A Word Embeddings Informed Focused Topic Model

He Zhao HE.ZHAO@MONASH.EDU
Lan Du LAN.DU@MONASH.EDU
Wray Buntine WRAY.BUNTINE@MONASH.EDU
Faculty of Information Technology, Monash University, Melbourne, Australia

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

In natural language processing and related fields, it has been shown that the word embeddings can successfully capture both the semantic and syntactic features of words. They can serve as complementary information to topics models, especially for the cases where word co-occurrence data is insufficient, such as with short texts. In this paper, we propose a focused topic model where how a topic focuses on words is informed by word embeddings. Our models is able to discover more informed and focused topics with more representative words, leading to better modelling accuracy and topic quality. With the data argumentation technique, we can derive an efficient Gibbs sampling algorithm that benefits from the fully local conjugacy of the model. We conduct extensive experiments on several real world datasets, which demonstrate that our model achieves comparable or improved performance in terms of both perplexity and topic coherence, particularly in handling short text data.

Keywords: Topic Models, Word Embeddings, Short Texts, Data Augmentation

1. Introduction

With the rapid growth of the internet, huge amounts of text data are generated everyday in social networks, online shopping and news websites, etc. Probabilistic topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are popular approaches for text analysis, by discovering latent topics from text collections.

Recently, word embeddings generated by GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013), have attracted a lot of attention in Natural Language Processing (NLP) and related fields. It has been shown that trained on large corpus, word embeddings can capture both the semantic and syntactic features of words so that similar words are close to each other in the embedding space. Therefore, if word embeddings can be used in topic models, it should improve modelling accuracy and topic quality. Moreover, as conventional topic models usually require a sufficient amount of word co-occurrences to learn meaningful topics, they can suffer from a large performance degradation over short texts (e.g., tweets and news headlines) because of insufficient word co-occurrence information. In such cases, the semantic and syntactic features of words encoded in embeddings can play a more important role, serving as complementary information.

On the other hand, sparse topics are preferred in topic modelling, which means most topics should be specific and are encouraged to focus on a small subset of the vocabulary. The topic sparsity is first implemented by using a small concentration parameter of the Dirichlet prior over topics (0.01 usually works well) (Wallach et al., 2009a). Recently, sparsity-enforcing priors such as the Indian Buffet Process compound Dirichlet (Williamson et al., 2010) have been proposed on topics (Wang and Blei, 2009), which place a "hard" constraint on the words that a topic can focus on. These sparsity-enforcing priors lead to compression as well as an easier interpretation of topics.

Inspired by those two lines of work, in this paper, we propose a Word Embedding Informed Focused Topic Model (WEI-FTM). This allows improved model accuracy and topic quality, especially for the cases where word co-occurrence data is poor, such as with short texts. Specifically, WEI-FTM applies a sparsity-enforcing prior on topics, allowing a subset of words to describe a topic. Unlike conventional FTM, where the focusing is learnt purely on the word co-occurrences, the focusing in WEI-FTM is additionally informed by external word embeddings. In this way, the proposed model encourages the topics to focus on the words that are more semantically or syntactically related, which is preferred in topic modelling.

WEI-FTM has the following key properties:

1. Compared to FTM, our model is able to discover more informed focused topics with more representative words, which leads to better model accuracy and topic quality.

2. Unlike most models incorporating word embeddings, our model does so by using them as prior knowledge, which we argue is a more coherent approach.

3. With the data augmentation technique, the inference of WEI-FTM can be done by an efficient and closed-form Gibbs sampling algorithm that benefits from local conjugacy of the model.

4. Finally, besides the word distribution of topics, our model also offers an alternative topic presentation over words, which can be obtained from word embeddings. It gives us a new way of interpreting topics and even better topic quality.

We conduct extensive experiments with several real datasets including both regular and short texts in various domains. The experimental results demonstrate that WEI-FTM achieves improved performance in terms of perplexity, topic coherence, and running time.

2. Related Work

In this section, we review three lines of related work: focused topic models, topic models with word embeddings, and short text topic models.

Focused Topic Models Focusing in topic models is first introduced on documents, allowing a document to focus on a subset of topics. Williamson et al. (2010) proposed the Focused Topic Model (FTM) on the document side with the Indian Buffet Process compound Dirichlet prior, where topics are selected by the IBP (Ghahramani and Griffiths, 2006). Zhou et al. (2012a) proposed a focused Poisson factorisation model with the negative binomial distribution, which can be viewed as a generalisation of FTM. Recently, Gan et al. (2015a) introduced a deep focused Poisson factorisation model. Instead of using IBP, document focusing in the model is constructed by stacking multiple layers of binary latent variables, connected by Gaussian weights. With the augmentation of the Pólya Gamma distribution (Polson et al., 2013), the model can be sampled with full conjugacy.

Unlike most of the previous approaches, the focussing in our model is applied to topics, not documents. The closest work to ours is the Sparse-Smooth Topic Model (Wang and Blei, 2009) which applied the IBP compound Dirichlet prior on topics, allowing a topic to focus on a subset of words. Teh and Gorur (2009) proposed the Stable-Beta IBP, a generalised IBP with a discount parameter. The Stable-Beta IBP can be used to model the power law behaviour in word occurrences. Furthermore, Archambeau et al. (2015) introduced the Latent IBP Dirichlet Allocation (LIDA), which uses the Stable-Beta IBP compound Dirichlet prior for both document focusing and topic focusing.

