JSPS Grant-in-Aid for Scientific Research (S)

Central Bank Communication Design

working paper series

No.017 (March 2020)

Gaussian Hierarchical Latent Dirichlet Allocation: Bringing Polysemy Back

Takahiro Yoshida
Ryohei Hisano
Takaaki Ohnishi

Research Project on Central Bank Communication
702 Faculty of Economics, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Tel: +81-3-5841-5595 E-mail: watlab@e.u-tokyo.ac.jp
http://www.centralbank.e.u-tokyo.ac.jp/en/

Gaussian Hierarchical Latent Dirichlet Allocation: Bringing Polysemy Back

Takahiro Yoshida*
Graduate School of Information Science and Technology
The University of Tokyo

Ryohei Hisano*
Graduate School of Information Science and Technology
The University of Tokyo

Takaaki Ohnishi
Graduate School of Information Science and Technology
The University of Tokyo

February 26, 2020

Abstract

Topic models are widely used to discover the latent representation of a set of documents. The two canonical models are latent Dirichlet allocation, and Gaussian latent Dirichlet allocation, where the former uses multinomial distributions over words, and the latter uses multivariate Gaussian distributions over pre-trained word embedding vectors as the latent topic representations, respectively. Compared with latent Dirichlet allocation, Gaussian latent Dirichlet allocation is limited in the sense that it does not capture the polysemy of a word such as "bank." In this paper, we show that Gaussian latent Dirichlet allocation could recover the ability to capture polysemy by introducing a hierarchical structure in the set of topics that the model can

*Equal contribution

use to represent a given document. Our Gaussian hierarchical latent Dirichlet allocation significantly improves polysemy detection compared with Gaussian-based models and provides more parsimonious topic representations compared with hierarchical latent Dirichlet allocation. Our extensive quantitative experiments show that our model also achieves better topic coherence and held-out document predictive accuracy over a wide range of corpus and word embedding vectors.

1 Introduction

Topic models are widely used to identify the latent representation of a set of documents. Since latent Dirichlet allocation (LDA) [4] was introduced, topic models have been used in a wide variety of applications. Recent work includes the analysis of legislative text [24], detection of malicious websites [33], and analysis of the narratives of dermatological disease [23]. The modular structure of LDA, and graphical models in general [17], has made it possible to create various extensions to the plain vanilla version. Significant works include the correlated topic model (CTM), which incorporates the correlation among topics that co-occur in a document [6]; hierarchical LDA (hLDA), which jointly learns the underlying topic and the hierarchical relational structure among topics [3]; and the dynamic topic model, which models the time evolution of topics [7].

LDA uses multinomial distributions over words, whereas Gaussian LDA (GLDA) [11] uses multivariate Gaussian distributions over a pre-trained word embedding to represent the underlying topics. Using the word embedding vector space representation, GLDA has the added benefit of incorporating semantic regularities in a language, which results in increasing coherency [21, 29, 10] of topics [11]. Recent developments of this line of research include correlated Gaussian topic models (CGTM) [35], which add a correlational structure to the topics used in a document; the work of [2], which replaces the Gaussian distribution with a von Mises–Fisher distribution; and the latent concept topic model [15], which redefines each topic as the distribution over latent concepts, where the latent concept is modeled as a multivariate Gaussian distribution over the word embeddings.

A crucial discrepancy of GLDA and CGTM is that they fail to detect the polysemy of a term, such as "bank," which LDA and hLDA capture well [30]. LDA is a mixed membership model with no mutual exclusivity

constraint that restricts the assignment of words to one topic only [30]. As we show in the current paper, the delicate balance between a term that captures the probability of a word under a topic, and the probability of a topic given a document, in the collapsed Gibbs sampler of LDA [14], makes it possible to capture polysemy. However, although GLDA and CGTM are mixed membership models with no mutual exclusivity constraint, the probability of a word under a topic is characterized by a multivariate $\mathcal{T}$ distribution that outweighs the term that reflects the likelihood of a topic given a document. Hence, mutual exclusivity is likely to be unintentionally recovered, and the ability to detect polysemy is lost.

In this paper, we show that the ability to capture polysemy in GLDA-type models can be recovered by restricting the set of topics that can be used to represent a given document. One parsimonious implementation of such a restriction can be achieved by incorporating a hierarchical topic structure, as in hLDA [3, 5]. In our Gaussian hLDA, topics that can be used in a document are restricted by a path of topics that are learned jointly from the data. Instead of assigning a topic to each word position in a document, we assign levels that describe the position of the path from which the word was sampled.

At first glance, our model may seem to have a price to pay in terms of time complexity because of the added complexity of the model. However, because we do not need to sample from the entire set of topics for each word position in a document, the time complexity of our model does not necessarily worsen compared with GLDA and CGTM. Moreover, our model has the benefit of capturing polysemy in addition to being able to learn a compact hierarchical structure that shows the relationships among topics. Additionally, as in hLDA [3], Bayesian nonparametric techniques can also be used, thereby making it possible to determine the hierarchical tree structure more flexibly.

