A Semantic Embedding Enhanced Topic Model for User-Generated Textual Content Modeling in Social Ecosystems

Peng Zhang*[1], Baoxi Liu*[1], Tun Lu†[1], Hansu Gu[2], Xianghua Ding[1,3] and Ning Gu[1]

[1]*School of Computer Science & Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China*
[2]*Seattle, WA, USA*
[3]*University of Glasgow, Glasgow, United Kingdom*

The development of ICT (Information and Communication Technologies) and Web 2.0 promotes the emergence of diverse social ecosystems like social IoT (Internet of Things), social media and online communities. User-generated textual content (UGTC), which consists of unstructured texts, is the most important and common type of user-generated content in social ecosystems. UGTC in social ecosystems is generated according to two types of context information - global context (topics) and local context (semantic regularities). For UGTC modeling, topic models just consider global context but ignore semantic regularities, while semantic embedding models are on the opposite. So only utilizing topic models or semantic embedding models to model UGTC suffers from some drawbacks. For this problem, we propose a semantic embedding enhanced topic model named SEE-Twitter-LDA for accurately modeling UGTC in social ecosystems. The core of SEE-Twitter-LDA is that words are generated according to mutual semantic information of topics and semantic regularities. So global context and local context are jointly considered for UGTC modeling. By utilizing 553,098 tweets sampled from Twitter and 211,233 posts sampled from Weibo, we validate SEE-Twitter-LDA's better performance on perplexity, topic divergence and topic coherence versus existing related models.

Keywords: Social Ecosystems; User-generated Textual Content; Topic Model; Semantic Embedding; Twitter; Weibo

Received 00 January 2009; revised 00 Month 2009

1. INTRODUCTION

The development of ICT (Information and Communication Technologies) and Web 2.0 promotes the emergence of diverse social ecosystems like social IoT (Internet of Things)[1, 2], social media and online communities. These social ecosystems generate a large amount of user-generated content like texts, photos and videos. User-generated textual content (UGTC), which consists of unstructured texts, is the most important and common type of user-generated content. Considering Weibo as an example, users share a large amount of textual posts related to news, opinions or general interests with others on that. The number of words published by users everyday is over 130 millions [§]. UGTC of social ecosystems is characterized with large size, good instantaneity and high publicity, which provides a valuable corpus for mining users' interests, preferences, sentiments and opinions. Thus, in recent years, lots of research collects social posts from social ecosystems to make studies on user demographic inference [3, 4], human activity prediction [5, 6, 7], sentiment analysis [8], personalized recommendation [9], etc.

The basis of user-generated textual content mining is UGTC modeling. Existing UGTC modeling methods can be classified into two categories: sparse representation models and dense representation models. Sparse representation models represent each post as

*These authors contributed equally to this work.
†Corresponding author.

[§]https://data.weibo.com/report/reportDetail?id=433&display=0&retcode=6102

a bag of words. Term Frequency-Inverse Document Frequency (TF-IDF) [10] and one-hot encoding [11] are two common-used sparse representation methods but with different representation strategies. In the TF-IDF model, each word in a social post can be represented as a weight that is calculated by the product of TF and IDF, while one-hot encoding collects all words in the corpus into a dictionary and describes each word as a high-dimensional vector in which its corresponding position is set as 1, and the other positions are set to 0. Sparse representation models have the advantage of easy to implement. But such models generally suffer from problems like dimensionality curse and synonym confusion, which limits their use in UGTC modeling and some other NLP tasks. For these problems, dense representation models project words, sentences or documents into latent topic or semantic spaces. It can not only eliminate dimensionality curse but also obtain the relevance among words in terms of topics or semantics. Nowadays the prevalent dense representation approaches are topic models and semantic embedding models [12, 13, 14]. Topic models discover latent topics from corpus based on word co-occurrence and describe each document with a probability distribution over the discovered topics. As UGTC in social ecosystems is a special kind of short text, many nowadays topic models designed for short text modeling like BTM (Biterm Topic Model) [15] and DMM (Dirichlet Multinomial Mixture model) [16] can be utilized for UGTC modeling. Such models are based on the Latent Dirichlet Allocation (LDA) [17] model and adopt new strategies to handle the data sparsity problem existing in short text modeling. However, compared with common short texts like Web page titles, short news and text advertisements, UGTC is associated more closely with users, while these models ignore the role of users in topic modeling. In social ecosystems, users essentially have topic preference and language styles [18]. Incorporating the concept "users" into topic models and associating that with topics and words can characterize UGTC more accurately. For this problem, the Twitter-LDA model which considers the role of users is proposed for modeling social posts of Twitter [19]. It has been proved more efficient for social post modeling and utilized for many different tasks like user behavior and emotion mining [20, 21], user interest inference and personalized recommendation [22, 23], etc. Different from topic models, semantic embedding models conduct projection based on semantic regularities. For example, in one of the most common-used word embedding methods - Word2Vec [24], semantic embedding is performed by predicting the occurrence probability of words given their window-based context or predicting context words based on a given word, through which each word can be represented as a low-dimension vector that is highly correlated with the real semantics.

UGTC in social ecosystems is generated according to two types of context information - global context and local context. Global context carries topical information, while local context reflects semantic regularities. Previous research has suggested that users have platform preference in terms of topics [20, 25], which implies when publishing a post, users essentially have one or some topics in mind, and then choose the most appropriate platform. For example, when sharing a life event, a user might first choose Facebook as the expected platform as it has been found to be a more common-used medium for sharing personal lives. After then, the user writes the post according to the topics of the event and semantics. If the current word is "European", the next word may be "football", "airline" or "movie" by considering semantics. Jointly considering topics, if the post focuses on the sport topic, the next word would be more likely to be "football" versus "airline" and "movie". Such user-generated textual content driven by topics and semantics are difficult to model by nowadays topic models or semantic embedding models independently. Topic models are generative models, and the extracted content of each document is quite interpretable by humans. However, the basic assumption of topic models is words occur independently, and each document is treated as bag-of-words. Thus, semantic regularities are not reflected in such models, which is not consistent with real content generation procedure. On the contrary, semantic embedding models like Word2Vec consider semantics but ignore global context of a document, and the results generated by such models are generally difficult to interpret by humans. Thus only utilizing topic models or semantic embedding models to model UGTC suffers from some drawbacks.

Incorporating semantic embedding into topic models to build semantic embedding enhanced topic models can combine the topic representation capability and good interpretability of topic models with semantic representation capability of semantic embedding models, which is an effective approach to solve the aforementioned drawbacks and has been a fresh research topic in natural language processing areas. Many studies [26, 27, 28, 29, 30, 31, 32] attempt to improve LDA by incorporating semantic embedding trained by Word2Vec or matrix factorization into the generation procedure of words, and some research [33, 34, 35, 36, 37, 38] also focuses on designing semantic embedding enhanced topic models for modeling short texts. However, as mentioned above, UGTC in social ecosystems is characterized with stronger association with users and more noisy words compared with other short texts, which results that the existing semantic embedding enhanced topic models cannot obtain good performance on that. Thus building semantic embedding enhanced topic models for accurately modeling UGTC in social ecosystems deserves deep exploration and study. For this problem, in this paper, we incorporate semantic embedding into Twitter-LDA and propose a Semantic Embedding

Enhanced Twitter-LDA (SEE-Twitter-LDA) model for UGTC modeling in social ecosystems. The major characteristics of SEE-Twitter-LDA lie in: 1) Words are generated according to mutual semantic information of topics and semantics, through which global context and local context are jointly considered when modeling user-generated textual content; 2) Each user corresponds to a certain topic distribution, and each tweet is associated with one topic to alleviate the sparsity problem; 3) A distribution representing background words is reserved to address the noisy nature of UGTC. To validate the performance of our proposed model, we collected 211,233 Weibo posts and 553,098 Twitter tweets as two data sets and made experiments by setting existing common-used short text modeling methods and their corresponding semantic embedding enhanced models as baselines. The results show that our proposed model is superior to the baselines in terms of perplexity, topic divergence and topic coherence. To be specific, our main contributions are:

- A new topic model SEE-Twitter-LDA, which incorporates semantic embedding into Twitter-LDA, is proposed for modeling UGTC in social ecosystems. To the best of our knowledge, this is the first work on semantic embedding enhanced topic model building for UGTC.
- We present a strategy to characterize mutual semantic information of topics and semantic regularities, which provides insights for integrating global context with local context in text modeling.
- We validate SEE-Twitter-LDA's better performance versus related models utilizing two data sets sampled from Weibo and Twitter.

The rest of this paper is organized as below. In Section 2, we review related research on topic models, semantic embedding and semantic embedding enhanced topic models. Our proposed model is shown in Section 3, and the model inference procedure is given in Section 4. In Section 5, we show the experimental results. Finally, conclusion is given in Section 6.

2. RELATED WORK

2.1. Topic Models

Most of existing topic models derive from Latent Dirichlet Allocation (LDA) [17]. LDA assumes that each document is modeled as mixtures of topics, and each topic is associated with words. A document is generated by choosing a topic based on document-topic distribution and choosing a word according to topic-word distribution iteratively. As LDA's advantages of simplified parameter tuning, overfitting avoidance and good interpretability, it has been widely used for handling different tasks like text categorization [39], sentiment analysis [40], recommendation [41], etc.

