequency effectively limits stopwords, it is not a targeted form of restriction—it equally penalizes a term that occurs a few times in many documents and a term that occurs repeatedly in only a few documents (which is a signal of topic-relevance). We propose instead a prior penalization for the domain-relevant topics proportional to the TF-IDF score [20] of the word (again, the overall model can achieve low perplexities because frequent terms can be explained by stopword topics). Our *TF-IDF Prior* for K topics in an LDA model with I stopword topics and $K-I$ non-relevant topics is

$$\beta_1, \dots, \beta_I \sim Dir(1, 1, \dots, 1)$$
$$\beta_{I+1}, \dots, \beta_K \sim Dir(c_1\overline{\mathrm{TI}(w_1)}, \dots, c_1\overline{\mathrm{TI}(w_V)})$$

where $\overline{\mathrm{TI}(w_v)}$ represents the average TF-IDF score of word v across the documents in the corpus and c_1 is an arbitrary scaling constant used to appropriately size the TF-IDF scores. We use the common TF-IDF score for word v in document d of TF(v, d)logIDF(v). This prior shrinks the posterior probability of words with small TF-IDF scores, for example, common words that consistently appear across the corpus, in the domain topics.

Choices for domain topics (η_1): Keyword seeding

The term weighting approach relies only on patterns of word usage within the documents to create the prior. However, in many situations, domain experts may have additional knowledge about the corpus vocabulary that the TF-IDF score does not take into account. In particular, many domains have publicly available, curated lists such as key terms for article abstracts, lists of medications, or categories of accidents. We incorporate such information using *keyword* topics with a prior that reduces shrinkage on those prespecified vocabulary words. Similar to the TF-IDF prior, we set a K topic LDA model to have I stopword topics and J TF-IDF weighted topics, but then set the remaining keyword topics to promote domain specific words:

$$\beta_1, \dots, \beta_I \sim Dir(1, 1, \dots, 1)$$
$$\beta_{I+1}, \dots, \beta_{I+J} \sim Dir(c_1\overline{\mathrm{TI}(w_1)}, \dots, c_1\overline{\mathrm{TI}(w_V)})$$
$$\beta_{I+J+1}, \dots, \beta_K \sim Dir(c_2\gamma_1, \dots, c_2\gamma_V)$$

where $\gamma_i = c$ with $c \gg 1$ if w_i is a keyword and 1 otherwise, and c_2 is a scaling constant akin to c_1. The presence of the TF-IDF weighted topics serves a similar purpose as the stopword topics—providing topics for nonkeywords to fill discourages word intrusion into the keyword topics. The large prior setting on relevant domain keywords act as pseudo-counts that counteract their lower corpus frequency compared to stopwords.

We emphasize that these domain-specific keywords—words to *include* rather than words to exclude—are much distinct from domain-specific stopword lists. The keyword lists used in our experiments are large, generic, and downloaded off the Internet. For example, for the ASD keywords, we take the entire list of symptoms and diseases from the unified medical language system. We find that these very general lists of domain terminology, when used as keywords, significantly reduce the number of stopwords in those topics. Generating lists of domain-specific terminology often does not require an expert; it is easy to point other researchers to these sources of generic terminology for reproducing experiments. In contrast, domain-specific stopword lists—words to *exclude*—often require, for a given corpus, an expert to engage in an iterative process of pruning based on repeated LDA model runs.

Examined Variations

In our evaluation we use several priors: (a) the *Word Frequency Prior*, consisting of a stopword topic and the rest word frequency prior topics, (b) the *TF-IDF Prior*, containing a stopword topic and the rest TF-IDF prior topics, and (3) the *Keyword Seeding Prior*, containing a stopword topic, some TF-IDF topics, and some keyword seeding topics. We also have a *Keyword Topics Baseline*, where we have all Keyword Topic priors with no predesignated stopword topic or TF-IDF topics.

4.2 | Evaluation metrics

Following standard practice, we take the n most probable words in each topic as the ones that define the topic. We use n = 30. As our work is focused on reducing the effect of stopwords on topic quality, we consider two axes for evaluation: (a) the proportion of top words identifiable as stopwords and (b) the number of top words identifiable as domain-relevant. We use both automatic and human evaluation.

To measure the number of stopwords, we report the percentage of NLTK canonical stopwords appearing in the top n = 30 most probable words across all topics. To verify that the topics contained domain-relevant content, we asked two domain experts each in the medical autism and labor law domains to independently identify terms deemed important to generate keyword whitelists. For the NIPS corpus, we used the paper titles as whitelist words (canonical stopwords were removed) under the assumption that titles are concise signals of content. The average percentage of these whitelist words in the top 30 words of each topic are reported as the *Expert Words*. This expert whitelist evaluation is related to the studies presented in Chang et al. [5], which used Mechanical Turk to identify topic words that did not belong. We quantify the opposite—words predesignated to belong by domain experts—for three main reasons. First, expert whitelists are a more scalable evaluation method compared to Turk. Second, our corpora require more specific domain knowledge for accurate topic evaluation, making our topics less accessible for the average Turk worker. Finally, unlike the generic keyword lists, the experts were very selective in choosing important words from the corpora. Thus, we also report the co-occurrence of the top topic words with the expert-identified terms within documents (*Codocument Appearance*) as a measure of whether our top words tend to co-occur with the expert-produced lists.

