Information Sciences 570 (2021) 185–199

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Author topic model for co-occurring normal documents and short texts to explore individual user preferences

Yang Yang [a], Feifei Wang [b,c,*]

[a] School of Electronic Engineering and Computer Science, Peking University, China
[b] Center for Applied Statistics, Renmin University of China, China
[c] School of Statistics, Renmin University of China, China

ARTICLE INFO

*Article history:*
Received 16 February 2020
Received in revised form 2 April 2021
Accepted 11 April 2021
Available online 16 April 2021

*Keywords:*
Authorship
Short texts
Topic model
User preference

ABSTRACT

The investigation of user preferences through user comments has attracted significant attention. Although topic models have been verified as useful tools to facilitate the understanding of textual contents, they cannot be directly applied to accomplish this task because of two problems. The first problem is the severe data sparsity suffered by user comments because they are generally short. The second problem is the mixture of opinions from both user comments and the original documents the users commented on. To simultaneously solve the data sparsity problem and explore clean user preferences, we propose an author co-occurring topic model (AOTM) for normal documents and their short user comments. By considering authorship, AOTM allows each author of short texts to have a probability distribution over a set of topics represented only short texts. Accordingly, the individual user preferences can be investigated based on these author-level distributions. We verify the performance of AOTM using two news article datasets and one e-commerce dataset. Extensive experiments demonstrate that the AOTM outperforms several state-of-the-art methods in topic learning and topic representation of documents. The potential usage of AOTM in exploring individual user preferences is further illustrated by drawing user portraits and predicting user posting behaviors.



## 1. Introduction

Investigating user preferences from web service users has always been a valuable task in marketing research, consensus observation, and many other scenarios. Topic models have been verified as useful tools to facilitate the understanding of text contents by extracting topics underlying the textual documents [1,2]. However, they cannot be directly applied to investigate the user preferences frow user comments because of two problems. The first problem is that online user comments are generally very short. Then traditional topic models would suffer from a severe data sparsity problem [3,4]. The second problem is that, online user comments are often mixtures of topics, not only discussing the basic facts in the original documents the users commented, but also expressing individual user preferences [5]. Therefore, it is hard for traditional topic models to extract clean and interpretable user preferences for web service users.

To exploit clean user preferences from short user comments, a topic model for co-occurring normal documents and short texts (COTM) [5] was proposed. In COTM, the normal documents represent news articles, product descriptions and other

* Corresponding author.
*E-mail addresses:* yyang1988@pku.edu.cn (Y. Yang), feifei.wang@ruc.edu.cn (F. Wang).

https://doi.org/10.1016/j.ins.2021.04.060
0020-0255/

original documents that users would comment on; while the corresponding short texts represent the short user comments. The basic facts discussed in the normal documents are characterized by a type of *formal topics*; while the user preferences described only by users in the short texts are characterized by a type of *informal topics*. The COTM tries to capture the inner correlation between the normal documents and the short texts by assuming: (1) each normal document has a probability distribution over all formal topics, and (2) each short text has a probability distribution over two topics; one from the formal topics, and the other from the informal topics. Through these assumptions, COTM can extract clean contextual topics (i.e., the informal topics) discussed by users in the short texts. In the meanwhile, by taking advantage of the combination of normal documents and short texts, COTM can discover more prominent, coherent, and comprehensive topics than other baseline methods.

Although COTM discovers user preferences by separating informal topics from formal topics, it ignores the authorship of short texts and assumes that all short texts are written by the same author or share the same topic distribution over the informal topics. As a consequence, COTM fails to capture the author-level topic preferences for individual users. To alleviate this assumption and explore individual user preferences, we incorporate the authorship information into COTM and propose an author topic model for co-occurring normal documents and short texts, which we call AOTM hereafter. In AOTM, we assume that each author of short texts has its own probability distribution over all informal topics. It is notable that, the definition of authorship can be very flexible. Therefore, the AOTM can be generalized to deal with datasets whose short texts are assigned with any class labels.

Compared with the existing author topic model [6] and labeled topic model [7], the authorship extension in the AOTM is much simpler because there is only one author for each short text. Specifically, in the author topic model (AuTM) [6], each document has multiple authors. The authorship of each word is chosen uniformly from the group of authors, and each author is associated with a probability distribution over topics. In labeled latent Dirichlet allocation [7], a one-to-one correspondence between labels and topics is defined. Next, the words in each text are only chosen from topics in accordance with its label set. However, in AOTM, we assume that there is only one author for each short text, which is the most common case for short texts in real-world scenarios.

To summarize, we propose in this study an AOTM for the co-occurring normal documents and short texts to explore user preferences. The contributions of this study are as follows:

- In AOTM, the authorship extension helps relax the "same author" assumption in COTM and makes it more flexible to model labeled short texts in real-world scenarios.
- By incorporating authorship information, AOTM can explore individual user preferences better. We provide a practical methodology for drawing user portraits and predicting user posting behaviors to further illustrate its abilities in exploring user preferences.
- We evaluate the proposed model on two news article datasets and one e-commerce dataset. The experimental results demonstrate that AOTM can outperform COTM and most state-of-the-art methods in topic learning (measured using coherence score) and the topic representation of documents (measured using perplexity).

The rest of this paper is organized as follows. Section 2 presents a review of the related works. Section 3 introduces the AOTM and discusses its model estimation algorithm. Section 4 conducts a variety of experiments to evaluate the performance of AOTM in topic learning and illustrates its potential usage in exploring user preferences. Section 5 concludes the paper with a brief discussion.

## 2. Related works

### *2.1. Topic models for short texts*

Three major strategies have been adopted to address the sparsity of word co-occurrence in short texts. The first one involves customizing topic models to strategically magnify the co-occurrence of morphemes under topics. An intuitive method under this strategy was to add strong assumptions that all words in a short text belonged to the same topic [8–10]. Moreover, the biterm topic model (BTM) and its extensions heuristically constructed word-pairs from texts, turned documents into biterm sets, and assumed that biterms were drawn independently from topics [4,11]. To make topics more focused on representative words rather than the entire corpus, Lin et al. [12] and Chen et al. [10] replaced the symmetric Dirichlet prior with the "Spike and Slab" prior, which acted as a binary switcher for selecting focused words for topics. He et al. [13] assigned only one targeted topic to each short text, which could help group similar words to achieve topic homogeneity. However, the additional information considered in these studies is not enough, and the sparsity problem is not solved completely [14].

Another strategy involves aggregating short texts into longer pseudo-documents. For example, Weng et al. [15] and Mehrotra et al. [16] aggregated tweets based on authorships and hashtags. Kou et al. [17] proposed a concept named special region and utilized location information to aggregate short texts into long texts. When auxiliary information was not available, Hong et al. [18] aggregated tweets containing similar words. Bicalho et al. [19] organized pseudo-documents based on the similarities of morphemes, which were evaluated through co-occurrence and the distance between vector representa-

tions. Standard topic models were then applied to learn more prominent topics from the enriched contexts. This paradigm was further improved by Quan et al. [20] and Zuo et al. [14], who jointly modeled the process of short texts aggregation and topics generation in a more sophisticated and elegant way.

The third strategy involves exploring semantic relations from external large datasets to improve the topic inference of short texts. For example, Phan et al. [21,22] trained topic models on a collection of long texts that were in the same domain with short texts. Jin et al. [23] gained knowledge from topic-related long texts to infer the topics of short texts. Li et al. [24] utilized pre-trained word embeddings as global word relatedness knowledge to promote the semantically related words under similar topics. Chen et al. [10] and Qiang et al. [25] exploited auxiliary word embeddings to infer the number of topics and aggregated short texts into long pseudo-texts. Although these methods provide generic solutions to facilitate the topic learning of short texts, organizing the external datasets requires additional labor and could bring in new uncertainties if not properly conducted.

To avoid the uncertainties in organizing the external datasets, Yang et al. [5] focused on a special type of short texts, which often co-occurred with normal documents, and proposed a method of COTM. Typical examples of the co-occurring normal documents and short texts included: news articles with reader comments, product descriptions with consumer reviews, and blog posts with reader feedback. The intrinsic semantic relations among the co-occurring documents could then be utilized to facilitate the learning of topics for both normal documents and short texts. Specifically, the proposed COTM assumed each short text to discuss two topics; one formal topic and one informal topic. The formal topics were represented by both normal documents and short texts, while the informal topics were represented only by the short texts.

Although the informal topics discovered by COTM represent the preferences of all short texts, COTM fails to capture individual user preferences. This is because it implicitly assumes that all short texts are written by the same user or share the same topic probability distribution. Consequently, it can only extract coarse-grained topics shared by all users, such as rude words and mutual judgments [5]. To distinguish between individual preferences, we propose the AOTM as an extension of COTM by incorporating authorships of short texts into the topic generation process. Specifically, AOTM assumes that each author of a short text has a topic distribution over all informal topics. Through this assumption, AOTM naturally learns the preferences of each author, and it also gains improved representations of topics and documents. The superiority of AOTM also lies in its potential usage in a wide range of applications, such as drawing user portraits and predicting user posting behaviors.

