Topic Modeling of Short Texts: A Pseudo-Document View

Yuan Zuo[1,2], Junjie Wu[3*], Hui Zhang[1,2], Hao Lin[3], Fei Wang[1,2], Ke Xu[1], Hui Xiong[4]

[1]School of Computer Science, Beihang University, China
[2]National Science and Technology Resources Sharing Services Engineering Research Center, Beijing, China
[3]School of Economics and Management, Beihang University, China
[4]Rutgers Business School, Rutgers University, USA
[*]*corresponding author: wujj@buaa.edu.cn*

ABSTRACT

Recent years have witnessed the unprecedented growth of online social media, which empower short texts as the prevalent format for information of Internet. Given the nature of sparsity, however, short text topic modeling remains a critical yet much-watched challenge in both academy and industry. Rich research efforts have been put on building different types of probabilistic topic models for short texts, among which the self aggregation methods without using auxiliary information become an emerging solution for providing informative cross-text word co-occurrences. However, models along this line are still rarely seen, and the representative one Self-Aggregation Topic Model (SATM) is prone to overfitting and computationally expensive. In light of this, in this paper, we propose a novel probabilistic model called Pseudo-document-based Topic Model (PTM) for short text topic modeling. PTM introduces the concept of *pseudo document* to implicitly aggregate short texts against data sparsity. By modeling the topic distributions of latent pseudo documents rather than short texts, PTM is expected to gain excellent performance in both accuracy and efficiency. A Sparsity-enhanced PTM (SPTM for short) is also proposed by applying Spike and Slab prior, with the purpose of eliminating undesired correlations between pseudo documents and latent topics. Extensive experiments on various real-world data sets with state-of-the-art baselines demonstrate the high quality of topics learned by PTM and its robustness with reduced training samples. It is also interesting to show that *i*) SPTM gains a clear edge over PTM when the number of pseudo documents is relatively small, and *ii*) the constraint that a short text belongs to only one pseudo document is critically important for the success of PTM. We finally take an in-depth semantic analysis to unveil directly the fabulous function of pseudo documents in finding cross-text word co-occurrences for topic modeling.

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939880

CCS Concepts

•Computing methodologies → Topic modeling; •Applied computing → Document analysis;

Keywords

Short Texts; Topic Modeling; Latent Dirichlet Allocation; Pseudo Document

1. INTRODUCTION

With the spur of various kinds of Web applications, especially the explosively growth of online social media such as Twitter and Facebook, short texts have become the prevalent format for information of Internet. For instance, around 250 million active users in Twitter can generate nearly 500 million tweets everyday. This huge volume of short texts contain sophisticated information that can hardly be found in traditional information sources [30]. Hence accurately discovering knowledge behind these short texts has been recognized as a challenging and promising research problem.

Probabilistic topic models have been widely used to automatically extract thematic information from large archive of documents [1]. Standard topic models [2, 4] assume a document is generated from a mixture of topics, where a topic is a probabilistic distribution of words. As one of the most typical probabilistic topic model, LDA [2] has achieved great success in modeling text collections like news articles, research papers and blogs. However, the results are mixed when LDA is applied directly to short texts such as tweets, instant messages and forum messages [5, 19]. The reason is mainly due to the lack of word co-occurrence information in each short text as compared to regular-sized documents [24].

Many research efforts have been devoted to tackle the incompetence of standard topic models in modeling short texts. One straightforward strategy, widely adopted for short texts in social media, is to utilize plentiful auxiliary contextual information to aggregate short texts into regular-sized pseudo documents before applying a standard topic model. For example, tweets contain not only textual content but also contextual information such as authorship, hashtags, time and locations. These contextual information can be leveraged to aggregate tweets before performing topic modeling [5, 25, 14]. Besides of the direct aggregation of short texts, some studies take auxiliary information into the generative process of their models [20, 8]. They actually regularize learned topics in accordance with some auxiliary information.

Generally speaking, the aggregation solution mentioned above can bring in additional useful word co-occurrence information across short texts, and therefore may boost the performance of standard topic models when applied to short texts. The problem lies in that auxiliary information is not always available or just too costly for deployment. In light of this, several customized topic models for short texts have been proposed. For instance, Yan et al. [27] propose a biterm topic model to directly model word pairs extracted from short texts. Lin et al. [11] propose the dual sparse topic model, which learns focused topics of a document and focused terms of a topic by replacing symmetric Dirichlet priors with Spike and Slab priors [7]. The problem of these methods lies in that they bring in little additional word co-occurrence information and therefore still face data sparsity problem. Recently, Quan et al. [18] propose a self-aggregated topic model named SATM, which can aggregate short texts into latent pseudo documents according to their own topics rather than auxiliary information. However, the number of parameters of SATM grows with the size of data, which makes it prone to overfitting. Moreover, the time complexity of SATM is also unacceptably high. Both weaknesses prevent it from being widely applied in practice.

Motivated by the promising potential of aggregation methods in dealing with data sparsity, we propose a Pseudo-document-based Topic Model (PTM) for short texts without using auxiliary information. To our best knowledge, our work is among the earliest studies in this interesting direction. The key of PTM is the introduction of pseudo documents for implicit aggregation of short texts against data sparsity. In this way, the modeling of topic distributions of tremendous short texts is transformed into the topic modeling of much less pseudo documents, which could be beneficial for parameter estimation in terms of both accuracy and efficiency. To further eliminate undesired correlations between pseudo documents and latent topics, we also propose a Sparsity-enhanced PTM (SPTM) by applying Spike and Slab prior to topic distributions of pseudo documents.

Extensive experiments on four real-world data sets with classic as well state-of-the-art baselines demonstrate the high quality of topics learned by PTM. Its robustness is also testified using reduced training samples in the comparative study with various baselines. Besides, two interesting observations are also noteworthy. First, SPTM outperforms PTM only when the number of pseudo documents is relatively small. Second, the constraint that a short text belongs to one and only one pseudo document is critically important for the success of PTM. Finally, an in-depth semantic analysis is conducted and unveils the merit of pseudo documents in finding cross-text word co-occurrences for topic modeling.

The remainder of this paper is organized as follows. In Section 2, we propose our models PTM and SPTM and give the inference details. In Section 3, we present experimental results. We give related work in Section 4 and finally conclude our work in Section 5.

