Dirichlet Multinomial Mixture with Variational Manifold Regularization: Topic Modeling over Short Texts

Ximing Li, Jiaojiao Zhang, Jihong Ouyang
College of Computer Science and Technology, Jilin University, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
liximing86@gmail.com

Abstract

Conventional topic models suffer from a severe sparsity problem when facing extremely short texts such as social media posts. The family of Dirichlet multinomial mixture (DMM) can handle the sparsity problem, however, they are still very sensitive to ordinary and noisy words, resulting in inaccurate topic representations at the document level. In this paper, we alleviate this problem by preserving local neighborhood structure of short texts, enabling to spread topical signals among neighboring documents, so as to correct the inaccurate topic representations. This is achieved by using variational manifold regularization, constraining the close short texts should have similar variational topic representations. Upon this idea, we propose a novel Laplacian DMM (LapDMM) topic model. During the document graph construction, we further use the word mover's distance with word embeddings to measure document similarities at the semantic level. To evaluate LapDMM, we compare it against the state-of-the-art short text topic models on several traditional tasks. Experimental results demonstrate that our LapDMM achieves very significant performance gains over baseline models, e.g., achieving even about 0.2 higher scores on clustering and classification tasks in many cases.

Introduction

Short texts, such as text advertisements and social media posts, are becoming more and more prevalent on the Internet. With the emerging large-scale collections of short texts, discovering the hidden topic structure from them is important for many content analysis applications. However, short texts, as suggested by the name, often contain very few words. For example, in *StackOverFlow* of question titles, each title sample has only about 4.9 word tokens averagely after a removal of the meaningless stopwords. Therefore, there must be very limited valuable information for short text collections at the document level, resulting in the so-called **sparsity problem**.

Conventional topic models, such as probabilistic latent semantic indexing (PLSI) (Hofmann 1999) and latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), also suffer from the sparsity problem. That is because with statistical methods, the topic inference is mainly depended on the document-level word co-occurrence information (Wang and McCallum 2006) that the short texts lack. This raises up a significant challenge to topic modeling of short texts.

Recently, many efforts have been made to handle the sparsity problem. Borrowing the taxonomic hierarchy from multi-label learning (Zhang and Zhou 2014), we organize the existing works on topic modeling of short texts into two categories, i.e., *problem transformation method* and *adaptation method*. The problem transformation method (Mehrotra et al. 2013; Quan et al. 2015; Zuo et al. 2016; Li et al. 2018b) tackles the task by aggregating short texts into long pseudo-documents and then applying a well-established topic model, e.g., LDA. For this category of methods, the short texts can be aggregated using side information, e.g., user ID (Mehrotra et al. 2013), or adaptive scenarios (Quan et al. 2015; Zuo et al. 2016; Li et al. 2018b). However, a drawback of them is that any long pseudo-document may consist of many irrelevant short texts, making the topic inference less effective. The adaptation method (Nigam et al. 2000; Cheng et al. 2014; Sridhar 2015; Zuo, Zhao, and Xu 2016; Xin et al. 2011; Yin and Wang 2014; Li et al. 2016; 2017; 2018c; 2018d) directly modifies traditional topic models by enriching word co-occurrences, so as to remedy the sparsity problem. A straightforward methodology is to model the global word co-occurrences at the corpus level (Cheng et al. 2014; Sridhar 2015; Zuo, Zhao, and Xu 2016). For example, the biterm topic model (BTM) (Cheng et al. 2014) learns topics by modeling word co-occurrence pairs over the entire corpus; the word network topic model (WNTM) (Zuo, Zhao, and Xu 2016) refers to each word type as a pseudo-document following a global word co-occurrence network. These models can alleviate the sparsity problem to some extent. However, they may create many meaningless word co-occurrences without any word pair filtering process, and more importantly they lose document-specific topic structure. Additionally, another methodology indirectly enriches document-level word co-occurrences by supposing that each short text covers a small subset of topics. The representative methods include Dirichlet multinomial mixture (DMM) (Nigam et al. 2000; Yin and Wang 2014) and its variants (Li et al. 2016;

2017), which are widely used on analysis tasks of short texts (Xin et al. 2011).

Orthogonal to BTM and WNTM, the family of DMM can maintain document-specific topic structure, and the recent variants (Li et al. 2016; 2017) have empirically shown very competitive performance. However, a salient problem is that they are sensitive to ordinary and noisy words, therefore the document-level topic representation can be easily miscalculated. We refer to this as the **sensitivity problem**. For example, if the majority of words in a short text are without any topic-inclination or even noises, the topic of this document can be probably miscalculated. Unfortunately, this often happens as short texts contain very few words.

Our Model

To break the limitation of DMM, we develop a novel **Lap**lacian **DMM** (**LapDMM**) topic model for short texts. Our basic idea is to extend DMM by preserving local neighborhood structure of short texts using manifold regularization, which has been successfully used for topic models (Cai et al. 2008; Cai, Wang, and He 2009; Huh and Fienberg 2010; Du et al. 2015; Hu et al. 2017). The manifold regularization implies that the learned manifolds should be smooth, which here constrains nearby document pairs have similar latent topic representations. This can indirectly spread topical signals among neighboring documents, enabling to correct the miscalculated topic representations, so as to remedy the sensitivity problem of DMM.