### 2.2. Topic models with side information

Side information, such as authorships and tags, are often utilized to enhance topic learning by constructing relationships among documents. For example, the labeled LDA [7] assumed a label distribution for each document. Then it used these label distributions as topic presence/absence indicators to ensure that all topic assignments in a document were limited within its label set. The AuTM [6] associated each document with a group of authors. For each word in the document, it first chose an author and then chose a topic from the corresponding topic distribution for this author. The tag-weighted topic model [26] extended LDA by assuming that each document was associated with a group of observed tags and the weights of the tags constituted the topic proportions of the document.

The side information is also used to divide the whole corpus into document groups, so that the consistency of topics in the whole corpus is broken. Typical examples are the temporal tags, which divide the whole corpus into different time slices, and the geographical tags, which divide the whole corpus into spatially distributed blocks. Existing works in this stream include Cheng et al. [4] and Yang et al. [5], who discovered periodic topics based on dynamic documents. They first split documents into different time slice, and then fitted their original static models on each time slice for online topic learning by transferring hyperparameters from time to time. Ahmed et al. [27] and Qiang et al. [28] incorporated spatial information into a stochastic process to discover location-aware topics. Guo et al. [29,30] assumed that the location of a topic was sampled from a geographical distribution over a set of regions, after which they detected topics in a joint geographical and temporal space.

Finally, the existing works also accommodate the side information (such as user ratings, popularities and class labels) into the generative process of topic models to achieve better performance against many discriminant models. For instance, Blei et al. [31] demonstrated that, the supervised LDA, which utilized the class labels of documents, could outperform the penalized method Lasso in predicting user movie ratings and votes of senators. Wang et al. [32] also took advantage of the class labels. They proposed a class specific topic model to deal with the document classification task in both supervised and semi-supervised framework. Yang et al. [33] considered the named entities in documents, and proposed a named-entity topic model to predict the popularity of news articles.

## 3. AOTM

### 3.1. Model description

We follow the notations used in COTM to describe the newly proposed AOTM. Table 1 summarizes the notations used in AOTM. Specifically, we assume that there are $K$ formal topics underlying $D$ normal documents. Each formal topic $k(k = 1, \ldots, K)$ has a vector of word probabilities $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kV})^\top$ over a dictionary with size $V$, and each document

**Table 1**
Summary of key notations used in the AOTM.

| | |
|---|---|
| | *For the Whole Corpus* |
| $V$ | the number of unique words appearing in all documents |
| $K$ | the number of formal topics |
| $J$ | the number of informal topics |
| $A$ | the number of authors writing short texts |
| $D$ | the number of normal documents |
| $\boldsymbol{\phi}_k$ | the vector of word probabilities for formal topic $k(1 \leqslant k \leqslant K)$ |
| $\boldsymbol{\psi}_j$ | the vector of word probabilities for informal topic $j(1 \leqslant j \leqslant J)$ |
| $\boldsymbol{\xi}_a$ | the vector of informal topic probabilities for author $a(1 \leqslant a \leqslant A)$ |
| | *For the dth Normal Document* |
| $\boldsymbol{\theta}_d$ | the vector of formal topic probabilities in document $d$ |
| $N_d$ | the number of words in document $d$ |
| $C_d$ | the number of short texts under document $d$ |
| $w_{dn}$ | the $n$th observed word in document $d$ |
| $z_{dn}$ | the formal topic represented by the $n$th word in document $d$ |
| | *For the cth Short Text under the dth Normal Document* |
| $a_{dc}$ | the author of the $c$th short text under the $d$th normal document |
| $x_{dc}$ | the formal topic represented by the $c$th short text under the $d$th normal document |
| $y_{dc}$ | the informal topic represented by the $c$th short text under the $d$th normal document |
| $p_{dc}$ | the percentage of formal topic in the $c$th short text under the $d$th normal document |
| $M_{dc}$ | the number of words in the $c$th short text under the $d$th normal document |
| $w_{dcm}$ | the $m$th observed word in the $c$th short text under the $d$th normal document |
| $b_{dcm}$ | the topic of the $m$th word in the $c$th short text under the $d$th normal document |

$d(d = 1, \ldots, D)$ has a vector of topic probabilities $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})^\top$ over $K$ formal topics. Here the normal documents can be news articles or product descriptions, which mainly describe basic facts and have less emotional tendencies or author preferences. Therefore, we do not assume any authors associated with them. On the contrary, the short texts often provide personal opinions based on the facts discussed in the normal documents. Therefore, we consider the authorships of such short texts.

Specifically, we assume that there are $J$ informal topics underlying all short texts, and each informal topic $j(j = 1, \ldots, J)$ has a vector of word probabilities $\boldsymbol{\psi}_j = (\psi_{j1}, \ldots, \psi_{jV})^\top$ over the dictionary. We assume that, in total, $A$ authors are writing the corpus of short texts. Each author $a(a = 1, \ldots, A)$ has its own vector of topic probabilities $\boldsymbol{\xi}_a = (\xi_{a1}, \ldots, \xi_{aJ})^\top$ over all $J$ informal topics. We assume that for the $c$th short text following a normal document $d$, we can observe its author $a_{dc}$. Following the assumption of COTM, we also assume that one formal topic $x_{dc}$ and one informal topic $y_{dc}$ are used to characterize the semantic information of the short text, where $x_{dc}$ is related to $\boldsymbol{\theta}_d$ and $y_{dc}$ is related to $\boldsymbol{\xi}_{a_{dc}}$. Finally, a probability $p_{dc}$ is used to describe the percentage of formal topics discussed in the short text.

Based on the above assumptions, we describe the generative process for normal documents and short texts as follows: The graphical representation of AOTM is illustrated in Fig. 1.

1. For each formal topic $k \in \{1, \ldots, K\}$:
   (a) Generate $\boldsymbol{\phi}_k$ independently from a homogeneous Dirichlet distribution with parameter $\beta : \boldsymbol{\phi}_k \sim Dir(\beta)$.
2. For each normal document $d \in \{1, 2, \ldots, D\}$:
   (a) Generate topic probabilities $\boldsymbol{\theta}_d$ from a homogeneous Dirichlet distribution with parameter $\alpha : \boldsymbol{\theta}_d \sim Dir(\alpha)$;
   (b) For the $n$th word in a normal document $d, n \in \{1, 2, \ldots, N_d\}$:
      i. Choose a topic $z_{dn}$ from the $K$ formal topics with probabilities given by $\boldsymbol{\theta}_d : z_{dn} \sim Multi(\boldsymbol{\theta}_d)$;
      ii. Choose a word $w_{dn}$ from the dictionary with probabilities given by $\boldsymbol{\phi}_{z_{dn}} : w_{dn} \sim Multi\left(\boldsymbol{\phi}_{z_{dn}}\right)$.
3. For each informal topic $j \in \{1, \ldots, J\}$:
   (a) Generate $\boldsymbol{\psi}_j$ independently from a homogeneous Dirichlet distribution with parameter $\omega : \boldsymbol{\psi}_j \sim Dir(\omega)$.
4. For each author $a \in \{1, \ldots, A\}$:
   (a) Generate topic probabilities $\boldsymbol{\xi}_a$ from a homogeneous Dirichlet distribution with parameter $\epsilon : \boldsymbol{\xi}_a \sim Dir(\epsilon)$;
5. For the $c$th short text associated with a normal document $d, c \in \{1, 2, \ldots, C_d\}$, we can observe its author $a_{dc}$:
   (a) Choose the association probability $p_{dc}$ from a beta distribution with parameter $\gamma : p_{dc} \sim Beta(\gamma, \gamma)$;
   (b) Choose a topic $x_{dc}$ from $K$ formal topics with probabilities given by $\boldsymbol{\theta}_d : x_{dc} \sim Multi(\boldsymbol{\theta}_d)$;
   (c) Choose a topic $y_{dc}$ from $J$ informal topics with probabilities given by $\boldsymbol{\xi}_{a_{dc}} : y_{dc} \sim Multi\left(\boldsymbol{\xi}_{a_{dc}}\right)$;
   (d) For the $m$th word in the short text, $m \in \{1, 2, \ldots, M_{dc}\}$:
      i. Generate a topic indicator $b_{dcm}$ with probability given by $p_{dc} : b_{dcm} \sim Bernoulli(p_{dc})$;
      ii. If $b_{dcm} = 1$, the word is chosen with probabilities under the formal topic: $w_{dcm} \sim Multi\left(\boldsymbol{\phi}_{x_{dcm}}\right)$.