2. MODEL AND INFERENCE

In this section, we propose a Pseudo-document-based Topic Model (PTM) for extremely short texts. PTM assumes huge volume of short texts are generated from much less yet regular-sized latent documents, called *pseudo documents*. By learning topic distributions of pseudo documents rather than short texts, PTM has fixed number of parameters and gains ability in avoiding overfitting when training corpus is in relative shortage. The sparsified version of PTM (SPTM) is also proposed to enhance the topical representation of pseudo documents if needed. We finally give the inference method and discuss the possible extension of PTM.

2.1 Basic Model

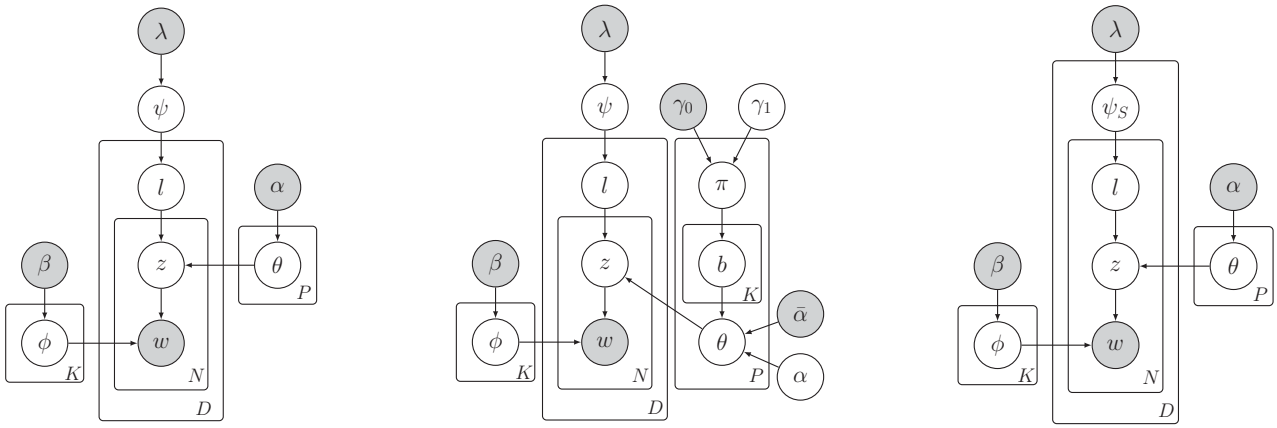
Now we give formal descriptions for PTM. We assume there are K topics $\{\phi_z\}_{z=1}^{K}$, each is a multinomial distribution over a vocabulary of size V. There are D short texts $\{d_s\}_{s=1}^{D}$ and P pseudo documents $\{d'_l\}_{l=1}^{P}$. The short texts are observed documents and the pseudo documents are latent ones. A multinomial distribution ψ is introduced to model the distribution of short texts over pseudo documents. We further assume each short text belongs to *one and only one* pseudo document. Each word in a short text is generated by first sampling a topic z from topic distribution θ of its pseudo document, and then sampling a word $w \sim \phi_z$.

The plate notation of PTM is illustrated in Fig. 1a. We give the generative process as follows:

1. Sample $\psi \sim Dir(\lambda)$
2. For each topic z:
   (a) Sample $\phi_z \sim Dir(\beta)$
3. For each pseudo document d'_l:
   (a) Sample $\theta_l \sim Dir(\alpha)$
4. For each short text d_s:
   (a) Sample a pseudo document $l \sim Multi(\psi)$:
   (b) For each word w_i in d_s:
      i. Sample a topic $z \sim Multi(\theta_l)$:
      ii. Sample the ith word $w_i \sim Multi(\phi_z)$

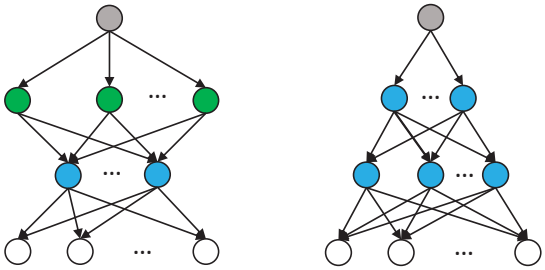
Remark 1. The introduction of pseudo documents in PTM is the critical factor against the negativity of data sparsity. To better understand this, assume there are D short texts and each one has averagely N tokens. It has been proven that when N is too small, LDA can not learn topics accurately even though D is extremely large [19]. This is because the shortage of co-occurrent words scattered in different short texts for topic learning is no better under such situation. However, PTM finds topics from P pseudo documents rather than D short texts, with $P \ll D$ in general. Therefore, we can roughly estimate that each pseudo document has N' tokens on average, $N' = DN/P \gg N$, which implies the potential improvement of word co-occurrence. In a nutshell, PTM runs on much denser pseudo documents, which by [19] could be very beneficial for topic modeling.

Remark 2. To our best knowledge, the self aggregation methods like PTM are still hardly seen in the literature, except for the Self-Aggregate Topic Model (SATM) [18]. While PTM and SATM both aggregate short texts into pseudo documents, their generative processes are substantially different. SATM assumes a two-phase generative process of short texts. The first phase follows the standard LDA to generate regular-sized pseudo documents, and in the second phase each short text will be generated from its pseudo document via the process of mixture of unigram [17]. The first phase implies that sampling a word will cost $O(PK)$ time,



(a) PTM (b) SPTM (c) EPTM

Figure 1: Plate notations.



(a) PTM (b) PAM

Figure 2: Model comparison.

which is very intensive. The second phase means the inference procedure has to estimate the probability distribution of pseudo documents on short texts independently, and the number of parameters thus grows linearly with the size of corpus, which might lead to serious overfitting problem when training samples are in shortage. As a sharp contrast, given the only pseudo document a short text belongs to, PTM generates the short text according to the process of LDA. This means to sample a word only costs $O(K)$ time, and the number of parameters is fixed to avoid overfitting.

Remark 3. It is also interesting to discuss the similarities and differences between PTM and the so-called Pachinko Allocation Model (PAM) [10]. PAM was proposed to capture arbitrary correlations between topics using a directed acyclic graph, and thus thought as a more general version of LDA [10]. Therefore, even though the four-level hierarchical PAM (see Fig. 2b) shows similar model structure as PTM (see Fig. 2a), they are different in nature. In Fig. 2b, the second level of PAM consists of super-topics capturing the commonness between sub-topics in the third level (all in blue). In this sense, we can obtain a reduced number of topics from the third level to the second level. In contrast, the nodes in the second level of PTM represent pseudo documents (in green), which are therefore more in number than the topic nodes in the third level (in blue), and should be better regarded as combined topics that can generate specific topics of short texts.