We would like to notice that the manifold regularization can not be directly applied to DMM, because it supposes that each document only covers a single topic. To solve this, we train LapDMM following the spirit of collapsed variational inference (Teh, Newman, and Welling 2006), a more accurate inference method for topic models (Chi et al. 2018) than Gibbs sampling used in other DMM variants (Li et al. 2016; 2017). We then incorporate a manifold regularization term with respect to variational distributions into the original variational objective of DMM, such that the close short texts should have similar variational topic representations. LapDMM is optimized by maximizing the regularized variational objective. To better capture similarities between short texts, we employ the word mover's distance (WMD) (Kusner et al. 2015) with word embeddings (Mikolov, tau Yih, and Zweig 2013), which describes document similarities at the semantic level. We employ a regularized version of WMD with an entropic regularizer (Cuturi 2013) for efficient computations. Empirical results indicate that our LapDMM significantly outperforms the state-of-the-art baselines.

The main contributions of this paper are summarized as follows:

1 We develop a novel LapDMM model by incorporating a variational manifold regularization term.

2 We use the WMD with word embeddings to measure (semantic) similarities between short text pairs.

3 Empirical results on popular benchmark datasets demonstrate that our LapDMM significantly outperforms the state-of-the-art baselines on topic quality, clustering and classification tasks. Specifically, the performance gain achieves even above 0.2 in many cases.

Model

In this section, we give a brief introduction to Dirichlet multinomial mixture (DMM) (Nigam et al. 2000; Yin and Wang 2014), and then describe the proposed LapDMM topic model for short texts.

Dirichlet Multinomial Mixture

DMM is a generative topic model with the assumption that each document covers only a single topic. Actually, this assumption can indirectly enrich word co-occurrences at the document level, making the model more effective for short texts.

Formally, DMM consists of (1) *K* topic distributions ϕ over the vocabulary of *V* words, drawn from a Dirichlet prior β and (2) a corpus-level distribution θ over topics, drawn from a Dirichlet prior α. For each document *d*, it first draws a topic indicator z_d from θ, and then draws each word token w_{dn} from the selected topic ϕ_{z_d}. Given a corpus of *D* short texts, the generative process of DMM can be described as follows:

1. Draw a distribution over topics: $\theta \sim \mathbf{Dir}\left(\alpha\right)$
2. For each topic *k*
   a. Draw a distribution over words $\phi_k \sim \mathbf{Dir}\left(\beta\right)$
3. For each document *d*
   a. Draw a topic : $z_d \sim \mathbf{Multinomial}\left(\theta\right)$
   b. For each of the N_d words w_{dn}
      i. Draw a word: $w_{dn} \sim \mathbf{Multinomial}\left(\phi_{z_d}\right)$

LapDMM with Variational Manifold Regularization

We propose an extension of DMM with variational manifold regularization, namely LapDMM, to preserve local neighborhood structure of short texts.

Manifold regularization In the context of topic modeling, the manifold regularization constrains that the latent topic representations of document pairs should be similar to each other if they are nearest neighbors in document manifolds.

Formally, consider a directed graph with *D* vertices, where each vertex corresponds to a document in the corpus. Each component of the edge weight matrix *W* is defined by:

$$W_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } \ d_i \in \Omega(d_j) \text{ or } d_j \in \Omega(d_i) \\ 0 & \text{otherwise} \end{array} \right. , \qquad (1)$$

where $\Omega(d)$ is a document set containing *R* nearest neighbors of document *d*. Specifically, let θ_d denote a latent *K*-dimensional topic representation of document *d*. We define a least square manifold regularization term as follows:

$$\mathcal{R} = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j=1}^{D} \left(\theta_{ik} - \theta_{jk}\right)^2 W_{ij} \qquad (2)$$

Variational objective of LapDMM Note that we can not directly incorporate the manifold regularization term of Eq.2 into DMM inference. Because in DMM each document is supposed to be drawn from a single topic, there are no explicit *K*-dimensional topic representations θ for documents.

To address this issue, we resort to the collapsed variational inference optimization (Teh, Newman, and Welling 2006), and propose a variational manifold regularization term instead.