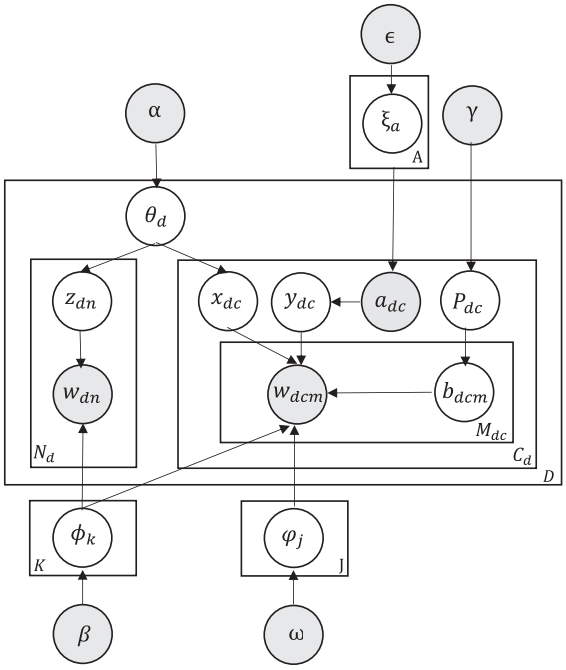


**Fig. 1.** Graphical representation of the AOTM.

iii. If $b_{dcm} = 0$, the word is chosen with probabilities under the informal topic: $w_{dcm} \sim Multi\left(\boldsymbol{\psi}_{y_{dcm}}\right)$;

From the perspective of the generative process, AOTM and COTM have two main differences. First, AOTM considers the authorship of each short text and assumes that each author can have a vector of topic probabilities $\xi_a (1 \leqslant a \leqslant A)$ over all informal topics. However, in COTM, it assumes that all short texts share the same probability distribution $\xi$ over all informal topics. Second, AOTM ensures that the informal topic represented by each short text is author-related. In other words, the informal topic $y_{dc}$ is generated from a multinomial distribution determined by its author's topic probabilities $\xi_{a_{dc}}$. However, in COTM, $y_{dc}$ is related to the general topic probability $\xi$. This flexible assumption in AOTM makes topic representations in short texts written by the same author more focused.

It is also noteworthy that when all short texts are written by the same author, AOTM degenerates into COTM. However, in practice, this is seldom the case. In fact, we often encounter short texts from a large number of authors. For example, the numbers of authors in three real datasets used in this study are all larger than 100,000; see Table 2 for more details. Another possible solution is to build multiple COTMs, with each model considering all normal documents but short texts from one single author. However, given the fact that the number of authors is often huge, this solution is computationally expensive. Therefore, it becomes infeasible in practice. Even if the number of authors is small, building multiple COTMs separately cuts off the possible content similarities among users. Moreover, with fewer short texts used in each COTM, the data sparsity problem would become more severe. On the contrary, AOTM utilizes all short texts in topic modeling, and it characterizes individual preferences by incorporating the authorship information. This makes AOTM more flexible and closer to reality. The good performance of AOTM in topic learning and topic representations in documents is also empirically verified through extensive numerical experiments. More importantly, AOTM serves as a good starting point for the further investigation of individual user preferences, such as drawing user portraits and predicting user posting behaviors.

### 3.2. Model estimation

Given the generative process shown in Fig. 1, we present a Gibbs sampling algorithm for model estimation. We first define some notations used in AOTM. For normal documents, let $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dN_d})^\top$ denote the collection of topic indicators in document $d$, and $\boldsymbol{z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_D\}$. For all the $C_d$ short texts associated with a normal document $d$, let $\boldsymbol{b}_{dc} = (b_{dc1}, \ldots, b_{dcM_{dc}})^\top$ denote the collection of formal topic indicators, and let $\boldsymbol{b}_d = \{\boldsymbol{b}_{d1}, \ldots, \boldsymbol{b}_{dC_d}\}$. Let $\boldsymbol{P}_d = \left(p_{d1}, \ldots, p_{dC_d}\right)^\top$ denote the collection of percentages of formal topics, and $\boldsymbol{A}_d = (a_{d1}, \ldots, a_{dC_d})^\top$ denote the collection of authors. Let $\boldsymbol{x}_d = (x_{d1}, \ldots, x_{dC_d})^\top$ and $\boldsymbol{y}_d = \left(y_{d1}, \ldots, y_{dC_d}\right)^\top$ denote the collection of selected formal topics or informal topics, respectively. Next, let $\boldsymbol{b} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_D\}, \boldsymbol{P} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_D\}, \boldsymbol{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_D\}, \boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D\}$, and $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_D\}$ for all short texts. Moreover, let

**Table 2**
Basic statistics of NetEase and JD datasets

| Dataset | NetEase | Tencent | JD |
|---|---|---|---|
| # of docs ($D$) | 476,940 | 87,083 | 6,741 |
| Avg. doc len ($\bar{N}$) | 333.62 | 360.38 | 681.54 |
| # of comments | 49,351,902 | 977,859 | 1,731,849 |
| Avg. comm len ($\bar{M}$) | 7.19 | 8.95 | 22.70 |
| # of words ($V$) | 161,045 | 103,427 | 117,731 |
| # of authors ($A$) | 1,343,249 | 792,091 | 1,006,153 |

$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}, \boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}, \boldsymbol{\Psi} = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K\}$, and $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_A\}$. Finally, let $\boldsymbol{w}$ represent all words in normal documents and short texts.

Given the observed data $(\boldsymbol{w}, \boldsymbol{A})$ and all the hyperparameters $(\alpha, \beta, \gamma, \omega, \epsilon)$, we can derive the full posterior distribution according to the generative process of AOTM presented in Fig. 1 as follows:

$$
\begin{aligned}
f(\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{P}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Xi} | \boldsymbol{w}, \boldsymbol{A}, \alpha, \beta, \gamma, \omega, \epsilon) \propto & \left\{ \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\beta-1} \right\} \left\{ \prod_{j=1}^{J} \prod_{v=1}^{V} \psi_{jv}^{\omega-1} \right\} \left\{ \prod_{a=1}^{A} \prod_{j=1}^{J} \xi_{aj}^{\epsilon-1} \right\} \times \left\{ \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1} \right\} \left\{ \prod_{d=1}^{D} \prod_{n=1}^{N_d} \theta_{d,z_{dn}} \phi_{z_{dn},w_{dn}} \right\} \\
& \times \left\{ \prod_{d=1}^{D} \prod_{c=1}^{C_d} p_{dc}^{\gamma-1} (1-p_{dc})^{\gamma-1} \right\} \left\{ \prod_{d=1}^{D} \prod_{c=1}^{C_d} \theta_{d,x_{dc}} \xi_{a_{dc} y_{dc}} \right\} \times \left\{ \prod_{d=1}^{D} \prod_{c=1}^{C_d} \prod_{m=1}^{M_{dc}} \left( p_{dc} \phi_{x_{dc},w_{dcm}} \right)^{b_{dcm}} \right\} \\
& \times \left\{ \prod_{d=1}^{D} \prod_{c=1}^{C_d} \prod_{m=1}^{M_{dc}} \left[ (1-p_{dc}) \psi_{y_{dc},w_{dcm}} \right]^{1-b_{dcm}} \right\}.
\end{aligned}
$$

From the full posterior distribution presented in (1), we can easily derive the conditional posterior distributions for $\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Xi}$, and $\boldsymbol{P}$, which are all Dirichlet and conjugate with their priors. Therefore, we can first integrate out these variables, after which we can develop a collapsed Gibbs sampling algorithm for the other variables $(\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{y})$. The Gibbs sampling procedure for most variables is similar to that in COTM, except for those associated with $\boldsymbol{\Xi}$. Below, to save space, we only discuss the sampling steps associated with $\boldsymbol{\Xi}$.

For $\boldsymbol{\xi}_a (a = 1, \ldots, A)$, its full conditional posterior distribution is expressed as follows:

$$
f(\boldsymbol{\xi}_a | \cdot) \propto \prod_{j=1}^{J} (\xi_{aj})^{h_{aj}+\epsilon-1}. \tag{2}
$$

Next, by integrating out $\boldsymbol{\xi}_a (a = 1, \ldots, A)$, we can obtain the following result:

$$
\int f(\boldsymbol{\Xi} | \cdot) d\boldsymbol{\Xi} \propto \prod_{a=1}^{A} \frac{\prod_{j=1}^{J} \Gamma(h_{aj} + \epsilon)}{\Gamma\left\{ \sum j = 1^{J} (h_{aj} + \epsilon) \right\}}. \tag{3}
$$

We can integrate out $\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$, and $\boldsymbol{P}$ similarly. The full posterior distribution of $(\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{y})$ can then be simplified as follows:

$$
\begin{aligned}
f(\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{y} | \boldsymbol{w}, \boldsymbol{A}, \alpha, \beta, \gamma, \omega, \epsilon) = & \int f(\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{P}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Xi} | \boldsymbol{w}, \alpha, \beta, \gamma, \epsilon) d\boldsymbol{\Theta} d\boldsymbol{\Phi} d\boldsymbol{\Psi} d\boldsymbol{\Xi} d\boldsymbol{P} \\
\propto & \left\{ \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma\left( l_{kv}^{(2)} + g_{kv}^{(2)} + \beta \right)}{\Gamma\left\{ \sum_{v=1}^{V} \left( l_{kv}^{(2)} + g_{kv}^{(2)} + \beta \right) \right\}} \right\} \left\{ \prod_{j=1}^{J} \frac{\prod_{v=1}^{V} \Gamma\left( g_{jv}^{(3)} + \omega \right)}{\Gamma\left\{ \sum_{v=1}^{V} \left( g_{jv}^{(3)} + \omega \right) \right\}} \right\} \\
\propto & \left\{ \prod_{d=1}^{D} \frac{\prod_{i=1}^{K} \Gamma\left( l_{dk}^{(1)} + g_{dk}^{(1)} + \alpha \right)}{\Gamma\left\{ \sum_{k=1}^{K} \left( l_{dk}^{(1)} + g_{dk}^{(1)} + \alpha \right) \right\}} \right\} \left\{ \prod_{a=1}^{A} \frac{\prod_{j=1}^{J} \Gamma(h_{aj} + \epsilon)}{\Gamma\left\{ \sum_{j=1}^{J} (h_{aj} + \epsilon) \right\}} \right\} \times \left\{ \prod_{d=1}^{D} \prod_{c=1}^{C_d} \frac{\Gamma\left( s_{dc}^{(1)} + s_{dc}^{(2)} + \gamma \right)}{\Gamma\left( s_{dc}^{(1)} + \gamma \right) \Gamma\left( s_{dc}^{(2)} + \gamma \right)} \right\}.
\end{aligned} \tag{4}
$$