4.3 | Parameter Settings

For each dataset, we performed a grid-search over the number of topics (5 to 50 topics in increments of 5), settings for prior weights c_1 (100, 10, 1, $\frac{1}{10}$, $\frac{1}{100}$) and c_2 (100, 10, 1, $\frac{1}{10}$, $\frac{1}{100}$), number of TF-IDF topics (1, 5, 10, 19), number of keyword seeding topics (1, 5, 10, 18, 19), weight of keyword seeding c (10, 50, 100, 1000), and number of Gibbs Sampling iterations (100, 200, 500, 1000). Our models were largely insensitive to these choices: the number of stopwords and number of expert words deviated little. The most important parameter setting was that the total prior weight on the stopword topics (η_0) should be larger than the total prior weight placed on the TF-IDF topics. This encourages separation between stopword and domain topics by ensuring that stopwords are sufficiently penalized in domain topics.

We present results for $K = 20$ topics, $c_1 = c_2 = 1$, and $c = 100$. Word Frequency Prior and TF-IDF Prior models were trained with $I = 1$ stopword topic. Keyword Prior models were trained with $I = 1$ stopword topic, 9 TF-IDF topics, and 10 keyword seeding topics. For the ASD data set, keywords for keyword seeding models were downloaded from the unified medical language system. The keyword seeds used for the OSHA corpus were pretagged in the data set as one-word descriptors of the accident (eg, “ship” to indicate the accident occurred on a ship). For NIPS, we used the list of 2015 NIPS submission category keywords. We emphasize that all of these keyword lists were produced automatically, without any expert curation.

4.4 | Results

Figure 1 shows our primary results. The top row shows how the different quality metrics correlate with percent of words not marked as canonical stopwords. Coherence and PMI are generally negatively correlated or relatively flat. The second and bottom rows correspond to the percent of words marked by experts as relevant, and the co-document occurrence of these expert words, respectively. A well-performing metric would be positively correlated for all these measures. These two standard quality metrics for LDA do not correlate well with our human evaluations. By contrast, log lift has consistent positive association with nonstopword rate, positive association with co-document occurrence, and, other than the OSHA set, positive association with the percent of expert words. We next discuss our findings in more detail.

4.4.1 | Traditional topic quality measures do not correctly correlate with human measures of quality

As stated above, Figure 1 shows that coherence and PMI—two standard quality metrics for LDA—do not correlate with our human evaluations when the presence of

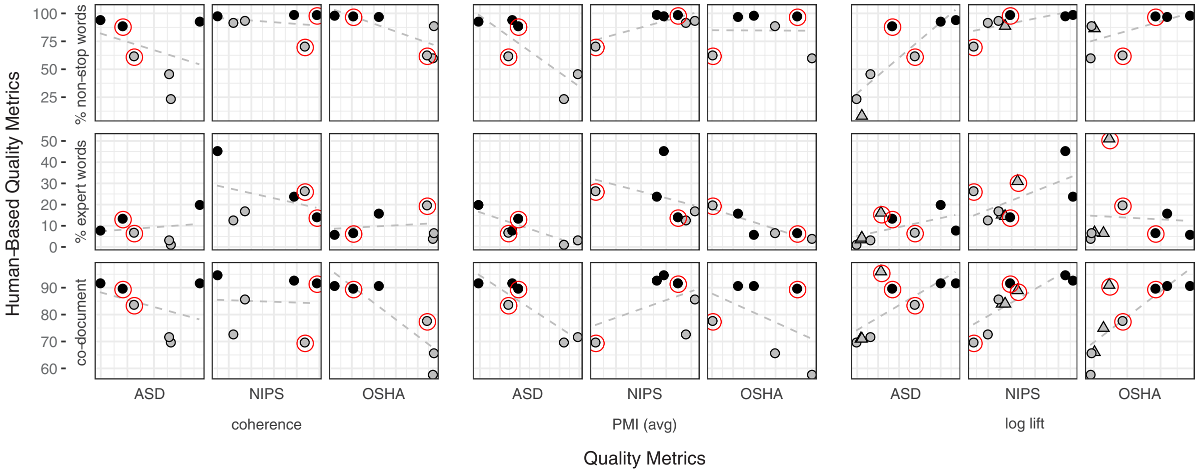

FIGURE 1 Evaluation results on the top 30 words of each topic. Each point represents a single method run on a given corpus, with quality averaged across topics. Scatterplots of machine-based quality metrics with human-based quality metrics show log lift is generally correlated with human metrics but coherence and PMI are not. For example, the correlation of model coherence with percentage of nonstopwords is negative for the ASD data set, whereas log-lift is correctly associated with this metric across all three datasets. Solid points indicate prior-based approach, gray indicates baseline. State-of-the-art baseline (both hyperparameter optimization baselines) and favored prior method (keyword prior) circled with red. Methods involving manual deletion marked with triangles. Methods with incomparable PMI or coherence due to differing vocabulary and methods with forced 0 stopword rate due to deletion are dropped