Thanks to the conjugate Dirichlet-multinomial design in DMM, the two distributions θ and ϕ can be marginalized out. We then define a mean-field variational distribution with respect to the topic assignment *z* of documents,

$$q(z) = \prod_{d=1}^{D} q\left(z_d | \gamma_d\right), \tag{3}$$

where each $q(z_d|\gamma_d)$ is a multinomial distribution with a *K*-dimensional variational parameter vector γ_d, i.e., $\sum_{k=1}^{K} \gamma_{dk} = 1$. Given a short text collection *S*, we train DMM by maximizing the following variational objective with respect to γ:

$$\mathcal{L}(\gamma) = \mathrm{E}_{q(z)}\left[\log p\left(S, z | \alpha, \beta\right) - \log q\left(z\right)\right] \tag{4}$$

Since each document-specific variational distribution $q(z_d|\gamma_d)$ is used as an approximation to the latent topic representation of the current document, we can define a variational manifold regularization term on $q(z)$ to achieve manifold constraints. That is, we re-write the manifold regularization (i.e., Eq.2) by replacing θ with the *K*-dimensional variational parameter γ as follows:

$$\mathcal{R}(\gamma) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j=1}^{D} \left(\gamma_{ik} - \gamma_{jk}\right)^2 W_{ij} \tag{5}$$

By combining Eq.4 and Eq.5, we reach the final regularized variational objective of LapDMM:

$$\widehat{\mathcal{L}}(\gamma) = \frac{1}{D} \mathcal{L}(\gamma) - \lambda \mathcal{R}(\gamma), \tag{6}$$

where $\lambda \in [0, 1]$ is a regularization parameter.

Optimization

We use a double-loop optimization procedure to maximize the objective of LapDMM $\widehat{\mathcal{L}}(\gamma)$ of Eq.6. In the outer iteration, we optimize γ by maximizing the first term of Eq.6, i.e., the original variational objective of DMM; in the inner iteration, we use the Newton-Raphson method to update γ by minimizing the second term of Eq.6, i.e., the variational manifold regularization, until the value of $\widehat{\mathcal{L}}(\gamma)$ decreases. Due to the space limitation, we omit derivation details, and directly show the update equations.

Outer iteration Actually, the outer update is a standard step of collapsed variational inference for DMM. Following (Bishop 2006), the optimum of γ is given by:

$$\begin{aligned} \gamma_{dk} &\propto \exp\left(\mathrm{E}_{q(z^{\neg d})}\left[\log p(S, z^{\neg d}, z_d = k | \alpha, \beta)\right]\right) \\ &\propto \exp\Bigg(\mathrm{E}_{q(z^{\neg d})}\Bigg[\sum_{v \in d} \sum_{n=1}^{N_{dv}} \log\left(\beta + N_{kv}^{\neg d} + n - 1\right) \\ &\quad + \log\left(\alpha + \widehat{N}_k^{\neg d}\right) - \sum_{n=1}^{N_d} \log\left(V\beta + N_k^{\neg d} + n - 1\right)\Bigg]\Bigg), \end{aligned} \tag{7}$$

where N_{dv} is the number of times word *v* occurring in document *d*; $\widehat{N}_k$ is the number of documents assigned to topic *k*; N_{kv} and N_k are the number of word *v* assigned to topic *k* and total number of words assigned to topic *k*, respectively; the superscript "$\neg d$" means the corresponding variables and counts with document *d* excluded.

We can efficiently compute an approximation of Eq.7 using the first-order Taylor expansion at the expectation values of number counts in Eq.7:

$$\begin{aligned} \gamma_{dk} &\propto \left(\alpha + \mathrm{E}_{q(z^{\neg d})}\left[\widehat{N}_k^{\neg d}\right]\right) \\ &\times \frac{\sum_{v \in d} \sum_{n=1}^{N_{dv}} \left(\beta + \mathrm{E}_{q(z^{\neg d})}\left[N_{kv}^{\neg d}\right] + n - 1\right)}{\sum_{n=1}^{N_d} \left(V\beta + \mathrm{E}_{q(z^{\neg d})}\left[N_k^{\neg d}\right] + n - 1\right)}, \end{aligned} \tag{8}$$

where for example the expectation of $\widehat{N}_k^{\neg d}$ is $\sum_{i \neq d}^{D} \gamma_{ik}$, and the other two expected number counts are similar. We refer readers to (Asuncion et al. 2009) for more details of this Taylor expansion approximation.

Inner iteration In the inner iteration, we focus on minimizing $\mathcal{R}(\gamma)$. We continue updating γ using Newton-Raphson iterations as long as the value of the overall objective $\widehat{\mathcal{L}}(\gamma)$ does not decrease (Cai et al. 2008). The update equation is as follows:

$$\begin{aligned} \gamma_{dk} &\leftarrow \gamma_{dk} - \rho \frac{\mathcal{R}'(\gamma_{dk})}{\mathcal{R}''(\gamma_{dk})} \\ &\leftarrow (1 - \rho)\gamma_{dk} + \rho \frac{\sum_{i=1}^{D} \gamma_{ik} W_{di}}{\sum_{i=1}^{D} W_{di}}, \end{aligned} \tag{9}$$

where $\rho \in [0, 1]$ is the learning rate. Note that this update equation guarantees $\sum_{k=1}^{K} \gamma_{dk} = 1$ for any document *d*.
Remark: The learning rate ρ can be roughly considered as a tuning parameter used to balance the two terms in Eq.9. When $\rho = 0$, LapDMM is downgraded to the standard DMM without manifold constraints.