The notations used in (4) simply follow those used in COTM. Based on (4), a collapsed Gibbs sampling algorithm is developed by updating $(\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{y})$ one by one. The detailed sampling procedure can be found in the work of Yang et al. [5]. After model convergence, we can derive the posterior estimates for $(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Xi}, \boldsymbol{P})$ as follows:

$$\begin{aligned} \hat{\theta}_{dk} &= \frac{l_{dk}^{(1)}+g_{dk}^{(1)}+\alpha}{l_{d\cdot}^{(1)}+g_{d\cdot}^{(1)}+K\alpha}, \hat{\phi}_{kv} = \frac{l_{kv}^{(2)}+g_{kv}^{(2)}+\beta}{l_{k\cdot}^{(2)}+g_{k\cdot}^{(2)}+V\beta}, \\ \hat{\psi}_{jv} &= \frac{g_{jv}^{(3)}+\beta}{g_{j\cdot}^{(3)}+V\beta}, \hat{\xi}_{aj} = \frac{h_{aj}+\epsilon}{h_{a\cdot}+J\epsilon}, \hat{p}_{dc} = \frac{s_{dc}^{(1)}+\gamma}{s_{dc}^{(1)}+s_{dc}^{(2)}+2\gamma}. \end{aligned} \tag{5}$$

## 4. Evaluating the performance of AOTM

In this section, we present a variety of experiments to evaluate the performance of AOTM in exploring individual user preferences. The experiments are conducted on three large real-world co-occurring normal documents and short texts to answer four questions:

- How effective is AOTM in learning topics?
- How effective is AOTM in representing documents?
- How effective is AOTM in extracting individual user preferences?
- How can AOTM be further utilized to predict user behaviors?

By answering the first two questions, we can validate the abilities of AOTM in extracting semantic representations from texts, which serve as foundations for downstreaming tasks to answer the last two questions.

### 4.1. Data description and experimental setup

**Dataset.** The effectiveness of AOTM is evaluated using real-world datasets parsed from two news article platforms and one e-commerce website. The basic statistics of the three datasets are listed in Table 2, including the numbers of normal documents and short texts, the average lengths of normal documents and short texts, the number of unique words, and the number of unique authors. Below, we introduce each dataset in details.

The *NetEase* (NASDAQ listed) collection includes news articles, corresponding user comments, and authorships of comments crawled from one of the most popular Chinese news publishing platforms.[1] The authorship of each comment is tagged using a unique user ID, which is assigned by the platform to every website user. All the texts were published between December 1, 2016 and July 1, 2018.

The *Tencent* (HKEX listed) corpus is collected from another Chinese news platform,[2] including the same type of data as in the NetEase collection. All the texts were published between June 1, 2018 and March 1, 2019.

The *JD* (NASDAQ listed) collection includes product descriptions for electronic products and consumer reviews crawled from a well-known Chinese e-commerce website.[3] Because the product advertisements and detailed descriptions in JD are all posted as pictures, we downloaded all the advertising pictures for each product. Then we parsed text from the pictures using a commercial optical character recognition product.[4] All the products were on sale before November 1, 2018, and we crawled the latest 1500 comments of each product, if there were enough. The authorships of the comments are also tagged using unique IDs assigned by the website.

The raw texts in all datasets are mainly written in Chinese, and standard preprocessing procedures were performed to obtain a clean text corpus. First, we erased non-Chinese characters and punctuations, after which we converted traditional Chinese characters to simplified Chinese characters. Second, we performed word segmentation using an open source package *Jieba*.[5] Third, we removed stop words, low frequency words, and normal documents with no following short texts. Finally, we erased short texts whose authors posted fewer than 10 short comments in NetEase, 5 short comments in Tencent, and 2 short comments in JD.

**Baselines.** To evaluate the performance of AOTM, we first compare AOTM with COTM [5] because COTM has been verified to outperform many well-accepted baselines, including LDA [1], BTM [4], and the pseudo-document topic model [14]. We then compare AOTM with several state-of-the-art methods. They are, the embedding-based topic model (ETM) [25], the generalized Pólya urn (GPU) based and poisson-based Dirichlet multinomial mixture model (GPU-PDMM) [24], the word network topic model (WNTM) [34], and the aggregated topic model (AggTM) [35]. We also carry out comparisons with the AuTM [36] because both AuTM and AOTM incorporate authorship as side information. Finally, the LDA is also considered as a baseline. The specific details of all compared methods are listed as follows:

[1] http://news.163.com/
[2] http://news.qq.com
[3] http://www.jd.com/
[4] http://ai.baidu.com/tech/ocr/general
[5] https://pypi.org/project/jieba/

- *COTM* [5]: The *COTM* is applied to co-occurring normal documents and short texts. The model directly exploits the co-occurring structure and has been verified to outperform many baseline methods in learning more prominent and comprehensive topics and getting better topic representations of documents.
- *ETM* [25]: The *ETM* is applied to short texts. The model incorporates auxiliary semantic knowledge from word embeddings to aggregate short texts and improves topic inference.
- *GPU-PDMM* [24] The *GPU-PDMM* is applied to short texts. The model first incorporates the Poisson distribution of the number of topics to relax one topic assumption in the Dirichlet multinomial mixture, after which it utilizes the Pólya urn model to incorporate word embeddings as background knowledge to improve topic inference.
- *AggTM* [35]: The *AggTM* is applied to short texts. The model employs Jensen-Shannon divergence to aggregate the captured topics.
- *WNTM* [34]: The *WNTM* is applied to short texts. The model infers the latent topics of short texts explicitly based on a word co-occurrence network.
- *AuTM* [36]: The *AuTM* is applied to short texts. There is only one author for each short text.
- *LDA* [1]: The *LDA* is applied to normal documents as a baseline for evaluating document representations.

For ETM, GPU-PDMM, and WNTM, we utilize their Java implementation presented by Qiang et al. [37]. For methods that utilize pre-trained word embeddings to incorporate external word correlation knowledge, we choose the Tencent AILab ChineseEmbedding trained by Song et al. [38] as their auxiliary inputs. This embedding provides 200-dimension vector representations for 8 million Chinese words and phrases. However, it is noteworthy that there are 9.31%, 6.33%, and 4.20% words from the NetEase, Tencent, and JD datasets, respectively, that do not appear in the embedding space. We then omit them from the original dataset, which may exert negative influences on the performance of corresponding methods. COTM and AOTM are implemented in C++.

**Hyperparameter settings.** For AOTM and COTM, all hyperparameters are selected using grid search to achieve the best performance. For the other competitiors, we choose the hyperparameters according to their original papers. For AOTM and COTM, we try different combinations of the number of formal topics and informal topics. The coherence score is then calculated for formal topics only, informal topics only, as well as all topics. For ETM, GPU-PDMM, AggTM, and WNTM, because they are only applied on short texts, we set their number of topics as the number of informal topics in AOTM and COTM and then calculate the coherence score. The perplexity is computed on normal documents or short texts under different methods in the test set.

### *4.2. Experimental design and evaluation metrics*

**Experimental design.** We design the following four tasks corresponding to the four questions mentioned above to demonstrate the effectiveness of AOTM:

- *Topic quality:* The quality of topics is evaluated through the coherence score (CS).
- *Document representation:* The quality of document representations is evaluated through perplexity (PP).
- *User portraits:* We explore topic semantics and user-specific topic distributions to evaluate the performance of AOTM in drawing user portraits.
- *Posting behavior prediction:* We evaluate the performance of AOTM by predicting the posting behavior of users given a normal document.

**Evaluation metrics.** All methods are compared from two perspectives: the quality of learned topics, which is measured by topic coherence, and the quality of the topic representation of documents, which is measured by perplexity. Below, we shall first describe the two measures and then present the corresponding results.

*Coherence Score (CS).* The coherence score is a commonly used measure for evaluating the quality of learned topics [39]. It focuses on the word probabilities under each topic, i.e., the $\phi_k$'s and $\psi_j$'s in AOTM. The basic idea of the coherence score is that if a topic is well-generated, words with high frequencies under this topic tend to co-occur within the same document. To formulate this idea, for any formal topic $k$ or informal topic $j$ and its top $L$ words $(w_1, w_2, \ldots, w_L)^\top$ ordered by the corresponding word probabilities, the coherence score is defined as follows:

$$CS = \sum_{l=2}^{L} \sum_{l'=1}^{l-1} \log \frac{F(w_l, w_{l'}) + 1}{F(w_{l'})}, \tag{6}$$

where $F(w)$ represents the number of documents including word $w$, and $F(w, w')$ represents the number of documents including both words $w$ and $w'$. A higher coherence score implies a better topic model.