Other works also exist that combine topic modeling and word embeddings. [26] used information from the word similarity graph to achieve more coherent topics. [22] modified the likelihood of the model by combining information from pre-trained word embeddings with a log-linear function. Instead of using pre-trained word embedding vectors, some works have attempted to learn word embeddings and topics from the corpus jointly. The embedded topic model[13] uses the inner product between a word embedding and topic embedding as the natural parameter that governs the multinomial distribution and learns the two representations from the corpus simultaneously. In

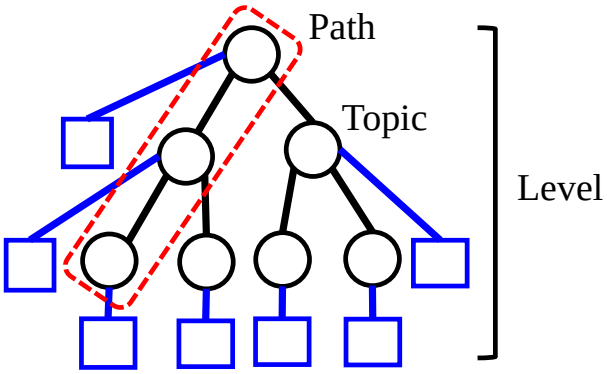

Figure 1: Hierarchy among topics: a circle represents a topic, dashed rectangle represents a path, each layer represents a level, and rectangular nodes represent the candidate branch that might be added by the nCRP.

[12], the model was further extended to incorporate the time evolution of the topic embeddings. The Wasserstein topic model[34] unifies topic modeling and word embedding using the framework of Wasserstein learning. Compared with these models, we leave the word embedding vectors as it is and enrich the topic co-occurrence structure of a document to adapt to the corpus of interest.

Our contributions are summarized as follows:

- We propose the Gaussian hLDA, which significantly improves the capture of polysemy compared with GLDA and CGTM.
- Our model jointly learns the topics, in addition to the hierarchical structure, and characterizes the relationship among the topics. The hierarchical structure can also be used to analyze the correlation structure among topics.
- The hierarchical tree structure can be estimated in a flexible manner using the nested Chinese restaurant process [3].
- Even though our model is far more expressive than GLDA and CGTM, the time complexity does not necessarily worsen compared with that of those two models.
- We show that our model exhibits a more parsimonious representation of topics than hLDA.

- Using three real-world corpora and three different pre-trained word embedding vectors, we show that our model outperforms state-of-the-art models both in terms of the held-out predictive likelihood and topic coherence.

2 Notation

We briefly summarize the mathematical notation used throughout the paper. D denotes the number of documents in a corpus, V denotes the number of unique words in the corpus, K denotes the number of topics, M denotes the dimension of the word embedding vector, and L denotes the depth of the maximum level of the hierarchy. Lower-case letters (e.g., d, v, and k) denote a specific document, word, or topic. $\theta_{d,k}$ denotes the probability of topic k for document d and, $\phi_{k,v}$ denotes the probability of word v in topic k. N_d denotes the number of words in a document. For each word position n in document d, $z_{d,n}$ denotes either the topic or level assignment for that word position and $w_{d,n}$ denotes the word that appears in word position n for document d. Furthermore, c_d denotes a path assignment to document d and l_d denotes the level distribution in d. For path and level assignments, a topic is uniquely defined as shown in Fig. 1. $N^{i}_{d(-n)}$ denotes the number of word positions in d that are assigned to either topic (LDA, GLDA, CGTM) or level i (hLDA, GhLDA), excluding $z_{d,n}$. $N^{>i}_{d(-n)}$ is defined similarly counting the number of word positions above level i. $N^{kv}_{-(d,n)}$ denotes the number of word positions in the entire corpus with word v and topic k, excluding $z_{d,n}$. $GEM(m, b)$ denotes the Griffiths, Engen, and McCloskey distribution [27], which is used to define a prior level distribution among a path, and m and b denote hyperparameters that control the stick-breaking process. $nCRP(\gamma)$ represents the nested Chinese restaurant process [3], where γ denotes a hyperparameter that controls the probability of a new branch that emerges in the current tree (i.e., the parameter that controls the likelihood of the blue rectangle being chosen in Fig. 1). $Dir(\alpha)$ represents a Dirichlet distribution and $Mult$ represents a multinomial distribution, where α denotes a hyperparameter vector. $N(\mu, \Sigma)$ and $\mathcal{T}_v(\mu, \Sigma)$ denote a normal distribution and multivariate $\mathcal{T}$ distribution with mean vector μ and covariance matrix Σ, respectively. $\mathcal{NIW}(u, \Psi, v, \kappa)$ denotes a normal inverse Wishart distribution with hyperparameters u, Ψ, v, κ, where u denotes a vector, Ψ denotes a matrix, and v and κ denote positive real values. Furthermore, $\kappa_k = \kappa + |s_k|$,

$v_k = v + |s_k|$, $\Psi_{s_k} = \Psi + \frac{\kappa|s_k|}{\kappa_k}(\bar{x}_{s_k} - u)(\bar{x}_{s_k} - u)^T + \Sigma_{i \in s_k}(x_i - \bar{x}_{s_k})(x_i - \bar{x}_{s_k})^T$, $u_k = \frac{\kappa\mu_0 + |s_k|\bar{x}_{s_k}}{\kappa_k}$, where s_k denotes the set of indicators of word positions that is assigned to topic k and $\bar{x}_{s_k}$ denotes the mean vector among the indicators in s_k.

3 Related Work

3.1 Gaussian Latent Dirichlet Allocation

The generative process of LDA and GLDA can be written similarly, and we focus on the GLDA case. GLDA uses word embedding vectors to characterize words in a document. We define $D, N_d, K, \theta_d, z_{d,n}$ precisely, as summarized in the previous section. Instead of considering $w_{d,n}$ as an indicator that denotes a word, as in LDA, we consider it as a vector from a pre-trained word embedding. The generative process is summarized as follows:

(1) For all topics k, sample $\mu_k, \Sigma_k \sim \mathcal{NIW}(u, \Psi, \kappa, v)$.
(2) For each document d,
  (a) sample topic proportion $\theta_d \sim Dir(\alpha)$; and
  (b) for each word position in d, sample topic assignments $z_{d,n} \sim Mult(\theta_d)$ and words from $w_{d,n} \sim \mathcal{N}(\mu_{z_{d,n}}, \Sigma_{z_{d,n}})$.