Topic Models with Word Embeddings Recently, there is growing interest in incorporating word features in topic models. For example, DF-LDA (Andrzejewski et al., 2009) incorporates word must-links and cannot-links using a Dirichlet forest prior in LDA; MRF-LDA (Xie et al., 2015) encodes word correlations in LDA with a Markov random field; WF-LDA (Petterson et al., 2010) extends LDA to model word features with the logistic-normal transform. As word embeddings have gained great success in NLP, they have been used as popular word features for topic models. LF-LDA (Nguyen et al., 2015) integrates word embeddings into LDA by replacing the topic-word Dirichlet multinomial component with a mixture of a Dirichlet multinomial component and a word embedding component. Instead of generating word types (tokens), Gaussian LDA (GLDA) (Das et al., 2015) directly generates word embeddings with the Gaussian distribution. MetaLDA (Zhao et al., 2017b) is a topic model that incorporates both document and word meta information. However, in MetaLDA, word embeddings have to be binarised, which will lose useful information. Despite the exciting applications of the above models, their inference is usually less efficient due to the non-conjugacy and/or complicated model structures. Moreover, to our knowledge, most of the existing models with word embeddings are extensions of a full LDA model, and neither use the embeddings as information for the prior, like WF-LDA, nor do they use the embeddings with topic focusing.

Short Text Topic Models Analysis of short text with topic models has been an active area with the development of social networks. One popular approach is to aggregate short texts into larger groups, for example, Hong and Davison (2010) aggregates tweets by the corresponding authors and Mehrotra et al. (2013) shows that aggregating tweets by their hashtags yields superior performance over other aggregation methods. Recently, PTM (Zuo et al., 2016) aggregates short texts into latent pseudo documents. Another approach is to assume one topic per short document, known as mixture of unigrams or Dirichlet Multinomial Mixture (DMM) such as Yin and Wang (2014); Xun et al. (2016). Closely related to ours are short text models that use word feature like embeddings. For example, Xun et al. (2016) introduced an extension of GLDA on short texts which samples an indicator variable that chooses to generate either the type of a word or the embedding of a word and GPU-DMM (Andrzejewski et al., 2011) extends DMM with word correlations for short texts. Although existing models showed improved performance on short texts, there still exist some challenges. For aggregation-based models, it is usually hard to choose which meta information to use for aggregation. The "single topic" assumption makes DMM models lose the flexibility to capture different topic ingredients of a document. The incorporation of word embeddings in the existing models is usually less efficient.

3. Model Details

Now we introduce the details of the proposed model. In general, WEI-FTM is a focused topic model where the focusing of topics are learnt from the target corpus and informed by external word embeddings. Specifically, suppose a collection of D documents with a vocabulary of V tokens is denoted as $\mathcal{D}$ and the L dimensional embeddings of the tokens are stored in a matrix $\mathbf{F} \in \mathbb{R}^{V \times L}$. Similar to LDA, WEI-FTM generates document $d \in \{1, \cdots, D\}$ with a mixture of K topics. Unlike LDA, where a topic is a distribution over all the tokens in the vocabulary, WEI-FTM allows a topic $k \in \{1, \cdots, K\}$ to focus on fewer tokens. We introduce a binary matrix $\mathbf{B} \in \{0, 1\}^{K \times V}$ where $b_{k,v}$ indicates whether topic k focuses on token v. Given $\boldsymbol{b_{k,:}}$, topic k is a distribution over a subset of the tokens, drawn from the Dirichlet distribution:

$$\boldsymbol{\phi_k} | \boldsymbol{b_{k,:}} \sim \text{Dirichlet}_V(\beta_0 \boldsymbol{b_{k,:}}) \tag{1}$$

where $\phi_{k,v} = 0$ iff $b_{k,v} = 0$.

To get informed by word embeddings, $b_{k,v}$ is drawn from the Bernoulli distribution whose parameter is constructed with word v's embeddings $\boldsymbol{f_{v,:}}$:

$$b_{k,v} \sim \text{Bernoulli}\left(\sigma(\pi_{k,v})\right) \tag{2}$$

$$\pi_{k,v} = \boldsymbol{f_{v,:}} \boldsymbol{\lambda_{k,:}}^T + c_k \tag{3}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{K \times L}$, $\boldsymbol{c} \in \mathbb{R}^K$, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function.

If we view $\boldsymbol{\lambda_{k,:}}$ as the embeddings of topic k, the intuition of our model is that if the closer the semantic/syntactic meanings (encoded in the embeddings) of tokens v to topic k, the larger the probability of k being described by v. Acting as the bias of topic k, c_k captures the information irrelevant to the embeddings. Gaussian prior is then used for both $\boldsymbol{\lambda_{k,:}}$ and $\boldsymbol{c}$:

$$\boldsymbol{\lambda_{k,:}}, \boldsymbol{c} \sim \mathcal{N}(\mathbf{0}, (\sigma_0)^2 \mathbf{I}) \tag{4}$$

where $(\sigma_0)^2$ is a hyper-parameter that controls the Gaussian variance.

Figure 1 shows the graphical model of WEI-FTM and the generative process is as follows:

1. For each topic k:
   - (a) Draw $\boldsymbol{\lambda_{k,:}}$ according to Eq. (4)
   - (b) For each token v: Draw $b_{k,v}$ according to Eq. (2)
   - (c) Draw $\boldsymbol{\phi_{k,:}}$ according to Eq. (1)
2. For each document d:
   - (a) Draw $\boldsymbol{\theta_{d,:}} \sim \text{Dirichlet}_K(\alpha_0 \mathbf{1}_K)$
   - (b) For the i^{th} word $w_{d,i}$ in document d:
     - i. Draw topic $z_{d,i} \sim \text{Categorical}_K(\boldsymbol{\theta_{d,:}})$
     - ii. Draw word $w_{d,i} \sim \text{Categorical}_V(\boldsymbol{\phi}_{z_{d,i},:})$

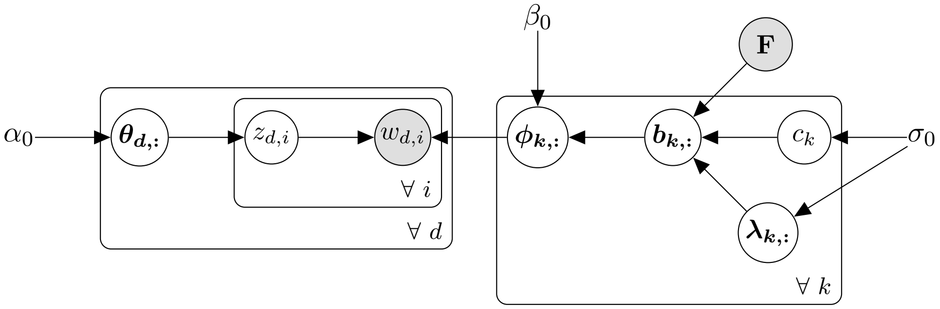


Figure 1: The graphical model of WEI-FTM. The i^{th} of N_d words in document d is $w_{d,i}$ and the corresponding topic is $z_{d,i}$. $\boldsymbol{\theta_{d,:}}$ is the topic distribution of d. α_0, β_0, and σ_0 are hyper-parameters of the model.