LDA reveals latent topics by capturing the document-level word co-occurrence patterns. While in reality, there are large amounts of short texts that cannot supply sufficient co-occurrence patterns like Web page titles, short news, text advertisements, image captions as well as social posts studied in this research. So directly applying LDA on such short texts generally suffers from severe data sparsity problem. For this problem, lots of research focuses on methods for short text modeling. We summarize these studies into three categories: short text aggregation, Biterm Topic Model (BTM) and models assuming each document is mainly associated with one topic. The details of these models are given below.

First, aggregating short texts into lengthy pseudo-documents is a simple strategy to handle short texts. By taking advantage of such a strategy, conventional topic models like LDA and ATM (Author Topic Model) can be utilized to process the corpus. According to [42], there are several short text pooling schemes like author-wise pooling (aggregating texts corresponding to the same author into a document), burst-score wise pooling (aggregating texts according to burst scores), temporal pooling (e.g. aggregating texts posted within the same hour into a document) and hashtag-based pooling (aggregating texts corresponding to the same hashtag into a document). For example, [43] aggregates tweets published by the same user into a pseudo-document and then performs topic analysis utilizing the standard LDA model, [44] proposes several schemes like author-wise pooling and entity pooling (aggregating texts that contain a specific term into a document) to train a standard topic model and compares their quality, etc.

Second, Biterm Topic Model (BTM) is based on a novel scheme to model short texts. In traditional topic models, word co-occurrence patterns are extracted implicitly from the document level, which limits their use in short text corpus. Different with that, BTM directly models the word co-occurrence patterns based on observed biterms (word-word co-occurrence patterns) in texts, which can alleviate the document-level word sparsity [15]. In recent years, some research attempts to improve BTM for different short text mining tasks. [45] proposes a Bursty Biterm Topic Model (BBTM) by incorporating the burstiness of biterms as prior knowledge to mine bursty topics from microblogs, [46] builds a novel bilingual topic model named Bilingual Biterm Topic Model (BiBTM) for cross-lingual taxonomy alignment, [47] proposes FastBTM to reduce the sampling time of BTM, etc.

Third, in traditional topic models, a document is modeled as mixtures of topics, while for short texts, it is reasonable to assume that each document is associated with just one topic. Based on this idea, some models assuming each document corresponds to one topic have been proposed for short text modeling. Dirichlet Multinomial Mixture (DMM) is such a model that obtains a lot of attention from researchers [16]. In [48], an improved model named GSDMM which focuses

on Gibbs Sampling algorithm of DMM is proposed for short text clustering. The authors in [49] propose an adaptive Dirichlet Multinomial Mixture model that considers time slices, and a collapsed Gibbs sampling algorithm named e-GSDMM algorithm is also designed for model inference. In [50], the authors thought close short texts should have similar variational topic representations and improve DMM by spreading topics among neighboring documents.

As mentioned above, user-generated textual content in social ecosystems is associated more closely with users compared with common short texts like Web page titles, short news and text advertisements. Incorporating the concept "users" into topic models and associating that with topics and words can represent user-generated textual content more accurately. Based on this idea, the Twitter-LDA model is proposed for tweet modeling [19]. The major characteristics of Twitter-LDA lie in: 1) It considers users' topic preference by assuming each user corresponds to a certain topic distribution; 2) Similar to DMM, each tweet is only associated with one topic to alleviate the sparsity problem; 3) It introduces the concept of background words to address the noisy nature of tweets. As its better performance for user-generated textual content modeling, Twitter-LDA has been utilized for processing diverse tasks. [20, 21] utilize Twitter-LDA to mine users' characteristics of behavior and emotion in social media platforms like Twitter and Instagram. [22, 23] propose novel recommendation methods by utilizing Twitter-LDA to extract users' interested topics from tweets. [51] applies Twitter-LDA to extract topics from tweets and proposes a context-sensitive topical PageRank method for keyphrase extraction. [52] proposes a method to identify the right audience from the massive amount of social media users based on tweet analysis using Twitter-LDA, and [53] utilizes Twitter-LDA to obtain topic words to help chatbots generate informative and interesting responses. Besides the above research, few studies aim to improve Twitter-LDA to solve difficult tasks. For example, for users' dynamic interests and dynamic topics, [54] proposes an improved model that considers the time sequence of tweets and has the capability of online inference, and [55] presents the Multi-Faceted Topic Model (MfTM) to jointly capture the temporal characteristics of each topic and model latent semantics among terms and entities.

2.2. Semantic Embedding

With the developments of neural networks and deep learning [56, 57], semantic embedding has been a research focus in recent years. The core of semantic embedding is to project words into a low-dimension semantic space, wherein items that are associated in semantics should be adjacent. Nowadays common-used semantic embedding methods for word embedding can be categorized into two types: probabilistic models and matrix factorization models [58].

Probabilistic models perform semantic embedding by predicting the occurrence probability of words given their window-based context or predicting context words based on a given word. As the rise of deep learning, neural networks are adopted by a few studies to learn the occurrence probabilities [59]. Word2Vec, which was launched by Google in 2013, is the most famous model among neural network probabilistic models [24]. It learns semantics by treating each sentence in corpus as a training sample. Word2Vec contains two models: CBOW (Continuous Bag Of Words) and skip-gram. The former predicts each word by using contexts of its surroundings, while the latter uses each word to predict its neighboring words. CBOW is generally faster than skip-gram, while skip-gram performs better to process uncommon words.

Different from probabilistic models, matrix factorization models generate a low-dimension semantic space by matrix factorization. For example, in a recent matrix factorization model - GloVe (Global Vectors) [60], a large word-context matrix R is first initialized, wherein each element represents the co-occurrence of each pair of two words. After then, stochastic gradient descent (SGD) algorithm is utilized to transform R into product of two low-dimension matrices: word-feature matrix P and feature-context matrix Q. These two matrices are finally utilized for word representation and document representation. Besides SGD, many other methods like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) can be utilized for matrix factorization.

2.3. Semantic Embedding Enhanced Topic Models

Incorporating semantic embedding into topic models to construct semantic embedding enhanced topic models has been a research focus in recent text modeling studies. Most of these studies leverage probabilistic models especially Word2Vec as semantic embedding methods, while little research is based on matrix factorization models.

From the perspective of conventional topic models, many studies [26, 27, 28, 29, 30] propose semantic embedding enhanced LDA models based on probabilistic models, and some research [31, 32] improves LDA based on matrix factorization models. For example, [27, 29] propose WS-TSWE (Weakly Supervised Topic Sentiment joint model with Word Embeddings) and CGTM (Correlated Gaussian Topic Model) by using Word2Vec as word embedding methods, [28, 30] incorporate word embedding into LDA by modeling words or terms as Gaussian distribution or Mises-Fisher distribution over the embedding space, [31] obtains word embeddings by PCA and builds an enhanced model named WELDA (Word Embeddings with Latent Dirichlet Allocation),

etc. In [61], the authors propose EETM (Embedding Enhanced Topic Model) and utilize multiple methods including CBOW, Skip-gram and GloVe to train word embedding vectors. Experiments suggest that the two Word2Vec models obtain better performance versus GloVe on most settings.

From the perspective of short text topic models, some semantic embedding enhanced models are proposed based on the BTM and DMM. For the problem that BTM ignores the inside relationship between words, some improved models like RIBSTM (RNN-IDF based Biterm Short-text Topic Model), NPMM (NonParaMetric Model) and SeedBTM are proposed [33, 34, 35, 62]. RIBSTM utilizes RNN for relationship learning among words, while NPMM and SeedBTM are based on GloVe. [26, 36, 37, 38] are based on DMM to build semantic embedding enhanced short text topic models. [36, 37] present two enhanced DMM models named GPU-DMM (Generalized Polya Urn model enhanced DMM) and GPU-PDMM (Generalized Polya Urn model enhanced Poisson-based DMM) which learn word embeddings by a probabilistic model named Polya urn model, while in [26, 38], word embeddings are obtained by Word2Vec. Besides the above research, [63, 64] propose semantic embedding enhanced models including KGNMF (Knowledge-Guided Non-negative Matrix Factorization) and SeaNMF (Semantics-assisted Non-negative Matrix Factorization) for short text topic mining using non-negative matrix factorization as the semantic embedding models, [65, 66] focused on the neural topic model - Variational Auto-Encoder Topic Model and proposed the corresponding semantic embedded enhanced models, etc.

Above all, many studies have paid attention to conventional topic models and short text topic models, and some research also improves them by introducing semantic embedding. However, a semantic embedding enhanced topic model for UGTC modeling in social ecosystems has not been studied, which motivates us to conduct the research in this paper.

3. SEE-TWITTER-LDA MODEL

SEE-Twitter-LDA is a semantic embedding enhanced topic model which combines the topic representation capability and good interpretability of topic models with semantic representation capability of semantic embedding models. The graphical model for SEE-Twitter-LDA using plate notation is described in Figure 1, and the description of its symbols is listed in Table 1 .

