A Prior Setting that Improves LDA in both Document Representation and Topic Extraction

Juncheng Ding
Department of Computer Science and Engineering
University of North Texas
Denton, Texas 76201
junchengding@my.unt.edu

Wei Jin
Department of Computer Science and Engineering
University of North Texas
Denton, Texas 76201
Wei.Jin@unt.edu

***Abstract*—Latent Dirichlet Allocation (LDA), as the most popular topic model, models documents as mixtures of topics and topics as mixtures of words. Topic mixture well represents documents while words mixture extracts meaningful topics from a corpus. The nature of LDA makes it a powerful tool in documents organizing and corpus summarizing. One limitation of LDA is that its performance depends heavily on the priors. Researchers show priors matters in LDA and propose methods to learn the priors for better modeling, regardless of using symmetric priors. However, LDA modeling ability does not necessarily consent with the performance of LDA in documents representation and topic extraction. In this paper, we propose a novel prior setting for LDA. The setting improves LDA in both documents representation and topic extraction performance. We experiment to compare our setting with symmetric priors and previously proposed priors that enhances modeling ability. Experiments on the topic quality show that LDA with our prior setting extracts better topics than LDA with other kinds of prior settings. We compare LDA document representation ability through tasks such as document clustering and document classification. These experiments demonstrate LDA with our proposed priors represents document better. Moreover, our analyses also reveal that better modeling does not necessarily lead to better performance in documents representation and topic extraction.**

***Index Terms*—topic model, LDA, prior**

I. Introduction

The number of digit text documents grows explosively in modern society due to the rapid development of technologies. The vast amount of text records provides people with numerous information covering various fields. Nevertheless, people find it hard to find information or documents they need in such ocean of documents. The conflict leads to the problem of organizing these documents in different categories while explicitly describing the categories. The number of classes in large corpora is considerable, which makes it hard to organize the documents through a supervised way because it is too expensive to label the data. Meanwhile, clustering the documents in an unsupervised manner provides no explicit description of the documents categories, which compromises the benefit of clustering in this application.

Topic models provide a solution to the above problem. They represent documents as mixtures of topics in the format of fix-length real-value vectors, which is inherently a categorizing. Meanwhile, topics in the model are mixtures of words, which describes the topic itself. Figure 1 is an example of topic analysis of a paragraph from New York Times[1]. In Figure 1, we learn the topics out of a documents collection and infer the topic distribution of the example paragraph. A rank of words describes each topic according to the words' probabilities appearing in the respective topic (we only choose the top three words in each topic in Figure 1). The text representation is the probabilities of all the topics appearing in it. By using the topic model, we can characterize the documents using topics which have explicit expressions as well. Moreover, it is evident that the performance of topic models lays in two aspects: topic quality and document representation quality.

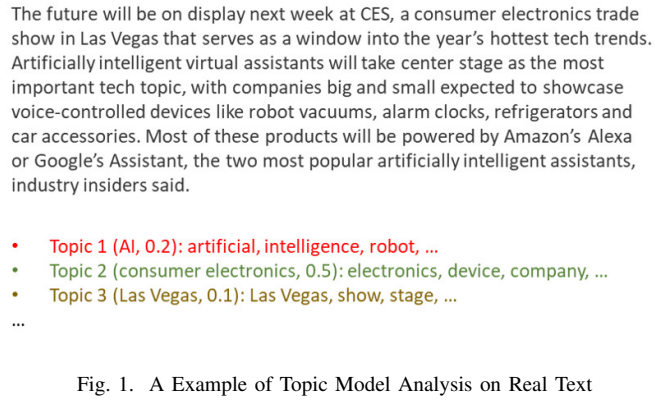

Fig. 1. A Example of Topic Model Analysis on Real Text

There are currently many existing topic models with different characteristics, the most popular one of which is Latent Dirichlet Allocation (LDA) [1]. LDA is prevalent in data mining, machine learning, and statistical natural language processing research societies. The model assumes a generative process of two steps, i.e., topics generation and documents generation. The generative nature makes LDA depend heavily on the hyper-parameters (priors) setting, especially on the priors for topics distribution of documents, regarding its performance and modeling ability. Researchers show priors matters and propose a prior optimization method during posterior inference [2] to gain better modeling ability (i.e., after training, LDA with their prior setting achieves higher likelihood when predicting previously unseen documents in training dataset).