It is noteworthy that both $\boldsymbol{\pi_{k,:}} \in \mathbb{R}^V$ and $\boldsymbol{\phi_{k,:}}$ describe the weights of the words in topic k. Unlike $\phi_{k,v}$ which is obtained from the statistics of topic allocation of words, $\pi_{k,v}$ models the "similarity" of the embeddings of word v and topic k, obtained with the word embeddings. Therefore, π in our model can serve as an alternative presentation of topic k. This will be studied later in Section 5.3.

4. Inference

Unlike most existing models with word embeddings, our model facilitates the derivation of an efficient Gibbs sampling algorithm. With a data augmentation technique, WEI-FTM admits local conjugacy and a closed-form Gibbs sampling algorithm can be derived.

According to the generative process of WEI-FTM, the complete model likelihood is:

$$\prod_{d,i}^{D,N_d} p(w_{d,i}|z_{d,i}, \boldsymbol{\phi_{z_{i,n},:}})p(z_{i,n}|\boldsymbol{\theta_{d,:}}) \cdot \prod_d^D p(\boldsymbol{\theta_{d,:}}|\alpha_0) \cdot \prod_k^K p(\boldsymbol{\phi_{k,:}}|\boldsymbol{b_{k,:}}, \beta_0)$$
$$\cdot \prod_{k,v}^{K,V} p(b_{k,v}|\boldsymbol{\lambda_{k,:}}, c_k, \boldsymbol{f_{v,:}}) \cdot \prod_{k,l}^{K,L} p(\lambda_{k,l}|\sigma_0) \cdot \prod_k^K p(c_k|\sigma_0) \tag{5}$$

Sampling $z_{d,i}$ The sampling of a topic $z_{d,i}$ for a word $w_{d,i} = v$ is similar to LDA, while the candidate topics are limited to the topics that v describes:

$$p(z_{d,i} = k) \propto (\alpha_0 + m_{d,k}^{\neg i}) \frac{\beta_0 + n_{k,v}^{\neg d,i}}{\beta_0 V + n_{k,\cdot}^{\neg d,i}} \mathbb{I}_{(b_{k,v}=1)} \tag{6}$$

where $n_{k,v}^{\neg d,i} = \sum_{d',i'}^{D,N_{d'}} \mathbb{I}_{(d',i')\neq(d,i),w_{d',i'}=v,z_{d',i'}=k)}$, $m_{d,k}^{\neg i} = \sum_{i'}^{N_d} \mathbb{I}_{(i'\neq i,z_{d,i'}=k)}$, $n_{k,\cdot}^{\neg d,i} = \sum_v^V n_{k,v}^{\neg d,i}$, and $\mathbb{I}_{(\cdot)}$ is the indicator function.

Sampling $b_{k,v}$ Recall that $b_{k,v}$ indicates whether token v describes topic k. Therefore, if $n_{k,v} > 0$, which means there are words of v allocated to k, we do not need to sample $b_{k,v}$ (i.e., $p(b_{k,v}|n_{k,v} > 0) = 1$). When $n_{k,v} = 0$, the following Gibbs sampling for $b_{k,v}$ is used:

$$p(b_{k,v} = 1|n_{k,v} = 0) \propto \frac{\mathcal{B}(b_{k,\cdot}^{\neg v}\beta_0 + n_{k,\cdot}, \beta_0)}{\mathcal{B}(b_{k,\cdot}^{\neg v}\beta_0, \beta_0)} \sigma(\pi_{k,v}) \tag{7}$$

$$p(b_{k,v} = 0|n_{k,v} = 0) \propto 1 - \sigma(\pi_{k,v}) \tag{8}$$

where $\mathcal{B}(\cdot,\cdot)$ is the beta function and $b_{k,\cdot}^{\neg v} = \sum_{v' \neq v}^{V} b_{k,v'}$.

Sampling $\boldsymbol{\lambda_{k,:}}$ and $\boldsymbol{c}$ Recall that the likelihood of $b_{i,k}$ from Equation (2) is:

$$\frac{(e^{\pi_{k,v}})^{b_{k,v}}}{1+e^{\pi_{k,v}}} \tag{9}$$

The above likelihood can be augmented by introducing an auxiliary variable: $\gamma_{k,v} \sim \mathrm{PG}(1,0)$ (Gan et al., 2015b), where PG denotes the Pólya Gamma distribution (Polson et al., 2013). The augmentation works as following:

$$\frac{(e^{\pi_{k,v}})^{b_{k,v}}}{1+e^{\pi_{k,v}}} = \frac{1}{2} e^{(b_{k,v}-1/2)\pi_{k,v}} \int_0^\infty e^{-\gamma_{k,v}(\pi_{k,v})^2/2} p(\gamma_{k,v}) \mathrm{d}\gamma \tag{10}$$

Augmented in this way, the likelihood on $\pi_{k,v}$ has a Gaussian form, which means the likelihood of $\boldsymbol{\lambda_{k,:}}$ and $\boldsymbol{c}$ after the augmentation will be in Gaussian form as well. Given their Gaussian prior, one can sample $\boldsymbol{\lambda_{k,:}}$ as:

$$\boldsymbol{\lambda_{k,:}} \sim \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{11}$$

$$\boldsymbol{\mu_k} = \boldsymbol{\Sigma_k} \left(\textstyle\sum_v^V (b_{k,v} - 1/2 - c_k \gamma_{k,v}) \boldsymbol{f_{v,:}}^T \right) \tag{12}$$

$$\boldsymbol{\Sigma_k} = \left(\textstyle\sum_v^V \gamma_{k,v} \boldsymbol{f_{v,:}}^T \boldsymbol{f_{v,:}} + (\sigma_0)^{-2} \mathbf{I} \right)^{-1} \tag{13}$$

Also, $\boldsymbol{c}$ can be sampled similarly. Note that the Cholesky factorization can be applied to $\boldsymbol{\Sigma_k}$ to reduce the sampling complexity of $\boldsymbol{\lambda_{k,:}}$.