As is shown in Figure 1, the core of our model is a word $W_{u,s,n}$ (n >1) is generated based on ϕ_t or ϕ_B and the prior word $W_{u,s,n-1}$. User-generated textual content in social ecosystems generally contains many noisy items like semantic auxiliary words. Such words appear frequently, which brings bias to topic analysis. In order to solve this problem, in the Twitter-LDA model, a controller $Y_{u,s,n}$ is introduced to represent

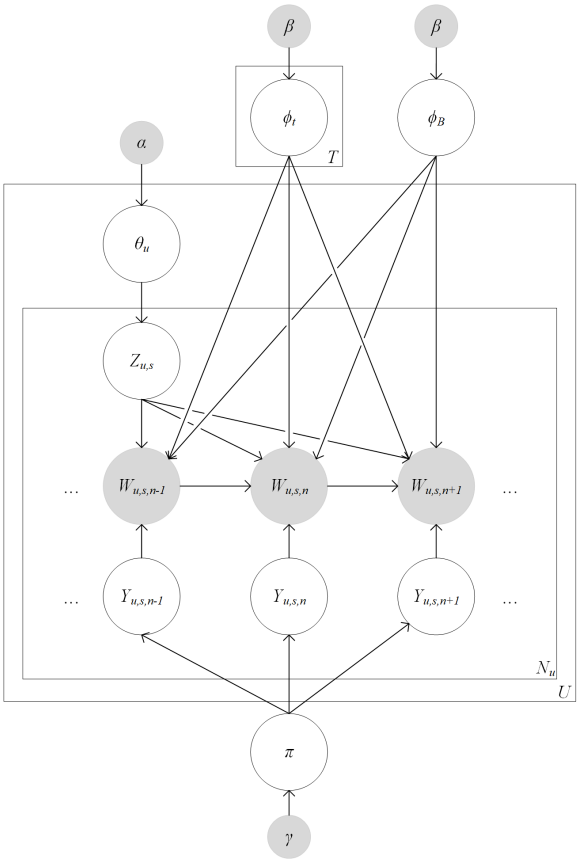


FIGURE 1. Graphical model of SEE-Twitter-LDA using plate notation.

whether $W_{u,s,n}$ is a topic word or background word (noisy word). We follow such a strategy in our model. When $Y_{u,s,n} = 1$, $W_{u,s,n}$ is generated based on ϕ_B and $W_{u,s,n-1}$, while if $Y_{u,s,n} = 0$, $W_{u,s,n}$ will be generated based on ϕ_t and $W_{u,s,n-1}$. So given the parameters α, β and γ for $Z_{u,s} = t$ ($t \in \{1, 2, \cdots, T\}$), the controller $Y_{u,s,n} = c$ ($c \in \{0, 1\}$), and the word $W_{u,s,n} = v$ ($v \in \{1, 2, \cdots, V\}$), if $Y_{u,s,n}$=0, the joint distribution is:

$$\begin{aligned} &p(Z_{u,s} = t, Y_{u,s,n} = 0, W_{u,s,n} = v|\alpha, \beta, \gamma) \\ &= p(Z_{u,s} = t|\theta_u, \alpha) \\ &\quad \cdot p(Y_{u,s,n} = 0|\pi, \gamma) \cdot p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_t, \beta), \end{aligned} \tag{1}$$

and if $Y_{u,s,n}$=1, the joint distribution is:

$$\begin{aligned} &p(Z_{u,s} = t, Y_{u,s,n} = 1, W_{u,s,n} = v|\alpha, \beta, \gamma) \\ &= p(Z_{u,s} = t|\theta_u, \alpha) \\ &\quad \cdot p(Y_{u,s,n} = 1|\pi, \gamma) \cdot p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta). \end{aligned} \tag{2}$$

In Eq. 1 and Eq. 2, $p(Z_{u,s} = t|\theta_u, \alpha)$, $p(Y_{u,s,n} = 0|\pi, \gamma)$ and $p(Y_{u,s,n} = 1|\pi, \gamma)$ are evaluated by:

$$\begin{aligned} p(Z_{u,s} = t|\theta_u, \alpha) &= \frac{n_t^u + \alpha_t}{\sum_{t=1}^{T}(n_t^u + \alpha_t)}, \\ p(Y_{u,s,n} = 0|\pi, \gamma) &= \frac{n_0 + \gamma_0}{n_0 + \gamma_0 + n_1 + \gamma_1}, \\ p(Y_{u,s,n} = 1|\pi, \gamma) &= \frac{n_1 + \gamma_1}{n_0 + \gamma_0 + n_1 + \gamma_1}, \end{aligned} \tag{3}$$

where n_t^u represents the number of times that the tth topic occurs in the uth user's posts, n_0 is the number

TABLE 1. Description of symbols in SEE-Twitter-LDA.

Symbol	Description
U	Number of social ecosystem users in corpus
T	Number of topics
V	Number of words in dictionary
N_u	Number of posts of the uth ($u \in \{1, 2, \cdots, U\}$) user
$N_{u,s}$	Number of words of the uth user's sth ($s \in \{1, 2, \cdots, N_u\}$) post
$W_{u,s,n}$	The nth ($n \in \{1, 2, \cdots, N_{u,s}\}$) word in the uth user's sth post
$Z_{u,s}$	The topic of the uth user's sth post
$Y_{u,s,n}$	Controller of word $W_{u,s,n}$
θ_u	Topic distribution of the uth user
ϕ_t	Word distribution of the tth ($t \in \{1, 2, \cdots, T\}$) topic
ϕ_B	Word distribution for background words
π	Bernoulli distribution that governs the choice between background words and topic words
α	Dirichlet prior hyperparameter for topics
β	Dirichlet prior hyperparameter for words
γ	Dirichlet prior hyperparameter for controllers

of times that topic words occur in corpus, and n_1 is the number of times that background words occur in corpus.

$p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_t, \beta)$ and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta)$ describe the generation of a word $W_{u,s,n}$ based on ϕ_t and $W_{u,s,n-1}$ or ϕ_B and $W_{u,s,n-1}$, which means $W_{u,s,n}$ is generated based on the mutual semantic information of topics and semantics. According to Standard Cross Entropy Loss [61, 67], we formulate $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_t, \beta)$ and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta)$ as the exponential form of binomial cross entropy loss. When $Y_{u,s,n}$=0, it can be evaluated as:

$$\begin{aligned} &p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_t, \beta) \\ &= p(W_{u,s,n} = v|\phi_t, \beta)^{p(W_{u,s,n}=v|W_{u,s,n-1}=v')} \\ &\quad \cdot (1 - p(W_{u,s,n} = v|\phi_t, \beta))^{(1-p(W_{u,s,n}=v|W_{u,s,n-1}=v'))}, \end{aligned} \tag{4}$$

where $p(W_{u,s,n} = v|W_{u,s,n-1} = v')$ is calculated using softmax function shown in Eq. 5, and $p(W_{u,s,n} = v|\phi_t, \beta)$ is formulated as Eq. 6.

$$p(W_{u,s,n} = v|W_{u,s,n-1} = v') = \frac{exp(\vec{W}_v \cdot \vec{W}_{v'})}{\sum_{v=1}^{V} exp(\vec{W}_v \cdot \vec{W}_{v'})}, \tag{5}$$

where $\vec{W}_v$ and $\vec{W}_{v'}$ are the word vectors generated by Word2Vec model corresponding to the vth word and the v'th word respectively in dictionary.

$$p(W_{u,s,n} = v|\phi_t, \beta) = \frac{n_v^t + \beta_v}{\sum_{v=1}^{V} (n_v^t + \beta_v)}, \tag{6}$$

where n_v^t represents the number of times that the vth word in dictionary occurs in the tth topic, and the definitions of other symbols are same as Table 1.

When $Y_{u,s,n}$=1, $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta)$ can be evaluated as:

$$\begin{aligned} &p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta) \\ &= p(W_{u,s,n} = v|\phi_B, \beta)^{p(W_{u,s,n}=v|W_{u,s,n-1}=v')} \\ &\quad \cdot (1 - p(W_{u,s,n} = v|\phi_B, \beta))^{(1-p(W_{u,s,n}=v|W_{u,s,n-1}=v'))}, \end{aligned} \tag{7}$$

where $p(W_{u,s,n} = v|W_{u,s,n-1} = v')$ is also obtained by Eq. 5, and $p(W_{u,s,n} = v|\phi_B, \beta)$ is formulated as:

$$p(W_{u,s,n} = v|\phi_B, \beta) = \frac{n_v^B + \beta_B}{\sum_{v=1}^{V} (n_v^B + \beta_v)}, \tag{8}$$

where n_v^B is the number of times that the vth word in dictionary occurs in background words, and the meanings of the other symbols are same as earlier definitions.