stopwords is not addressed. In fact, as shown in Table 2, we generally see the coherence and PMI scores are highest for the No Deletion Baseline, even though it contains some of the largest percentages of canonical stopwords appearing in top topic words for all three data sets.[1] Similarly, the Keyword Topics Baseline falsely appears to perform well, despite containing both more canonical stopwords and less domain words than the full informative prior models. These results confirm our mathematical analysis in Section 3: our standard quality measures systematically produce counterintuitive results when faced with irrelevant words.

We stress that these issues are not solved by stopword deletion; they plague topic model evaluation even for stopword deletion models, as these scoring mechanisms inevitably prefer topics composed of common words and domain-specific stopwords. For example, for the ASD corpus the Hyperparameter Opt Baseline (circled gray dots in Figure 1) appears to be a worse model when only looking at Coherence and PMI metrics, but clearly produces better topics compared to the No Deletion Baseline (Table 1). Standard metrics, while sensible in the absence of stopwords, produce results that prefer stopword-laden topics, and do not correlate with our human evaluation studies or expert topic evaluation.

4.4.2 | Informative priors have superior quantitative performance

Across the three data sets, the models with informative priors generally produce topics with (a) more domain-specific keywords deemed important by experts and (b) fewer stopwords. In Figure 2 generally the informative prior models have small stopword rates, high co-document ratings, and generally fair to good proportions of expertly marked words. Further, as Table 2 shows, these models outperform baselines with a hard trimming threshold such as the TF-IDF Deletion Baseline [10,14]. The stopwords that remain in topics with informative priors are almost all present in the predesignated stopword topics. Informative priors increase the number of expert-designated domain-relevant words even though those words were *not* used for the keyword seeding; our keyword seeds came from large, generic online lists. Most domain content appears in the domain-relevant topics, with the predesignated stopword topic containing very few expert words. Additionally, the co-occurrence scores reveal that topic words from the informative prior model correlate more strongly with the independently produced expert keywords. The Keyword Seeding Prior is most effective at producing topics that contain more expert words compared to the other informative prior methods as seen in Table 2. This suggests the other informative priors are effective at producing topics robust to stopword appearance in the top topic words, but addition of keyword seeding is important for producing domain-specific content. In contrast, even though the deletion-based methods reduced the number of canonical stopwords present they fail to capture as much domain content and do not remove domain-specific stopwords.

We analyze how the different methods and baselines rank against each other for the human quality metrics of % nonstop words, % expert words, and co-document appearance with expert words. The ranking is calculated by ranking each method for each metric for each corpus, then averaging the

[1] For numerical comparability, we have to leave out the Stopword Deletion Baseline and TF-IDF Deletion Baseline, as their vocabulary sizes differ from our other baselines and proposed informative prior models.

TABLE 1 Sample illustrative topics

	Qualitative topic evaluation	
	Model	**Topic**
ASD	No Deletion Baseline	Social diagnosis as an or only are autism that child
	Stopword Deletion Baseline	Schools lea information need special son statement parents support class
	TF-IDF Deletion Baseline	Had just get school of will very not me out
	Keyword Topics Baseline	Special lea it need has statement support needs school to
	Hyperparameter Opt Baseline	The to school needs support statement we permit chairman he
	Word Frequency Prior	Reading paed attention helpful short communication cope aba diagnosed system
	TF-IDF Prior	Mobility improvements treatment preschool responsible expected friends panic professionals speak
	Keyword Seeding Prior	Learning attention symptoms similar problem development negative disorder positive school
	Example Stopword Topic	Child autism or on you it parent as son have
OSHA	No Deletion Baseline	from approximately fell his hospitalized is him falling injured in
	Stopword Deletion Baseline	report trees surface backing inc degree determined forks fork board
	TF-IDF Deletion Baseline	his hospital due while work him in death pronounced tree
	Keyword Topics Baseline	head at injured falls for an balance fractures slipped lost
	Hyperparameter Opt Baseline	the employee lift number operator operating approximately jack to by
	Word Frequency Prior	mower limb top operator chain trees cutting log ground fell
	TF-IDF Prior	collapses street trees lacerations wooden laceration construction chipper facility tree
	Keyword Seeding Prior	work rope tree landing protection caught lift edge open story
	Example Stopword Topic	hospitalized employee by for at when ft fall his fell
NIPS	No Deletion Baseline	determinantal progress coordinate real learned cases theta ll super arms
	Stopword Deletion Baseline	includes top analytically margin framework incurs parameterizations normal confirmed ucsd
	TF-IDF Deletion Baseline	optimizations sgld finding others brownian strings logs generation recognize neurophysiological
	Keyword Topics Baseline	kappa keyword weights integrating similarly geometric dependence spatial definiteness either
	Hyperparameter Opt Baseline	the model learning bounded show algorithm algorithms feature optimization results
	Word Frequency Prior	mlfre vanilla wold validation inexactness benchmark gumbel bckw newton generalized
	TF-IDF Prior	scalability variations index parameter parametric calibration versions condition infinite generalize
	Keyword Seeding Prior	nonlinear scalable newton optimization hyperparameter stochastic optimality outliers epoch control
	Example Stopword Topic	based problem algorithms method from show be can learning data