The collapsed Gibbs sampler of GLDA can be written as

$$p(z_{d,n} = k | w, z_{-(d,n)}) \propto \frac{\alpha_k + N^k_{d(-n)}}{\sum_{k'}(\alpha_{k'} + N^{k'}_{d(-n)})} \cdot \mathcal{T}_{v_k - M + 1}\left(w_{d,n} | u_k, \frac{\kappa_k + 1}{\kappa_k(v_k - M + 1)}\Psi_{s_k}\right). \quad (1)$$

LDA is recovered by replacing "$\mu_k, \Sigma_k \sim \mathcal{NIW}(u, \Psi, \kappa, v)$" in (1) with "$\phi_k \sim Dir(\beta)$," "$w_{d,n} \sim \mathcal{N}(\mu_{z_{d,n}}, \Sigma_{z_{d,n}})$" in (2)(b) with "$w_{d,n} \sim Mult(\phi_{z_{d,n}})$," and the second term in the sampler with "$\frac{\beta_v + N^{kv}_{-(dn)}}{\sum_{v'}(\beta_{v'} + N^{kv'}_{-(dn)})}$."

3.2 Correlated Gaussian Topic Model

CGTM [35] is an extension of GLDA that incorporates correlation among topics used in a document, similar to CTM [6]. The generative process is summarized as follows:

(1) For all topics k, sample $\mu_k, \Sigma_k \sim \mathcal{NIW}(u, \Psi, \kappa, v)$.
(2) To model the correlation among topics, sample
$\mu_a, \Sigma_a \sim \mathcal{NIW}(u_a, \Psi_a, \kappa_a, v_a)$.
(3) For all documents d,
(a) sample $\eta_d \sim \mathcal{N}(\mu_a, \Sigma_a)$;
(b) transform η_d to a topic proportion vector θ_d
using a softmax function $\theta_d = \frac{\exp(\eta_d)}{\sum_k \exp(\eta_{d,k})}$; and
(c) for all word positions in d, sample topic assignments
$z_{d,n} \sim Mult(\theta_d)$ and resulting words from
$w_{d,n} \sim \mathcal{N}(\mu_{z_{d,n}}, \Sigma_{z_{d,n}})$.

CGTM can be estimated by alternatively sampling η_d and topic assignments for each word position $z_{d,n}$. The sampling of η_d is rather involved, and includes additional auxiliary variable λ_d and sampling from a Polya–Gamma distribution[28, 18]. After η_d (and therefore θ_d) is sampled, the topic assignments $z_{d,n}$s are sampled using

$$p(z_{d,n} = k|w, z_{-(d,n)}) \propto \frac{exp(\eta_d^k)}{\Sigma_i exp(\eta_d^i)} \cdot \mathcal{T}_{v_k - M + 1}\left(w_{d,n}|u_k, \frac{\kappa_k + 1}{\kappa_k(v_k - M + 1)}\Psi_{s_k}\right). \tag{2}$$

3.3 Hierarchical Latent Dirichlet Allocation

The goal of hLDA is to identify topics and hierarchical relationships among the topics simultaneously from the corpus. Words in a document are drawn from the restricted set of topics that are characterized using paths from the hierarchical topic structure. Because of the hierarchical tree structure, topics in the upper level are used more frequently and thus capture more general terms than the lower level. To learn the hierarchical structure more flexibly, hLDA[3] uses the nested Chinese restaurant process as the prior distribution that defines the hierarchy over topics. The generative process is summarized as follows:

(1) For all topics k, sample $\phi_k \sim Dir(\beta)$.
(2) For each document d,
(a) sample a path assignment $c_d \sim nCRP(\gamma)$;
(b) sample a distribution over levels in the path, $l_d \sim GEM(m, b)$; and
(c) for all word positions in d, first choose the level assignments $z_{d,n} \sim Mult(l_d)$ and then the resulting words from the topic at that level in the path, $w_{d,n} \sim Mult(\phi_{c_d[z_{d,n}]})$.

In hLDA, we need to sample both the path assignments for all documents and level assignments for all word positions. The Gibbs sampling algorithm is similar to those used in GhLDA, so we omit it here.

4 Gaussian Hierarchical Latent Dirichlet Allocation

4.1 Mutual Exclusivity

The problem with GLDA and CGTM can be clarified by considering the sampling equations of GLDA (i.e., Eq. 1) and CGTM (i.e., Eq. 2). Two observations are worth mentioning. First, the only difference between Eq.1 and Eq.2 is the first term on the right-hand side of each equation, which corresponds to the probability of a topic given a document (Eq. 1) and the probability of a topic given a document with correlation (Eq. 2).

Second, although the first term on the right-hand side of the sampling equation can vary at most in the order of $O(10^{-N_d})$ among the topics, the second term is a multivariate probability density function that can vary much more widely. The order of variability of the $\mathcal{T}$ distribution among the topics widens when the data points in the word embedding that we want to cluster are multimodal, thereby ensuring each centroid of the Gaussian mixture to be placed in distinct positions in the word embedding space. Similar words in a word embedding space tend to cluster together, which makes word embeddings far from unimodal. This condition results in the second term outweighing the first term, and mutual exclusivity is likely to be unintentionally recovered in GLDA and CGTM.