Algorithm 1 Generation scheme of the SEE-Twitter-LDA model

Input: α, β, γ, U, T, V, N_u and $N_{u,s}$
Output: A collection of user-generated textual content in social ecosystems

Draw ϕ_B from Dir(β) and π from Dir(γ)
for each $t \in \{1, 2, \cdots, T\}$ **do**
    Draw topic-word distribution ϕ_t from $Dir(\beta)$
end for
for each $u \in \{1, 2, \cdots, U\}$ **do**
    Draw user-topic distribution θ_u from $Dir(\alpha)$
    for each $s \in \{1, 2, \cdots, N_u\}$ **do**
        Draw topic $Z_{u,s}$ from $Multi(\theta_u)$
        for each $n \in \{1, 2, \cdots, N_{u,s}\}$ **do**
            Draw word controller $Y_{u,s,n}$ from $Multi(\pi)$
            if $Y_{u,s,n}$=0 **then**
                if n=1 **then**
                    Draw word $W_{u,s,n}$ based on $p(W_{u,s,n} = v|\phi_t, \beta)$
                end if
                if n >1 **then**
                    Draw word $W_{u,s,n}$ based on $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_t, \beta)$
                end if
            end if
            if $Y_{u,s,n}$=1 **then**
                if n=1 **then**
                    Draw word $W_{u,s,n}$ based on $p(W_{u,s,n} = v|\phi_B, \beta)$
                end if
                if n >1 **then**
                    Draw word $W_{u,s,n}$ based on $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta)$
                end if
            end if
        end for
    end for
end for

Based on the joint distribution, each post in social ecosystems can be generated according to the generation scheme exhibited in Algorithm 1. The procedure is initialized by drawing a word distribution for each of the T topics and a word distribution for background words. After then, for each user, the

corresponding topic distribution θ_u is sampled, based on which draw a topic $Z_{u,s}$ for the sth post to be generated. To obtain each word of this post, a controller $Y_{u,s,n}$ is sampled to represent whether the word is a topic word or background word. If it is a topic word, it will be sampled based on $p(W_{u,s,n} = v|\phi_t, \beta)$ (for the first word) and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_t, \beta)$ (for the remainders), while if the word is a background word, it will be sampled based on $p(W_{u,s,n} = v|\phi_B, \beta)$ (for the first word) and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_B, \beta)$ (for the remainders).

4. MODEL INFERENCE

Gibbs Sampling is a prevalent method for topic model inference. Thus we utilize Gibbs Sampling to infer π, ϕ_t, ϕ_B and θ_u. The procedure is listed below.

- Initialization: Randomly assign a topic number for each post, and assign a controller for each word.
- Review the corpus, sample a topic number for each post according to topic update formula, and sample a controller number for each word according to controller update formula.
- Iterate sampling until reaching convergence.
- Calculate the co-occurrence matrixes π, ϕ_t, ϕ_B and θ_u.

In the above procedure, the topic update formula and controller update formula are the key points. Considering the aforementioned joint distribution, for the uth user's sth post, the topic update formula is exhibited as Eq. 9.

$$\begin{aligned} &p(Z_{u,s} = t|Z_{\neg(u,s)}, W, Y) \\ &\propto p(Z_{u,s} = t, W_{u,s}|Z_{\neg(u,s)}, W_{\neg(u,s)}, Y_{\neg(u,s)}) \\ &= p(Z_{u,s} = t|\theta_{u,\neg(u,s)}, \alpha) p(W_{u,s,1} = v''|\phi_{t,\neg(u,s)}, \beta) \\ &\quad \cdot \prod_{n=2}^{N_{u,s}} p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{t,\neg(u,s)}, \beta). \end{aligned} \tag{9}$$

In Eq. 9, $\neg(u,s)$ means ignoring the current post (the uth user's sth post), and $p(Z_{u,s} = t|\theta_{u,\neg(u,s)}, \alpha)$, $p(W_{u,s,1} = v''|\phi_{t,\neg(u,s)}, \beta)$ and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{t,\neg(u,s)}, \beta)$ are formulated as Eq. 10, Eq.11 and Eq.12 respectively.

$$p(Z_{u,s} = t|\theta_{u,\neg(u,s)}, \alpha) = \frac{n^u_{t,\neg(u,s)} + \alpha_t}{\sum_{t=1}^{T} (n^u_{t,\neg(u,s)} + \alpha_t)}, \tag{10}$$

where $n^u_{t,\neg(u,s)}$ is the number of times that the tth topic occurs in the uth user's posts ignoring the sth post.

$$p(W_{u,s,1} = v''|\phi_{t,\neg(u,s)}, \beta) = \frac{n^t_{v'',\neg(u,s)} + \beta_v}{\sum_{v=1}^{V} (n^t_{v,\neg(u,s)} + \beta_v)}, \tag{11}$$

where $n^t_{v'',\neg(u,s)}$ means the number of times that the v''th word in dictionary occurs in the tth topic without considering the sth post.

$$\begin{cases} p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{t,\neg(u,s)}, \beta) \\ \quad = p(W_{u,s,n} = v|\phi_{t,\neg(u,s)}, \beta)^{p(W_{u,s,n}=v|W_{u,s,n-1}=v')} \\ \quad \cdot (1 - p(W_{u,s,n} = v|\phi_{t,\neg(u,s)}, \beta))^{(1-p(W_{u,s,n}=v|W_{u,s,n-1}=v'))} \\ p(W_{u,s,n} = v|\phi_{t,\neg(u,s)}, \beta) = \frac{n^t_{v,\neg(u,s)} + \beta_v}{\sum_{v=1}^{V} (n^t_{v,\neg(u,s)} + \beta_v)} \\ p(W_{u,s,n} = v|W_{u,s,n-1} = v') = \frac{exp(\vec{W}_v \cdot \vec{W}_{v'})}{\sum_{v=1}^{V} exp(\vec{W}_v \cdot \vec{W}_{v'})} \end{cases}, \tag{12}$$

where $n^t_{v,\neg(u,s)}$ is the number of times that the vth word in dictionary occurs in the tth topic ignoring the sth post, and the meanings of the other symbols are same as aforementioned definitions.

Based on the results of topic sampling, a controller (0 or 1) is sampled for each word $W_{u,s,n}$ according to Eq. 13 and Eq. 16.

$$\begin{aligned} &p(Y_{u,s,n} = 0|Y_{\neg(u,s,n)}, Z, W) \\ &\propto p(Z_{u,s} = t, Y_{u,s,n} = 0, W_{u,s,n} = v|Z_{\neg(u,s)}, Y_{\neg(u,s,n)}, W_{\neg(u,s,n)}) \\ &\propto \begin{cases} p(Y_{u,s,n} = 0|\pi_{\neg(u,s,n)}, \gamma) \\ \quad \cdot p(W_{u,s,n} = v|\phi_{t,\neg(u,s,n)}, \beta), \text{ if } n = 1 \\ p(Y_{u,s,n} = 0|\pi_{\neg(u,s,n)}, \gamma) \\ \quad \cdot p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{t,\neg(u,s,n)}, \beta), \text{ OTW} \end{cases} \end{aligned}, \tag{13}$$

where $\neg(u,s,n)$ means ignoring the current word $W_{u,s,n}$, $p(Y_{u,s,n} = 0|\pi_{\neg(u,s,n)}, \gamma)$ is formulated as Eq. 14, and $p(W_{u,s,n} = v|\phi_{t,\neg(u,s,n)}, \beta)$ and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{t,\neg(u,s,n)}, \beta)$ are formulated in Eq. 15 according to the aforementioned joint distribution.

$$\begin{aligned} &p(Y_{u,s,n} = 0|\pi_{\neg(u,s,n)}, \gamma) \\ &= \frac{n_{0,\neg(u,s,n)} + \gamma_0}{n_{0,\neg(u,s,n)} + \gamma_0 + n_{1,\neg(u,s,n)} + \gamma_1}. \end{aligned} \tag{14}$$

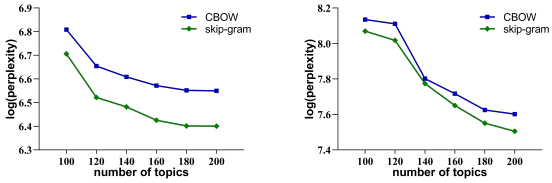
$$\begin{cases} p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{t,\neg(u,s,n)}, \beta) \\ \quad = p(W_{u,s,n} = v|\phi_{t,\neg(u,s,n)}, \beta)^{p(W_{u,s,n}=v|W_{u,s,n-1}=v')} \\ \quad \cdot (1 - p(W_{u,s,n} = v|\phi_{t,\neg(u,s,n)}, \beta))^{(1-p(W_{u,s,n}=v|W_{u,s,n-1}=v'))} \\ p(W_{u,s,n} = v|\phi_{t,\neg(u,s,n)}, \beta) = \frac{n^t_{v,\neg(u,s,n)} + \beta_v}{\sum_{v=1}^{V} (n^t_{v,\neg(u,s,n)} + \beta_v)} \\ p(W_{u,s,n} = v|W_{u,s,n-1} = v') = \frac{exp(\vec{W}_v \cdot \vec{W}_{v'})}{\sum_{v=1}^{V} exp(\vec{W}_v \cdot \vec{W}_{v'})} \end{cases} \tag{15}$$

In Eq. 14 and Eq. 15, $n_{0,\neg(u,s,n)}$ is the number of times that topic words occur in corpus ignoring $W_{u,s,n}$, $n_{1,\neg(u,s,n)}$ is the number of times that background words occur in corpus ignoring $W_{u,s,n}$, $n^t_{v,\neg(u,s,n)}$ means the number of times that the vth word occurs in the tth topic ignoring $W_{u,s,n}$, and the other symbols are with aforementioned meanings.

$$\begin{aligned} &p(Y_{u,s,n} = 1|Y_{\neg(u,s,n)}, Z, W) \\ &\propto p(Z_{u,s} = t, Y_{u,s,n} = 1, W_{u,s,n} = v|Z_{\neg(u,s)}, Y_{\neg(u,s,n)}, W_{\neg(u,s,n)}) \\ &\propto \begin{cases} p(Y_{u,s,n} = 1|\pi_{\neg(u,s,n)}, \gamma) \\ \quad \cdot p(W_{u,s,n} = v|\phi_{B,\neg(u,s,n)}, \beta), \text{ if } n = 1 \\ p(Y_{u,s,n} = 1|\pi_{\neg(u,s,n)}, \gamma) \\ \quad \cdot p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{B,\neg(u,s,n)}, \beta), \text{ OTW} \end{cases} \end{aligned}, \tag{16}$$