Notes: We present the top 10 words of a tree accident topic for OSHA, a school difficulty topic for ASD, and a hyperparameter tuning topic for NIPS. Informative prior model topics are more specific and contain fewer stopwords. Stopword topics from the informative prior models contain both domain specific and canonical stopwords.

ranks across the corpus to rate each method by the human metric. This is displayed in Table 3. We find that the Keyword Seeding Prior and TF-IDF Prior have the lowest overall rankings, indicating they perform the best compared to other models in the human metric evaluations. The Hyperparameter Optimization with Deletion baseline performs well for % expert words and co-document appearance, but contains many stop words compared to the informative prior methods.

In a separate study of the ASD corpus, we had human evaluators identify the number of low-information words in the produced topics for three of the models: No Deletion baseline, Stopword Deletion Baseline, and TF-IDF Prior. 10 evaluators were presented with the task of identifying low-information words, each assessed 2 runs of each model. No examples were given in order to avoid priming the identification of canonical stopwords. In this experiment, human evaluators marked 71% of the words in the No Deletion Baseline as stopwords, 26% of words in the Stopword Deletion Baseline as stopwords, but only 17% of words in the TF-IDF Prior model as stopwords. Furthermore, for this model the 19 domain topics contained only 13% of marked stopwords, again indicating that the predesignated stopword topic can effectively sequester

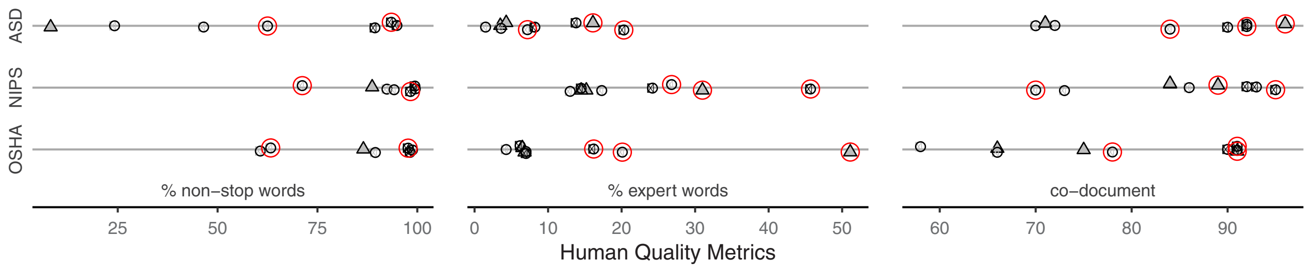

FIGURE 2 Ranking of prior-based (black) vs baseline (gray) approaches. See caption for Figure 1 for details on further annotations. Generally the solid points (prior-based methods) are high for percent non-stopword (stopword rate has been reversed to nonstopword rate for clarity), percent expert word, and co-document score, indicating superior performance for prior-based methods, in general

stopwords and prevent stopword intrusion into the domain topics.

These results again emphasize that canonical stopword deletion does not create topics that humans judge to be stopword-free. In contrast, the prior-based models can create more readable, domain-specific topics with no vocabulary removal.

Lastly, simply seeding keywords as a prior without having topic types (Keyword Topics Baseline) is not effective at reducing the stopword effect or generating domain relevant topics (Table 2). The combination of penalizing priors and topic types is required for interpretable topics.