Full algorithm Given the optimum of γ, the point estimates of ϕ and θ can be computed by:

$$\phi_{kv} = \frac{\mathrm{E}_{q(z)}[N_{kv}] + \beta}{\mathrm{E}_{q(z)}[N_k] + V\beta} \tag{10}$$

$$\theta_k = \frac{\mathrm{E}_{q(z)}[\widehat{N}_k] + \alpha}{D + K\alpha} \tag{11}$$

We outline the full optimization process of LapDMM in *Algorithm 1*.

Algorithm 1 Optimization for LapDMM

```
Set model and training parameters, including the number of nearest neighbors R, the regularization parameter $\lambda$ and Newton-Raphson learning rate $\rho$
Construct the document graph
Initialize $\gamma$ randomly and then expected number counts
For $t = 1, 2, \ldots,$ MaxIter
    Update $\gamma$ using Eq.8
    $\widehat{\gamma} \leftarrow \gamma$
    While $\left(\widehat{\mathcal{L}}(\gamma) \leq \widehat{\mathcal{L}}(\widehat{\gamma})\right)$ Do
        $\gamma \leftarrow \widehat{\gamma}$
        Update $\widehat{\gamma}$ using Eq.9
    End While
    Update expected number counts with the current $\gamma$
End for
Compute $\phi$ and $\theta$ using Eqs.10 and 11
```

Graph Construction

Before LapDMM training, we need to construct a document graph, i.e., finding *R* nearest neighbours for each document (ref. Eq.1). In this paper, we exploit two ways to measure distances between document pairs detailed as below:

Measuring document distances in the original term space Straightforwardly, we employ the popular cosine distance of documents' term frequency vectors.

Measuring document distances in a latent semantic space with word embeddings (Mikolov, tau Yih, and Zweig 2013) In the sparse short text context, semantically related documents may not contain any same word, so that they seem far away in the term space. To alleviate this, we employ the word mover's distance (WMD) (Kusner et al. 2015) to measure document distances at the semantic level. The formulation of WMD of a document pair (d_i, d_j) with an entropic regularization term (Cuturi 2013) is given by:

$$W_c(d_i, d_j) = \inf_{P \in \Pi(d_i, d_j)} \langle P, C \rangle - \frac{1}{\lambda'} H(P) \tag{12}$$

where d_i denotes the normalized term frequency vector of document *i* that can be considered as a multinomial distribution; $\Pi(d_i, d_j)$ is the set of the joint distributions of d_i and d_j; $H(\cdot)$ denotes the entropy; λ' is a regularization parameter[1]; and *C* is the distance matrix measured by the cosine distances of word embedding pairs (i.e., semantic distances between words measured by the corresponding word embeddings). In summary, the WMD actually measures the optimal (i.e., cheapest) transport from one document to any other in a semantic space with word embeddings. We can use the method proposed in (Cuturi 2013) to efficiently optimize Eq.12 and then obtain the WMD values.

For clarity, we refer to LapDMM using the two document distances as LapDMM$_\mathrm{T}$ and LapDMM$_\mathrm{W}$, respectively.

Discussion Actually, the document graph construction is independent of LapDMM, so it can be done off-line. However, it computes the distances of all document pairs, requiring $O(D^2)$ time. This is computationally expensive, especially for big datasets and streaming data. Such limitation not only arises in LapDMM, but also other models with manifold regularization. We will attempt to alleviate this problem in the future work.

Related Work

We review recent related works on topic models for short texts and topic modeling with manifold regularization.

Topic Models for Short Texts

Conventional topic models, such as PLSI and LDA, suffer from the sparsity problem when facing short texts, because they are lack of word co-occurrences at the document level. The models proposed in (Cheng et al. 2014; Zuo, Zhao, and Xu 2016; Lu et al. 2017) address the sparsity problem by directly using word co-occurrences at the corpus level. For example, BTM considers a corpus as a single big document, and models all the biterms, i.e., word co-occurrence patterns, extracted from documents. DMM assumes that each document is drawn from a single topic. Given the sparse content of short texts, this assumption is more reasonable, making DMM more effective than traditional topic models (Xin et al. 2011). GPU-DMM and GPU-PDMM (Li et al. 2016; 2017), two extensions of DMM, incorporate a generalized Pólya urn process into the topic inference process, so that similar words measured by word embeddings should be clustered in topics. In contrast to GPU-DMM and GPU-PDMM, our LapDMM not only captures the semantic information of word embeddings, but further considers document similarities with manifold constraints.

Besides, many other models (Mehrotra et al. 2013) address the sparsity problem of short texts by aggregating them into long pseudo-documents before applying LDA. Some recent extensions (Quan et al. 2015; Zuo et al. 2016; Li et al. 2018b) can adaptively aggregate short texts without using side information, e.g., user ID. Roughly, this kind of adaptive aggregation-based models is equivalent to an EM-like iteration process, i.e., clustering short texts (E-step) and LDA optimization (M-step). In some sense, our LapDMM is aggregating short texts by linking neighboring documents. In contrast, LapDMM is safer since it learns topics with the help of the neighboring document graph, instead of short text clusters that may be inaccurate.