*Perplexity (PP).* perplexity is used to measure the predictive power of topic models. It is based on the joint likelihood of appearing words. In this study, 80% normal documents and corresponding short texts are taken as the training set. Models trained on the training set are then used to calculate the likelihood of the rest 20% hold-out test set. The perplexity of each model on the test set can then be calculated accordingly. The definition of perplexity is described as follows:

$$PP = \exp\left\{-\frac{\sum_{d\in\boldsymbol{C}_{\text{test}}}\log p(\boldsymbol{w}_d)}{\sum_{d\in\boldsymbol{C}_{\text{test}}}N_d}\right\}, \tag{7}$$

where $\boldsymbol{C}_{\text{test}}$ denotes the collection of test documents and $\boldsymbol{w}_d$ denotes the collection of all words appearing in document $d$. A lower perplexity implies a better topic model.

### 4.3. Experimental results

#### 4.3.1. Results of coherence score (for Q1)

Table 3 presents the results of the coherence score under different models for the NetEase, Tencent, and JD datasets. We mainly focus on the results of the NetEase dataset, which are similar to those of the Tencent and JD datasets. We first compare AOTM with COTM by evaluating the coherence scores for formal topics only, informal topics only, as well as all topics. The results show that AOTM achieves higher coherence scores in all these comparisons. These findings imply the advantages of AOTM in learning topics of higher quality.

The other competitors, including ETM, GPU-PDMM, AggTM, WNTM and AuTM, are all applied to short texts only. As for these methods, GPU-PDMM achieves the closest performance to AOTM. It is notable that GPU-PDMM borrows external knowledge of word correlations from auxiliary embeddings. Therefore, it outperforms both AOTM and COTM in the learning of informal topics. However, the formal topics in AOTM can be improved by taking advantage of information in short texts. Thus we observe AOTM-F performs better than GPU-PDMM. Additionally, AOTM outperforms WNTM in most experimental settings. The only exception is in the JD dataset, in which the coherence score of WNTM is larger than that of AOTM-I. However, even in this case, when evaluating on all topics, AOTM still achieves higher coherence scores than WNTM. Finally, AOTM outperforms AggTM and AuTM in all experimental settings. All the above results suggest that by modeling the co-occurring structure and incorporating author information together, AOTM is superior to most methods that only consider partial information or utilize auxiliary word embeddings.

#### 4.3.2. Results of perplexity (for Q2)

Besides the coherence score, we also compare model performance using perplexity to evaluate the quality of document representation and the predictive power of different models. Table 4 presents the results of perplexity under different models for the NetEase, Tencent, and JD datasets. From these results, we draw the following conclusions. First, when evaluating perplexity on normal documents, both AOTM and COTM significantly outperform LDA. This demonstrates that by gaining knowledge from normal documents and short texts, the proposed co-occurring structure successfully enhances topic inference. Additionally, AOTM achieves better performance than COTM, which can be attributed to the efficiency of AOTM in incorporating authorships into the model.

When evaluating perplexity on short texts, we omit the method WNTM because it does not model the generative process of documents. Therefore, it is not suitable for calculating perplexity [34]. Among all the methods, AOTM and GPU-PDMM achieve better perplexity performance than the other methods. For GPU-PDMM, its good performance in representing documents results from the high quality of topics learned by this model; see the results of the coherence score in Table 3 for evidence. For AOTM, each short text is represented by both a formal topic and an informal topic. Although AOTM does not obtain a significantly better quality of informal topics than GPU-PDMM, the quality of formal topics achieved by AOTM has been improved. Therefore, when compared with GPU-PDMM, AOTM shows comparable representation performance in short texts. Besides, compared with COTM, AOTM incorporates side information for modeling short texts. This mechanism provides an advantage for AOTM in perplexity measurements, especially when texts are short or have just a few observed words [6].

#### 4.3.3. Drawing user portraits (for Q3)

In AOTM, we assume that each author has a vector of probabilities on informal topics (i.e., the $\xi_a$). Individual author preferences can then be further investigated. Specifically, by analyzing the word probabilities under each informal topic (i.e., the $\psi_j$), we can obtain the meaning of each topic. Next, by observing the distribution of $\xi_{aj}$ on a specific informal topic $j$ among all authors, we can find authors that mainly focus on this topic. Then we label these authors using this topic.

To illustrate this idea, consider the NetEase dataset as an example. Table 5 shows four example informal topics extracted from the NetEase dataset when setting $K = 100$ and $J = 20$. For each informal topic, we report the top ten words with high probabilities. Based on these probabilities, we summarize the meanings of these topics as "Destitution", "Responsibility", "Encouragement", and "Conspiracy".

To investigate individual user preferences on these four topics further, Fig. 2 presents the histograms of author probabilities $\xi_{aj}$ among all $A$ authors for each informal topic $j$. As shown in Fig. 2, in general, the topics "Destitution" and "Conspiracy" have been discussed more by authors in the NetEase website than the topics "Responsibility" and "Encourage". Specifically, most authors discuss the topic "Destitution" with probabilities between 0.05 and 0.2, and there also exist authors who discuss this topic extensively with probabilities higher than 0.35. The probability distribution for the topic "Responsibility" shows a right-skewed pattern, indicating that the majority of authors talk about this topic with relatively lower probability.

**Table 3**
Coherence scores under different models for NetEase, Tencent, and JD datasets. "AOTM-F" and "COTM-F" indicate the coherence scores of formal topics only, "AOTM-I" and "COTM-I" indicate the coherence scores of informal topics only, and "AOTM" and "COTM" indicate the coherence scores of both formal topics and informal topics. "GPDMM" represents the method "GPU-PDMM". Methods with top-3 performances are marked with asterisks, i.e., "*", "**", and "***" indicate the "first", "second", and "third" ranks, respectively.

| Model | $K = 100, J = 20$ | | $K = 100, J = 50$ | | $K = 200, J = 100$ | |
|---|---|---|---|---|---|---|
| | Top5 | Top10 | Top5 | Top10 | Top5 | Top10 |
| *NetEase* | | | | | | |
| AOTM | −63.88** | −226.80** | −73.28** | −252.52 | −97.33 | −270.24*** |
| AOTM-F | −60.27* | −220.46* | −64.72* | −232.75* | −85.24* | −243.31* |
| AOTM−I | −81.98 | −258.50 | −90.39 | −292.08 | −121.52 | −324.11 |
| COTM | −71.49 | −238.15 | −86.38 | −259.61 | −104.87 | −297.30 |
| COTM-F | −68.63 | −229.21*** | −74.22*** | −239.31** | −94.09** | −264.53*** |
| COTM-I | −85.76 | −282.87 | −110.71 | −300.19 | −126.43 | −362.84 |
| ETM | -96.31 | −294.67 | −118.22 | −314.79 | −137.24 | −325.47 |
| GPDMM | −67.21*** | −238.53 | −75.28 | −249.80*** | −97.21*** | −299.64 |
| AggTM | −113.48 | −347.28 | −133.05 | −407.43 | −162.92 | −401.98 |
| WNTM | −96.83 | −341.80 | −127.33 | −382.56 | −136.75 | −370.10 |
| AuTM | −157.87 | −461.99 | −172.61 | −499.83 | −224.17 | −579.14 |
| *Tencent* | | | | | | |
| AOTM | −74.27 | −229.39*** | −79.17 | −260.66 | −100.76 | −307.61 |
| AOTM-F | −71.26** | −219.53* | −70.02* | −244.62* | −91.29** | −296.12** |
| AOTM-I | −89.32 | −278.68 | −97.46 | −292.74 | −119.70 | −330.59 |
| COTM | −76.08 | −246.04 | −89.01 | −273.50 | −106.61 | −331.12 |
| COTM-F | −73.39*** | −233.81 | −78.89*** | −245.32** | −96.97*** | −302.32*** |
| COTM-I | −89.55 | −307.19 | −109.26 | −329.87 | −125.90 | −388.72 |
| ETM | −88.06 | −327.96 | −109.78 | −296.65 | −137.01 | −413.89 |
| GPDMM | −65.91* | −221.34** | −74.92** | −246.35*** | −88.34* | −290.09* |
| AggTM | −102.46 | −331.80 | −114.93 | −355.16 | −152.70 | −364.93 |
| WNTM | −94.29 | −320.17 | −105.39 | −326.69 | −139.82 | −358.73 |
| AuTM | −136.65 | −413.65 | −155.19 | −468.37 | −202.54 | −561.53 |
| *JD* | | | | | | |
| AOTM | −73.47*** | −246.10*** | −82.06 | −271.35*** | −90.98 | −318.10*** |
| AOTM-F | −73.21** | −245.48** | −80.39** | −262.81** | −85.36** | −317.51** |
| AOTM-I | −87.08 | −269.20 | −85.40 | −288.42 | −102.21 | −327.81 |
| COTM | −85.39 | −297.55 | −84.27 | −293.20 | −95.4 | −358.62 |
| COTM-F | −83.53 | −295.74 | −81.30*** | −283.46 | −87.73*** | −348.21 |
| COTM-I | −94.71 | −306.61 | −90.22 | −312.69 | −110.74 | −379.44 |
| ETM | −87.03 | −303.78 | −105.91 | −284.06 | −131.53 | −349.57 |
| GPDMM | −61.20* | −227.91* | −70.58* | −234.22* | −79.03* | −249.38* |
| AggTM | −99.03 | −313.57 | −102.00 | −324.82 | −133.77 | −349.12 |
| WNTM | −75.52 | −256.35 | −82.50 | −282.38 | −97.62 | −320.94 |
| AuTM | −125.87 | −378.84 | −143.12 | −431.19 | −193.60 | −489.86 |

As for the topic "Encourage", authors can be divided into two groups, centering on probabilities 0.01 and 0.05, respectively. Finally, the topic "Conspiracy" has been discussed a lot as its probability distribution shows a left-skewed pattern.