Finally, according to (Polson et al., 2013), we can sample $\gamma_{k,v}$ from its Pólya Gamma posterior: $\gamma_{k,v} \sim \mathrm{PG}(1, \pi_{k,v})$. One can approximate samples from the Pólya Gamma distribution by using a truncated sum of Gamma variables (Zhou et al., 2012b). In practice, a truncation level of 20 works well, so the sampling will be efficient.

Hyper-parameter Sampling We use a Gamma prior on $\beta_0 \sim \mathrm{Gamma}(\mu_0, \nu_0)$. The likelihood of β_0 is:

$$\prod_k^K \frac{\Gamma(b_{k,\cdot}\beta_0)}{\Gamma(b_{k,\cdot}\beta_0 + n_{k,\cdot})} \prod_v^V \frac{\Gamma(\beta_0 + n_{k,v})}{\Gamma(\beta_0)} \mathbb{I}_{(b_{k,v}=1)} \tag{14}$$

Two auxiliary variables are then introduced: $q_k \sim \mathrm{Beta}(b_{k,\cdot}\beta_0, n_{k,\cdot})$ and $t_{k,v} \sim \mathrm{CRP}(\beta_0, n_{k,v})$, which is the probability on the partition size of a Chinese Restaurant Process (Lemma 16 Buntine and Hutter, 2012) with β_0 and $n_{k,v}$ as the concentration and the number of customers respectively. The posterior becomes augmented as: $\prod_{k,v}^{K,V} (q_k)^{\beta_0} (\beta_0)^{t_{k,v}} \mathbb{I}_{b_{k,v}=1}$, which is conjugate to the Gamma prior of β_0 (Zhao et al., 2017b,a).

Similarly, α_0 can be sampled as well. However, as we are more interested in studying the word side, for fair comparison with other models, α_0 is not sampled in the experiments.

5. Experiments

In this section, we evaluate the proposed WEI-FTM against several recent advances including focus topic models, models with word embeddings, and short text topic models. The experiments were conducted on five real datasets including both regular and short texts. We report the performance in terms of perplexity, topic coherence, and running time per iteration. We also qualitatively compare the focusing and the topic quality of models.

5.1. Datasets, Compared Models, and Parameter Settings

In the experiments, two regular text datasets and three short text datasets were used:

- **Reuters** is extracted from the Reuters-21578 dataset[1] where documents without any labels are removed. There are 11,367 documents, the vocabulary size is 8,817, and the average document length is 73.
- **KOS** is obtained from the UCI Machine Learning Repository[2], which is used by Archambeau et al. (2015). It has 3,430 documents and the vocabulary size is 6,677. A document has 100 words on average.
- **WS**, Web Snippets, contains 12,237 web search snippets, used by Li et al. (2016). The vocabulary contains 10,052 tokens and there are 15 words in one snippet on average.
- **TMN**, Tag My News, consists of 32,597 English RSS news snippets from Tag My News, used by Nguyen et al. (2015). Each snippet contains a title and a short description. There are 13,370 tokens in the vocabulary and the average length of a snippet is 18.
- **Twitter**, is extracted in 2011 and 2012 microblog tracks at Text REtrieval Conference (TREC)[3], preprocessed in Yin and Wang (2014). It has 11,109 tweets in total. The vocabulary size is 6,344 and a tweet contains 21 words on average.

All the datasets were tokenised by Mallet[4] and we removed the words that exist in less than 5 documents and more than 95% documents. For word embeddings, we used the 50-dimensional GloVe word embeddings[5] pre-trained on Wikipedia for all the models that incorporate word embeddings. We further removed the words that are not in the vocabulary of GloVe embeddings in all the datasets. Note that besides word embeddings, our model can be used to incorporate other kinds of word features such as word correlations as well.

We evaluate the performance of the following models:

- **WEI-FTM**: The proposed model, the hyper-parameter σ_0 was set to 1.0 and β_0 was sampled according to Eq. (14). To comprehensively exam the effect of word embeddings, we further compare with WEI-FTM without any word embeddings, named "WEI-FTM-no", which only samples the bias $\boldsymbol{c}$. WEI-FTM and WEI-FTM-no were implemented in Matlab.
- **LDA** (Blei et al., 2003): the baseline model. A LDA variant with β_0 sampled according to Eq. (14) is also in comparison, named as "LDA-sym". LDA and LDA-sym were implemented in Matlab.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2] https://archive.ics.uci.edu/ml/datasets/bag+of+words
[3] http://trec.nist.gov/data/microblog.html
[4] http://mallet.cs.umass.edu
[5] https://nlp.stanford.edu/projects/glove/

- **WF-LDA**, Word Feature LDA (Petterson et al., 2010): a model that incorporates word features on the prior of ϕ. As no code is publicly available, we implemented it in Matlab, where the optimisation part was done by LBFGS with the default parameter settings. Following Mimno and McCallum (2008), the Gaussian variance was set to 10 for the default word feature and 0.05 for the other features. Note that we adopt the idea of WF-LDA for incorporating word embeddings while it was not originally proposed for that.

- **LF-LDA**, Latent Feature LDA (Nguyen et al., 2015): a model that incorporates word embeddings. The original implementation[6] was used. Following the paper, we used 1500 and 500 MCMC iterations for initialisation and sampling respectively and set λ to 0.6, and used the same word embeddings in WEI-FTM.