[1]https://www.nytimes.com/2019/01/03/technology/personaltech/tech-2019-overhyped.html

978-1-7281-2009-6/$31.00

However, the performance of LDA in real tasks does not necessarily agree with its modeling ability, i.e., good modeling ability does not always lead to better performance in real tasks [3].

In this paper, we propose a prior setting for LDA, which improves LDA in both document representation and topic extraction. Different from the previous approach that optimizes the priors in the process of posteriors inference, our proposed parameter setting learns the parameter before inference. We also conduct experiments to verify the effectiveness of our prior setting. Experiments on the topic quality show that LDA with our prior setting gains topics with better quality. To demonstrate the improvement in documents representation quality, we further compare the document's representation of LDA with our proposed priors, LDA with previous prior optimization and LDA with symmetric priors by conducting standard tasks such as documents clustering and document classification. The results validate that our proposed prior settings improve the documents representation power of LDA. We also experiment to show better modeling ability does not always align with better performance in other tasks.

The rest of this paper is structured as follows. Section 2 introduces LDA, its priors and previous prior settings. The proposed prior setting method is in Section 3. The experiments on documents representation and topic description together with discussion are in Section 4. Section 5 is the conclusion.

II. Related Work

There are currently many topic models, of which Latent Dirichlet Allocation (LDA) is the most prevalent and recognized one [1]. LDA assumes a two-stage generative process of a documents collection, i.e., topics generation and documents generation. The generative nature of LDA makes its performance depend heavily on the priors definition in its various applications. Therefore, it is necessary to set the priors properly. Researchers show that prior setting affects LDA in modeling ability and propose a method to optimize the prior during posterior inference for better modeling ability [2]. In this section, we will introduce the LDA model, why prior setting matters in LDA, and the previous parameter optimization method, as the background of our work.

A. LDA and Its Priors

The topic model is a family of models [4]–[6] that assume a topic layer between document layer and word layer, i.e., a document is a mixture is topics, and a topic is a mixture of words in a topic model. A topic model provides useful information about a corpus in two aspects: semantically meaningful topics information and fixed-length vector documents representations. The topics information is the description of topics as a distribution of words. Specifically, the representation of a topic is a vocabulary-length real-value vector of which each entry is the probability of a word appearing in the topic (for example, "currency" appears in topic "economics" with a high probability, so the respective entry of "currency" in "economics" vector is a relatively large value.). The representation of the topic thus provides semantically meaningful information. In a specific topic model, the topic number is a fixed value. The document representation is a real-valued vector of the topic number length, of which each entry is the probability of each topic appearing in the document. A quantity of experiments on real tasks has proven the effectiveness of this representation.

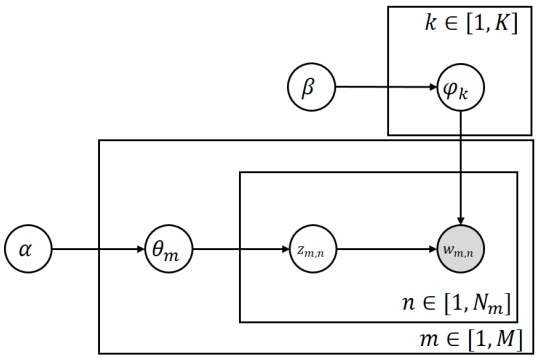

Fig. 2. The Generative Process of LDA

LDA, which is the most used topic model and was firstly proposed by David Blei et al. [6] in 2003, is a three-level hierarchical generative probabilistic model for collections of discrete data such as text corpora. LDA assumes a generative process of a corpus as in Figure 2.

In the generative process of a corpus, an author samples K topics $\phi_k (k = 1, 2, ..., K)$, which are specific multinomial word distributions, according to a Dirichlet distribution (defined by **document priors** α, which is a topic number length vector) before generating any document as in Equation 1.

$$\phi_k \sim Dirichlet(\alpha) \tag{1}$$

When generating a document, the author samples a multinomial topic distribution $\theta_m (m = 1, 2, ..., M)$ for document m, which is the respective document representation, according to another Dirichlet distribution (defined by **topic priors** β, which is a vocabulary number length vector) as in Equation 2.

$$\theta_m \sim Dirichlet(\beta) \tag{2}$$

According to the multinomial topic distribution, the author samples a topic. According to the multinomial word distribution respective to the topic, the author then samples a word. The author repeats this process as in Equation 3 to generate documents.

$$\begin{aligned} z &\sim Multinomial(\theta_m) \\ w &\sim Multinomial(\phi_t) \end{aligned} \tag{3}$$

where z is the topic index, and w is the word index. Two groups of priors or hyper-parameters, which are topic priors and document priors respectively, defines the generative process of a corpus.