4.4.3 | Informative prior topics are more readable

The informative prior models generate more interpretable topics (Table 1). For example, in the OSHA data set the baseline topics were overly general (eg, domain-specific stopwords such as "report" from phrases such as "an accident report was filed" in the deletion baseline) while the informative prior models captured greater specificity (eg, one topic on tree-related accidents from the Word Frequency prior shows accidents often occur when "cutting" "log[s].") In the ASD corpus, the informative prior models captured specific concerns about "learning," "reading," and "mobility" for ASD patients entering primary education. In contrast, the deletion-baseline topics included the domain-specific stopwords "son" and "parent" and addressed school concerns only vaguely. In the NIPS data set, the more concise writing and technical terminology allow the baseline models' topics to contain far fewer stopwords. However, the topic words do not form a coherent theme with each other. In contrast, the topics for the Keyword Seeding Prior and the TF-IDF Prior are much clearer as a grouping. For example, the words "optimization," "hyperparameter," and "epoch" reference tuning various model parameters.

The learned stopword topics capture both canonical and domain-specific stopwords. In the OSHA case, we see the words "employee," "ft," and "hospitalized," as well as "by," "for," and "at." For ASD, we see "child," "autism," "son," and "parent" as domain-specific stopwords. In the NIPS data set, the words "problem," "algorithms," "method," "data," and "learning" are domain-specific stopwords.

4.4.4 | The lift-score predicts quality topics

As shown on the rightmost nine panels of Figure 1, the lift-score correlates with the human-assessed performance metrics. The informative prior models perform better overall than all baselines, with the TF-IDF Prior and Keyword Seeding Prior generally the best. Unlike Coherence and PMI, lift can be calculated across LDA models of varying vocabulary sizes and is not easily maximized by topics full of frequent words.

We analyze the correlation of the lift-score metric with the three human metrics: percentage of nonstop words, percentage of expert words, and co-document appearance with expert words. Results are shown in Table 4. To understand the strength of the association between automatic and human metrics, we conduct a Spearman rank test. We use a permutation test to calculate a *P*-value, using the corpus as a blocking factor to reduce the variability between datasets. We conduct the analysis in the blocked form to aggregate the data, which provides greater power to detect trends. The results indicate log lift has a statistically significant and positive association with all three of the human metrics: percentage of nonstop words and co-document appearance with expert words. In contrast, both coherence and PMI display either no strong relationship or a negative relationship with the human metrics. As a modeling sensitivity check, we did two further analyses. We first fit a linear model of the human metric onto machine metric, using heteroskedastic-robust standard errors; the *P*-value for percent expert words changed to a marginally significant 0.10 but overall findings remained. We also conducted independent analyses for each combination of corpus, metric, and human metric; results were again broadly similar with significant relationships for most of the log-lift correlations and not the others. Results available upon request.

We analyze how the different methods rank against each other when scored using the human quality metrics of % nonstop words, % expert words, and co-document appearance with expert words. The ranking is calculated by, for each

TABLE 2 Full table of results

		Coherence		Average	log	%	%	Co-document
Corpus	Model	10 words	30 words	PMI	Lift	Stopword	Expert	Appearance
ASD	No Deletion Baseline	−45.5	−554.2	−1.56	1.94	76	2	70
	Stopword Deletion Baseline				2.17	0	4	71
	TF-IDF Deletion Baseline				2.22	92	4	71
	Keyword Topics Baseline	−48.2	−580.1	−1.42	2.61	54	4	72
	Deletion + Hyp. Opt.				3.13	0	16	96
	Hyperparameter Opt.	−105.8	−1107.9	−2.12	4.73	38	7	84
	Word Frequency Prior	−115.2	−1278.3	−2.02	3.65	15 (11)	14 (14)	90
	TF-IDF Prior	−143.3	−1611.8	−2.08	6.71	10 (5)	9 (8)	92
	Keyword Seeding Prior	−102.8	−119.6	−2.42	5.98	9 (6)	20 (20)	92
NIPS	No Deletion Baseline	−71.2	−790.7	−2.06	2.96	8	13	73
	Stopword Deletion Baseline				3.58	0	15	84
	TF-IDF Deletion Baseline				3.72	11	14	84
	Keyword Topics Baseline	−71.0	−765.2	−1.97	3.42	6	17	86
	Deletion + Hyp. Opt.				4.25	0	31	89
	Hyperparameter Opt.	−72.7	−633.2	−2.96	2.35	29	27	70
	Word Frequency Prior	−76.5	−606.5	−2.14	3.91	3 (1)	16 (14)	92
	TF-IDF Prior	−86.7	−656.8	−2.35	6.60	4 (0)	24 (24)	93
	Keyword Seeding Prior	−87.1	−825.7	−2.28	6.27	3 (2)	48 (46)	95
OSHA	No Deletion Baseline	−68.2	−831.9	−2.66	2.89	39	4	58
	Stopword Deletion Baseline				3.29	0	6	75
	TF-IDF Deletion Baseline				3.02	14	7	66
	Keyword Topics Baseline	−68.5	−819.9	−3.01	2.91	10	7	66
	Deletion + Hyp. Opt.				3.46	0	51	91
	Hyperparameter Opt.	−74.8	−899.1	−3.60	3.85	37	20	78
	Word Frequency Prior	−154.4	−1738.6	−2.80	4.83	6 (2)	8 (7)	90
	TF-IDF Prior	−171.9	−1951.2	−3.21	5.87	5 (1)	7 (6)	91
	Keyword Seeding Prior	−129.8	−1447.4	−3.36	5.18	5 (2)	17 (16)	91