- **LIDA-topic**, Latent IBP Dirichlet Allocation (Archambeau et al., 2015) with topic focusing only. Reviewed in Section 2, LIDA applies the Stable-Beta IBP for both document and topic focusing. As our intent here is on topic focusing, the Stable-Beta IBP is only applied on the topic side. Unlike the proposed WEI-FTM, its focusing is not informed by external word features. Because the code of LIDA is not public available, we implemented it in Matlab, according to the paper. The MH sampling details of ξ are not given in the paper, we adopted LBFGS to optimise ξ in terms of the likelihood shown in Eq. (43) in the paper. β_0 was sampled by Eq. (14), the same as WEI-FTM. For the other sampling algorithms and settings, we followed the paper, where ζ was set to 0.25.

- **SSTM**, Sparse-Smooth Topic Model (Wang and Blei, 2009): a focused topic model that allows a topic to focus on a subset of words. Discussed by Archambeau et al. (2015), SSTM can be viewed as a special case of LIDA (by fixing ξ and ζ to 1.0 and 0.0 respectively).

- **GPU-DMM**, Generalized Pólya Urn DMM (Li et al., 2016): a model that incorporates word correlations. The original implementation[7] was used. The word correlations were generated from the distances of the word embeddings. Following the paper, we set the hyper-parameters μ and ϵ to 0.1 and 0.7 respectively, and the symmetric document Dirichlet prior to $50/K$.

- **PTM**, Pseudo document based Topic Model (Zuo et al., 2016): a model for short text analysis. The original implementation[8] was used. Following the paper, we set the number of pseudo documents to 1000 and λ to 0.1.

All the models, except where noted, the Dirichlet parameter of the document-topic distribution (α_0) and of the topic-word distribution (β_0) were set to 0.1 and 0.01 respectively. Our intent is to fairly compare just the topic-word aspect of the models, so we keep the document-topic aspect equivalent. Also, 2000 MCMC iterations were used to train the models.

[6] https://github.com/datquocnguyen/LFTM

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

[8] http://ipv6.nlsde.buaa.edu.cn/zuoyuan/

In summary, WEI-FTM (the proposed model), WF-LDA, LF-LDA, and GPU-DMM are models with word embeddings; LDA, LDA-sym, WEI-FTM-no, LIDA-topic, and SSTM are models without word embeddings; GPUDMM and PTM are models particularly for short texts.

5.2. Perplexity Evaluation

Perplexity is a measure that is widely used (Wallach et al., 2009b) to evaluate the modelling accuracy of topic models. The lower the score, the higher the modelling accuracy. To get unbiased perplexity, we randomly selected some documents in a dataset as the training set and the remaining as the test set. We first trained a topic model on the training set to get the word distributions of each topic k ($\boldsymbol{\phi_{k,:}}$). Each test document d was split into two halves containing every first and every second words respectively. We then fixed the topics and trained the models on the first half to get the topic proportions ($\boldsymbol{\theta_{d,:}}$) of test document d and computed perplexity for predicting the second half. We ran all the models 5 times with different random number seeds and report the average perplexity scores and the standard deviations. Note that GPU-DMM and PTM provided no code for inference on new documents so no corresponding perplexity results are given.

Table 1: Perplexity on regular texts. The best and second results are in boldface and underline respectively.

Dataset	Reuters			KOS		
#Topics	*50*	*100*	*200*	*50*	*100*	*200*
LDA	672±2	634±1	627±1	1488±4	1395±5	1315±2
LDA-sym	672±2	631±1	631±3	1461±4	1384±4	1327±4
LF-LDA	841±4	771±4	634±1	1707±16	1637±8	1636±10
WF-LDA	**651**±3	621±2	618±1	1426±10	1357±4	1306±3
SSTM	670±4	633±1	629±1	1462±4	1384±2	1324±4
LIDA-topic	671±1	638±3	-	1462±5	1385±4	1340±5
WEI-FTM-no	666±1	629±2	628±1	1445±3	1377±7	1322±1
WEI-FTM	656±4	**616**±2	**610**±3	**1416**±4	**1335**±2	**1284**±6

Tables 1 and 2 show the perplexities of the compared models[9]. The results indicate that the proposed WEI-FTM performed best on nearly all the datasets. In regular text datasets, it can be observed that WF-LDA was the second best, approaching WEI-FTM closely. However, our model had a clear win in short text datasets, especially on TMN and WS, which indicates that our incorporation of word embeddings is more effective than WF-LDA. Moreover, our model runs much faster than WF-LDA, which will be studied later in Section 5.5. Not informed by the word embeddings, WEI-FTM-no performed similarly to vanilla LDA. The comparison between WEI-FTM-no and WEI-FTM shows that the benefit of using the information encoded in the word embeddings. While using word embeddings, LF-LDA did not get better results than LDA in terms of perplexity, which is in line with

[9] The experiment of LIDA-topic on Reuters with 200 topics did not finish in a reasonable time due to the failure of optimising ξ with LBFGS.

Table 2: Perplexity on short texts. The best and second results are in boldface and underline respectively.

Dataset	WS		TMN		Twitter	
#Topics	*50*	*100*	*50*	*100*	*50*	*100*
LDA	957±6	875±4	1956±14	1855±14	580±2	497±2
LDA-sym	955±8	886±6	1951±8	1880±6	579±3	498±2
LF-LDA	1164±6	1039±17	2415±35	2393±11	849±16	685±6
WF-LDA	888±8	829±8	1881±9	1833±11	582±9	507±10
SSTM	956±8	881±4	1932±5	1858±10	578±6	496±5
LIDA-topic	964±7	882±6	1951±11	1875±15	578±1	488±2
WEI-FTM-no	948±5	877±6	1943±13	1874±12	573±2	497±2
WEI-FTM	**885**±11	**819**±8	**1845**±6	**1747**±12	**559**±5	**479**±5

the report in Fu et al. (2016). While introducing focusing as well, LIDA-topic and SSTM were not observed to have clear improvements in terms of perplexity.