B. Why Prior Matters?

The priors play an essential role in the posterior inference process, which is to infer the document's representation Θ

and topic description Φ given the corpus and priors. There are many inference methods for LDA. We choose Gibbs sampling to explain the importance of priors.

Gibbs sampling infers the posterior in two steps: randomly sampling a topic for each word and estimate the posteriors accordingly [7]. The first step samples the topic number of a word according to Equation 4.

$$p(z_i = j|\mathbf{z_{-i}}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\beta} \tag{4}$$

where i means the i^{th} word in document d, W is the vocabulary number, T is the topic number, and $n_{-i}^{(\cdot)}$ is a count that does not include the current assignment of z_i. The second step is the estimation of posteriors according to the sampled topics as in Equation 5.

$$\begin{aligned} \hat{\phi}_j^{(w)} &= \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \\ \hat{\theta}_j^{(d)} &= \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + T\alpha} \end{aligned} \tag{5}$$

Note that α and β in the above process are constant values, which assumes that all the entries in α or β are of the same value respectively. This prior setting is smoothing prior setting or symmetric prior setting.

In the Gibbs sampling process, we first randomly assign a topic number to each word and go through all the words in all the documents to sample the topic number. The process iterates several times till convergence. The final step is to estimate the posteriors according to the sampling.

According to the sampling process described in Equation 4 and Equation 5, we can get different documents representation and topic description by changing α or β before sampling, which shows that the performance of LDA depends heavily on priors setting. The same goes for other sampling methods because the corpus and priors are the evidence of all sampling methods. Previous research also shows that the selection of priors matters in LDA [2]. Especially, Wallach et al. also show that the document prior α contributes much greater than topic prior β [2].

C. Current Prior Settings

Although priors matters in LDA, the most common practice in the prior setting is to use symmetric prior setting, in which all the entries are the same in α or β.

Notably, researchers also proposed possible asymmetric prior setting that reports significantly improving the modeling ability of LDA [2]. In their proposed parameter setting, they integrate the priors out and learn the priors during inference implicitly. The topic sampling probability after integrating out prior is in Equation 6.

$$P(z_{N_d+1}^{(d)}|\mathcal{Z}, \alpha, \alpha' u) = \frac{N_{t|d} + \alpha \frac{\hat{N}_t + \frac{\hat{\alpha}}{T}}{\sum_t \hat{N}_t + \alpha'}}{N_d + \alpha} \tag{6}$$

where α is the sum of all entries in the document prior. The sampling process introduces global sampling and document sampling to integrate the entries of the document prior out, thus making it possible to learn asymmetric priors according to the corpus. This proposed prior setting improves the modeling ability of LDA, which is measured by predicting the probability of previous unseen document in the training corpus.

However, researchers show better modeling ability does not necessarily lead to better (more consistent with human interpretation) document representation or topic extraction [3], which compromises the effectiveness of the previous proposed prior setting. In this paper, we propose a novel prior setting to LDA that improves the performance of LDA in real tasks such as documents representation and topic extraction. We will compare the proposed prior setting to the above two in both document representation and topic extraction in our experiments.

III. The Proposed Prior Setting

Document priors α affects the performance the LDA significantly [2], which are the Dirichlet priors when sampling topic distribution for each document. The fact provides us with a chance to improve the model performance by learning more appropriate document priors. In this section, we will propose a new document prior setting method for LDA. Different from the previous study that acquires the "optimal parameters" during the iterative posteriors inference process [2], our proposed method learns the prior before posteriors inference.

Document priors are the prior confidence of topic distribution of a corpus in LDA. We assume a specific topic appears with a higher probability by setting the respective entry in the document prior α with a higher value (i.e., if topic k is more likely to appear in a corpus, α_k is larger). In other words, document priors describe the membership information of a corpus. The above view of document priors leads to a new prior setting. We can infer the membership information for each document first and integrate the documents' information out to gain the general membership information.