Notes: First columns are quality metrics with average topic coherence and average pointwise mutual information (closer to 0 is better) and average log lift calculated on the top 30 words of all models (large is good). Remaining columns are percent stopwords and percent content words in the top topic words, and co-document appearance of marked content words and top topic words in the documents. Numbers in parenthesis are percentages for domain topics only. We omit coherence and PMI for models with canonical stopword removal as they are not comparable due to different vocabulary sets.

human metric, ranking the methods within each corpus, and then averaging each method's ranks across the corpora. We conservatively gave the stopword deletion methods perfect scores (and thus the best ranks) for percent nonstop words. We finally calculated the overall score for each method by averaging their three human metric scores. Results are in Table 3. Of the nonmanual deletion methods, TF-IDF prior scores best for percent nonstopwords. The Keyword Seeding Prior is strongest for the co-document measure and ties with hyperparameter with deletion for percent expert words. Hyperparameter optimization with deletion is best overall, partially due to the manual deletion, with the keyword prior coming in second. The TF-IDF Prior and Word Frequency Prior also score well, further suggesting the utility of the prior-based approaches.

5 | DISCUSSION

The problem of stopwords is systemic—while LDA has been empirically useful, it often picks up on spurious word co-occurrences as a result of lingual structure. For example, researchers may wish to model important nouns, but these

TABLE 3 Ranking of models by human evaluation metrics

Model	Overall Rank	% nonstopwords	% Expert words	Co-document
Keyword Seeding Prior	2.7	4.7	1.7	1.8
TF-IDF Prior	3.5	3.0	5.3	2.2
Word Frequency Prior	4.3	4.3	5.0	3.7
Deletion + Hyp. Opt.	1.8	1.5	1.7	2.3
Hyperparameter Opt.	5.8	7.7	3.3	6.3
Keyword Topics Baseline	6.0	6.3	5.5	6.2
TF-IDF Deletion Baseline	7.2	8.0	6.5	7.2
Stopword Deletion Baseline	5.1	1.5	7.0	6.7
No Deletion Baseline	8.6	8.0	9.0	8.7

Notes: The first column is the overall rank of the method compared to the other models, averaged across corpora (closer to 1 is better, averaged across corpora) and across the three human metrics. The subsequent columns display the average rank for that human evaluation metric.

TABLE 4 Correlation of Automatic Evaluation Metrics with Human Metrics

Metric	Human metric	Correlation	*P* value
Coherence	% Nonstopwords	−0.46	0.08
Coherence	% Expert words	0.00	0.71
Coherence	Co-document	−0.43	0.08
PMI (avg)	% Nonstopwords	−0.01	0.34
PMI (avg)	% Expert words	−0.70	0.00*
PMI (avg)	Co-document	−0.30	0.18
log Lift	% Nonstopwords	0.72	0.00*
log Lift	% Expert words	0.36	0.02*
log Lift	Co-document	0.79	0.00*

Notes: The correlation from a Spearman rank test is provided in the third column, and the *P* value calculated from a permutation test blocked on the corpus is shown in the fourth column. The * indicates statistical significance.

are often preceded by articles such as “the.” LDA’s bag of words assumption treats these co-occurrences as important indicators of words that appear together, allowing stopwords to have undue influence. Much prior work that has focused on improving the quality of topics does not focus on the presence of stopwords due to the widespread usage of canonical deletion methods. Our work surfaces the relevant concern that domain-specific stopwords and other high frequency words reduce topic quality, even when using techniques such as hyperparameter optimization.

We expose an important gap in topic quality evaluation—even if deletion methods are used to remove generic stopwords, human evaluators still judge the resulting topics to contain large quantities of low-information words. Furthermore, traditional topic quality measures did not reveal these trends. Our proposed lift-score, however, correlates to both human stopword evaluation and domain expert topic assessment, and can be used to assess topic quality in the presence of stopwords. However, more generally, an important question is to define an appropriate constellation of metrics that capture different factors such as concentration, uniqueness, coherence, and relevance, all of which are relevant to evaluating topic quality. Previous work has indicated that using individual metrics alone struggle to capture holistic topic quality [18], suggesting instead an evaluation of multiple metrics together. We believe assessing the presence of stopwords, domain or otherwise, is an important part of this overall evaluation strategy.