Table 1: Summary of qualitative characteristics

Model	Pruning	Polysemy	Correlation	Embedding
LDA	×	◯	×	×
hLDA	◯	◯	◯	×
GLDA	◯	×	×	◯
CGTM	◯	×	◯	◯
GhLDA	◯	◯	◯	◯

4.2 Gaussian Hierarchical Latent Dirichlet Allocation

To create a mixed membership model with no mutual exclusivity constraint, even in cases that consider multivariate Gaussian distributions, we need to go beyond merely sampling topic assignments for each word position in the corpus and restrict the set of topics that can be used to represent a given document. By doing so, when a topic such as "finance, bank, loan" appears in a document, we can only use a particular topic such as "banks, ratio, interest" without being able to sample from all the available topics. This restriction guarantees that there is no restriction on mutual exclusivity and, as a bonus, can be used to capture the correlation among topics. One straightforward approach to add this constraint is via hierarchical topic modeling, as in [3, 5]. In the hierarchical construction, topics are ordered according to the level of abstraction from top to bottom. Path c_d is used to characterize the topics that can be used in a document d, and each word position in a document has level assignments $l_{d,n}$s that capture the level at which the word is sampled.

The generating process of GhLDA is as follows:

(1) For all topics k, sample $\mu_k, \Sigma_k \sim \mathcal{NIW}(u, \Psi, \kappa, v)$.
(2) For each document d,
(a) sample a path assignment $c_d \sim nCRP(\gamma)$;
(b) sample a distribution over level of the path:
$l_d \sim GEM(m, b)$; and
(c) for all word positions in d, first choose the level assignments $z_{d,n} \sim Mult(l_d)$ and the resulting words from the topic at level $z_{d,n}$ in the path, $w_{d,n} \sim \mathcal{N}(\mu_{c_d[z_{d,n}]}, \Sigma_{c_d[z_{d,n}]})$.

4.3 Gibbs Sampling Algorithm

We need to sample both the path assignments for all documents d and level assignments for all word positions $w_{d,n}$. The Gibbs sampling algorithm is as follows;

(1) For each document d, first sample path assignment
$c_d \sim p(c_d|w, c_{-d}, z, H)p(w_d|c, w_{-d}, z, H)$; and
(2) for all word positions in d, sample level assignments
$p(z_{d,n}|z_{-(d,n)}, c, w, H) \propto p(z_{d,n}|z_{d,-n}, H)$
$p(w_{d,n}|z, c, w_{-(d,n)}, H)$,

where H is the set of hyperparameters in the model. The probability of a path is the product of the prior on paths defined by $nCRP(\gamma)$ (i.e., $p(c_d|w, c_{-d}, z, H)$) [3], and the probability of a word given a specific path, which is

$$p(w_d|c, w_{-d}, z, H) = \prod_{l=1}^{L} \frac{1}{\pi^{tM/2}} \frac{\Gamma_M(\frac{v+|s_{c[l]}|+|t_l|}{2})}{\Gamma_M(\frac{v+|s_{c[l]}|}{2})} \cdot \frac{|\Psi_{|s_{c[l]}|}|^{\frac{v+|s_{c[l]}|}{2}}}{|\Psi_{|s_{c[l]}|+|t_l|}|^{\frac{v+|s_{c[l]}|+|t_l|}{2}}} (\frac{\kappa + |s_{c[l]}|}{\kappa + |s_{c[l]}|+|t_l|})^{M/2}, \tag{3}$$

where N^c_{-d} denotes the number of documents assigned to path c, excluding d, $s_{c[l]}$ denotes the set of word positions assigned to topic $c[l]$, t_l denotes the set of word positions assigned to level l in d, and Γ_d denotes the multivariate gamma function. The probability of a level is defined as

$$p(z_{d,n} = l|c, z_{-(d,n)}, w) = \frac{mb + N^l_{d(-n)}}{b + N^{\geq l}_{d(-n)}} \prod_{i=1}^{i=l-1} \frac{(1-m)b + N^{>l}_{d(-n)}}{b + N^{\geq l}_{d(-n)}} \cdot \mathcal{T}_{v_k - M + 1}\left(w_{d,n}|u_k, \frac{\kappa_k + 1}{\kappa_k(v_k - M + 1)}\Psi_{s_{c[l]}}\right). \tag{4}$$

The qualitative characteristics of LDA, hLDA, GLDA, CGTM, and GhLDA are summarized in Table 1. Pruning implies the necessity to prune highly frequent words, such as stop words, from the corpus. Whereas LDA fails to

Table 2: Running time complexities

Model	Complexity
LDA	$O(N_d K)$
hLDA	$O(K + N_d L)$
GLDA	$O(N_d K M^2)$
CGTM	$O(K^3 + N_d K M^2)$
GhLDA	$O(K M^2 + N_d L M^2)$

provide interpretable topics without pruning, all the other models handle this with ease. Polysemy implies the ability to capture polysemy. The manner in which GLDA and CGTM fail is described in Section 5. Correlation implies capturing the co-occurrence of topics in a document and embedding means the use of pre-trained word embedding vectors.

4.4 Complexity Analysis

We compare the running time complexity of all the models. Because hLDA, CGTM, and GhLDA include steps that require us to sample document-level parameters using all the words that appear in a document, we focus on the running time complexity to sample all assignments for a given document d. Table 2 summarizes the time complexities. Each sampling step in GLDA requires us to evaluate the determinant and inverse of the posterior covariance matrix, which is cubic. However, as indicated by [11], this can be reduced to $O(M^2)$ using the Cholesky decomposition of a covariance matrix. Because each word position has K topics to consider, and there are N_d words in a document, the total time complexity of GLDA is $O(N_d K M^2)$. LDA does not require us to calculate the inverse of the posterior covariance matrix, which makes the time complexity $O(N_d K)$. For each document, CGTM requires the sampling of document-level parameters η_d and λ_d. This step adds another $O(K^3)$ to the complexity.