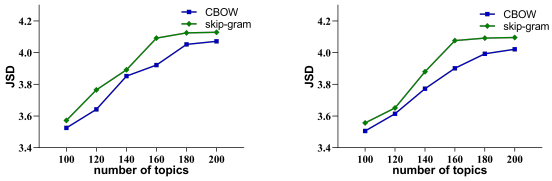
where $p(Y_{u,s,n} = 1|\pi_{\neg(u,s,n)}, \gamma)$ is formulated as Eq. 17, and $p(W_{u,s,n} = v|\phi_{B,\neg(u,s,n)}, \beta)$ and $p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{B,\neg(u,s,n)}, \beta)$ are formulated in Eq. 18 according to the aforementioned joint distribution.

$$\begin{aligned} &p(Y_{u,s,n} = 1|\pi_{\neg(u,s,n)}, \gamma) \\ &= \frac{n_{1,\neg(u,s,n)} + \gamma_1}{n_{0,\neg(u,s,n)} + \gamma_0 + n_{1,\neg(u,s,n)} + \gamma_1}. \end{aligned} \tag{17}$$



(a) On the Twitter corpus. (b) On the Weibo corpus.

FIGURE 2. Comparison of CBOW and skip-gram in terms of perplexity.



(a) On the Twitter corpus. (b) On the Weibo corpus.

FIGURE 3. Comparison of CBOW and skip-gram in terms of JSD.

$$\begin{cases} p(W_{u,s,n} = v|W_{u,s,n-1} = v', \phi_{B,\neg(u,s,n)}, \beta) \\ = p(W_{u,s,n} = v|\phi_{B,\neg(u,s,n)}, \beta)^{p(W_{u,s,n}=v|W_{u,s,n-1}=v')} \\ \cdot (1 - p(W_{u,s,n} = v|\phi_{B,\neg(u,s,n)}, \beta))^{(1-p(W_{u,s,n}=v|W_{u,s,n-1}=v'))} \\ p(W_{u,s,n} = v|\phi_{B,\neg(u,s,n)}, \beta) = \frac{n^B_{v,\neg(u,s,n)} + \beta_v}{\sum_{v=1}^{V}(n^B_{v,\neg(u,s,n)} + \beta_v)} \\ p(W_{u,s,n} = v|W_{u,s,n-1} = v') = \frac{exp(\bar{W}_v \cdot \bar{W}_{v'})}{\sum_{v=1}^{V} exp(\bar{W}_v \cdot \bar{W}_{v'})}. \end{cases} \tag{18}$$

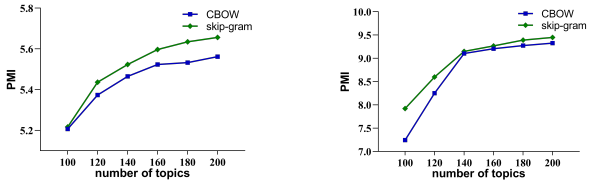
In Eq. 17 and Eq. 18, $n^B_{v,\neg(u,s,n)}$ is the number of times that the vth word occurs in background words without considering $W_{u,s,n}$, and the meanings of the other symbols are same as earlier definitions.

5. EXPERIMENTS

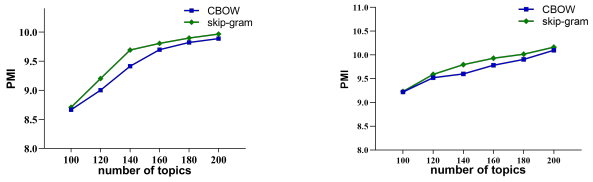
Data set. We adopted two post sets sampled from Weibo (211,233 posts) and Twitter (553,098 posts) as our data sets for evaluation. These posts were all published publicly by users. The reasons why we utilize these posts as corpus are three-fold. First, Weibo and Twitter have been popular social network sites. Weibo now has been one of the top 10 biggest social media platforms among the world. In January 2022, 573 million users are active in Weibo to obtain latest news and share content with friends [¶]. Twitter is also very popular among the world. According to the report released in 2022 [‖], 500 million tweets are published each day, and 350,000 tweets are posted every minute. The large user bases of these two platforms and the active content sharing behaviors on that provide us a large volume of social posts to make experiments. Second, the majority of Weibo posts are published in Chinese, while that of Twitter tweets are written in English. By utilizing such two data sets, we can evaluate our model's capability to process different languages. Third, although there exist some short text experimental data sets that are shared by previous research, most of them are limited in size and lack of user information, which motivated us to construct data sets by sampling posts from nowadays popular social media platforms. In our data sampling, we ignored the posts with less than five words as they were meaningless for topic modeling. For the obtained posts, we first filtered out the noisy items like emojis and special symbols. The hashtags were remained since

[¶] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[‖] https://www.websiterating.com/research/twitter-statistics/

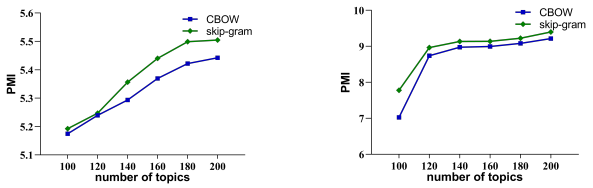


(a) M=5. (b) M=10.

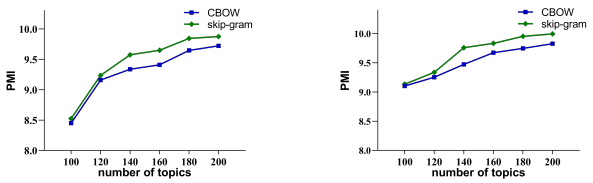


(c) M=15. (d) M=20.

FIGURE 4. Comparison of CBOW and skip-gram in terms of PMI on the Twitter corpus.



(a) M=5. (b) M=10.



(c) M=15. (d) M=20.

FIGURE 5. Comparison of CBOW and skip-gram in terms of PMI on the Weibo corpus.

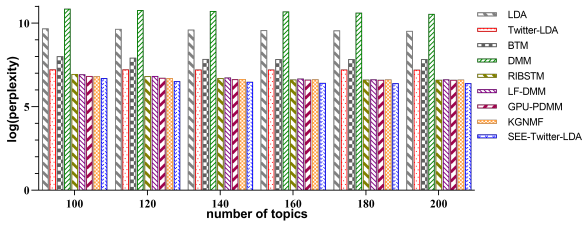


FIGURE 6. Comparison with baselines in terms of perplexity on Twitter corpus.

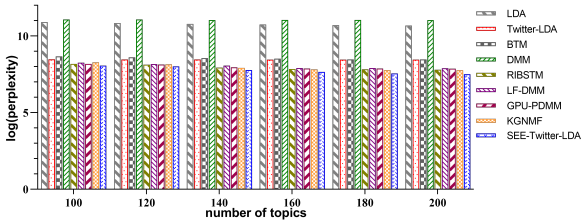


FIGURE 7. Comparison with baselines in terms of perplexity on Weibo corpus.

these tokens were the keywords within posts. After then, the Weibo posts were segmented into tokens by utilizing Jieba **, and the Twitter posts were segmented by spaces and punctuations. These token-based posts were input into different models to generate topics.

Metrics. We evaluate the performance of SEE-Twitter-LDA from three perspectives: perplexity [68], topic divergence and topic coherence [36, 63]. First, as is shown in Eq. 19, perplexity is an indicator of uncertainty. The lower the perplexity is, the better the performance of a topic model will be. Second, a better topic model should be capable of generating independent topics. Jensen–Shannon Divergence (JSD) exhibited in Eq. 20 is a measurement to evaluate the independence and divergence among topics. The higher the value is, the better a topic model will be. The last, topic coherence evaluates the word consistency within each topic. We utilize the Pointwise Mutual Information (PMI) exhibited in Eq. 21 to evaluate topic coherence. A higher PMI value indicates a better

**https://github.com/fxsjy/jieba

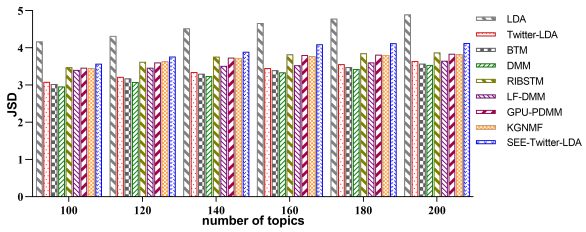


FIGURE 8. Comparison with baselines in terms of JSD on Twitter corpus.

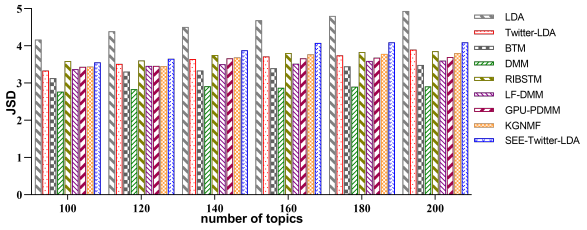


FIGURE 9. Comparison with baselines in terms of JSD on Weibo corpus.

performance of topic coherence.

$$perplexity = exp\{-\frac{\sum_{u=1}^{U}\sum_{s=1}^{N_u} log(p(W_{u,s}))}{\sum_{u=1}^{U}\sum_{s=1}^{N_u} N_{u,s}}\}. \quad (19)$$

$$JSD(\phi) = H(\frac{1}{T}\sum_{t=1}^{T}\phi_t) - \frac{1}{T}\sum_{t=1}^{T}H(\phi_t), \quad (20)$$

where $H(\cdot)$ means Shannon entropy.