Table 6 shows four example topics extracted from the JD dataset with $K = 100$ and $J = 20$. By summarizing the top ten words with high probabilities, the four informal topics are named as "After-sale", "Logistics", "Price", and "Positive". Fig. 3 presents the corresponding histograms of author probabilities $\xi_{aj}$ under each informal topic. It is obvious that different authors talk about the same topic with different probabilities. In general, the authors on the JD website care more about "After-sale" and "Logistics" than "Price". The topic "Positive" describes the users' attitude toward the products and services in JD. As shown in Fig. 3, most authors discuss their positive attitude with probabilities lying between 0.08 and 0.18.

With the probability distributions among all authors shown in Figs. 2 and 3, we can further draw portraits for each individual user. Specifically, for each informal topic, we can find the authors that heavily discuss this topic and then label these authors using the topic name. For each specific topic $j$, we can first rank all authors based on their topic probabilities $\xi_{aj}$'s, and then label the top 20% authors using the topic name. Under this definition, an author may have multiple topic labels or no topic label. Fig. 4 shows the distributions of the number of labels per author in the NetEase and JD datasets. We can find that in both datasets, more than 50% authors have one or two labels. Additionally, the percentages of authors with more than two labels decrease drastically. Finally, in both datasets, there also exist authors with no labels, suggesting that these authors have no special preferences toward any specific informal topic.

#### *4.3.4. Predicting user posting behaviors (for Q4)*

Investigating and characterizing user online posting behaviors is a popular research topic in both academic and industrial fields [40–42]. In this subsection, we illustrate the potential usage of predicting user posting behaviors from the learned formal topics and individual author preferences in AOTM.

**Table 4**
Perplexities under different models for the NetEase, Tencent, and JD datasets. "GPDMM" represents the "GPU-PDMM" method. Methods with the top 3 performances in short texts are marked with asterisks, i.e., "*", "**", and "***," indicating "first", "second," and "third" ranks, respectively.

| Data | | Model | $K = 100$ $J = 20$ | $K = 100$ $J = 50$ | $K = 200$ $J = 100$ |
|---|---|---|---|---|---|
| NetEase | Normal | AOTM | 4352.0 | 4272.3 | 4026.1 |
| | | COTM | 4787.2 | 4422.9 | 4197.4 |
| | | LDA | 4958.5 | | 4814.3 |
| | Short | AOTM | 4585.8** | 4408.4** | 4182.5* |
| | | COTM | 5083.5 | 4947.5*** | 5025.3 |
| | | ETM | 4780.6*** | 5161.7 | 4932.7*** |
| | | GPDMM | 4428.5* | 4260.9* | 4681.7** |
| | | AggTM | 5735.4 | 5978.5 | 6158.6 |
| | | AuTM | 5642.9 | 5517.8 | 5928.4 |
| Tencent | Normal | AOTM | 1895.8 | 1746.2 | 1807.9 |
| | | COTM | 2148.2 | 2026.3 | 1939.6 |
| | | LDA | 2364.8 | | 2147.4 |
| | Short | AOTM | 2583.5* | 2426.8** | 2254.5** |
| | | COTM | 3062.8*** | 2687.0*** | 2615.4*** |
| | | ETM | 3294.0 | 3176.8 | 3235.9 |
| | | GPDMM | 2731.0** | 2212.0* | 2183.5* |
| | | AggTM | 4248.3 | 3991.8 | 4405.7 |
| | | AuTM | 3830.7 | 3717.8 | 4012.5 |
| JD | Normal | AOTM | 3076.4 | 3032.9 | 2856.1 |
| | | COTM | 3386.2 | 3192.4 | 3141.3 |
| | | LDA | 3704.0 | | 3647.4 |
| | Short | AOTM | 3264.7* | 3147.5** | 3014.6** |
| | | COTM | 3920.1*** | 4057.8 | 3832.9 |
| | | ETM | 4208.7 | 3872.6*** | 3354.1*** |
| | | GPDMM | 3528.9** | 3069.7* | 2910.5* |
| | | AggTM | 5379.2 | 5016.5 | 5276.5 |
| | | AuTM | 5475.1 | 5227.5 | 5149.1 |

**Table 5**
Four example informal topics extracted from the NetEase dataset

| No. | Topic Meaning | Top Words with High Probabilities |
|---|---|---|
| 1 | Destitution | mortgage, poor, life, cold blood, human rights, privilege, money, reality, illegal, justice |
| 2 | Responsibility | courage, responsibility, helpful, society, punish, duty, law, onlooking, morality, mandatory |
| 3 | Encouragement | cheer up, support, bless, justice, congratulate, solute, endorse, thumbs-up, greetings, grateful |
| 4 | Conspiracy | official, rigged, ulterior, intrigues, block, suspect, government, deception, conspiracy, forbidden |

In AOTM, normal documents and short texts are modeled in a unified framework. Therefore, we can find a relationship between the formal topics underlying both normal documents and short texts with the informal topics underlying only short texts. This relationship can describe the co-occurrence probability of a formal topic paired with an informal topic, which we call "closeness degree" in the following. On the other hand, AOTM helps investigate user preferences on informal topics by using $\xi_a$. Next, based on $\xi_a$ and treating the closeness degree as a bridge between formal topics and informal topics, we can exploit user preferences toward formal topics and further predict user posting behaviors. Below, we would first describe the predicting procedure in details, and then we present some experimental results.

Specifically, the predicting procedure consists of three steps.

1. **Calculate the closeness degree.** For a normal document $d$, recall $\theta_d$ describes its distribution among $K$ formal topics. For the $c$th short text, $1 - p_{dc}$ describes the probability of informal topic $y_{dc}$. Next, by summarizing all $C_d$ short texts, we calculate $\zeta_{dj} = \frac{1}{C_d}\sum_{c=1}^{C_d}(1 - p_{dc})I(y_{dc} = j)$ to describe the averaged probability of informal topic $j$ associated with normal document $d$. Then we use $\theta_{dk}\zeta_{dj}$ to measure the closeness of formal topic $k$ and informal topic $j$ in document $d$. Finally, by summarizing all normal documents, we define $\omega_{kj} = \frac{1}{D}\sum_{d=1}^{D}(\theta_{dk}\zeta_{dj})$ as the closeness degree between formal topic $k$ and informal topic $j$.
2. **Calculate preference toward formal topics.** For each author $a$, recall $\xi_{aj}$ describes its distribution on the informal topic $j$. By combining $\xi_{aj}$ and $\omega_{kj}$, we can then compute author $a$'s probability on formal topic $k$, i.e., $\pi_{ak} = \xi_{aj}\omega_{kj}$. Then $\pi_{ak}$ can reflect the preference of author $a$ toward the formal topic $k$.

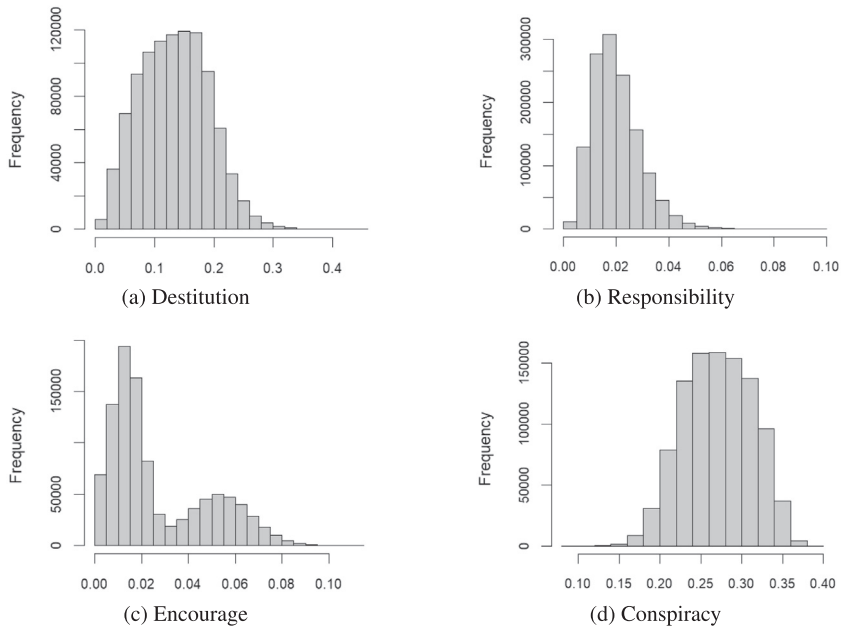


**Fig. 2.** Histograms of author probabilities for four example informal topics in the NetEase dataset.

**Table 6**
Four example informal topics extracted from the JD dataset

| No. | Topic Meaning | Top Words with High Probabilities |
|---|---|---|
| 1 | After-sale | seller, official website, customer service, after-sale, receipt, service, warranty, bad, value, fulfill |
| 2 | Logistics | package, deliver, express, fast, logistics, Shunfeng, online shopping, vacuum, sealed, courier |
| 3 | Price | on sale, cost, value insurance, economic, bargain, expensive, quality, cheap, great deal, value |
| 4 | Positive | good, enthusiastic, support, attitude, brand, fine, cheer up, very well, goodness, OK |