5.3. Coherence Evaluation

To evaluate the coherence of the learnt topics, we used Normalised Pointwise Mutual Information (NPMI) (Lau et al., 2014) to calculate topic coherence score for topic k with top T words: $\text{NPMI}(k) = \sum_{j=2}^{T} \sum_{i=1}^{j-1} \log \frac{p(w_j,w_i)}{p(w_j)p(w_i)} / -\log p(w_j, w_i)$, where $p(w_i)$ is the probability of word i, and $p(w_i, w_j)$ is the joint probability of words i and j that co-occur together within a sliding window. Those probabilities are computed on an external large corpus, i.e., a 5.48GB Wikipedia dump in our experiments. The NPMI score of each topic in the experiments is calculated with top 10 words ($T = 10$) by the Palmetto package[10]. Again, we report the average scores and the standard deviations over 5 random runs of all the models.

Table 3: NPMI averaged over all the 100 topics on short text datasets. The best and second results are in boldface and underline respectively.

Datasets	WS	TMN	Twitter
LDA	-0.0044±0.0028	0.0343±0.0026	-0.0110±0.0064
LF-LDA	0.0130±0.0052	0.0397±0.0026	0.0008±0.0026
WF-LDA	**0.0289**±0.0060	0.0463±0.0015	-0.0074±0.0033
SSTM	-0.0012±0.0064	0.0381±0.0023	-0.0065±0.0040
LIDA-topic	-0.0063±0.0027	0.0420±0.0021	-0.0042±0.0036
WEI-FTM-ϕ	0.0043±0.0038	0.0417±0.0036	-0.0096±0.0017
WEI-FTM-π	-0.0092±0.0074	**0.0567**±0.0081	**0.0392**±0.0083
GPU-DMM	-0.0934±0.0106	-0.0970±0.0034	-0.1458±0.0104
PTM	-0.0029±0.0048	0.0355±0.0016	-0.0078±0.0008

[10] http://palmetto.aksw.org

Table 4: NPMI averaged over the top 20 topics. The best and second results are in boldface and underline respectively.

Datasets	WS	TMN	Twitter
LDA	0.1175±0.0122	0.1462±0.0036	0.0923±0.0042
LF-LDA	0.1230±0.0153	0.1456±0.0087	0.0972±0.0024
WF-LDA	**0.1499**±0.0131	0.1390±0.0527	0.0881±0.0090
SSTM	0.1163±0.0168	0.1476±0.0020	0.1002±0.0059
LIDA-topic	0.1147±0.0048	0.1553±0.0010	0.0964±0.0022
WEI-FTM-ϕ	0.1271±0.0015	0.1536±0.0041	0.0893±0.0026
WEI-FTM-π	0.1298±0.0079	**0.1832**±0.0172	**0.1615**±0.0120
GPU-DMM	0.0836±0.0105	0.0968±0.0076	0.0367±0.0164
PTM	0.1033±0.0081	0.1527±0.0052	0.0882±0.0037

It is known that conventional topic models directly applied to short texts suffer from low quality topics, caused by the insufficient word co-occurrence information. Here we study whether the focusing informed by word embeddings helps WEI-FTM improve topic quality, compared with other topic models that can also handle short texts. Table 3 and 4 show the NPMI scores for the compared models trained with 100 topics on the three short text datasets. Higher scores indicate better topic coherence. Besides the NPMI scores averaged over all the 100 topics, we also show the scores averaged over the top 20 topics with highest NPMI (Table 4), where "rubbish" topics are eliminated, following Yang et al. (2015). Recall that in WEI-FTM, the top words of the topics can be obtained by ranking either ϕ (WEI-FTM-ϕ) or π (WEI-FTM-π). We report both of them here.

Shown in Tables 3 and 4, it can be seen that in TMN and Twitter, WEI-FTM-π outperformed the others significantly. It indicates that the word embeddings successfully inform our model to learn better topics. It is also noteworthy that WEI-FTM-ϕ still got better NPMI than other models except WF-LDA in WS and TMN in general, although the word embeddings do not directly affect the top word ranking with ϕ.

To qualitatively analyse the topic qualities, we show the top 10 words of the topics of WEI-FTM in Table 5. The words in each topic were ranked by ϕ and π. It can be seen that in general, the coherence of the words ranked by π is better than that ranked by ϕ, which is in line with the overall NPMI scores shown in Table 3 and 4.

5.4. Focusing Analysis

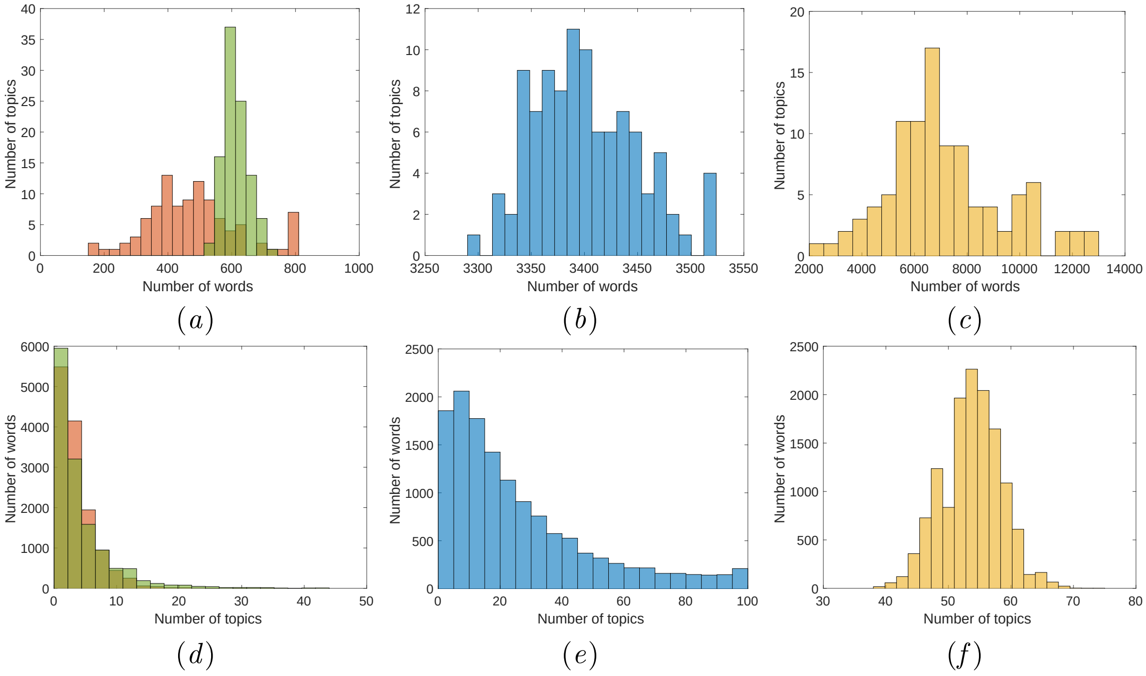
To compare the focusing in the focused topic models (WEI-FTM, WEI-FTM-no, SSTM, LIDA-topic), we show the histograms of the number of words per topic and the number of topics per word of two datasets in Figure 2 and 3. It can be observed that in WEI-FTM, the topics focused on fewer words than the others and the words described less topics. Compared to LIDA-topic, our model discovered more focused topics and the topics trend to be more diverse. It is also interesting to see how the word embeddings informed in the focusing: the topics in WEI-FTM-no are much less focused than those in WEI-FTM. This phenomenon helps explain why WEI-FTM gives better performance in the quantitative evaluations.