Compared with these models, GhLDA first evaluates the posterior predictive probability for all paths. The straightforward calculation results in $O(PLM^2)$, where P denotes the number of paths and L denotes the maximum depth among all paths. However, exploiting the tree structure, we can reduce the calculation to $O(KM^2)$. After sampling the path, GhLDA pro-

Table 3: Selected topics related to "Rivers" and "Banks/Financial" in the Wikipedia dataset

Model	Topic and Top 5 Words
LDA	0 [the,in,creek,is,it]
(K=20)	2 [the,in,is,financial,for]
GLDA	10 [bank, financial,banks,banking,central]
(K=20)	13 [creek,de,lake,water,french]
GLDA	0 [river,bank,creek,flows,group]
(K=40)	21 [police,financial,banking,market,management]
CGTM	14 [bank,financial,university,mathematical,theory]
(K=20)	17 [police,creek,air,services,lake]

ceeds to sample levels for each word position in a document. Because each path only has a most L topics, sampling-level assignment for all words in a document takes $O(N_d L M^2)$. Adding both steps leads to $O(KM^2 + N_d L M^2)$ in total. Similar arguments can be used to calculate the time complexity of hLDA, which is $O(K + N_d L)$.

A few points are worth mentioning. All the models that use word embedding vectors are much slower than their plain counterparts because of the additional step of computing the Cholesky decomposition. However, comparing GLDA and GhLDA, we can see that GhLDA does not necessarily increase the time complexity compared with GLDA. If $N_d K \leq K + N_d L$, the time complexity of GhLDA is lower than that of GLDA[1]. Surely enough, this argument does not take into account the number of iterations required for collapsed Gibbs sampling to converge. However, it still highlights the fact that the time complexity of GhLDA is not necessarily worse than that of GLDA.

[1]This is indeed a reasonable scenario. For instance, assume that there are 100 words in a document d (i.e., $N_d = 100$). Whereas GLDA with $K = 20$ leads to $N_d \times K = 2,000$, GhLDA with the branch structure of $[1, 1, 4, 4]$ (i.e., $K = 22$ and $L = 4$) results in $K + N_d L = 422$.

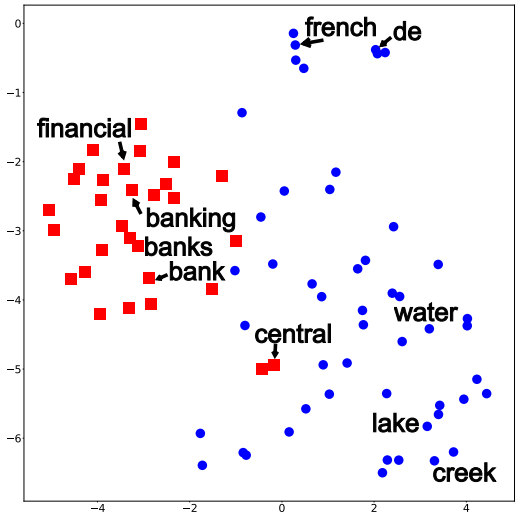

Figure 2: Low-dimensional representation of words and their topic assignments using GLDA

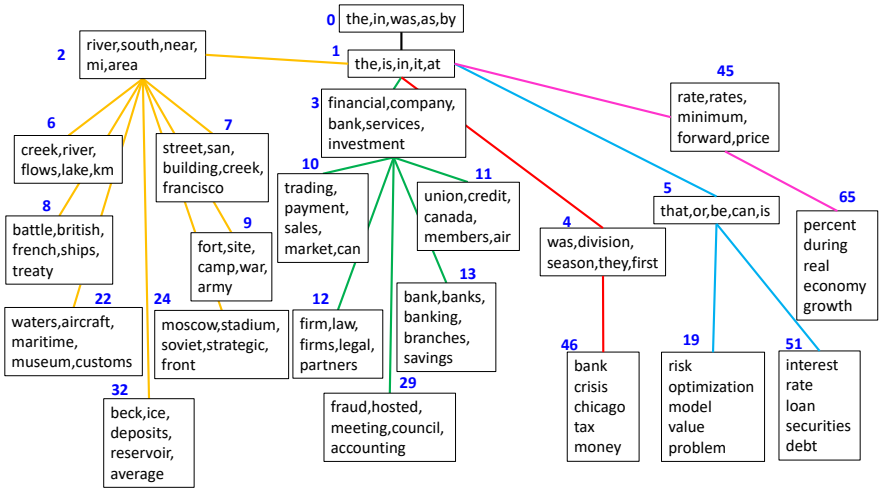

Figure 3: Partial depiction of the topic hierarchy estimated using hLDA

5 Experiments

5.1 Datasets

We conducted experiments using three open datasets, which were all included in our source code. One of the datasets (i.e., Wikipedia) was assembled particularly for the bank polysemy capturing task. We summarize the datasets below.