Soft clustering solves the membership learning problems [8]. We assume there are m clusters. For a set of data points $\{X|x_1, x_2, ..., x_n\}$, soft clustering learns a cluster probability vector $(c_{1j}, c_{2j}, ..., c_{mj})$ of each point x_j, of which the entry c_{kj} is the probability that the point x_j belongs to the cluster k. We can do soft clustering first to learn a clusters distribution of each document. The clustering provides membership information for each document. The membership information of the corpus is the integration of the membership information of all documents. We can regard the learned corpus-level membership information as the documents priors we need.

We use fuzzy c-means (FCM) [9] to do clustering in our approach which assigns each item a membership grade to each cluster thus fits our purpose well. Because FCM faces serious problem when the feature dimension is too high [10] and our current documents representations are vectors with thousands

of dimensions (which is equal to unique words numbers in the vocabulary), we introduce Principle Component Analysis (PCA) [11] to reduce the dimensions while preserving the key differences before doing soft clustering.

The learning process includes three procedures in general: documents representation, soft clustering, and integration as in Algorithm 1. We use five steps to explain the process:

Step1: In this step, we formulate a word frequency vector W of vocabulary size length for each document, and each entry of a document vector is the respective word occurrence times in the document (e.g., w_{ij} is the i^{th} word occurrence times in document j).

Step2: According to the word occurrence times vectors $W_1, W_2, ..., W_m$, we represent each document in TF-IDF format as in Equation 7.

$$r_{ij} = w_{ij} * \log \frac{n}{\sum_{j=1}^{n} w_{ij}} \tag{7}$$

The representation assigns each term a weight according to its importance in the corpus [12]. Each document representation is a vector R of vocabulary size length. Note that we can also replace TF-IDF with other representation such as entropy [12] in this step. The R vectors are features of documents in the following steps.

Step3: Because FCM will lose its effect in high dimensional space, we employ PCA on all the documents to reduce the feature space into two dimensions. After conducting PCA, each document representation is a two-dimensional vector $R2D$.

Step4: We use FCM [9] algorithm to do soft clustering in this step. The number of clusters is the number of topics m. FCM minimize the following function:

$$C_1, C_2, ..., C_n = \underset{C_1, C_2, ..., C_n}{\operatorname{argmin}} \sum_{k=1}^{m} \sum_{j=1}^{n} c_{kj}^{q} * \|R2D_j - CEN_k\|^2 \tag{8}$$

$$c_{kj} = \frac{1}{\sum_{o=1}^{m} (\frac{\|R2D_k - CEN_j\|}{\|R2D_k - CEN_o\|})^{\frac{2}{q-1}}} \tag{9}$$

$$CEN_k = \frac{\sum_{j=1}^{n} (c_{kj})^q x_j}{\sum_{j=1}^{n} (c_{kj})^q} \tag{10}$$

subject to:

$$\begin{aligned} 0 \leq c_{kj} \leq 1 \\ \sum_{k=1}^{m} c_{kj} = 1 \\ 0 < \sum_{j=1}^{n} < n \end{aligned} \tag{11}$$

where q is the fuzzifier and $0 < q < \infty$. CEN is the center a cluster. We define $q = 2$ in this algorithm.

After soft clustering in this step, each document has a membership vector $p(cluster|document)$ in the format of length m vector C, each entry of which is the probability that the document belongs to the respective cluster. We regard the membership information as the "topic information". Therefore, we can describe each document with "topic" distribution C.

Step5: To achieve the belief of unbalanced "topic" distribution as $p(cluster)$, we integrate the cluster membership information through Equation 12.

$$\begin{aligned} p(cluster) &= \sum_{documents} p(cluster|document)p(document) \\ &= \sum_{j=1}^{n} c_{kj} * p_j \end{aligned} \tag{12}$$

where p_j is $p(document)$ as in Equation 13.

$$p(document) = \frac{\sum_{i=1}^{V} w_{ij}}{\sum_{j=1}^{n} \sum_{i=1}^{V} w_{ij}} \tag{13}$$

Finally, we get the document prior α as the cluster probability $p(cluster)$. In the view of the discriminative model, $\alpha = p(cluster)$ is the maximum posterior estimation of the real document priors [13]. By setting priors in this way, we corporate the advantage of the discriminative model to improve the performance in a generative model.