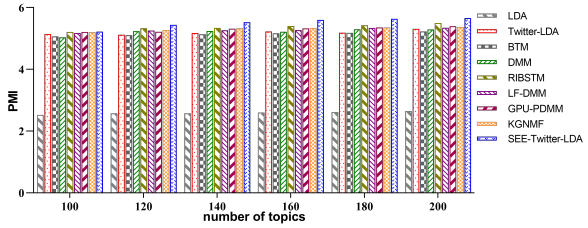
$$PMI(t, V^{(t)}) = \frac{2}{M(M-1)}\sum_{m=2}^{M}\sum_{l=1}^{m-1} log(\frac{p(v_m^{(t)}, v_l^{(t)})+1}{p(v_m^{(t)})p(v_l^{(t)})}), \quad (21)$$

where $V^{(t)} = \{v_1^{(t)}, v_2^{(t)}, \cdots, v_M^{(t)}\}$ is a list of the M most probable words in the tth topic, $p(v_m^{(t)})$ or $p(v_l^{(t)})$ is the probability that the word $v_m^{(t)}$ or $v_l^{(t)}$ occurs in corpus, and $p(v_m^{(t)}, v_l^{(t)})$ denotes the probability that these two words $v_m^{(t)}$ and $v_l^{(t)}$ appear in the same post.

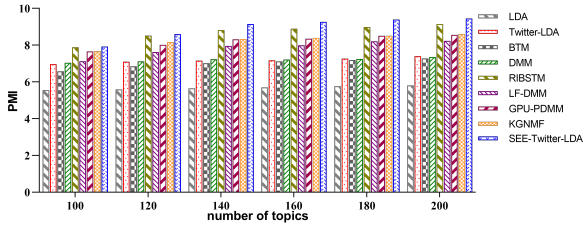
Baselines and parameter settings. First, as we propose the semantic embedding enhanced topic model based on Twitter-LDA, LDA and Twitter-LDA are first considered as baselines. Second, as mentioned above, BTM and DMM are two common-used short text modeling methods. We also set these two models as baselines. Third, as elaborated in Section 2, previous research has proposed some semantic embedding enhanced topic models for short text modeling. These models can be utilized for UGTC modeling. We analyzed these models and chose four state-of-the-art methods as baselines. To be specific, our baseline methods and corresponding parameter settings are listed below. Besides these parameters, the iteration times is set as 2000 for all methods.

- LDA. It is the classical topic model which sets α=0.1, and β=0.01.
- Twitter-LDA [19]. We set α=50/T, β=0.01, and γ=0.01, according to the authors' recommendations.
- BTM [15]. We set α=50/T, and β=0.01 according to the authors' recommendations.
- DMM [16]. We set α=50/T, and β=0.01 according to the authors' recommendations.
- RIBSTM [33, 35]. RIBSTM is a semantic embedding enhanced model of BTM which uses RNN for word relationship learning. Its parameter setting is same as that of BTM.

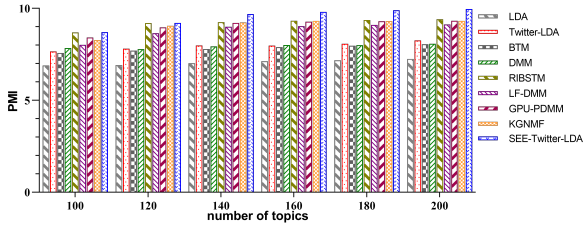
- LF-DMM [26]. LF-DMM is a semantic embedding enhanced model of DMM which uses Word2Vec for word embedding. We set λ=0.6, α=0.1, and β= 0.01 according to the authors' recommendations.
- GPU-PDMM [36, 37]. GPU-PDMM is a semantic embedding enhanced model of DMM which uses Polya urn model for word embedding. As suggested by the authors, we set μ=0.1, ι= 2, and λ=1.5.
- KGNMF [63]. KGNMF combines LDA and non-negative matrix factorization for short text modeling. We set α=1, β=0.01, and λ=5 according to the authors' recommendations.



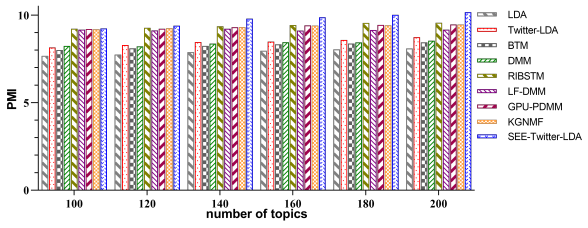
(a) M=5.



(b) M=10.

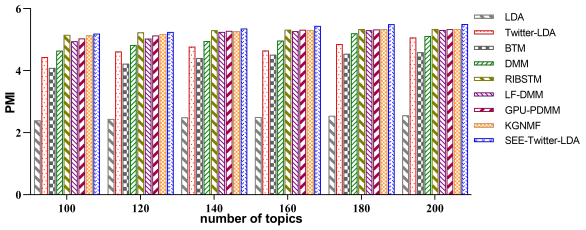


(c) M=15.

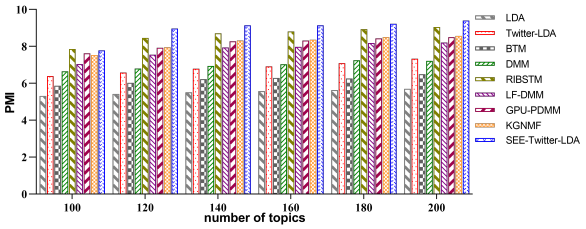


(d) M=20.

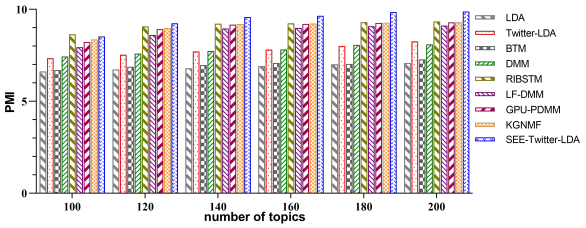
FIGURE 10. Comparison with baselines in terms of PMI on the Twitter corpus.



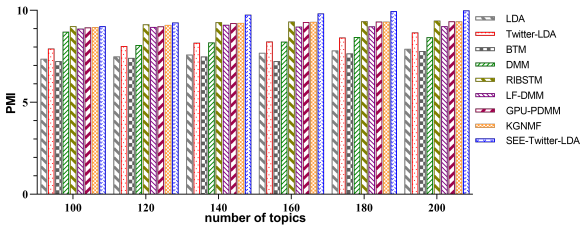
(a) M=5.



(b) M=10.



(c) M=15.



(d) M=20.

FIGURE 11. Comparison with baselines in terms of PMI on the Weibo corpus.

5.1. Comparison of CBOW and skip-gram

As mentioned above, Word2Vec contains two models: CBOW and skip-gram. CBOW predicts each word by using contexts of its surroundings, while skip-gram uses each word to predict its neighboring words. As we utilize Word2Vec for semantic embedding, we first make experiments to evaluate the performances of these two models. The results are shown in Figure 2, Figure 3, Figure 4 and Figure 5. In these figures, the horizontal axis gives six topic size settings increasing from 100 to 200, the vertical axis in Figure 2 exhibits the perplexity values (the values are log transformed for ease of comparison), the vertical axis in Figure 3 exhibits the JSD values, and the vertical axis in Figure 4

and Figure 5 exhibits the PMI values. In Figure 4 and Figure 5, we choose four settings for M including 5, 10, 15 and 20 to compare CBOW and skip-gram with different sizes of most probable words.

From the results, we first can see perplexity decreases with topic size and M, while JSD and PMI all increase with that. These measurements of CBOW and skip-gram all have exhibited convergence trends in observation windows, which verifies the reasonableness of our settings for topic size and M. Second, skip-gram obtains better performances in all settings. Previous research and practice have suggested that CBOW is faster than skip-gram, while skip-gram performs better to process uncommon words. In reality, many buzzwords originate in social media context. We thought that is why skip-gram obtains better performances versus CBOW. These results indicate skip-gram is a better semantic embedding method for our SEE-Twitter-LDA model. So in the following experiments, we all implement SEE-Twitter-LDA based on skip-gram. Third, we can see the experimental results especially the perplexity values on the Twitter corpus are overall better than that on the Weibo corpus. It is mostly as modeling Chinese texts is generally more complicated than processing English texts. For example, for English texts, words are naturally segmented by punctuations and space. While to handle Chinese texts, we need to segment each sentence into words by a tokenizer, and the results of word segmentation can significantly impact the following processing tasks. Such procedures increase the complexity of Chinese language processing.

5.2. Comparison with baselines

The experimental results comparing with baselines are shown in Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11. Same as the foregoing experiments, we also set topic size increasing from 100 to 200 in all experiments and M varying from 5 to 20 in PMI analysis. Based on these figures, we can obtain the following results.