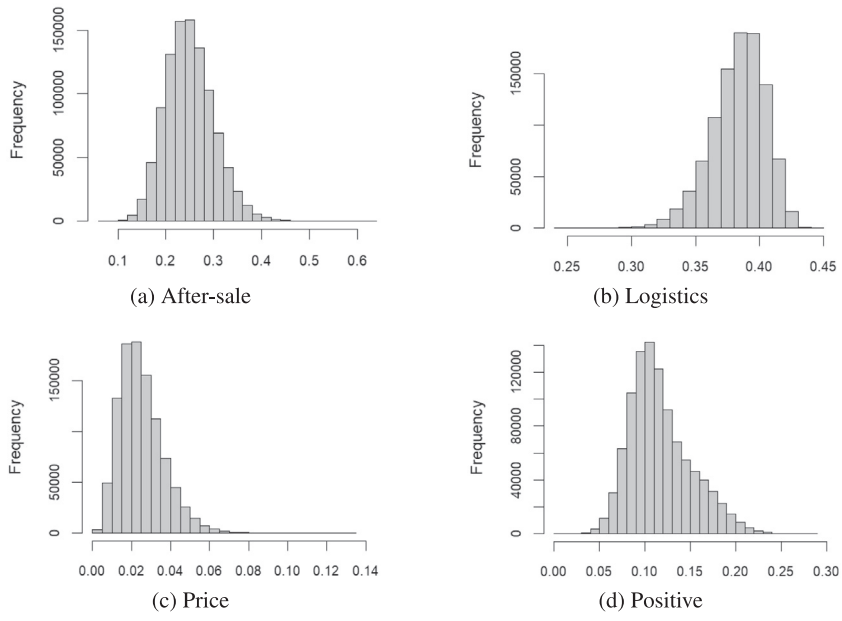


**Fig. 3.** Histograms of author probabilities for four example informal topics in JD dataset.

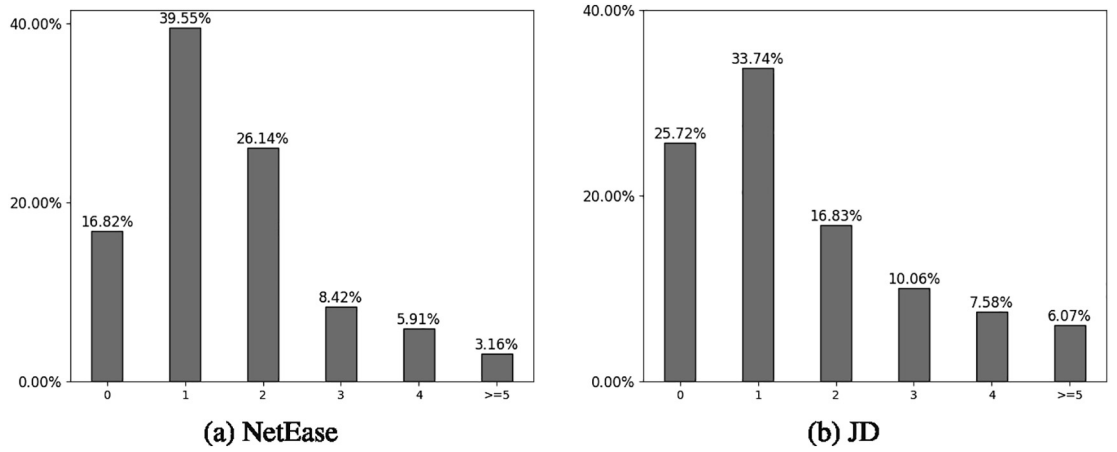


**Fig. 4.** Distributions of the number of labels per author in the NetEase dataset and the JD dataset, respectively.

3. **Predict posting behavior.** For a new normal document $d'$, we assume it has a probability $\theta_{d'}$ among $K$ formal topics. Because $\pi_{ak}$ can reflect author $a$'s preference on formal topic $k$, we can further explore author $a$'s total preference toward the new document $d'$ by calculating $\tau_{d'a} = \sum_k \theta_{d'k}\pi_{ak}$. A higher value of $\tau_{d'a}$ indicates a higher probability that author $a$ is interested in document $d'$ and would post/comment on it.

To evaluate the performance of predicting user posting behaviors, we conduct experiments on the NetEase, Tencent, and JD datasets. For each dataset, AOTM is first established on the 80% training set. The experimental settings are similar to those described in Section 5.1. Next, the user posting behaviors are predicted on the 20% test set. Specifically, for normal document $d'$ in the test set, the top $q$ authors with the highest $\tau_{d'a}$ values would be predicted to post short texts on this document. We try different values of $q$ in the experiments, i.e., $q = 10, 50, 100, 200$. We then calculate the percentages of the correctly cached authors to evaluate the prediction performance of AOTM.

We compare the prediction performance of AOTM with a simple baseline method. In the baseline method, a standard LDA model with $K$ topics is first implemented on all normal documents in the training set. Next, for a specific author $a$, we would calculate its preference $\widetilde{\pi}_{ak}$ on topic $k(k = 1, \ldots, K)$. To this end, we could first find all normal documents that author $a$ has commented on. Then we regard the averaged topic probabilities of these normal documents as author $a$'s preference. In other words, we define $\widetilde{\pi}_{ak} = \frac{1}{M}\sum_{d\in \boldsymbol{C}_{\text{train}}}\left\{\theta_{dk}\sum_{c=1}^{C_d} I(a_{dc} = a)\right\}$, where $\boldsymbol{C}_{\text{train}}$ denotes the training set, $C_d$ denotes the number of short texts associated with normal document $d$, and $M = \sum_{d\in \boldsymbol{C}_{\text{train}}}\sum_{c=1}^{C_d} I(a_{dc} = a)$. Given $\widetilde{\pi}_{ak}$, author $a$'s total preference $\widetilde{\tau}_{d'a}$ toward a normal document $d'$ in the test set can be calculated similarly, as described in Step (3).

**Table 7**
Percentages of correctly predicted authors using AOTM and the LDA model on NetEase, Tencent, and JD datasets.

| | $K = 100, J = 20$ | | $K = 100, J = 50$ | | $K = 200, J = 100$ | |
|---|---|---|---|---|---|---|
| | LDA | AOTM | LDA | AOTM | LDA | AOTM |
| *NetEase* | | | | | | |
| Top 10 | 0.373 | 0.453 | 0.394 | 0.461 | 0.402 | 0.473 |
| Top 50 | 0.401 | 0.513 | 0.413 | 0.531 | 0.426 | 0.541 |
| Top 100 | 0.421 | 0.538 | 0.437 | 0.553 | 0.450 | 0.562 |
| Top 200 | 0.457 | 0.564 | 0.470 | 0.581 | 0.483 | 0.592 |
| *Tencent* | | | | | | |
| Top 10 | 0.342 | 0.410 | 0.352 | 0.432 | 0.369 | 0.449 |
| Top 50 | 0.360 | 0.438 | 0.372 | 0.454 | 0.382 | 0.471 |
| Top 100 | 0.378 | 0.448 | 0.391 | 0.469 | 0.402 | 0.489 |
| Top 200 | 0.392 | 0.457 | 0.408 | 0.491 | 0.418 | 0.503 |
| *JD* | | | | | | |
| Top 10 | 0.217 | 0.358 | 0.229 | 0.371 | 0.240 | 0.382 |
| Top 50 | 0.241 | 0.381 | 0.259 | 0.393 | 0.269 | 0.405 |
| Top 100 | 0.269 | 0.402 | 0.279 | 0.420 | 0.287 | 0.431 |
| Top 200 | 0.281 | 0.426 | 0.294 | 0.438 | 0.443 | 0.454 |

Table 7 presents the percentages of correctly predicted authors using AOTM and the simple LDA model. In general, AOTM outperforms the LDA model under different experimental settings and for different datasets. Specifically, in the NetEase dataset, the percentages of correctly predicted authors by AOTM are approximately 55%, whereas those predicted by LDA are only approximately 45%. In the Tencent dataset, the percentages of correctly predicted authors decrease for both methods, which may be mainly due to its smaller sample size. As for the JD dataset, author posting behaviors are closely related to their purchasing behaviors, noting that the website allows only buyers to make comments. Therefore, the posting behaviors should be affected by several complicated factors, and thus, the correct prediction percentages become smaller than those for news articles in the NetEase and Tencent datasets. Finally, in all three datasets, as the number of topics increases, we observe an increasing trend in the correctly predicted percentages. These results suggest that with the assumption of more topics, both AOTM and LDA achieve improved model fitting performance.