2.2 Sparsification

As mentioned above, a pseudo document in PTM is essentially a hybrid topic combined from the specific topics of various short texts. Along this line, it is a natural conjecture that when the number of pseudo documents gets smaller, their topical representations tend to be ambiguous. To deal with this, we here propose SPTM, a sparsified version of PTM applying the Spike and Slab prior for the topic distribution of pseudo documents.

The "Spike and Slab" prior [7] is a well established method in mathematics. It can decouple the sparsity and smoothness of a distribution. In details, auxiliary Bernoulli variables are introduced into the prior, which are used to indicate the "on" or "off" status of particular variables. Therefore, a model can determine whether corresponding variables appear or not. In our case, this indicates whether or not a topic is selected to appear in a particular pseudo document.

Note that the Spike and Slab prior may have empty selection, which will cause the probabilistic distribution to be ill-defined. Wang and Blei [23] introduce never-appearing-terms into the distributions of topics, which could impose greater difficulty into the inference procedure. We therefore apply the weak smoothing prior and smoothing prior proposed by Lin et al. [11], which can avoid the ill-defined distributions by the direct application of Spike and Slab prior. Moreover, it results in a much simpler inference procedure, which ensures the scalability of our model. In order to better describe our sparse-enhanced model, we first give definitions of *topic selectors*, *smoothing prior* and *weak smoothing prior*.

Definition 1. For pseudo document d'_l, a *topic selector* $b_{l,k}$, $k \in \{1, \cdots, K\}$, is a binary variable that indicates whether topic k is relevant to d'_l. $b_{l,k}$ is sampled from Bernoulli(π_l), where π_l is a Bernoulli parameter for d'_l.

Definition 2. The *smoothing prior* is Dirichlet hyperparameter α used to smooth topics selected by the topic selector. The *weak smoothing prior* is another Dirichlet hyperparameter $\bar{\alpha}$ used to smooth topics not selected. Since $\bar{\alpha} \ll \alpha$, the hyperparameter $\bar{\alpha}$ is called *weak smoothing prior*.

The topic selectors are referred as "Spikes", while the smoothing prior and weak smoothing prior correspond to "Slabs".

In this way, sparseness and smoothness of topic proportion of pseudo documents are decoupled. Given topic selectors $\vec{b_l} = \{b_{l,k}\}_{k=0}^{K}$, the topic proportion of pseudo document d'_l is sampled from $Dir(\alpha\vec{b_l} + \bar{\alpha}\vec{1})$. The introduction of $\bar{\alpha}$ fixes the ill-definition of distributions while maintaining the effect of sparsity.

Fig. 1b illustrates the plate notation of SPTM. The complete generative process of pseudo documents is given as follows:

1. For each pseudo document d'_l:
   (a) Sample $\pi_l \sim Beta(\gamma_0, \gamma_1)$
   (b) For each topic z:
      i. Sample topic selector $b_{l,k} \sim Bernoulli(\pi_l)$, $\vec{b_l} = \{b_{l,k}\}_{k=0}^{K}$.
   (c) Sample $\theta_l \sim Dir(\alpha\vec{b_l} + \bar{\alpha}\vec{1})$

2.3 Inference

Exact posterior inference is intractable in our model, so we turn to a collapsed Gibbs sampling algorithm [3] for approximate posterior inference, which is simple to derive, comparable in speed to other estimators, and can approximate a global maximum. Due to the space limit, we omit the derivation details and only present the sampling formulas.

We give details about the inference of SPTM in the following, and describe the inference of PTM at the end of this section. Integrating out θ, ϕ, ψ and π analytically, the latent variables needed by the sampling algorithm are pseudo document assignment l, topic assignment z and topic selector b. We also sample Dirichlet hyperparameter α and Beta hyperparameter γ_1, and fix $\bar{\alpha}$ equal to 10^{-7} and γ_0 equal to 1.

Sampling pseudo document assignments l. Given rest variables, sampling l is similar to sampling approach for Dirichlet Multinomial mixtures [29]. That is,

$$p(l_{d_s} = l|rest) \propto \frac{M_{l,\neg d_s}}{D-1+P\lambda} \frac{\frac{\prod_{z\in d_s}\Gamma\left(N_l^z + b_{l,z}\alpha + \bar{\alpha}\right)}{\prod_{z\in d_s}\Gamma\left(N_{l,\neg d_s}^z + b_{l,z}\alpha + \bar{\alpha}\right)}}{\prod_{i=1}^{N_{d_s}}\left(N_{l,\neg d_s} + |A_l|\alpha + K\bar{\alpha} + i - 1\right)}$$
$$= \frac{M_{l,\neg d_s}}{D-1+P\lambda} \frac{\prod_{z\in d_s}\prod_{j=1}^{N_{d_s}^z}\left(N_{l,\neg d_s}^z + b_{l,z}\alpha + \bar{\alpha} + j - 1\right)}{\prod_{i=1}^{N_{d_s}}\left(N_{l,\neg d_s} + |A_l|\alpha + K\bar{\alpha} + i - 1\right)}, \quad (1)$$

where M_l is the number of short texts assigned to the lth pseudo document d'_l. N_{d_s} is the length of the sth short text d_s, and $N_{d_s}^z$ is the number of tokens assigned to topic z in d_s. N_l^z is the number of tokens assigned to topic z in d'_l, and N_l is the total number of tokens in d'_l. All counts with $\neg d_s$ mean excluding counting from d_s. $b_{l,z}$ is topic selector of pseudo document d'_l for topic z. $A_l = \{z : b_{l,z} = 1, z \in \{1, \cdots, K\}\}$ is the set of indices of $\vec{b_l}$ that are "on", and $|A_l|$ is the size of A_l.

Sampling topic assignments z. The approach to sample topic assignments z is similar to latent Dirichlet allocation [3]. The difference lies in that θ no longer belongs to original short texts, but rather belongs to pseudo documents. And θ is sampled from *Spike and Slab* prior instead of symmetric Dirichlet prior. That is,

$$p(z_{d_s,i} = z|rest) \propto (N_{l_{d_s}}^z + b_{l_{d_s},z}\alpha + \bar{\alpha})\frac{N_z^{w_{d_s,i}} + \beta}{N_z + V\beta}, \quad (2)$$

where N_z^w is the number of times w being assigned to topic z, and $N_z = \sum_{w=0}^{V} N_z^w$.