- The Wikipedia dataset, abbreviated as Wiki in the table, is a dataset particularly assembled for the bank polysemy capturing task. The corpus was created from DBpedia-2016 long abstract data [1]. Each long abstract in the DBpedia dataset has several labels that are attached to classify each article. We focused on the following six categories: "Rivers," "Banks/Financial," "Military," "Law," "Mathematical," and "Football." We sampled evenly from these categories to create a corpus of 6,000, of which 5,000 were used for training and 1,000 for testing. The main feature of this dataset is the inclusion of the "Rivers" and "Banks/Financial" categories. By randomly sampling from these categories, we created a corpus that used "bank" both as a financial institution and a steep place near a river. We used words that appeared more than 50 times in the corpus, and did not remove stop words, as in hLDA [3]. We further focused on words that appeared in all the pre-trained word embeddings described below.

- Amazon review data is a dataset of gathered ratings and review information [19][2]. We sampled evenly from the following five categories: "Electronics," "Video Games," "Home and Kitchen," "Sports and Outdoors," and "Movies and TV," and created a corpus of 6,000, of which 5,000 were used for training and 1,000 for testing. The other settings were the same as above.

- Reuters data is a news dataset web-scraped from Reuters news. We collected 6,000 news stories during the period Jan 2016 to Feb 2016, of which 5,000 were used for training and 1,000 for testing. The other settings were the same as above.

For pre-trained word embedding vectors, we used the GloVe (50 dimension) [25], word2vec (300 dimension) [20], and fasttext (300 dimension) [8]

[2]The entire dataset is available at `http://jmcauley.ucsd.edu/data/amazon/`

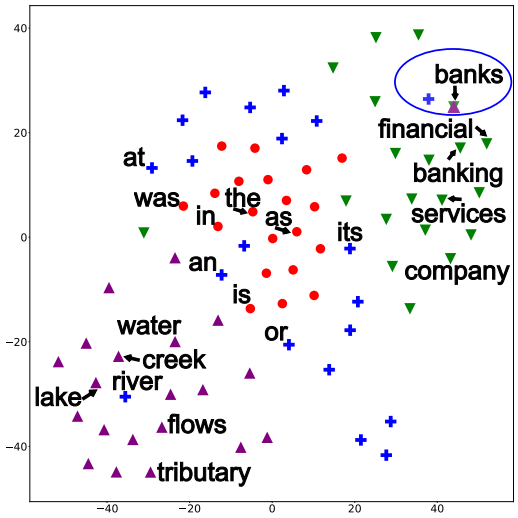

Figure 4: Low-dimensional representation of words and their topic assignments using GhLDA

word embedding vectors. Hence, in total, we had nine settings for models using word embeddings.

5.2 Settings

We compared GhLDA with LDA, hLDA [3], GLDA [11], and CGTM [35]. For the topic coherence and predictive held-out likelihood experiments, the number of topics for LDA, GLDA, and CGTM was fixed to 40. For our qualitative analysis, we also considered the case of 20 topics.

The hyperparameters that governed the topic distributions were set to $\alpha = 0.1, \beta = 0.1$ for LDA, and $v = 0.1, \kappa = 0.1, \Psi_{glove} = 50 * I, \Psi_{word2vec} = 40 * I, \Psi_{fasttext} = 20 * I$ for GLDA and CGTM, where I denotes an identity matrix. We ran the sampler for 50 epochs for these models, where one epoch was equal to sampling all the word positions in the corpus once. The hyperparameters controlling GEM and $nCRP$ were set to $m = 0.5, b = 100, \gamma = 0.1$ similar to [3]. The initial tree structure of hLDA and GhLDA was set to $[1, 1, 4, 4]$, where each number corresponds to the number of branches at each level. In hLDA, η was set to vary among the levels as $[2, 1, 0.5, 0.25]$. A similar strategy was used in GhLDA, where we adjusted Ψ to vary among the levels in the ratio $[1, 0.8, 0.6, 0.4]$, where the top level was identical to

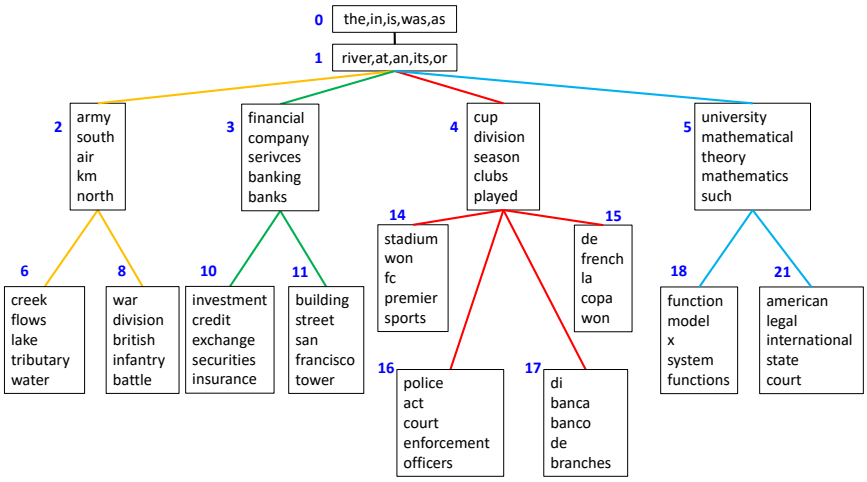

Figure 5: Topic hierarchy estimated using GhLDA

GLDA. We truncated the tree at level four, as in [3]. For GhLDA, we further ran the sampler without adding any leaves for five epochs. For the initial level assignments, half of the assignments were chosen by dividing the cumulative distribution function of word frequency into four segments and assigning from top to bottom according to the segments. The other half was chosen randomly. These additional steps were performed to stabilize the learning of the Gaussian mixture components. We ran the sampler for 100 epochs.