## 5. Conclusion and discussion

In this study, we propose AOTM to extend the original COTM, which can be used to investigate individual user preferences of short texts. Compared with COTM, AOTM relaxes the "same author" assumption in COTM by incorporating the authorship of each short text. Therefore, each author has its own probability distribution over all informal topics. The preferences of each author can then be naturally learned. To compare the performance of AOTM with COTM and other state-of-the-art methods, we evaluate each model from two perspectives: (1) the quality of learned topics, which is measured using the topic coherence score, and (2) the quality of the topic representation of documents, which is measured using perplexity. To demonstrate the performance of AOTM, extensive experiments have been conducted. From the experimental results, the following conclusions can be summarized. First, AOTM can outperform COTM in both formal topics and informal topics from the perspectives of topic coherence score and perplexity. Second, AOTM can outperform most state-of-the-art methods, with the exception of GPU-PDMM, which achieves similar or even better performance than AOTM in some setups. However, the advantage of AOTM lies in its ability to obtain author-specific topic probability distributions, which provide a good starting point for investigating user preferences. To illustrate this idea, we provide a practical methodology for drawing user portraits and predicting user posting behaviors. More potential usage for AOTM in exploring individual user preferences can be further studied.

However, this study also has some limitations, which can inspire further improvements in the future. First, we do not consider the time effect in AOTM, while user preferences are likely to change over time. Therefore, a dynamic model for both topics and user preferences is worth considering. Second, we assume that each short text should represent two topics; one formal topic and one informal topic. This assumption can be further relaxed by adding a latent indicator or using the Dirichlet process in the future. Finally, AOTM assumes that each topic has a probability distribution over the entire vocabulary. However, for short texts, the co-occurrence information is often sparse. To make the informal topics more focused on specific words, a sparsity extension of AOTM can be considered in future studies.

## CRediT authorship contribution statement

**Yang Yang:** Conceptualization, Investigation, Software. **Feifei Wang:** Methodology, Writing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This study was supported by the National Natural Science Foundation of China (72001205, 11971504, 61632002, 61872166, 61902005, 62002002), the National Key R\&D Program of China (2019YFA0706401), the fund for building world-class universities (disciplines) of Renmin University of China, the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (2021030047), Foundation from Ministry of Education of China (20JZD023), and the Ministry of Education Focus on Humanities and Social Science Research Base (Major Research Plan 17JJD910001).

## References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[2] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (suppl 1) (2004) 5228–5235.
[3] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: Proceedings of the 22nd International Conference on World Wide Web, Association for Computing Machinery, New York, NY, USA, 2013, pp. 1445–1456..
[4] X. Cheng, X. Yan, Y. Lan, J. Guo, Btm: Topic modeling over short texts, IEEE Transactions on Knowledge and Data Engineering 26 (12) (2014) 2928–2941, https://doi.org/10.1109/TKDE.2014.2313872.
[5] Y. Yang, F. Wang, J. Zhang, J. Xu, P.S. Yu, A topic model for co-occurring normal documents and short texts, World Wide Web 21 (2018) 487–513.

[6] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, ACM Transactions on Information Systems 28 (1) (2010) 1–38.
[7] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, USA, 2009, pp. 248–256.
[8] J. Yin, J. Wang, A model-based approach for text clustering with outlier detection, in: Proceedings of IEEE 32nd International Conference on Data Engineering (ICDE), 2016, pp. 625–636.
[9] J. Yin, D. Chao, Z. Liu, W. Zhang, X. Yu, J. Wang, Model-based clustering of short text streams, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2018, pp. 2634–2642..
[10] J. Chen, Z. Gong, W. Liu, A non-parametric model for online topic discovery with word embeddings, Information Sciences 504 (2019) 32–47.
[11] J. Chen, Z. Gong, W. Liu, A Dirichlet process biterm-based mixture model for short text stream clustering, Applied Intelligence 50 (2020) 1609–1619.
[12] T. Lin, W. Tian, Q. Mei, H. Cheng, The dual-sparse topic model: Mining focused topics and focused terms in short text, in: Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 539–550.
[13] J. He, L. Li, Y. Wang, X. Wu, Targeted aspects oriented topic modeling for short texts, Applied Intelligence 50 (2020) 2384–2399.
[14] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, H. Xiong, Topic modeling of short texts: A pseudo-document view, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 2105–2114.
[15] J. Weng, E.-P. Lim, J. Jiang, Q. He, Twitterrank: finding topic-sensitive influential twitterers, in: Proceedings of the third ACM International Conference on Web Search and Data Mining, ACM, 2010, pp. 261–270.
[16] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving LDA topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 889–892.
[17] F. Kou, J. Du, C. Yang, Y. Shi, M. Liang, Z. Xue, H. Li, A multi-feature probabilistic graphical model for social network semantic search, Neurocomputing 336 (2019) 67–78.
[18] L. Hong, B. D. Davison, Empirical study of topic modeling in Twitter, in: Proceedings of the First Workshop on Social Media Analytics, 2010, pp. 80–88..
[19] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda, G.L. Pappa, A general framework to expand short text for topic modeling, Information Sciences 393 (2017) 66–81.
[20] X. Quan, C. Kit, Y. Ge, S.J. Pan, Short and sparse text topic modeling via self-aggregation, in: Proceedings of the 24th International Conference on Artificial Intelligence, 2015, pp. 2270–2276.
[21] X.H. Phan, L.M. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections (2008) 91–100.
[22] X.H. Phan, C.T. Nguyen, D.T. Le, L.M. Nguyen, S. Horiguchi, Q.T. Ha, A hidden topic-based framework toward building applications with short web documents, IEEE Transactions on Knowledge and Data Engineering 23 (7) (2011) 961–976.
[23] O. Jin, N.N. Liu, K. Zhao, Y. Yu, Q. Yang, Transferring topical knowledge from auxiliary long texts for short text clustering, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 775–784.
[24] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, Z. Ma, Enhancing topic modeling for short texts with auxiliary word embeddings, ACM Transactions on Information Systems 36 (2) (2017) 1–30.
[25] J. Qiang, P. Chen, T. Wang, X. Wu, Topic modeling over short texts by incorporating word embeddings, in: Advances in Knowledge Discovery and Data Mining, 2017, pp. 363–374..
[26] S. Li, J. Li, R. Pan, Tag-weighted topic model for mining semi-structured documents, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 2855–2861.
[27] A. Ahmed, L. Hong, A.J. Smola, Hierarchical geographical modeling of user locations from social media posts, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 25–36.
[28] S. Qiang, Y. Wang, Y. Jin, A local-global lda model for discovering geographical topics from social media, in: Web and Big Data, 2017, pp. 27–40..
[29] J. Guo, Z. Gong, A non-parametric model for event discovery in the geospatial-temporal space, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 499–508.
[30] J. Guo, Z. Gong, A density-based nonparametric model for online event discovery from the social media data, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 1732–1738.
[31] D.M. Blei, J.D. McAuliffe, Supervised topic models, in: Proceedings of the 20th International Conference on Neural Information Processing Systems, 2007, pp. 121–128.
[32] F. Wang, L. J. Zhang, Y. Li, K. Deng, S. J. Liu, Bayesian text classification and summarization via a class-specified topic model, Journal of Machine Learning Research (22) (2021) 1–51..
[33] Y. Yang, Y. Liu, X. Lu, J. Xu, F. Wang, A named entity topic model for news popularity prediction, Knowledge-Based Systems 208 (2020) 106430.
[34] Y. Zuo, J. Zhao, K. Xu, Word network topic model: A simple but general solution for short and imbalanced texts, Knowledge and Information Systems 48 (2) (2016) 379–398.
[35] B. Stuart, B. Y, M. Maurice, Aggregated topic models for increasing social media topic coherence, Applied Intelligence 50 (2020) 138–156..
[36] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 487–494.
[37] J. Qiang, Q. Zhenyu, Y. Li, Y. Yuan, X. Wu, Short text topic modeling techniques, applications, and performance: A survey, IEEE Transactions on Knowledge and Data Engineering (2020) 1, https://doi.org/10.1109/TKDE.2020.2992485.
[38] Y. Song, S. Shi, J. Li, H. Zhang, Directional skip-gram: Explicitly distinguishing left and right context for word embeddings, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 175–180..
[39] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 262–272..
[40] Q. Li, Q. Wu, C. Zhu, J. Zhang, W. Zhao, Unsupervised user behavior representation for fraud review detection with cold-start problem, in: Advances in Knowledge Discovery and Data Mining, 2019, pp. 222–236..
[41] D. Kumar, Y. Shaalan, X. Zhang, J. Chan, Identifying singleton spammers via spammer group detection, in: Advances in Knowledge Discovery and Data Mining, 2018, pp. 656–667..
[42] W. Wang, W. Zhang, J. Wang, J. Yan, H. Zha, Learning sequential correlation for user generated textual content popularity prediction, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 1625–1631.