Topic Modeling with Manifold Regularization

The manifold regularization methodology has been successfully used for topic models (Cai et al. 2008; Cai, Wang, and He 2009; Huh and Fienberg 2010; Du et al. 2015; Hu et al. 2017; Li et al. 2018a). For example, the authors of (Cai et al. 2008) incorporate manifold structure information, i.e., a manifold regularization term with the Euclidean distance, into the log-likelihood objective of PLSI (Hofmann 1999). The locally-consistent topic model (Cai, Wang, and He 2009) uses a manifold term with Kullback-Leibler divergence, instead of the Euclidean distance. The discriminative topic model (Huh and Fienberg 2010) develops a manifold

[1] Following previous studies, in this work we set λ' to 10.

Table 1: Summary of the datasets. *D*: the number of documents. *V*: the number of unique words. *AvgD*: the average document length. *L*: the number of categories.

Dataset	*D*	*V*	*AvgD*	*L*
Trec	5952	8392	4.94	6
Snipptes	12340	30445	17,5	8
StackOverFlow	20000	17996	4.93	20

term that not only pulls neighboring document pairs closer together, but also separates non-neighboring document pairs from each other. However, those previous models mainly focus on modeling normal long texts, therefore they also suffer from the sparsity problem of extremely short texts.

Experiment

We now present the empirical results of LapDMM on topic quality, clustering and classification tasks.

Experimental Setup

Dataset We employ three datasets, including *Trec*[2], *Snippets*[3] and *StackOverFlow*[4]. The *Trec* is a question dataset, consisting of 6 question types. The *Snippets* dataset was selected from the results of web search transaction of 8 different domains. The *StackOverFlow* dataset is a collection of question titles from 20 different tags. For each one, we removed the standard stopwords. The statistics of these datasets are summarized in Table 1.

Baseline model We compare LapDMM against four existing baseline topic models of short texts. For all models, the Dirichlet priors α and β are set to 0.1 and 0.01, respectively. For LapDMM, the parameters are empirically set as: $R = 9$, $\lambda = 0.1$ and $\rho = 0.1$. The specific settings of baseline models are presented as follows:

- **DMM**. We use collapsed variational inference for model inference.
- **GPU-DMM** (Li et al. 2016; 2017): an extension of DMM with word embeddings. We use the code provided by its authors[5]. To compute word similarities, we employ pre-trained 100-dimensional *GloVe*[6] word embeddings, trained on *Wikipedia* + *Gigaword*. For LapDMM, we use the same word embeddings to compute WMD values. Additionally, we haven't shown the empirical results of GPU-PDMM, since it have performed almost at the same level with GPU-DMM in our early experiments.
- **Latent topic model (LTM)** (Li et al. 2018b): a LDA-based topic model by adaptively aggregating short texts. We tune its parameters following the suggestions in the original paper.
- **BTM** (Cheng et al. 2014): a topic model of biterms for short texts. We use the code provided by its authors[7].

For all models, we tune their parameters, and report the best scores in all evaluation tasks.

Evaluation by Topic Quality

This section shows the topic quality evaluation results. We quantitatively evaluate the topic quality using the topic coherence (TC) project[8] developed by (Roder, Both, and Hinneburg 2015). This project automatically computes TC scores by counting co-occurrences of topical top-*M* words on a big reference corpus. The intuition is that for a topic, more co-occurrences between its top words, more semantically coherent it is.

Table 2 shows the average TC scores of top-10 words of all models. We have several observations. First, LapDMM$_{\mathrm{W}}$ performs the best among all models, where it ranks the first in most (i.e., 4/6) settings. LapDMM$_{\mathrm{W}}$ beats LapDMM$_{\mathrm{T}}$ in all settings. We argue that is because the WMD can better capture similarities between short texts at the semantic level. Second, LapDMM$_{\mathrm{T}}$ is also capable of outputting coherent topics. TC scores of LapDMM$_{\mathrm{T}}$ are roughly competitive with those of BTM, GPU-DMM and LTM, and it significantly outperforms BTM on *Trec*. Finally, we observe that both versions of LapDMM outperform the standard DMM. This indicates that the manifold regularizer can effectively improve the quality of learned topics.

Evaluation by Clustering

We compare LapDMM against baseline models by clustering. Here, each topic corresponds to a cluster, and the number of topics is set to the true category number of datasets. For LapDMM, we assign a document *d* to the largest topic in its variational approximation $q(z_d|\gamma_d)$.