IV. Experiments and Discussion

The utility of LDA includes two aspects: the extraction of topics and the representation of documents as in Figure 1. To evaluate the utility of LDA, we can evaluate its performance in both topic quality and documents representation quality. In this section, we compare LDA with different prior settings in the ability of topics extraction and documents representation. The first part is the introduction of the dataset and its preprocessing. The experiments part contains topic quality evaluation, documents clustering, documents classification, and an analysis of modeling ability. Note that in our following experiments we conduct each one (especially learning the priors) several times to avoid causality of our proposed prior setting due to the stochastic nature of our method.

A. Dataset and Implementation

We used "Health Twitter" dataset [14] and part of "20 Newsgroups" [15] to experiment. The "HealthTwitter" are tweets from several popular health-related Twitter accounts, and we only preserve the text body. "20 Newsgroups" is a famous dataset containing 20 different newsgroups. We choose 4 categories from it, which are "alt.atheism", "talk.religion.misc", "comp.graphics", and "sci.space". Before preprocessing, we remove the "headers", "footers", and "quotes".

The preprocessing process includes tokenizing, changing the letters into lower case, removing stop words, remove words that are less than three letters or appears less than two times in the corpus, and lemmatizing. Moreover, we collect the possible bigrams in the text [16]–[19]. After the above preprocessing, for "Health Twitter" dataset, we got 63326 documents and 24536 different words, and the documents have an average

Algorithm 1 Learn LDA Priors

Input: Number of topics m, Preprocessed text documents $D_1, D_2, ..., D_n$, *Vocabulary*
Output: Prior vector α
procedure DOCUMENT REPRESENTATION($D_1, D_2, ..., D_n$)
Step1: Represent each document with word frequency as vectors $W_1, W_2, ..., W_n$ according to *Vocabulary*
Step2: Convert the vectors $W_1, W_2, ..., W_n$ into *TF-IDF* representation as vectors $R_1, R_2, ..., R_n$
procedure SOFT CLUSTERING($R_1, R_2, ..., R_n, m$)
Step3: Use *PCA* to reduce each R_k into two-dimension vector $R2D_k$, $k = 1, 2, ..., n$
Step4: Do soft clustering on $R2D_1, R2D_2, ..., R2D_n$ using *FCM*, and get $C_1, C_2, ..., C_n$
procedure INTEGRATION($W_1, W_2, ..., W_n, C_1, C_2, ..., C_n$)
Step5: Integrate to get cluster probability vector $p_1, p_2, ..., p_m$, and $\alpha_k = p_k$, $k = 1, 2, ..., m$

length of 5.9 words. For the "20 Newsgroups" dataset, we got 3285 documents and 13518 different words. The average documents length is 49.6 words. The statistics of the dataset is in Table I.

TABLE I
DATASET

Dataset	"Health Twitter"	"20 Newsgroups"
Documents Number	63326	3285
Different Words	24536	13518
Average Length	5.9	49.6

For model implementation, we employ Gibbs sampling [7]. The number of iteration is 50 in our following experiments. Since the model inference involves random initialization, we repeat each experiment which includes LDA inference ten times to eliminate the effect caused by random initialization. The conclusion of our experiment relies on multiple runs and statistical tests. We also change the number of topics in our experiments to make our results more reliable. For comparison, we choose three document prior settings: the symmetric priors (noted as "symm.") [7], the previous optimization (noted as "prev.") [2], and our proposed prior setting (noted as "prop."). The priors for topics are all symmetric priors for the three settings. The symmetric priors are priors with the same values for all entries in the vector. The previous optimization is using inference method as in Equation 6, which integrates the priors out and learns the asymmetric prior during inference. For our proposed prior setting, we learn the priors before inference.

B. Topic Evaluation

We describe topics in the topic model as the ranked list of words according to their importance (probability) in the topic. The quality of a topic is how semantically meaningful the topic is. There are two categories of methods to judge the quality of a topic: the extrinsic method and the intrinsic method. The extrinsic method involves human judgment, which is different for different people so it is not well defined (e.g., different people may have different interpretations of a topic). The intrinsic method measures the topic quality according to the word co-occurrence in a corpus. In intrinsic measures, we will get high scores if the words describing the topic co-occurs with a high probability in the corpus, and high scores indicate good topic quality. UMass score [20] is one representation of intrinsic topic coherence measure as in Equation 14, which considers the word order as well.

$$score_{umass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + 1}{D(w_i)} \quad (14)$$

where N is the number of words describing the topic, $D(w_i, w_j)$ is the number of documents containing both word w_i and w_j, and $D(w_i)$ is the number of documents containing w_i.