Sampling topic selectors b. In order to sample $\vec{b_l}$, we follow Wang et al. [23] to use π_l as auxiliary variable. Let $B_l \triangleq \{z : N_l^z > 0, z \in \{1, \cdots, K\}\}$ be the set of topics that have assignments in pseudo document d'_l. We give the joint conditional distribution of π_l and $\vec{b_l}$:

$$p(\pi_l, \vec{b_l}|rest) \propto \prod_z p(b_{l,z}|\pi_l)p(\pi_l|\gamma_0,\gamma_1)\frac{I[B_l \in A_l]\Gamma(|A_l|\alpha + K\bar{\alpha})}{\Gamma(N_l + |A_l|\alpha + K\bar{\alpha})}, \quad (3)$$

where $I[\cdot]$ is an indicator function. With this joint conditional distribution, we iteratively sample $\vec{b_l}$ conditioned on π_l and π_l on $\vec{b_l}$ to ultimately obtain a sample for $\vec{b_l}$. Note that Wang et al. [23] integrate out b and sample π in case of slow convergence of topics. Since V is large, searching optimal combinatorial topics is very costly. In our case, however, K is relative small compared to V, and sampling z conditioned on π is time-consuming. Based on the above consideration, we take opposite approach by integrating out π to sample b.

For the hyper-parameter α, we use Metropolis-Hastings with a symmetric Gaussian as proposal distribution. For concentration parameter γ_1, we use previously developed approaches for Gamma priors [21].

So far, we have illustrated the collapsed Gibbs sampling algorithm for SPTM. Now we briefly describe the inference of PTM. After integrating out θ, ϕ and ψ analytically, latent variables needed by the sampling algorithm are pseudo document assignment l and topic assignment z. By replacing $b_{l,z}\alpha + \bar{\alpha}$ with α and $|A_l|\alpha + K\bar{\alpha}$ with $K\alpha$ in Equation 1, we obtain sampling equation for l. Similarly, by replacing $b_{l,z}\alpha + \bar{\alpha}$ with α in Equation 2, we obtain sampling equation for z.

Sampling l in PTM costs $O(P)$ per document, and sampling z costs $O(K)$ per word. Hence, the total time complexity of sampling algorithm of PTM is roughly $O(P + K)$ for a word. Due to the extra sampling of b, the time complexity of SPTM is slightly larger than PTM.

Note that since PTM and SPTM learn topic proportions θ_l for each pseudo document d'_l, we use empirical estimation to obtain θ_s for short text d_s with any single sample of $\vec{z}$. For PTM, $\theta_{s,z} = \frac{N_{l_{d_s}}^z + \alpha}{N_{l_{d_s}} + K\alpha}$. For SPTM, $\theta_{s,z} = \frac{N_{l_{d_s}}^z + b_{l_{d_s},z}\alpha + \bar{\alpha}}{N_{l_{d_s}} + |A_{l_{d_s}}|\alpha + K\bar{\alpha}}$.

2.4 Extension and Discussion

As described in Section 2, PTM assumes each short text comes from a single pseudo document. In this section, we will modify PTM to relaxing this restriction, and briefly discuss the influence caused by such modification.

In PTM, each short text is assigned to one pseudo document by choosing $l \sim \psi$, where $\psi \sim Dir(\lambda)$ is the distribution of short texts over pseudo documents. By assuming that each short text has a distribution over all pseudo documents ψ_S, we can relax the restriction and obtain Enhanced PTM(EPTM) as shown in Fig. 1c.

EPTM is a more flexible model than PTM, since words in a short text can be assigned to different pseudo documents. However, EPTM might performs less comparable with PTM and SPTM, although it is more general and flexible. The introduce of ψ_S, can bring errors to the inference of θ. Recall in Fig. 1c, if we assume $\vec{z}$ are observed variables, then sampling l is in the same with sampling z in LDA, therefore, θ can be regarded as ϕ in LDA. According to Tang et al. [19], with small N, LDA can not learn coherence and precise top-

Table 1: Statistics of data sets.

Data set	# Documents	Vocabulary size	Avg. document length
News	29,200	11,007	12.4
DBLP	55,290	7,525	6.4
Questions	142,690	26,470	4.6
Tweets	182,671	21,480	8.5

ics. When EPTM is applied to short texts with small N, the learning of θ is unreliable, which limits the total performance of EPTM.

Besides the unreliable performance on short texts, training of EPTM is also very time-consuming. The sampling time of (l, z) for each word costs $O(PK)$. In order to reduce the sampling complexity of EPTM, we apply Alias sampling [9] but give no details here for concision.

3. EXPERIMENTS

In this section, we present extensive experimental results on four real-world data sets to evaluate our model. Four baseline methods are also included for a thorough comparative study .

3.1 Experimental Setup

3.1.1 Data Sets

Our method is tested on four real-world short-text corpora, with the statistics listed in Table 1. In the following, we give brief descriptions to them.

News. This data set[1] contains 29,200 English news articles extracted from RSS feeds of three popular newspaper websites (nyt.com, usatoday.com, reuters.com). Categories include sport, business, U.S., health, sci&tech, world and entertainment. We retain news descriptions since they are typical short texts.

DBLP. We collect titles of conference papers from six research areas: data mining, computer vision, database, information retrieval, natural language processing and machine learning. This data set contains 55,290 short texts and each is labeled as one of above six research areas.

Questions. This collection consists of 142,690 questions crawled from a popular Chinese Q&A website[2] by Yan et al. [27]. Each question has a label chosen from 35 categories by its author.

Tweets. A large set of tweets are collected and labeled by Zubiaga et al. [32]. They crawl tweets that contain URL and label them with the categories of web pages pointed by the URLs. The categories of web pages are defined by the Open Directory Project (ODP). This dataset contains 10 different categories and totally around 360k labeled tweets. We select 9 topic-related categories and sample 182,671 tweets in total under those categories.

3.1.2 Baseline Methods

Latent Dirichlet Allocation (LDA). Being one of the most classical topic models, LDA [2] can induce sparsity as its Dirichlet prior approaches zero. We use jGibbLDA package[3] with collapsed Gibbs sampling for the comparison.