5.3 Capturing Polysemy

We compare the models' ability to capture polysemy, paying particular attention to the term "bank(s)," using the Wikipedia dataset. We use GloVe as a case study; the other word embeddings provide similar results. First, as shown in Table. 3, in topics trained using GLDA with $K = 20$, topic 10 included terms related to finance, such as "financial," "banking," and "central," and topic 13 contained terms related to the river, such as "creek," "lake," and "water." However, not a single "bank" or "banks" that appeared in the corpus was assigned to the river topic (i.e., topic 13), and all these words were assigned to the finance topic (i.e., topic 10). Similar observations were made, even when K was increased to 40. In this case, we could see terms related to finance, such as "financial," "market," and "management," in topic 21, and terms related to river, such as "river," "creek," and "flows," in topic

0. However, topic 0 also contained financial terms, such as " investment," "credit," and "exchange;" hence, the topic was inappropriately mixed. This observation implies that although increasing the number of topics makes the constraint on mutual exclusivity to soften; it does not improve the ability to capture polysemy. Similar observations hold for CGTM as well.

By contrast, GhLDA can capture the polysemy of "bank(s)." As is shown in Fig. 5, path [0-1-2-6] is related to the river and path [0-1-3-10] is related to finance. Although all uses of "bank" in the Wikipedia dataset were assigned to topic 1 because of the high frequency of the word, the meaning could be discerned from the path assignment. Moreover, "banks" in the dataset were assigned to the correct topic (i.e., either topic 3 or 6) in terms of the label of the documents (i.e., we utilized "Rivers" and "Banks/Financial" categories explained in the dataset section). Hence, we observe that GhLDA can distinguish "bank(s)" polysemy.

The inability to capture polysemy in GLDA is further illustrated using low-dimensional representations. Fig. 2 shows each word's assignment of topics 10 and 13 in addition to their two-dimensional representation using T-sne [31]. We can see that "bank(s)" is far apart from terms related to the river, and "bank(s)" is never assigned to the topic about rivers. As a comparison, Fig. 4 shows the two-dimensional representation of GhLDA. We can see that although "banks" is surrounded by terms that relate to finance, "banks," which is located in the upper right of the figure, is also assigned to a path that refers to rivers, showing that GhLDA can capture polysemy. Similar observations hold for other words, such as "law" and "order" (i.e., paths [0-1-5-15] and [0-1-5-21]), as well.

We further note that, because our corpus does not exclude highly frequent terms as in [3], LDA cannot capture polysemy well (i.e., Table. 3) because the topics are contaminated with stop words, and it is difficult to distinguish the difference between topics.

5.4 Comparison with hLDA

In this section, we mainly focus on the difference between GhLDA and hLDA. As shown in Fig. 3, the main difference can be seen in the hierarchical structure learned between the two models. Whereas the numbers of paths and topics estimated in hLDA are 54 and 83, in GhLDA, they are 10 and 16, respectively, on the Wikipedia datasets, which shows that hLDA tends to have a higher number of paths and topics than GhLDA. Both GhLDA and

Table 4: Topic coherence

Corpus	Wiki	Amazon	Reuters
LDA	-3.32	-1.94	-3.58
hLDA	-1.05	-1.50	-1.55
GLDA-GloVe	-1.13	-1.65	**-1.17**
GLDA-word2vec	-1.75	-1.92	-1.80
GLDA-fasttext	-1.96	-1.88	-2.07
CGTM-GloVe	-0.93	-1.34	-1.41
CGTM-word2vec	-1.63	-1.76	-1.85
CGTM-fasttext	-1.87	-1.77	-1.94
GhLDA-GloVe	-0.79	-1.54	-1.53
GhLDA-word2vec	**-0.60**	-1.66	-1.54
GhLDA-fasttext	-1.06	**-1.23**	-2.16

hLDA have paths for finance (e.g., [0-1-3-10],[0-1-4-46],[0-1-45-65]) and the river (e.g., [0-1-2-6]), thus capturing the polysemy of words (i.e., Table. 1). However, too many paths in hLDA cause crucial redundancy. For instance, there are seven paths related to finance that sometimes have no apparent distinction between them (e.g., [0,1,45,65] and [0,1,5,51]). These redundancy hearts the coherency of topics, as we show in the next section.

5.5 Topic Coherence

We calculated the topic coherence score [21, 10] using Palmetto [29] to check how coherently each model generates topics. We computed the average topic coherence score using the basic pointwise mutual information (PMI) measure, focusing on the top 10 words. Table 4 summarizes the results. First, we see that compared to LDA and hLDA models, using word embedding tends to outperform the no word embedding counterparts. Among the models that use word embeddings, GhLDA was the best model, except on the Reuters dataset.

However, the topics learned from the Reuters dataset using GhLDA were not at all worse than the GLDA and CGTM counterpart. For instance, in GhLDA-word2vec, there were topics such as "trump, republican, coal, party, workers, house, debate, school, bill, bankruptcy," which indicate the news topic that Trump made a promise to coal miners during his campaign, and "vehicles, water, vw, flint, safety, cars, emissions, volkswagen, detroit, filed,"

which indicate the news topics of Volkswagen's diesel cars and the tap water problem of Flint. Although these news topics were widely reported during the period in which the news dataset was collected, the PMIs of the topics were -1.36 and -2.79, respectively, which shows the limitations of Palmetto for evaluating new combinations of words correctly.

Furthermore, even though they connect to real word news, neither Trump nor Volkswagen appeared in the top 15 words of the 40 topics learned from GLDA-word2vec and CGTM-word2vec. Topics in GLDA were much general, such as "rate, dollar, assets, buy, goal, drop" and "government, end, federal, chinese, countries," which do not take into the word co-occurrence patterns of the corpus that we wish to analyze. As the Reuters examples suggest, even when the underlying word embedding is not in line with the corpus, the added flexibility of our model identifies critical topics that both GLDA and CGTM fail to identify. This observation further highlights the benefit of our model.