We employ two clustering metrics, i.e., accuracy (ACC) and normalized mutual information (NMI). For a document *d*, its true label and estimated cluster are respectively denoted to be y_d and c_d. Then the ACC score can be computed by:

$$ACC = \frac{\sum_{d=1}^{D} \mathcal{I}\left(y_d, map\left(c_d\right)\right)}{D}, \tag{13}$$

where $\mathcal{I}\left(\cdot\right)$ denotes the indicator function; and $map\left(c_d\right)$ is the mapping function between c_d and y_d, computed by the Hungarian algorithm. Besides, let Y and C be the true label set and the estimated cluster label set of a given dataset, respectively. The NMI score can be computed by:

$$NMI\left(Y, C\right) = \frac{MI\left(Y, C\right)}{\sqrt{H\left(Y\right) H\left(C\right)}}, \tag{14}$$

where $MI\left(Y, C\right)$ denotes the mutual information of Y and C; and $H\left(\cdot\right)$ denotes the entropy.

Table 3 shows the average scores of 10 independent runs. Surprisingly, we see that LapDMM significantly outperforms the baseline models. For example, the performance

[2] http://cogcomp.cs.illinois.edu/Data/QA/QC/

[3] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz

[4] https://github.com/jacoxu/STC2

[5] https://github.com/NobodyWHU/GPUDMM

[6] https://nlp.stanford.edu/projects/glove/

[7] http://code.google.com/p/btm/

[8] https://github.com/AKSW/Palmetto/wiki/Coherences. The setting of "C_V" is used.

Table 2: Results of topic coherence (mean±std). "‡" means that the gain of LapDMM is statistically significant at 0.01 level.

Dataset	Topic	**LapDMM$_\text{T}$**	**LapDMM$_\text{W}$**	**DMM**	**BTM**	**GPU-DMM**	**LTM**
Trec	*K*=25	0.46±0.08	**0.48±0.06**	0.44±0.07‡	0.35±0.04‡	0.47±0.07	**0.48±0.07**
	K=50	0.47±0.08	**0.49±0.08**	0.46±0.07‡	0.36±0.06‡	0.47±0.07	0.48±0.07
Snippets	*K*=25	0.45±0.07	0.46±0.08	0.43±0.07‡	0.43±0.06‡	0.43±0.06‡	**0.47±0.07**
	K=50	0.45±0.09	**0.47±0.08**	0.43±0.09‡	0.45±0.06‡	0.43±0.08‡	0.45±0.08
StackOverFlow	*K*=25	0.37±0.07	0.39±0.09	0.37±0.08‡	**0.40±0.07**	0.38±0.07	0.36±0.08‡
	K=50	0.36±0.09	**0.37±0.08**	0.34±0.08‡	0.36±0.07	0.35±0.06‡	0.35±0.07‡

Table 3: Clustering results of NMI and ACC (mean±std). "‡" means that the gain of LapDMM is statistically significant at 0.01 level.

Dataset	Topic	**LapDMM$_\text{T}$**	**LapDMM$_\text{W}$**	**DMM**	**BTM**	**GPU-DMM**	**LTM**
Trec	NMI	**0.292±0.02**	0.288±0.04	0.125±0.06‡	0.109±0.05‡	0.127±0.04‡	0.114±0.06‡
	ACC	**0.499±0.05**	0.484±0.05	0.355±0.05‡	0.337±0.04‡	0.352±0.03‡	0.348±0.05‡
Snippets	NMI	0.634±0.04	**0.653±0.01**	0.526±0.05‡	0.521±0.02‡	0.544±0.02‡	0.539±0.02‡
	ACC	0.761±0.06	**0.793±0.03**	0.698±0.04‡	0.683±0.04‡	0.723±0.02‡	0.705±0.05‡
StackOverFlow	NMI	0.641±0.01	**0.645±0.02**	0.457±0.05‡	0.429±0.02‡	0.439±0.01‡	0.442±0.01‡
	ACC	**0.728±0.02**	0.710±0.05	0.494±0.03‡	0.472±0.01‡	0.482±0.03‡	0.498±0.02‡

gain of NMI is about 0.15∼0.17 on *Trec* and the gain of ACC is even about 0.23∼0.25 on *StackOverFlow*. The possible reason is that the manifold regularizer effectively remains the local manifold structures, i.e., pulling closer short texts together at the topic level. This is obviously benefit to unsupervised learning tasks such as clustering. Our result is consistent with the previous study of (Cai et al. 2008), where it has shown that the manifold regularization scheme significantly improved the clustering performance of PLSI.

Evaluation by Classification

We compare LapDMM against baseline models by classification. For all models, we train topical features (i.e., the SW representation described in (Li et al. 2016)) to represent short texts, and then feed them into SVMs.[9] The classification accuracy is computed by 5-fold cross validation. In the experiment, the topic number has been set to 40 and 60.

Table 4 shows the average classification accuracies of 10 independent runs. Fortunately, we again observe significant improvement of LapDMM just as observations in clustering experiments. For example, the accuracy scores of LapDMM are about 0.17 and 0.09 higher than those of baseline models on *Trec* and *StackOverFlow*, respectively. That is to say, our LapDMM can output more discriminative topical representations, leading to better classification performance. Besides, we observe that LapDMM$_\text{W}$ performs better than LapDMM$_\text{T}$, especially for *Trec* with 40 topics. This again implies that the WMD is a better distance measurement for short texts.