[1]http://acube.di.unipi.it/tmn-dataset/

[2]http://zhidao.baidu.com

[3]http://jgibblda.sourceforge.net

Mixture of Unigrams (MU). The most important feature of MU [17, 29] is that it assumes each document is generated by only one topic, which forces the topic representation of a document to adopt the largest sparsity. The assumption sounds unreasonable when documents are regular-sized, but it may be feasible for certain collections of short texts.

Dual Sparse Topic Model (DSTM). DSTM [11] is a recently published sparsity-enhanced topic model, which use Spike and Slab prior to learn *focused topics* of documents and *focused terms* of topics. Since each short text tends to contain only a few topics, and each topic tends to cover a subset of vocabulary, DSTM sounds reasonable for short texts modeling.

Self-aggregate Topic Model (SATM). Like our method, SATM [18] also aggregates short texts into pseudo documents without auxiliary information. However,its parameters grow in number with the size of a short text collection, which makes it prone to overfitting. Thus, it is interesting to compare SATM with our method by varying the amount of available data.

3.1.3 Evaluation Measures

Topic Coherence. Evaluation of topic models is still an open problem. The commonly used metric named *perplexity* has been proved less correlated to human interpretability, which means better perplexity does not indicate understandable topics. Furthermore, many customized topic models for short texts do not reveal topics from short texts directly, such as SATM and our model PTM, which makes perplexity no longer a general way for topic evaluation. Many methods thus turn to use *topic coherence* to evaluate topics, which is proved more correlated to human evaluations and has good generalization ability. However, it is reported that the UMass topic coherence [15] on short texts is also not a good indicator for quality of topics [18]. Computing UCI topic coherence [16] requires an appropriate external corpus, which is not easily available. Wikipedia can be used as a generalized external corpus. It is suitable when evaluating topic models on well-edited texts such as news and research articles, while less appropriate on user-generated content like tweets or questions.

Based on above considerations, we use UCI topic coherence to evaluate topic models on well-edited news and DBLP. UCI topic coherence uses the point-wise mutual information (PMI) to measure the coherence of topics. For a given topic z, we choose the top-N probable words $w_1, w_2, \cdots, w_N$, and calculate the average PMI score of each pair of these words:

$$PMI(z) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \qquad (4)$$

where $p(w_i, w_j)$ is the joint probability of word pair w_i and w_j co-occurring in the same sliding window, and $p(w_i)$ is the marginal probability of word w_i appearing in a sliding window. These probabilities are estimated from the latest dump of Wikipedia articles. The average topic coherence of all topics is used to evaluate a topic model. The default value of N is set to 10 in our experiments.

Classification measures. Topic models are often evaluated on external tasks such as text classification. Therefore, we also conduct short-text classification experiments to compare the latent semantic representations learned by

Table 2: Classification results of five-fold cross validation.

	News			DBLP			Questions			Tweets		
	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
PTM	*0.755*	*0.757*	*0.754*	**0.667**	**0.672**	**0.668**	**0.532**	**0.554**	**0.536**	0.561	*0.568*	*0.559*
SPTM	**0.760**	**0.761**	**0.759**	*0.661*	*0.667*	*0.663*	*0.530*	*0.552*	*0.532*	0.551	0.558	0.550
SATM	0.697	0.702	0.686	0.657	0.662	0.654	0.312	0.353	0.297	*0.599*	**0.605**	**0.594**
LDA	0.727	0.732	0.728	0.613	0.624	0.614	0.502	0.529	0.506	0.553	0.560	0.546
DSTM	0.720	0.724	0.720	0.619	0.628	0.620	0.489	0.515	0.491	0.539	0.547	0.535
MU	0.697	0.617	0.626	0.640	0.643	0.638	0.511	0.526	0.509	**0.634**	0.546	0.546

our methods and baselines. Macro averaged precision, recall and f-measure are used in classifications.

For all methods, we set the number of topics to 100. Unless otherwise specified, the number of pseudo documents in SATM and our methods is set to 1000. All methods, except SPTM and DSTM, run 2,000 iterations of sampling. For SPTM and DSTM, due to the sampling of additional binary variables, we perform 3,000 iterations of Gibbs sampling to guarantee its convergence.

We set $\alpha = 0.1$, $\beta = 0.01$ for LDA, since LDA with weak priors performs better in short texts. Similarly, we set $\alpha = 0.1$, $\beta = 0.01$ for MU. We set $\pi = 0.1$, $\gamma = 0.01$ for DSTM, and find it outperforms the setting $\pi = 1.0$, $\gamma = 1.0$ suggested by [11]. As to $\bar{\pi}$ and $\bar{\gamma}$, we retain its original setting. For SATM, we set the parameters the same as the ones in [18]. For PTM and EPTM, we set $\alpha = 0.1, \lambda = 0.1$ and $\beta = 0.01$. For SPTM, we set $\gamma_0 = 0.1$ and $\bar{\alpha} = 10^{-12}$. All results reported below are averaged on five runs.

3.2 Experimental Results

3.2.1 Topic Evaluation by Short Text Classification

We first compare all the topic models by performing document classifications. To this end, we take topic model as a method for dimension reduction, and characterize documents by a fixed set of topics as features for classification.

Results of Five-fold Cross Validation. For each trained topic model, we perform five-fold cross-validation on four data sets. LIBLINEAR[4] is adopted for classification. The resultant macro precision, recall and f-measure are shown in Table 2. We highlight the best results in bold and the second best in italic.

As can be seen in Table 2, both the best and the second best results are obtained by PTM and SPTM on News, DBLP and Questions. This demonstrates the outstanding performance of our methods against baselines in learning semantic representations of short texts. Specifically, the clear edge of our methods over LDA on all data sets suggests that aggregating short texts into regular-sized pseudo documents contributes greatly to topic learning from texts, which further boosts the performance of classification. Our methods also consistently outperform DSTM and MU on all data sets, which indicates modeling short texts by self-aggregation is more reliable than merely adding sparse constraints to topic models.