- Experimental results overview. On the whole, the perplexity, JSD and PMI of all models have significant changes from the beginning of each observation window to the end. However, the values change very slightly when topic size varies from 180 to 200, and M increases from 15 to 20, which suggests convergence trends. These results verify the reasonableness of our parameter settings for utilizing such models to process the two large data sets.
- Comparison of LDA, Twitter-LDA, BTM and DMM. First, we can see LDA obtains worse perplexity and PMI values but higher JSD values. DMM performs the worst in terms of perplexity and JSD, while its PMI values are overall the highest when topic size and M are smaller. However, when approaching to convergence (topic size is 200, and M equals 15 or 20), its PMI values are lower than that of Twitter-LDA. Second, the perplexity and JSD of BTM are moderate among these three models, while the PMI values are the lowest almost in all settings. Third, Twitter-LDA obtains better perplexity and JSD performances in all settings and higher PMI values when approaching to convergence. These experimental results suggest that Twitter-LDA, which considers the concepts of users and noisy words, can model UGTC in social media more accurately.
- Comparison of traditional short text modeling methods and semantic embedding enhanced topic models. Semantic embedding enhanced topic models (RIBSTM, LF-DMM, GPU-PDMM, KGNMF and SEE-Twitter-LDA) have prominent improvements on perplexity, JSD and PMI compared with the three traditional short text models, which means incorporating semantic embedding into topic models is helpful for short text modeling.
- Comparison of semantic embedding enhanced topic models. Among the five semantic embedding enhanced topic models, LF-DMM performs worse than the others, the performances of RIBSTM, GPU-PDMM and KGNMF are comparable, and SEE-Twitter-LDA obtains better performances in most settings especially the settings approaching to convergence. From the perspective of topic size, the superiority of SEE-Twitter-LDA becomes more significant when topic size becomes bigger. For example, in Figure 6 and Figure 7, the perplexity values of SEE-Twitter-LDA are slightly lower than that of RIBSTM when topic size is set as 100 and 120, while when it is raised to 140, 160, 180 and 200, the superiority of SEE-Twitter-LDA becomes prominent. Figure 8, Figure 9, Figure 10 and Figure 11 also depict similar patterns. SEE-Twitter-LDA and RIBSTM obtain similar JSD values and PMI values in topic size settings 100 and 120, while SEE-Twitter-LDA obtains significant higher values versus RIBSTM when topic is augmented. From the perspective of M, we can see the PMI differences between SEE-Twitter-LDA and the other semantic embedding enhanced topic models as well as the differences among the other models are not significant when M is 5, while in other settings (M equals 10, 15 or 20), the differences especially SEE-Twitter-LDA's improvements become more prominent. When M is set as 5, very few most probable words are utilized for topic coherence evaluation. These semantic embedding enhanced topic models can obtain comparable results in detecting the top probable words for topics. While when M increases, more words are utilized for evaluation, and the differences among different models become significant, which highlights the improvements of

our proposed model versus the other methods.

5.3. Complexity

From the perspective of time complexity, the complexity of each iteration of SEE-Twitter-LDA is the same as that of Twitter-LDA. According to Algorithm 1, in each iteration, SEE-Twitter-LDA goes through all users by sampling a topic for each of a user's post and generating the post word-by-word. So the time complexity of SEE-Twitter-LDA is $O(iterationNumber \cdot U \cdot \overline{S} \cdot T \cdot \overline{N})$, where $iterationNumber$ means the number of iterations, U means the number of social ecosystem users, $\overline{S}$ means the average number of posts of each user, T means the number of topics, and $\overline{N}$ is the average length of each post. From the perspective of space complexity, the complexity of Twitter-LDA is higher than that of Twitter-LDA. In Twitter-LDA, for each user, each word of a post is generated according to the user-topic distribution (θ_u) and topic-word distribution (ϕ_t or ϕ_B). So it needs to store the matrices θ and ϕ, and the corresponding space sizes are $U \cdot T$ and $(T + 1) \cdot V$. For SEE-Twitter-LDA, each word is generated according to both the topical context and the prior word. So besides θ and ϕ, the model also needs to evaluate $P(W_{u,s,n}|W_{u,s,n-1})$ which depends on a $V \cdot V$ matrix (V means the number of tokens in dictionary). This characteristic results in a higher space complexity compared with Twitter-LDA.

In summary, SEE-Twitter-LDA can model user-generated textual content more accurately in terms of perplexity, topic divergence and topic coherence compared with the baseline methods. When utilizing such a model, skip-gram is suggested as the semantic embedding method as it can obtain better performance on perplexity, topic divergence and topic coherence.

6. CONCLUSIONS

To accurately model user-generated textual content in social ecosystems, we propose the SEE-Twitter-LDA model which combines the topic model with the semantic embedding model in this paper. Extensive experiments on a large amount of Weibo posts and Twitter tweets exhibit our proposed model performs better than traditional short text modeling methods like Twitter-LDA, BTM and DMM as well as existing semantic embedding enhanced topic models including RIBSTM, LF-DMM, GPU-PDMM and KGNMF in terms of perplexity, topic divergence and topic coherence. The future research of SEE-Twitter-LDA will focus on the following three aspects. First, there emerge some new kinds of word embedding techniques like BERT (Bidirectional Encoder Representations from Transformers) and GloVe (Global Vectors for word representation). We will integrate these new methods into SEE-Twitter-LDA to validate their performance for UGTC modeling. Second, UGTC is one of the common types of user-generated content in social ecosystems, and there are other types like images and video. In future research, we will further extend SEE-Twitter-LDA to model multimedia UGC. Third, most nowadays natural language processing techniques are based on neural networks. So future research can investigate to implement SEE-Twitter-LDA via neural networks by referring to neural topic models.

DATA AVAILABILITY STATEMENT

The data underlying this article will be shared on reasonable request to the corresponding author.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) under the Grants nos. 61932007 and 61902075.

REFERENCES

[1] Gao, H., Xu, K., Cao, M., Xiao, J., Xu, Q., and Yin, Y. (2022) The deep features and attention mechanism-based method to dish healthcare under social iot systems: An empirical study with a hand-deep local-global net. *IEEE Transactions on Computational Social Systems*, **9**, 336–347.

[2] Gao, H., Qiu, B., Barroso, R. J. D., Hussain, W., Xu, Y., and Wang, X. (2022) TSMAE: A novel anomaly detection approach for internet of things time series data using memory-augmented autoencoder. *IEEE Transactions on Network Science and Engineering*, **10.1109/TNSE.2022.3163144**.

[3] Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., FlÃPck, F., and Jurgens, D. (2019) Demographic inference and representative population estimates from multilingual social media data. *Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA, May 13-17*, pp. 2056–2067. ACM.

[4] Fang, Q., Sang, J., Xu, C., and Hossain, M. S. (2015) Relational user attribute inference in social media. *IEEE Transactions on Multimedia*, **17**, 1031–1044.

[5] Wilson, S. and Mihalcea, R. (2019) Predicting human activities from user-generated content. *Proceedings of the 57th Conference of the Association for Computational Linguistics (Volume 1: Long Papers), Florence, Italy, July 28- August 2*, pp. 2572–2582. Association for Computational Linguistics.

[6] Fast, E., McGrath, W., Rajpurkar, P., and Bernstein, M. S. (2016) Augur: Mining human behaviors from fiction to power interactive systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12*, pp. 237–247. ACM.

[7] Wilson, S. and Mihalcea, R. (2017) Measuring semantic relations between human activities. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, November 27 - December 1*, pp. 664–673. Asian Federation of Natural Language Processing.

[8] Saif, H., He, Y., and Alani, H. (2012) Semantic sentiment analysis of twitter. *Proceedings of the 11th International Semantic Web Conference (Part I), Boston, MA, USA, November 11-15*, Lecture Notes in Computer Science, **7649**, pp. 508–524. Springer.

[9] She, J. and Chen, L. (2014) Tomoha: Topic model-based hashtag recommendation on twitter. *Proceedings of the 23rd International World Wide Web Conference (Companion Volume), Seoul, Republic of Korea, April 7-11*, pp. 371–372. ACM.

[10] Mee, A., Homapour, E., Chiclana, F., and Engel, O. (2021) Sentiment analysis using tf–idf weighting of uk mps' tweets on brexit. *Knowledge-Based Systems*, **228**, 107238.

[11] Liu, T., Wang, K., Sha, L., Chang, B., and Sui, Z. (2018) Table-to-text generation by structure-aware seq2seq learning. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7*, pp. 4881–4888. AAAI Press.

[12] Xun, G., Li, Y., Gao, J., and Zhang, A. (2017) Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17*, pp. 535–543. ACM.

[13] Li, S., Chua, T.-S., Zhu, J., and Miao, C. (2016) Generative topic embedding: A continuous representation of documents. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, August 7-12*. The Association for Computer Linguistics.

[14] Shi, B., Lam, W., Jameel, S., Schockaert, S., and Lai, K. P. (2017) Jointly learning word embeddings and latent topics. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11*, pp. 375–384. ACM.

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

[16] Rigouste, L., Cappé, O., and Yvon, F. (2007) Inference and evaluation of the multinomial mixture model for text clustering. *Information processing & management*, **43**, 1260–1280.