Table 5: Top 10 words of the topics discover by WEI-FTM on Twitter. Top 10 topics with the largest weights ($\sum_d^D \theta_{d,k}$) are selected. For each topic, the top words in the first row are ranked by ϕ and in the second are ranked by π respectively.

Topic	Top 10 words	NPMI
1	video bound kanye rogen west franco seth james kim kardashian	-0.0626
	starring movie sexy actress funny animated film comedy cartoon comedian	0.0680
2	china zone air japan east defense sea island disputed beijing	0.0141
	airspace diaoyu nato sovereignty territorial border resolution maritime military force	0.0543
3	watkins ian lostprophets guilty singer sex child baby rape pleaded	-0.0028
	murder convicted guilty sentence alleged rape imprisonment kidnapping trial sentenced	0.1291
4	swift taylor prince william bon jovi gala jon white winter	-0.0218
	sang concert sing singing princess dinner singer greeted danced tour	-0.0309
5	storm travel morning thanksgiving woman pill east plan winter weather	-0.0209
	weather mph rain crash sleet snow air jet storm coast	0.1036
6	scotland independence scottish paper white government independent salmond alex minister	0.0445
	parliament liberal prime election party democratic minister congress conservative vote	0.2352
7	nokia lumia phone window tablet smartphone inch device microsoft launch	0.1778
	playstation iphone ipad server smartphone gsm android xbox smartphones nintendo	0.1455
8	friday black thanksgiving shopping store day holiday retailer shopper year	0.0540
	dinner thanksgiving holiday shopping menu shop supermarket retail meal store	0.0491
9	patriot bronco manning brady peyton tom england sunday denver night	0.0389
	touchdown quarterback goalkeeper punt fumble defensive steelers coach nfl kickoff	0.1822
10	lakers bryant kobe los angeles washington extension year contract wizard	0.0770
	quarterback lakers nba coach knicks seahawks offseason postseason touchdown belichick	0.0520

Table 6: Running time (seconds per iteration) on 80% documents of each dataset

Dataset	Reuters		WS	
#Topics	*50*	*100*	*50*	*100*
LDA	6.6167	8.1239	1.1481	1.3904
LDA-sym	6.3883	8.2225	1.1255	1.3609
LF-LDA	2.6895	5.3043	2.4920	6.0266
WF-LDA	289.6488	636.8966	327.0750	724.7727
SSTM	10.7333	14.8040	4.0399	6.2999
LIDA-topic	23.4365	28.7910	3.9989	6.1942
WEI-FTM	24.6280	27.6666	6.4997	8.5074



(a) (b) (c)

(d) (e) (f)

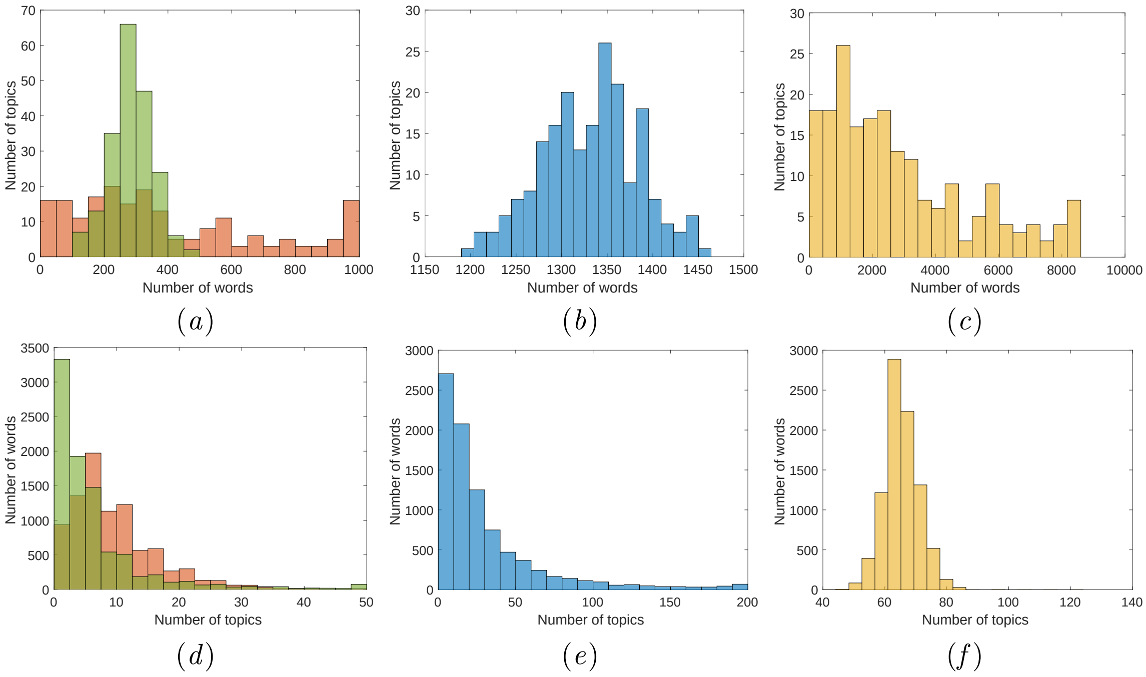
Figure 2: Histogram of the number of topics per word (a-c) and the number of words per topic (d-f) for the TMN dataset with 100 topics. Red: WEI-FTM, Green: LIDA-topic, Blue: SSTM, Yellow: WEI-FTM-no. The vocabulary size of TMN is 13,370. To show WEI-FTM and LIDA-topic in the same scale, we trimmed the topics and words with extremely low counts in (a).