It is also interesting to note that although SATM performs poorly on the rest three collections, it obtains the best result on Tweets. This unstable performance can be well understood by watching the average training size per category of each data set, which are 4,171, 9,215, 4,077, 20,297, respectively, for News, DBLP, Questions and Tweets. This implies that the performance of SATM relates positively to the training data per category, which in turn testifies our theoretical analysis in Section 2 that SATM is indeed prone to overfitting. In order to guarantee the performance of SATM, one may suggest feeding enough data to it. This simple treatment, however, will also encounter great difficulty since SATM is very time-consuming (with $O(PK)$ sampling complexity for a word), which prevents it from being used in practice. For instance, in our experiments, the training of SATM on Tweets takes more than one week, while PTM and SPTM consume less than one day. Note that we have used sparse Gibbs sampling introduced by Yao et al. [28] to boost SATM in implementation. If ordinary Gibbs sampling is used instead, its training on Tweets will cost more than one month!

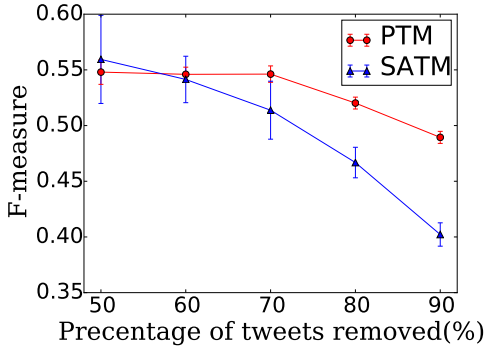
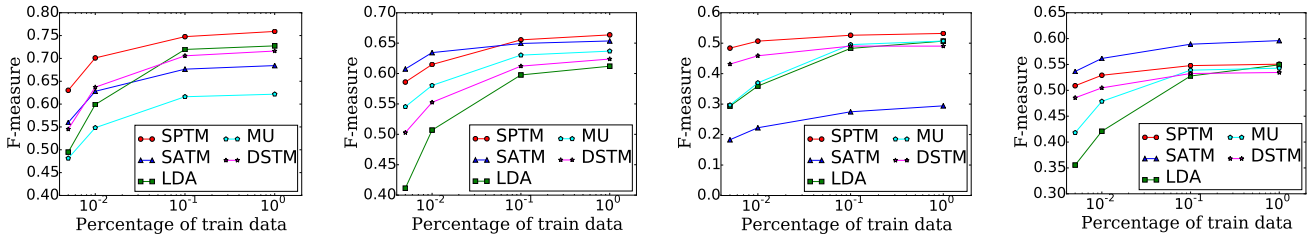


Figure 3: Illustration of overfitting of SATM.

To further validate the above finding on SATM, we design an interesting experiments by removing tweets in the original set from 50% to 90%, then training PTM and SATM on modified data sets. We report macro f-measure of five-fold cross validation in Fig. 3. From the results, we can see that SATM gets worse continuously with more tweets being removed, while PTM is quite stable when the removal ratio keeps smaller than 80%. This result again justifies our discussion above; that is, SATM is fragile to overfitting and performs poor when training set is inadequate.

Results of Varying Training Sizes. Topic representation of documents plays an important role when training examples are rare [13]. To compare how the latent semantic representation of documents learned by difference topic models enhances the classification performance when the training data is rare, we conduct this experiment. On all data sets, we use 80% short texts for training and the rest 20% for testing. With the topic representation, the short texts are classified by LIBLINEAR. To better understand the behavior of topic representations in classification, we

[4]http://www.csie.ntu.edu.tw/ cjlin/liblinear/



(a) News (b) DBLP (c) Questions (d) Tweets

Figure 4: Classification results of varying training sizes.

Table 3: Topic coherence (UCI) results on News and DBLP.

	PTM	SPTM	SATM	LDA	MU	DSTM
News	0.838	**0.910**	0.187	0.795	0.391	0.693
DBLP	**0.584**	0.514	-1.933	0.548	0.525	0.389

vary the ratio of labeled documents by sampling from the training set from the ratio of 0.5% to 100%.

Fig. 4 illustrates the f-measure under different sampling ratios of training data. For clarity, we omit the results of PTM since it performs closely to SPTM. From the results, we can find that SPTM consistently outperforms LDA, MU and DSTM on all data sets. This well demonstrates the robustness of our methods. SATM outperforms SPTM on Tweets, however, its results are again quite unstable due to overfitting.

3.2.2 Topic Evaluation by Topic Coherence

The UCI topic coherence results of our methods and all baselines on News and DBLP are presented in Table 3. As we have stated in Section 3.1.3, UMass topic coherence is not suitable for short texts [18]. UCI topic coherence is more appropriate, however, it requires external corpus. For well edited descriptions of news and titles of research papers, we can safely use Wikipedia as the reference corpus. As to user-generated contents, like Tweets and Questions, we omit the comparison of topic coherence due to absence of appropriate reference corpus.

From results in Table 3, we can find SPTM outperforms other methods on News and PTM outperforms other methods on DBLP. The superior performance of our methods as compared to LDA is in accordance with our understanding that learning topics from regular-sized pseudo documents can guarantee the quality of topics. On both data sets, LDA performs the best among baseline methods and SATM yields the worst coherence score. Surprisingly, LDA with weak prior produces higher topic coherence score than DSTM. The latter introduces sparse priors for both document-topic and topic-word distributions, which makes it more theoretically sounds than LDA. Possibly, having to inference large number of sparse priors causes DSTM faces with practical difficulty in learning a precise model. MU performs relative poor on news, which might indicates descriptions of news often cover more than one topic. On the contrary, MU yields relative good topic coherence score on DBLP, which might indicates titles in DBLP data set tends to cover fewer topics than descriptions of news. As a result of overfitting, SATM requires a large data set for training. The moderate size of news and DBLP prevents SATM from learning more coherence topics. However, the large time complexity of SATM's inference procedure in conflict with its needs of large training data. Note that, for a fair comparison with SATM, we set number of pseudo document P of PTM and SPTM as 1000. In fact, as shown in Section 3.2.4, PTM and SPTM can achieve better topic coherence with other P.

3.2.3 Topic Evaluation by Semantics

In this section, we show that the content of a pseudo document is semantic meaningful according to its topics as well as short texts assigned to it. After the training of PTM, short texts are grouped into pseudo documents, and topic proportion of each pseudo document is also learned. By looking into the most probably topics of a pseudo document, we can reveal its semantics.

We conduct this case study by performing PTM on DBLP. After the training, we first choose a pseudo document with a relative large number of short texts assigned to it, then obtain several most probably topics according to its topic proportion θ. At last, we show some short texts assigned to that pseudo document grouped by their significant topics. The result is illustrated in Fig. 5.