We also showed that simply adding specific informed priors that penalize uninteresting occurrence patterns and promote relevant words can create more interpretable topics by reducing the presence of domain-specific and canonical stopwords. In particular, the TF-IDF informative prior model not only drastically reduces the number of canonical stopwords appearing in the top 30 words of each topic, but also curtails the number of general, low information words. These informed priors are easily incorporated into existing software by simply changing the existing symmetric Dirichlet prior on the word-topic distribution to one of the proposed priors, with no other inference modifications. See Appendix for code snippets demonstrating this. This ease of approach is particularly noteworthy for the Keyword Seeding Prior, as many other LDA models that incorporate external information require custom inference methods that may not be accessible to all users. We also found that our prior parameter settings are also quite robust and require little modification. Despite the large structural differences between the corpora, the same parameters produced interpretable topics that performed well both quantitatively and qualitatively.

Interesting avenues for future work include assessing our lift-score metric with regards to additional human evaluations. It would also be interesting to see whether incorporating informative priors into much more complex topic models, such as supervised LDA models with correlational and time-varying structure, provides similar gains. Implementation-wise, the simplicity of setting priors is a strength. Informative priors could be easily incorporated into these more complex works. In fact, in these scenarios, more elaborate modeling of word frequencies might render the larger effort computationally infeasible, making prior-setting even more critical. More generally, it would be interesting to see whether these more interpretable topics show benefits in downstream prediction tasks.

ACKNOWLEDGMENTS

We thank Hanna Wallach for helpful discussions and the many volunteers who took part in the human evaluation studies. We also thank Robin and Janet Ackerman for the OSHA dataset.

ORCID

Angela Fan https://orcid.org/0000-0002-5478-3368

REFERENCES

1. D. Andrzejewski, X. Zhu, and M. Craven, *Incorporating domain knowledge into topic modeling via Dirichlet forest priors*. Proceedings of International Conference on Machine Learning, Montreal, 2009, 382(26).
2. V.P. Baradad, and A.M. Mugabushaka, *Corpus specific stop words to improve the textual analysis in scientometrics*, European Research Council Executive Agency, 2015.
3. J. Bischof, and E.M. Airoldi, *Capturing semantic content with word frequency and exclusivity*, Proceedings of the 29th International Conference on Machine Learning, Edinburgh, 2012.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet allocation*, J. Mach. Learn. Res. 3 (2003), 993–1022.
5. J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D.M. Blei, *Reading tea leaves: How humans interpret topic models*, Advances in Neural Information Processing Systems, Vancouver, Canada, 2009, pp. 288–296.
6. Y. HaCohen-Kerner, and S.Y. Blitz, *Initial experiments with extraction of stopwords in hebrew*, KDIR, Valencia, 2010, pp. 449–453.
7. J. Jagarlamudi, H. Daumé, and R. Udupa, *Incorporating lexical priors into topic models*, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, 2012a, pp. 204–213.
8. J.H. Lau, D. Newman, and T. Baldwin, *Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality*. EACL, Gothenburg, 2014, pp. 530–539.
9. M. Lee, and D. Mimno, *Low-dimensional embeddings for interpretable anchor-based topic inference*, Proceedings of Empirical Methods in Natural Language Processing, Citeseer, Doha, 2014.
10. R. T. W. Lo, B. He, and I. Ounis, *Automatically building a stopword list for an information retrieval system*, J. Digi. Inf. Manage. 5 (2005), 17–24.
11. M. Makrehchi and M. S. Kamel, Automatic extraction of domain-specific stopwords from labeled documents, in *Advances in information retrieval*, Springer, Glasgow, 2008, 222–233.
12. R. Mehrotra et al., *Improving LDA topic models for microblogs via tweet pooling and automatic labeling*, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Dublin, 2013, 889–892.
13. D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum, *Optimizing semantic coherence in topic models*, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011.
14. Z. Y. Ming, K. Wang, and T. S. Chua, *Vocabulary filtering for term weighting in archived question search*, in *Advances in knowledge discovery and data mining*, Springer, Hyderabad, 2010, 383–390.
15. D. Newman, J.H. Lau, K. Grieser, and T. Baldwin, *Automatic evaluation of topic coherence*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, 2010, pp. 100–108.
16. D. Newman, E.V. Bonilla, and W. Buntine, *Improving topic coherence with regularized topic models*. Advances in Neural Information Processing Systems, Granada, 2011.
17. M. J. Paul and M. Dredze, *Discovering health topics in social media using topic models*, PLoS One 9(8) (2014), e103408.
18. M. E. Roberts et al., *Structural topic models for open-ended survey responses*, Am. J. Polit. Sci. 58 (2014), 1064–1082.
19. H. Saif, M. Fernandez, H. Alani, *Automatic stopword generation using contextual semantics for sentiment analysis of twitter*, Proceedings of the 2014 International Conference on Posters & Demonstrations Track, CEUR-WS. org, 2014,1272, pp. 281–284.
20. G. Salton, *Developments in automatic text retrieval*, Science 253 (1991), 974–980.
21. A. Schofield, M. Magnusson, and D. Mimno, *Pulling out the stops: Rethinking stopword removal for topic models*. EACL, Valencia, Spain 2017. p. 432.
22. C. Sievert, and K.E. Shirley, *Ldavis: A method for visualizing and interpreting topics*, Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014, pp. 63–70.
23. M.P. Sinka and D. Corne, *Evolving better stoplists for document clustering and web intelligence*, HIS, Amsterdam, 2003, pp. 1015–1023.
24. M. Taddy, *On estimation and selection for topic models*, AISTATS, La Palma, 2012, pp. 1184–1193.