We use UMass score as the topic description quality measure in our following experiments. The final UMass score is an averaged one of all the topics in the model. The models are LDA with the symmetric priors (noted as "symm."), the previous optimization (noted as "prev."), and our proposed prior setting (noted as "prop."). Since the number of topics is one possible factor affecting the quality of topics, we use different numbers of topics in each dataset. The numbers of topics are different in the different dataset according to their respective scales. We reserve the most frequent 20 words to describe a topic for all models. For "Health Twitter", we use 80, 90, and 100 topics in this experiments. The topic numbers for "20 Newsgroups" are 5, 10 and 15.

The UMass scores achieved are in Table II and Table III. We record the data in the format of all the trials' $mean \pm stand\_deviation$. The bold text indicates significantly better results than its comparative models. Table II is the UMass scores of topics extracted from "Health Twitter" dataset. The table shows our proposed prior setting achieves topics of the highest Umass scores given different numbers of topics. Meanwhile, we can observe no significant difference of Umass scores between LDA with symmetric priors and LDA with the previous optimal setting. Table III is the UMass scores of topics extracted from "20 Newsgroup" dataset. We can see that LDA with our proposed prior setting extracts topics of the highest Umass scores when the number of topics is 5 and 10. When there are 15 topics, there is no significant difference in Umass scores between the three parameter settings. Moreover, the Umass scores from "20Newsgroups" are consistently better than that from "Health Twitter", which indicates that the

TABLE II
UMASS SCORES OF "HEALTH TWITTER"

Number of Topics	symm.	prev.	prop.
80	-17.7331 ± 0.2054	-17.6434 ± 0.2543	**-16.0915 ± 0.0619**
90	-17.9127 ± 0.1307	-18.0205 ± 0.1205	**-15.6297 ± 0.1146**
100	-17.4354 ± 0.0931	-17.452 ± 0.0451	**-16.3405 ± 0.0790**

TABLE III
UMASS SCORES OF "20 NEWSGROUPS"

Number of Topics	symm.	prev.	prop.
5	-1.9469 ± 0.1672	-1.9523 ± 0.1030	**-1.8466 ± 0.1277**
10	-2.1345 ± 0.1223	-2.1442 ± 0.1424	**-1.9826 ± 0.0855**
15	-2.3329 ± 0.1879	-2.3476 ± 0.2170	-2.4199 ± 0.2165

average length of documents may affect the Umass scores as well.

These experiments demonstrate that LDA with our proposed prior setting extracts topics of higher topic coherence than the other two settings. As in [21], [22], higher topic coherence scores indicate more semantically meaningful topic. We can conclude from our experiments that LDA with our proposed prior settings extracts better topics than current settings.

C. Documents Clustering

LDA provides fix-length vector documents representation in the form of the distribution of topics given specific document $p(topic|document)$, which is hard to evaluate. We cluster documents using the document's representation and evaluate the clustering performance. The clustering quality is a measure of documents representation quality.

In this documents clustering experiment, the document features are their respective topics distributions. The clustering algorithm is k-means. The inputs of k-means are the number of clusters and initialization of each cluster. We use "k-means++" to initialize in our case.

There is no ground truth for clustering. Therefore, we choose the metric of Calinski-Harabaz Index [23] as in 15 to evaluate the clustering quality, which is the ratio of the between-clusters dispersion mean and the within-cluster dispersion without knowing the ground truth.

$$
\begin{aligned}
s(k) &= \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1} \\
W_k &= \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \\
B_k &= \sum_{q} n_q (c_q - c)(c_q - c)^T
\end{aligned}
\tag{15}
$$

where N is the number of samples, C_q is the samples in cluster q, c_q is the center of cluster q, n_q is the number of samples in cluster q, and c is the center of all samples.

Higher Calinski-Harabaz Index indicates better clustering quality. Note that the score depends on the number of samples as in Equation 15. We can only compare the scores from the setting dataset, rather than the tendency of scores.