From the results we can find #371 pseudo document learned by PTM is mainly about classifier(in blue), combinatorial algorithm(in red) and optimization(in green). Obviously, above three topics are inherently correlated to each other, since they are closely related research areas. This convincing result suggests that the pseudo document learned by our method can aggregates similar short texts together according to their topics. Short texts listed below of those topics also support our observation, since their contents corresponds well to their topics.

3.2.4 Impact of Number of Pseudo Documents

PTM and SPTM both reveal topics from P pseudo documents, adjusting P is the key to ease the data sparsity problem faced with by traditional topic models like LDA. According to Tang et al. [19], number of documents D and length of documents N both are key limiting factors for LDA. Topic model can not learn topics accurately when training data has small D or small N. Since our methods learn topics from P pseudo documents, P is a key limiting factor for our methods. Intuitively, a small P will causes our model to produce less coherence topics.

Similar with Tang et al. [19], we use topic coherence to discuss the impacts of P. Specifically, we varying P from 50 to 2000, and study the performance of PTM and SPTM on news and DBLP according to UCI topic coherence. Results are reported in Fig. 6.

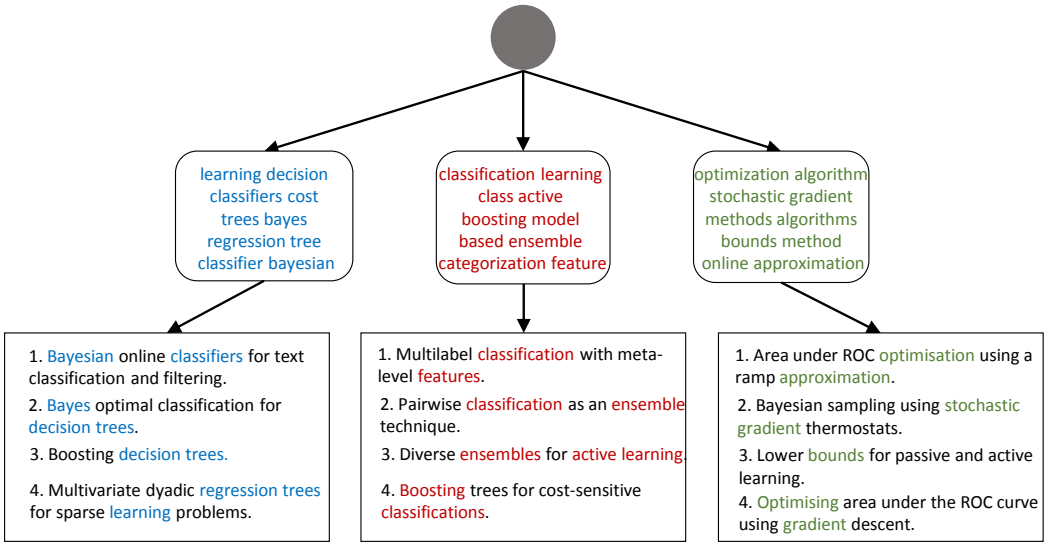
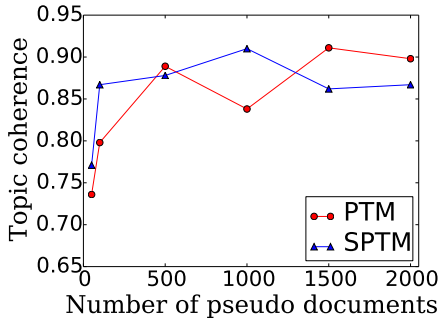
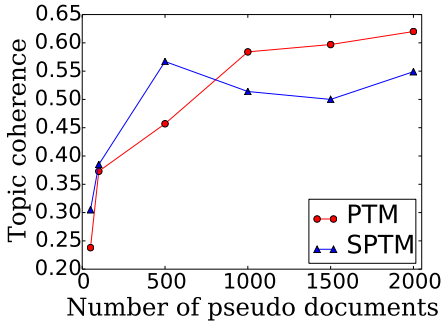


Figure 5: Semantic explanation of a sample pseudo document.



(a) News



(b) DBLP

Figure 6: Variation of topic coherence with the number of pseudo documents.

From the results, we can make two convincing observations. One observation is that topic model with small P results with less coherence topics. As illustrated in Fig. 6a, PTM produces less coherence topics when P equals to 50 and 100, and better topics since $P \geq 500$. SPTM also produces less coherence topics on news when $P = 50$. Such results are in accordance with theoretical findings about limiting factor D that topic model needs a large D to produce coherence topics. The other observation is sparse priors of pseudo documents can easy the topic errors when P is small. For instance, as shown in Fig. 6a, SPTM achieve large topic coherence score when $P = 100$, while PTM performs rather poor. Similar results in Fig. 6b, SPTM achieve large topic coherence score when $P = 500$, while PTM's topic coherence is relative small. Sparse prior can eliminate undesired correlation between pseudo documents and topics. When P is small, pseudo documents is highly likely to contain uncorrelated topics, therefore, sparse prior helps SPTM being more robust than PTM against small value of P.

Another interesting phenomenon is that when P is large enough, PTM consistently outperforms SPTM according to topic coherence. Which indicates adding sparse prior to document-topic distributions may weaken the performance of topic models when number of documents is large. This phenomenon is in accordance with topic coherence results in Table 3, where topic coherence of DSTM consistently smaller than LDA's.

Table 4: Classification results of PTM, SPTM and EPTM.

	News			DBLP		
	precision	recall	f-measure	precision	recall	f-measure
PTM	0.755	0.757	0.754	0.667	0.672	0.668
SPTM	0.760	0.761	0.759	0.661	0.667	0.663
EPTM	0.749	0.751	0.749	0.645	0.654	0.647

Table 5: Topic coherence of PTM, SPTM and EPTM.

	PTM	SPTM	EPTM
News	0.838	0.910	0.780
DBLP	0.584	0.514	0.489

3.2.5 *Impact of Membership Uniqueness to Topics*

Both PTM and SPTM assumes each short text comes from a single pseudo document, while EPTM relaxes such restriction. Although EPTM is more flexible than PTM and SPTM, it may performs worse than PTM and SPTM according to our discussion in Section 2.4. To study the performance of three models when applied to real word data sets,

we conduct this experiment. We compute UCI topic coherence of there models on DBLP and news. From Table 5, we can find EPTM performs consistently worse than other two models on both data sets. We also perform five-fold cross validation text classification on DBLP and news. From Table 4, we can find EPTM again performs consistently worse than other two models. Both results are in consistence with our theoretical discussion.