Parameter Evaluation

In this subsection, we empirically evaluate two crucial parameters of LapDMM, including the nearest neighbor number *R* and Newton learning rate ρ in the inner iteration. To this end, for each dataset we show the clustering and classification[10] scores of LapDMM$_\text{W}$.

We first evaluate the impact of different *R* values over the set $\{1, 2, \cdots, 10\}$. The results are shown in Figure 1. Roughly, the overall trend is that the performance becomes better as the value of *R* increases, e.g., the classification accuracy of *Trec*. The best scores are achieved at *R*=8 and 9 in most cases. The clustering ACC score seems a bit unsmooth, however *R*=9 also performs the best. The results tell us that using more nearest neighbors in manifolds is helpful. We thus fix *R*=9 in our experiments, and suggest to set a relatively larger value of *R* in practice.

Then, we evaluate the impact of ρ with different values over the set $\{0.1, 0.2, \cdots, 0.9\}$. The experimental results are shown in Figure 2. Overall, we argue that LapDMM is insensitive to ρ, and smaller values of ρ perform a little better. In some sense, the learning rate ρ describes the importance degree of the manifold regularizer during model training. A smaller value of ρ is safer when we cannot accurately find the nearest neighbors of short texts. Thus we use $\rho = 0.1$ as the default setting of LapDMM.

Conclusion

In this paper, we develop a novel LapDMM topic model for short texts. which incorporates a variational manifold regularization term into DMM. That is, we use collapsed variational inference to train DMM with a manifold regularization term with respect to variational distributions. To construct document graph for manifold constraints, we employ the WMD to measure semantic similarities between short texts. Extensive experiments show that LapDMM performs significantly better than the state-of-the-art baseline models.

[9] http://scikit-learn.org/

[10] For classification, we show results of $K = 60$.

Table 4: Results of classification accuracy (mean±std). "‡" means that the gain of LapDMM is statistically significant at 0.01 level.

Dataset	Topic	**LapDMM$_\mathrm{T}$**	**LapDMM$_\mathrm{W}$**	**DMM**	**BTM**	**GPU-DMM**	**LTM**
Trec	*K=40*	0.772±0.01	**0.844±0.01**	0.676±0.02‡	0.671±0.02‡	0.688±0.01‡	0.681±0.01‡
	K=60	**0.852±0.01**	**0.852±0.01**	0.696±0.02‡	0.676±0.01‡	0.703±0.01‡	0.712±0.01‡
Snippets	*K=40*	0.897±0.01	**0.901±0.01**	0.860±0.01‡	0.816±0.01‡	0.860±0.02‡	0.825±0.01‡
	K=60	0.901±0.01	**0.908±0.01**	0.867±0.01‡	0.828±0.02‡	0.868±0.01‡	0.833±0.01‡
StackOverFlow	*K=40*	0.801±0.01	**0.809±0.01**	0.714±0.01‡	0.692±0.01‡	0.702±0.01‡	0.711±0.01‡
	K=60	0.796±0.01	**0.805±0.01**	0.729±0.01‡	0.703±0.01‡	0.716±0.01‡	0.709±0.01‡

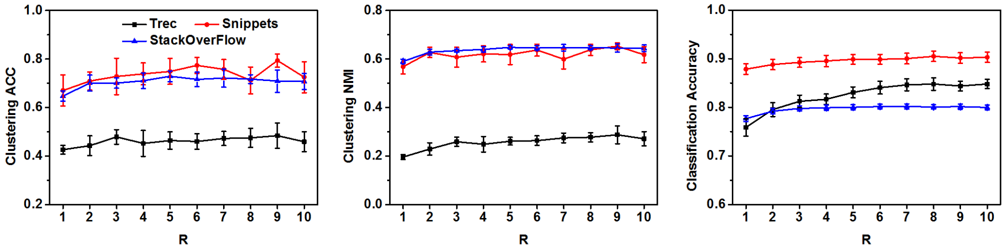

Figure 1: Evaluation results of different *R* values

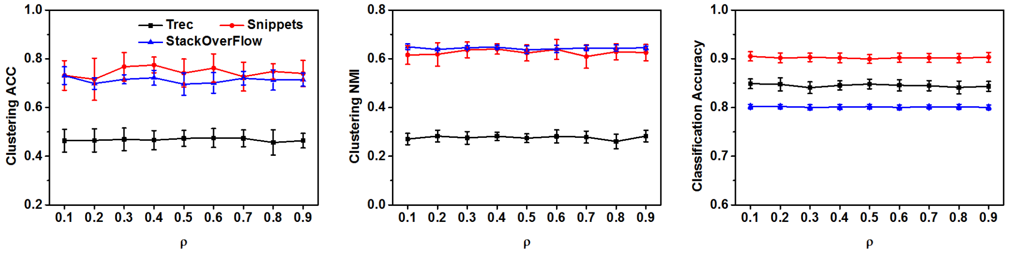

Figure 2: Evaluation results of different ρ values

An limitation of LapDMM is that the document graph construction may be time-consuming, especially for big datasets and streaming data. We plan to investigate this problem in the future work.