We use "Health Twitter" in clustering experiments. Since cluster number is another factor that affects clustering quality, we change the number of clusters in our experiments besides the number of topics. The cluster numbers are 5, 10, 15, and 20. The results are in Table IV.

The clustering results are in Table IV. The scores in the table are also in the format of all the trials' $mean \pm stand\_deviation$. The bold text indicates significantly better results. We can observe that LDA with our proposed prior setting achieves significantly higher CH scores than LDA with the other two settings. The table reveals that LDA with our proposed prior setting performs significantly better than the other two settings in documents clustering, which verifies that our proposed prior setting improves LDA in documents representation.

D. Documents Classification

For documents with ground truth (i.e., class labels), we evaluate the documents representation quality by conducting documents classification. Better documents representation performs better in the documents classification task. We use the distribution of topics given specific document $p(topic|document)$ as well in this experiments.

The classifier is Multi-Layer Perceptron with 10 hidden layers, "relu" as the activation function and "adam" as the solver [24]–[26]. We use precision, recall, and f1-score as the classification metrics. Since our task is a multi-labeled classification one, we used the weighted average score, i.e., the final score is the weighted average of scores from all categories [27]. Higher scores indicate better classification performance.

The dataset in this experiment is "20 Newsgroups". There are four categories in the dataset. We split the dataset into 80% and 20% as training data and testing data respective. To avoid the problem of unbalanced training data, we split each category of data separately and merge the documents from the four categories. The numbers of topics are 5 and 10 in documents classification.

The documents classification results are in Table V and Table VI. The records are in the format of all the trials' $mean \pm standard\_deviation$. The bold text indicates significantly better results. Table V is the classification result

TABLE IV
THE CALINSKI-HARABAZ INDEX IN DOCUMENTS CLUSTERING

Topic Number	Cluster Number	symm.	prev.	prop.
80	5	1275.02 ±141.57	1321.28 ±114.09	**7187.85 ±698.45**
80	10	1080.89 ±67.82	1122.70 ±71.50	**5105.98 ±477.29**
80	15	987.67 ±47.51	1019.63 ±44.54	**4215.69 ±401.73**
80	20	920.57 ±34.70	937.91 ±35.87	**3658.60 ±337.47**
90	5	1390.62 ±89.78	1447.15 ±156.85	**10452.20 ±1454.85**
90	10	1157.01 ±74.67	1225.97 ±71.09	**7354.25 ±1114.50**
90	15	1074.09 ±41.79	1129.76 ±42.63	**5961.38 ±1010.59**
90	20	980.27 ±33.59	1027.91 ±41.25	**5129.46 ±876.67**
100	5	1374.41 ±77.93	1363.24 ±117.40	**9448.92 ±921.52**
100	10	1150.42 ±60.97	1164.57 ±57.99	**6758.93 ±624.14**
100	15	1049.70 ±45.59	1045.50 ±49.68	**5492.19 ±519.69**
100	20	973.11 ±47.92	961.89 ±42.74	**4756.98 ±467.76**

TABLE V
"20 NEWGROUPS" DOCUMENTS CLASSIFICATION, 5 TOPICS

Metric	symm.	prev.	prop.
precision	0.4500 ± 0.0017	**0.5893 ± 0.0655**	0.5492 ± 0.0007
recall	0.5183 ± 0.0017	0.5502 ± 0.0078	**0.6536 ± 0.0007**
f1-score	0.4701 ± 0.0016	0.5082 ± 0.0124	**0.5922 ± 0.0007**

TABLE VI
"20 NEWGROUPS" DOCUMENTS CLASSIFICATION, 10 TOPICS

Metric	symm.	prev.	prop.
precision	0.5418 ± 0.0076	0.4777 ± 0.0288	**0.6678 ± 0.0049**
recall	0.5248 ± 0.0079	0.4990 ± 0.0029	**0.6364 ± 0.0032**
f1-score	0.5232 ± 0.0091	0.4539 ± 0.0066	**0.6217 ± 0.0060**

when using five topics in LDA. LDA with previous prior optimization performs better regarding precision, while our proposed prior setting performs better regarding recall and f1-score. Table VI is the classification result when using 10 topics in LDA. LDA with our proposed prior setting performs better in precision, recall, and f1-score. We can conclude that LDA with our proposed prior setting performs better than LDA with current prior settings. Experiments in documents classification validate that LDA with our proposed prior represents documents better in another aspect.