25. Y. Tan, and Z. Ou, *Topic-weak-correlated latent Dirichlet allocation*, Chinese 2010 7th International Symposium on Spoken Language Processing (ISCSLP), IEEE, Tainan, 2010, pp. 224–228.
26. H. M. Wallach, D. Mimno, and A. McCallum, *Rethinking lda: Why priors matter*, in *Advances in Neural Information Processing Systems*, Vancouver, 2009, pp. 1973–1981.
27. S. Wibisono and M. S. Utomo, *Dynamic stoplist generator from traditional Indonesian cuisine with statistical approach*, J. Theor. Appl. Inf. Technol. 87 (2016), 92–98.
28. P. Xie, D. Yang, and E.P. Xing, *Incorporating word correlation knowledge into topic modeling*. Conference of the North American Chapter of the Association for Computational Linguistics*,* 2015.
29. Y. Yang, D. Downey, J. Boyd-Graber, and J.B. Graber, *Efficient methods for incorporating knowledge into topic models*, Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015.
30. W.X. Zhao, J. Jiang, J. Weng, et al. *Comparing Twitter and traditional media using topic models*, European Conference on Information Retrieval*,* Springer, 2011, pp. 338–349.

How to cite this article: Fan A, Doshi-Velez F, Miratrix L. Assessing topic model relevance: Evaluation and informative priors. *Stat Anal Data Min: The ASA Data Sci Journal*. 2019;12:210–222. https://doi.org/10.1002/sam.11415

APPENDIX

We include some basic sample code (in Python) to illustrate how to create our priors.

```
import numpy as np
import math

text = ["Alice was beginning to get very tired of sitting", \
       "by her sister on the bank ,", \
       "and of having nothing to do :", \
       "once or twice she had peeped into the book her sister was reading", \
       "but it had no pictures or conversations in it", \
       "’ and what is the use of a book , ’", \
       "thought Alice", \
       "’ without pictures or conversations ? ’"]

keywords = [’book’, ’pictures’, ’conversations’]
vocab = sorted(list(set(" ".join(text).split())))

def priorData(data, keywords):
    dataDict = {}
    numDocuments = float(len(data))
    numWords = 0
    for index, words in enumerate(data):
        words = words.split(" ")
        docLength = float(len(words))
        numWords += docLength
        for word in set(words):
            wordCount = sum([word == i for i in words])
            if word not in dataDict:
                dataDict[word] = {"wordCount": 0, "tf": {}, "keyword": 0, "numDocAppearance": 0}
            dataDict[word]["wordCount"] += wordCount
            dataDict[word]["tf"][index] = wordCount / docLength
            dataDict[word]["numDocAppearance"] += 1

    for word in keywords:
        dataDict[word]["keyword"] = 1

    for key in dataDict:
        dataDict[key]["wf"] = dataDict[key]["wordCount"] / numWords
        dataDict[key]["idf"] = math.log(numDocuments / dataDict[key]["numDocAppearance"])
        dataDict[key]["tfidf"] = np.mean(list(dataDict[key]["tf"].values())) * dataDict[key]["idf"]

    return dataDict

def buildPrior(priorData, vocab, numStopwordTopics=0, numWFTopics=0, numTFTopics=10, numKeywordTopics=0, c1=1, c2=10):
    def buildStopwordTopic():
        return [1.0 for _ in vocab]

    def buildWFTopic():
        return [1.0 / priorData[word]["wf"] for word in vocab]

    def buildTFIDFTopic():
        return [c1 * priorData[word]["tfidf"] for word in vocab]

    def buildKeywordTopic():
        return [c2 * priorData[word]["keyword"] for word in vocab]

    prior = [buildStopwordTopic() for i in range(numStopwordTopics)] + \
            [buildWFTopic() for i in range(numWFTopics)] + \
            [buildTFIDFTopic() for i in range(numTFTopics)] + \
            [buildKeywordTopic() for i in range(numKeywordTopics)]

    return prior

# example usage:
priorStatistics = priorData(text, keywords)
modelPrior = buildPrior(priorStatistics, vocab, 1, 1, 1, 1)
```