5.5. Running Time

Here we empirically study the efficiency of the models. Table 5.5 shows the per-iteration running time of the compared models with different topics on the Reuters and WS datasets. Except LF-LDA (implemented in Java), the models were implemented in Matlab. All the models ran with the same settings on a cluster with a 14-core 2.6GHz CPU and 128GB RAM. It can be seen that although WF-LDA is comparable to WEI-FTM in some datasets, it runs much slower due to non-conjugacy.

6. Conclusion

In this paper, we have presented a focused topic model informed by word embeddings (WEI-FTM), which discovers more informed focused topics with more representative words, leading to better performance in terms of perplexity and topic quality. By leveraging the semantic and syntactic information encoded in word embeddings, our model is able to discover more focused and diverse topics with more representative words. In terms of inference, WEI-FTM enjoys full local conjugacy after augmentation, which facilitates an efficient Gibbs sampling algorithm for model inference. Without losing generality, WEI-FTM can work with both regular texts and short texts. The method of incorporating word



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: Histogram of the number of topics per word (a-c) and the number of words per topic (d-f) for the Reuters dataset with 200 topics. Red: WEI-FTM, Blue: SSTM, Green: LIDA-topic, Yellow: WEI-FTM-no. The vocabulary size of Reuters is 8,817. To show WEI-FTM and LIDA-topic in the same scale, we trimmed the topics and words with extremely low counts in (a).

embeddings introduced in this paper can also be applied to document features such as labels and authors, which is the subject of future work.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML*, pages 25–32, 2009.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into Latent Dirichlet Allocation using first-order logic. In *IJCAI*, pages 1171–1177, 2011.

Cedric Archambeau, Balaji Lakshminarayanan, and Guillaume Bouchard. Latent IBP compound Dirichlet allocation. *TPAMI*, 37(2):321–333, 2015.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *JMLR*, 3 (Jan):993–1022, 2003.

W. Buntine and M. Hutter. A Bayesian view of the Poisson-Dirichlet process. *arXiv preprint arXiv:1007.0296v2 [math.ST]*, 2012.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian LDA for topic models with word embeddings. In *ACL*, pages 795–804, 2015.

Xianghua Fu, Ting Wang, Jing Li, Chong Yu, and Wangwang Liu. Improving distributed word representation and topic model by word-topic mixture model. In *ACML*, pages 190–205, 2016.

Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, and Lawrence Carin. Scalable deep Poisson factor analysis for topic modeling. In *ICML*, pages 1823–1832, 2015a.

Zhe Gan, R. Henao, D. Carlson, and Lawrence Carin. Learning deep sigmoid belief networks with data augmentation. In *AISTATS*, pages 268–276, 2015b.

Zoubin Ghahramani and T.L. Griffiths. Infinite latent feature models and the Indian buffet process. In *NIPS*, pages 475–482, 2006.

Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proc. of the First Workshop on Social Media Analytics*, pages 80–88, 2010.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539, 2014.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *SIGIR*, pages 165–174, 2016.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*, pages 889–892, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionally. In *NIPS*, pages 3111–3119, 2013.

David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI*, pages 411–418, 2008.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313, 2015.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

James Petterson, Wray Buntine, Shravan M Narayanamurthy, Tibério S Caetano, and Alex J Smola. Word features for Latent Dirichlet Allocation. In *NIPS*, pages 1921–1929, 2010.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Yee W Teh and Dilan Gorur. Indian buffet processes with power-law behavior. In *NIPS*, pages 1838–1846, 2009.

Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *NIPS*, pages 1973–1981, 2009a.

H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In L. Bottou and M. Littman, editors, *ICML*, 2009b.

Chong Wang and David M Blei. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *NIPS*, pages 1982–1989, 2009.

Sinead Williamson, Chong Wang, Katherine A Heller, and David M Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, pages 1151–1158, 2010.

Pengtao Xie, Diyi Yang, and Eric Xing. Incorporating word correlation knowledge into topic modeling. In *NAACL*, pages 725–734, 2015.

Guangxu Xun, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li, Jing Gao, and Aidong Zhang. Topic discovery for short texts using word embeddings. In *ICDM*, pages 1299–1304, 2016.

Yi Yang, Doug Downey, and Jordan Boyd-Graber. Efficient methods for incorporating knowledge into topic models. In *EMNLP*, pages 308–317, 2015.

Jianhua Yin and Jianyong Wang. A Dirichlet multinomial mixture model-based approach for short text clustering. In *SIGKDD*, pages 233–242. ACM, 2014.

He Zhao, Lan Du, and Wray Buntine. Leveraging node attributes for incomplete relational data. In *ICML*, pages 4072–4081, 2017a.

He Zhao, Lan Du, Wray Buntine, and Gang Liu. MetaLDA: a topic model that efficiently incorporates meta information. *arXiv preprint arXiv:1709.06365*, 2017b.

Mingyuan Zhou, Lauren Hannah, David B Dunson, and Lawrence Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, pages 1462–1471, 2012a.

Mingyuan Zhou, Lingbo Li, David Dunson, and Lawrence Carin. Lognormal and Gamma mixed negative binomial regression. In *ICML*, volume 2012, page 1343, 2012b.

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. Topic modeling of short texts: A pseudo-document view. In *SIGKDD*, pages 2105–2114, 2016.