E. Modeling Ability

We compare the modeling ability of LDA with different prior settings here. Modeling ability of a model is its ability to predict previously unseen documents in the training data. Better modeling ability will achieve higher probability. We use the log-perplexity per word as the metric, which is the most popular measure in language modeling [28].

In this experiment, we split the dataset into the training set and the testing set by 90% and 10% respectively. As in previous experiments, we train LDA with three prior settings respectively. We then compute the log-perplexity per word on the testing dataset. The topic numbers are also 5, 10, and 15 for "20 Newsgroups", while the topic number for "Health Twitter" are 80, 90, and 100. We also repeat each process ten times and get the $mean$ and $standard\_deviation$ in the records.

The results of "Health Twitter" and "20 Newsgroups" are in Table VII and Table VIII respectively. We observe no significant different between LDA with different prior settings. For the same settings and dataset, we can observe significant differences in previous experiments evaluating topic quality and documents representation quality. The results are consistent with previous research [3] that better modeling ability does not necessarily lead to better models in real applications.

V. CONCLUSION

LDA is a useful tool in document organizing. Its effectiveness depends on the prior setting, especially document prior. Previous research proposed asymmetric prior setting that improves modeling ability. However, modeling ability does not necessarily consent with the performance in real tasks. In this paper, we proposed a novel prior setting method for LDA. We conduct experiments to compare the performance of LDA in different prior settings. Experiments validate that our proposed parameters setting improves the performance of LDA significantly in both topic description and document representation. Moreover, the experiments also show that better performance is not consistent with better modeling ability.

In the future, we will compare different features and different clustering algorithms in learning the priors to improve the prior setting in multiple aspects.

TABLE VII
LOG-PERPLEXITY PER WORD, "HEALTH TWITTER"

Topic Number	symm.	prev.	prop.
5	-20.3447 ± 0.2054	-20.2787 ± 0.1307	-20.1858 ± 0.0931
10	-30.7281 ± 0.2543	-30.6087 ± 0.1205	-30.3847 ± 0.0451
15	-32.9208 ± 0.0619	-32.7906 ± 0.1146	-32.5628 ± 0.0790

TABLE VIII
LOG-PERPLEXITY PER WORD, "20 NEWSGROUPS"

Topic Number	symm.	prev.	prop.
5	-11.1979 ± 0.4496	-11.2059 ± 0.4830	-11.2026 ± 0.4533
10	-12.4557 ± 0.7068	-12.4715 ± 0.7197	-12.4839 ± 0.6772
15	-14.5043 ± 1.0810	-14.4442 ± 1.0210	-14.4774 ± 1.0365

ACKNOWLEDGMENT

This research is supported by the National Science Foundation award IIS-1739095. The authors would like to thank the anonymous reviewers for their helpful and constructive comments.

REFERENCES

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," in *Advances in neural information processing systems*, 2009, pp. 1973–1981.

[3] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

[4] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.

[5] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[8] S. Z. Selim and M. A. Ismail, "Soft clustering of multidimensional data: a semi-fuzzy approach," *Pattern Recognition*, vol. 17, no. 5, pp. 559–568, 1984.

[9] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome biology*, vol. 3, no. 11, pp. research0059–1, 2002.

[10] R. Winkler, F. Klawonn, and R. Kruse, "Fuzzy c-means in high dimensional spaces," *International Journal of Fuzzy System Applications (IJFSA)*, vol. 1, no. 1, pp. 1–16, 2011.

[11] A. Mackiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers and Geosciences*, vol. 19, pp. 303–342, 1993.

[12] R. Baeza-Yates, B. d. A. N. Ribeiro *et al.*, *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,, 2011.

[13] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[14] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy approach topic discovery in health and medical corpora," *International Journal of Fuzzy Systems*, vol. 20, no. 4, pp. 1334–1345, 2018.

[15] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." Carnegie-mellon univ pittsburgh pa dept of computer science, Tech. Rep., 1996.

[16] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[17] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[20] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[21] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 952–961.

[22] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, 2015, pp. 399–408.

[23] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2004, pp. 22–30.

[28] C. D. Manning, C. D. Manning, and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.