Acknowledgements

We would like to acknowledge support for this project from the National Natural Science Foundation of China (NSFC) (No.61602204, No.61876071).

References

Asuncion, A.; Welling, M.; Smyth, P.; and Teh, Y. W. 2009. On smoothing and inference for topic models. In *Conference on Uncertainty in Artificial Intelligence*, 27–34.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cai, D.; Mei, Q.; Han, J.; and Zhai, C. 2008. Modeling hidden topics on document manifold. In *ACM Conference on Information and Knowledge Management*, 911–920.

Cai, D.; Wang, X.; and He, X. 2009. Probabilistic dyadic data analysis with local and global consistency. In *International Conference on Machine Learning*, 105–112.

Cheng, X.; Yan, X.; Lan, Y.; and Guo, J. 2014. BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26(12):2928–2941.

Chi, J.; Ouyang, J.; Li, X.; and Li, C. 2018. Empirical study on variational inference methods for topic models. *Journal of Experimental & Theoretical Artificial Intelligence* 30(1):129–142.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances on Neural Information Processing Systems*, 2292–2300.

Du, J.; Jiang, J.; Song, D.; and Liao, L. 2015. Topic modeling with document relative similarities. In *International Joint Conference on Artificial Intelligence*, 3469–3475.

Hofmann, T. 1999. Probabilistic latent semantic indexing.

In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57.

Hu, W.; Zhu, J.; Su, H.; Zhuo, J.; and Zhang, B. 2017. Semi-supervised max-margin topic model with manifold posterior regularization. In *International Joint Conference on Artificial Intelligence*, 1865–1871.

Huh, S., and Fienberg, S. E. 2010. Discriminative topic modeling based on manifold learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 653–662.

Kusner, M. J.; Sun, Y.; Kolkin, N. I.; and Weinberger, K. Q. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, 957–966.

Li, C.; Wang, H.; Zhang, Z.; Sun, A.; and Ma, Z. 2016. Topic modeling for short texts with auxiliary word embeddings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165–174.

Li, C.; Duan, Y.; Wang, N.; Zhang, Z.; Sun, A.; and Ma, Z. 2017. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems* 36(2):Article No.11.

Li, X.; Li, C.; Chi, J.; Ouyang, J.; and Li, C. 2018a. Dataless text classification: A topic modeling approach with document manifold. In *ACM International Conference on Information and Knowledge Management*, 973–982.

Li, X.; Li, C.; Chi, J.; and Ouyang, J. 2018b. Short text topic modeling by exploring original documents. *Knowledge and Information Systems* 56(2):443–462.

Li, X.; Wang, Y.; Zhang, A.; Li, C.; Chi, J.; and Ouyang, J. 2018c. Filtering out the noise in short text topic modeling. *Information Sciences* 456:83–96.

Li, X.; Zhang, A.; Li, C.; Guo, L.; Wang, W.; and Ouyang, J. 2018d. Relational biterm topic model: Short text topic modeling using word embeddings. *The Computer Journal.* bxy037, https://doi.org/10.1093/comjnl/bxy037.

Lu, H.-Y.; Xie, L.-Y.; Kang, N.; Wang, C.-J.; and Xie, J.-Y. 2017. Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery. In *AAAI Conference on Artificial Intelligence*, 1192–1198.

Mehrotra, R.; Sanner, S.; Buntine, W.; and Xie, L. 2013. Improving LDA topic models for microblogs via Tweet pooling and automatic labeling. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 889–892.

Mikolov, T.; tau Yih, W.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2292–2300.

Nigam, K.; Mccallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2):103–134.

Quan, X.; Kit, C.; Ge, Y.; and Pan, S. J. 2015. Short and sparse text topic modeling via self-aggregation. In *International Joint Conference on Artificial Intelligene*, 2270–2276.

Roder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *International Conference on Web Search and Data Mining*, 399–408.

Sridhar, V. K. R. 2015. Unsupervised topic modeling for short texts using distributed representations of words. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 192–200.

Teh, Y. W.; Newman, D.; and Welling, M. 2006. A collapsed variational bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 1353–1360.

Wang, X., and McCallum, A. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 424–433.

Xin, W.; Jiang, Z.; Shu, J.; He, W.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, 338–349.

Yin, J., and Wang, J. 2014. A Dirichlet multinomial mixture model-based approach for short text clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–242.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zuo, Y.; Wu, J.; Zhang, H.; Lin, H.; Wang, F.; Xu, K.; and Xiong, H. 2016. Topic modeling of short texts: A pseudo-document view. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2105–2114.

Zuo, Y.; Zhao, J.; and Xu, K. 2016. Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems* 48(2):379–398.