5.6 Quantitative Comparison

We further used the predictive held-out likelihood to quantitatively compare our models, as in [3]. We evaluated the probability of the held-out dataset using the 1,000 test documents described in the dataset section. [3] used the harmonic mean [16] to evaluate the held-out likelihood. However, [32, 9] showed that the harmonic mean method is biased. Hence, we used the left-to-right sequential sampler [9], which estimates the quantity:

$$p(w_{d,1:N_d}|\alpha,\theta) = \prod_{n \leq N_d} \Sigma_{k_n} \theta_{k_n,w_{d,n}} p(k_n|k_{1:n-1},\alpha,\theta), \tag{5}$$

and to make a fair comparison of the models, we evaluated θ_{k_n,j_n} for all the models using the topic assignments derived from each model and assessed the likelihood. Table 5 summarizes the results. First, models that use word embedding exhibited better results. Second, we see that CGTM beat GLDA significantly because of the additional correlation structure. Even without using word embedding, hLDA further beat the other two models by a large margin. However, GhLDA seemed to be the best model, outperforming all the other models.

Table 5: Predictive held-out likelihood

Corpus	Wiki	Amazon	Reuters
LDA	-1087.6	-1218.9	-1926.0
hLDA	-838.9	-1073.0	-1676.8
GLDA-glove	-1016.9	-1105.5	-1757.3
GLDA-word2vec	-1019.4	-1105.4	-1758.4
GLDA-fasttext	-1021.6	-1105.4	-1452.9
CGTM-glove	-946.7	-1046.8	-1667.6
CGTM-word2vec	-948.1	-1029.8	-1667.1
CGTM-fasttext	-943.3	-1049.3	-1665.7
GhLDA-glove	**-558.7**	**-659.1**	-1127.0
GhLDA-word2vec	-577.9	-664.7	**-1078.0**
GhLDA-fasttext	-578.9	-660.2	-1079.4

6 Conclusion

In this paper, we proposed Gaussian hLDA, which significantly improves the capture of polysemy compared with GLDA and CGTM. Our model learns the underlying topic distribution and hierarchical structure among topics simultaneously, which can be further used to understand the correlation among topics. The added flexibility of our model does not necessarily increase the time complexity compared with GLDA and CGTM, which makes our model a good competitor to GLDA. We demonstrated the validity of our approach using three real-world datasets.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC07/ASWC07, page 722735, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 537–542, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), February 2010.

[4] David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS03, page 1724, Cambridge, MA, USA, 2003. MIT Press.

[5] David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS03, page 1724, Cambridge, MA, USA, 2003. MIT Press.

[6] David M. Blei and John D. Lafferty. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS05, page 147154, Cambridge, MA, USA, 2005. MIT Press.

[7] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML 06, page 113120, New York, NY, USA, 2006. Association for Computing Machinery.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[9] Wray L. Buntine. Estimating likelihoods for topic models. In Zhi-Hua Zhou and Takashi Washio, editors, *ACML*, volume 5828 of *Lecture Notes in Computer Science*, pages 51–64. Springer, 2009.

[10] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic

models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009.

[11] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *ACL (1)*, pages 795–804. The Association for Computer Linguistics, 2015.

[12] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model, 2019.

[13] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces, 2019.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[15] Weihua Hu and Jun'ichi Tsujii. A latent concept topic model for robust topic inference using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 380–386, Berlin, Germany, August 2016. Association for Computational Linguistics.

[16] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[17] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[18] Enes Makalic and Daniel Schmidt. High-dimensional bayesian regularised regression with the bayesreg package. arXiv:1611.06649, 2016.

[19] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52. ACM, 2015.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.

[21] David Newman, Jey Han Lau, Karl Grieser, and Timothy 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*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[22] 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.

[23] N. Obot, L. OMalley, I. Nwogu, Q. Yu, W. S. Shi, and X. Guo. From novice to expert narratives of dermatological disease. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 131–136, March 2018.

[24] James O'Neill, Cécile Robin, Leona O'Brien, and Paul Buitelaar. An analysis of topic modelling for legislative texts. In *ASAIL@ICAIL*, 2016.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[26] James Petterson, Wray Buntine, Shravan M. Narayanamurthy, Tibério S. Caetano, and Alex J. Smola. Word features for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1921–1929. Curran Associates, Inc., 2010.

[27] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[28] Nicholas Polson, James Scott, and Jesse Windle. Bayesian inference for logistic models using polya-gamma latent variables. *Journal of the American Statistical Association*, 108, 05 2012.

[29] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM*

International Conference on Web Search and Data Mining, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM.

[30] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *In Latent Semantic Analysis: A Road to Meaning., Editors Landauer, T. and McNamara, D. and Dennis, S. and Kintsch, W.*, 2006.

[31] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[32] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 09, page 11051112, New York, NY, USA, 2009. Association for Computing Machinery.

[33] Senhao Wen, Zhiyuan Zhao, and Hanbing Yan. Detecting malicious websites in depth through analyzing topics and web-pages. In *Proceedings of the 2nd International Conference on Cryptography, Security and Privacy*, ICCSP 2018, page 128133, New York, NY, USA, 2018. Association for Computing Machinery.

[34] Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled wasserstein learning for word embedding and topic modeling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1716–1725. Curran Associates, Inc., 2018.

[35] Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. A correlated topic model using word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI17, page 42074213. AAAI Press, 2017.