4. RELATED WORK

Data sparseness has long been the "nightmare" of topic modeling of short texts. One intuitive way is to make use of auxiliary information when available. For example, tweets contain not only textual content but also contextual information such as authorship, hashtag, time, location and URL, which can serve as supplemental information for topic modeling. Research along this line can be further categorized into to two types. One type tries to aggregate short texts directly according to auxiliary information [5, 25, 14], and the other aims to build specific models with those information during short texts generation [20, 8]. The following two paragraphs give the respective details.

Short texts typically mean few word co-occurrences, which prevents traditional topic models from fitting well to data. A straightforward approach to increase term co-occurrences per document is to aggregate short texts into longer ones. For instance, Hong et al. [5] report that a better topic model can be trained on aggregated short texts. Weng et al. [25] aggregate tweets from a same user into a pseudo document, then feed them to standard topic models. Since they focus on topical interests of users rather than individual tweets, the aggregation makes sense. Mehrotra et al. [14] compare several ways of tweet aggregation using different auxiliary information, and find that the one with hashtags yields the best performances. Still more excellent research with similar treatments will not be covered here for space concern.

Besides of the direct aggregation methods, some variants of basic topic model have also been proposed, which take auxiliary information into modeling directly. For example, Tang et al. [20] propose a multi-context topic model, which generates both context-specific and consensus topics across contexts. Jin et al. [8] use the web pages pointed by URLs in tweets as auxiliary long documents to learn better topics in tweets. More work along this line can be found in [6, 31].

In practice, however, auxiliary information is not always available or just too costly for deployment. As a result, recent research efforts have been put more on designing customized topic models for short texts. To the best of our knowledge, the biterm topic model proposed by Yan et al. [27] is among the earliest work, which directly models word pairs (i.e. biterms) extracted from short texts. By switching from sparse document-word space to dense word-word space, the biterm topic model learns more coherent topics than LDA. Zuo et al. [33] propose the word network topic model to learn topics from word co-occurrence networks, which produces more coherent topics than the biterm topic model. However, both models are occasionally criticized for lacking of direct topical representation for documents. Lin et al. [11] propose the dual sparse topic model, which replaces symmetric Dirichlet priors of LDA with *Spike and Slab* [7, 23] sparse ones. They show that the new model can learn more coherent topics as well as better topical representation of documents. The work most similar to ours is the self-aggregation topic model proposed by Quan et al. [18], which assumes that short texts are extracted from pseudo documents generated by LDA. However, the parameters of this model grow with the number of short texts, which makes it prone to overfitting. Clustering based topic models [26, 22, 12] can also aggregate short texts into clusters. However, they assume topic distributions of documents in a same cluster share the same prior rather than the same topic distribution. This way of modeling, compared to direct short text aggregation, has little benefit to alleviating data sparsity of short text topic modeling for providing no additional word co-occurrence information.

Despite of rich studies mentioned above, short texts topic modeling remains an open problem calling for more accurate yet cost efficient solutions. Our study in this paper is just an attempt for this purpose.

5. CONCLUSIONS

In this paper, we propose a Pseudo-document-based Topic Model (PTM) for short texts. By leveraging much less pseudo documents to self aggregate tremendous short texts, PTM gains advantages in learning topic distributions without using auxiliary contextual information. A sparsified version of PTM (SPTM) is also proposed to improve the topical description of PTM when the pseudo documents is small in number. Extensive experiments on real-world data sets demonstrate the superiority of PTM and SPTM to some state-of-the-art methods. Various interesting observations regarded to the number of pseudo documents and the uniqueness of topic membership of a short text are also carefully touched. To our best knowledge, our work is among the earliest studies in using self aggregation method for short text topic modeling.

6. ACKNOWLEDGMENTS

Dr. Junjie Wu was supported in part by National Natural Science Foundation of China (71322104, 71531001, 71471009, 71490723, 71171007), National High Technology Research and Development Program of China (SS2014AA012303), and Fundamental Research Funds for the Central Universities. Dr. Hui Zhang was supported in part by National High Technology Research and Development Program of China (2014AA021504).

7. REFERENCES

[1] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, apr 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, mar 2003.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[4] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[5] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88, 2010.

[6] Y. Hu, A. John, F. Wang, and S. Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI*, pages 59–65, 2012.

[7] H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[8] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 775–784, 2011.

[9] A. Q. Li, A. Ahmed, S. Ravi, and A. J. Smola. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 891–900, 2014.

[10] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584, 2006.

[11] T. Lin, W. Tian, Q. Mei, and H. Cheng. The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings of the 23rd international conference on World wide web*, pages 539–550, 2014.

[12] X. Liu, B. Du, C. Deng, M. Liu, and B. Lang. Structure sensitive hashing with adaptive product quantization. *IEEE Transactions on Cybernetics*, PP(0):1–12, 2015.

[13] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: An empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203, apr 2011.

[14] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. 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*, pages 889–892, 2013.

[15] 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*, pages 262–272, 2011.

[16] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, 2010.

[17] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, may 2000.

[18] X. Quan, C. Kit, Y. Ge, and S. J. Pan. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2270–2276, 2015.

[19] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198, 2014.

[20] J. Tang, M. Zhang, and Q. Mei. One theme in all views: Modeling consensus topics in multiple contexts. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 5–13, 2013.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[22] H. M. Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.

[23] C. Wang and D. M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, pages 1982–1989. 2009.

[24] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[25] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270, 2010.

[26] P. Xie and E. P. Xing. Integrating document clustering and topic modeling. *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2013.

[27] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456, 2013.

[28] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946, 2009.

[29] J. Yin and J. Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, 2014.

[30] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349, 2011.

[31] X. W. Zhao, J. Wang, Y. He, J.-Y. Nie, and X. Li. Originator or propagator?: Incorporating social role theory into topic models for twitter content analysis. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1649–1654, 2013.

[32] A. Zubiaga and H. Ji. Harnessing web page directories for large-scale classification of tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 225–226, 2013.

[33] Y. Zuo, J. Zhao, and K. Xu. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, pages 1–20, 2015.