[17] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

[18] Lee, R. K.-W., Hoang, T.-A., and Lim, E.-P. (2017) On analyzing user topic-specific platform preferences across multiple social media sites. *Proceedings of the 26th International Conference on World Wide Web*, pp. 1351–1359.

[19] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011) Comparing twitter and traditional media using topic models. *Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, April 3-7*, pp. 1351–1359. ACM.

[20] Manikonda, L., Meduri, V. V., and Kambhampati, S. (2016) Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media. *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20*, pp. 639–642. AAAI Press.

[21] Manikonda, L. and De Choudhury, M. (2017) Modeling and understanding visual attributes of mental health disclosures in social media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11*, pp. 170–181. ACM.

[22] Deng, Z., Yan, M., Sang, J., and Xu, C. (2015) Twitter is faster: Personalized time-aware video recommendation from twitter to youtube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **11**, 1–23.

[23] Perera, D. and Zimmermann, R. (2018) LSTM networks for online cross-network recommendations. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, July 13-19*, pp. 3825–3833. ijcai.org.

[24] Wu, L., Yen, I. E.-H., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P., and Witbrock, M. J. (2018) Word mover's embedding: From Word2Vec to document embedding. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4*, pp. 4524–4534. Association for Computational Linguistics.

[25] Zhao, X. and Lindley, S. E. (2014) Curation through use: Understanding the personal value of social media. *Proceedings of the CHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, April 26 - May 01*, pp. 2431–2440. ACM.

[26] Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015) Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, **3**, 299–313.

[27] Fu, X., Sun, X., Wu, H., Cui, L., and Huang, J. Z. (2018) Weakly supervised topic sentiment joint model with word embeddings. *Knowledge-Based Systems*, **147**, 43–54.

[28] Hu, W. and Tsujii, J. (2016) A latent concept topic model for robust topic inference using word embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, August 7-12*. The Association for Computer Linguistics.

[29] Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. (2017) A correlated topic model using word embeddings. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, August 19-25*, pp. 4207–4213. ijcai.org.

[30] Batmanghelich, K., Saeedi, A., Narasimhan, K., and Gershman, S. (2016) Nonparametric spherical topic modeling with word embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, August 7-12*. The Association for Computer Linguistics.

[31] Bunk, S. and Krestel, R. (2018) Welda: Enhancing topic models by incorporating local word context. *Proceedings of the 18th ACM/IEEE on Joint Conference*

on Digital Libraries, Fort Worth, TX, USA, June 03-07, pp. 293–302. ACM.

[32] Zhao, H., Du, L., and Buntine, W. (2017) A word embeddings informed focused topic model. *Proceedings of The 9th Asian Conference on Machine Learning, Seoul, Korea, November 15-17*, Proceedings of Machine Learning Research, **77**, pp. 423–438. PMLR.

[33] Lu, H.-Y., Xie, L.-Y., Kang, N., Wang, C.-J., and Xie, J.-Y. (2017) Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, February 4-9*, pp. 1192–1198. AAAI Press.

[34] Chen, J., Gong, Z., and Liu, W. (2019) A nonparametric model for online topic discovery with word embeddings. *Information Sciences*, **504**, 32–47.

[35] Lu, H.-Y., Kang, N., Li, Y., Zhan, Q.-Y., Xie, J.-Y., and Wang, C.-J. (2019) Utilizing recurrent neural network for topic discovery in short text scenarios. *Intelligent Data Analysis*, **23**, 259–277.

[36] Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2017) Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, **36**, 1–30.

[37] Li, C., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2016) Topic modeling for short texts with auxiliary word embeddings. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, July 17-21*, pp. 165–174. ACM.

[38] Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J., and Zhang, A. (2016) Topic discovery for short texts using word embeddings. *Proceedings of the IEEE 16th International Conference on Data Mining, Barcelona, Spain, December 12-15*, pp. 1299–1304. IEEE Computer Society.

[39] Al-Salemi, B., Ab Aziz, M. J., and Noah, S. A. (2015) LDA-AdaBoost. mh: Accelerated adaboost. mh based on latent dirichlet allocation for text categorization. *Journal of Information Science*, **41**, 27–40.

[40] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007) Topic sentiment mixture: Modeling facets and opinions in weblogs. *Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, May 8-12*, pp. 171–180. ACM.

[41] Zhang, P., Gu, H., Gartrell, M., Lu, T., Yang, D., Ding, X., and Gu, N. (2016) Group-based latent dirichlet allocation (group-lda): Effective audience detection for books in online social media. *Knowledge-Based Systems*, **105**, 134–146.

[42] Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013) Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, July 28 - August 01*, pp. 889–892. ACM.

[43] Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010) Twitterrank: Finding topic-sensitive influential twitterers. *Proceedings of the Third International Conference on Web Search and Web Data Mining, New York, NY, USA, February 4-6*, pp. 261–270. ACM.

[44] Hong, L. and Davison, B. D. (2010) Empirical study of topic modeling in twitter. *Proceedings of the 3rd Workshop on Social Network Mining and Analysis, Paris, France, June 28*, pp. 80–88. ACM.

[45] Yan, X., Guo, J., Lan, Y., Xu, J., and Cheng, X. (2015) A probabilistic model for bursty topic discovery in microblogs. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, USA, January 25-30*, pp. 353–359. AAAI Press.

[46] Wu, T., Qi, G., Wang, H., Xu, K., and Cui, X. (2016) Cross-lingual taxonomy alignment with bilingual biterm topic model. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, February 12-17*, pp. 287–293. AAAI Press.

[47] He, X., Xu, H., Li, J., He, L., and Yu, L. (2017) FastBTM: Reducing the sampling time for biterm topic model. *Knowledge-Based Systems*, **132**, 11–20.

[48] Yin, J. and Wang, J. (2014) A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, August 24 - 27*, pp. 233–242. ACM.

[49] Duan, R. and Li, C. (2018) An adaptive dirichlet multinomial mixture model for short text streaming clustering. *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence, Santiago, Chile, December 3-6*, pp. 49–55. IEEE Computer Society.

[50] Li, X., Zhang, J., and Ouyang, J. (2019) Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, January 27 - February 1*, pp. 7884–7891. AAAI Press.

[51] Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P., and Li, X. (2011) Topical keyphrase extraction from twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 19-24 June*, pp. 379–388. The Association for Computer Linguistics.

[52] Lo, S. L., Chiong, R., and Cornforth, D. (2016) Ranking of high-value social audiences on twitter. *Decision Support Systems*, **85**, 34–48.

[53] Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W.-Y. (2017) Topic aware neural response generation. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, February 4-9*, pp. 3351–3357. AAAI Press.

[54] Sasaki, K., Yoshikawa, T., and Furuhashi, T. (2014) Online topic model for twitter considering dynamics of user interests and topic trends. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October 25-29*, pp. 1977–1985. ACL.

[55] Vosecky, J., Jiang, D., Leung, K. W.-T., and Ng, W. (2013) Dynamic multi-faceted topic discovery in twitter. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management,*

San Francisco, CA, USA, October 27 - November 1, pp. 879–884. ACM.

[56] Gao, H., Xiao, J., Yin, Y., Liu, T., and Shi, J. (2022) A mutually supervised graph attention network for few-shot segmentation: The perspective of fully utilizing limited samples. *IEEE Transactions on Neural Networks and Learning Systems*, **10.1109/TNNLS.2022.3155486**.

[57] Yin, Y., Huang, Q., Gao, H., and Xu, Y. (2020) Personalized apis recommendation with cognitive knowledge mining for industrial systems. *IEEE Transactions on Industrial Informatics*, **17**, 6153–6161.

[58] Li, S., Zhu, J., and Miao, C. (2015) A generative word embedding model and its low rank positive semidefinite solution. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, September 17-21*, pp. 1599–1609. The Association for Computational Linguistics.

[59] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pp. 3111–3119.

[60] Pennington, J., Socher, R., and Manning, C. D. (2014) Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October 25-29*, pp. 1532–1543. ACL.

[61] Zhang, P., Wang, S., Li, D., Li, X., and Xu, Z. (2019) Combine topic modeling with semantic embedding: Embedding enhanced topic model. *IEEE Transactions on Knowledge and Data Engineering*, **32**, 2322–2335.

[62] Yang, Y., Wang, H., Zhu, J., Wu, Y., Jiang, K., Guo, W., and Shi, W. (2021) Dataless short text classification based on biterm topic model and word embeddings. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Montreal-themed Virtual Reality, August 19 - 26*, pp. 3969–3975. ijcai.org.

[63] Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019) Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, **163**, 1–13.

[64] Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018) Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. *Proceedings of the 2018 World Wide Web Conference on World Wide Web, Lyon, France, April 23-27*, pp. 1105–1114. ACM.

[65] Lin, L., Jiang, H., and Rao, Y. (2020) Copula guided neural topic modelling for short texts. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China, July 25-30*, pp. 1773–1776. ACM.

[66] Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, C., and Zhuang, F. (2021) A neural topic model with word vectors and entity vectors for short texts. *Information Processing & Management*, **58**, 102455.

[67] Mou, L., Peng, H., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2015) Discriminative neural sentence modeling by tree-based convolution. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, September 17-21*, pp. 2315–2325. The Association for Computational Linguistics.

[68] Fang, Y., Si, L., Somasundaram, N., and Yu, Z. (2012) Mining contrastive opinions on political texts using cross-perspective topic model. *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, Seattle, WA, USA, February 8-12*, pp. 63–72. ACM.