Fast Supervised Topic Models for Short Text Emotion Detection

Jianhui Pang, Yanghui Rao, *Member, IEEE*, Haoran Xie, *Member, IEEE*, Xizhao Wang, *Fellow, IEEE*, Fu Lee Wang, *Senior Member, IEEE*, Tak-Lam Wong, *Member, IEEE*, and Qing Li, *Senior Member, IEEE*

***Abstract*—With the development of social network platforms, discussion forums, and question answering websites, a huge number of short messages that typically contain a few words for an individual document are posted by online users. In these short messages, emotions are frequently embedded for communicating opinions, expressing friendship, and promoting influence. It is quite valuable to detect emotions from short messages, but the corresponding task suffers from the sparsity of feature space. In this article, we first generate term groups co-occurring in the same context to enrich the number of features. Then, two basic supervised topic models are proposed to associate emotions with topics accurately. To reduce the time cost of parameter estimation, we further propose an accelerated algorithm for our basic models. Extensive evaluations using three short corpora validate the efficiency and effectiveness of the accelerated models for predicting the emotions of unlabeled documents, in addition to generate the topic-level emotion lexicons.**

***Index Terms*—Accelerated algorithm, emotion detection, short text analysis, topic model.**

Manuscript received January 17, 2019; revised May 16, 2019 and August 11, 2019; accepted August 26, 2019. Date of publication September 30, 2019; date of current version January 15, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61972426, in part by the Interdisciplinary Research Scheme of the Dean's Research Fund 2018-19 under Grant FLASS/DRF/IDS-3, in part by the Departmental Collaborative Research Fund 2019 under Grant MIT/DCRF-R2/18-19, in part by the Top-Up Fund (TFG-04) and Seed Fund (SFG-10) for General Research Fund/Early Career Scheme of the Dean's Research Fund (DRF) 2018-19, in part by the General Research Fund Proposal under Grant RG 39/2019-2020R, in part by the Internal Research Grant of the Education University of Hong Kong under Grant RG 90/2018-2019R, and in part by the Collaborative Research Fund by the Research Grants Council of the Hong Kong under Project C1031-18G. The work of X. Wang was supported by the National Natural Science Foundation of China under Grant 61732011. This article is an extended journal version of a conference paper published at DASFAA 2017 [50]. This article was recommended by Associate Editor S. Ozawa. *(Corresponding author: Yanghui Rao.)*

J. Pang and Y. Rao are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: raoyangh@mail.sysu.edu.cn).

H. Xie is with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong.

X. Wang is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.

F. L. Wang is with the School of Science and Technology, Open University of Hong Kong, Hong Kong.

T.-L. Wong is with the Department of Computing Studies and Information Systems, Douglas College, New Westminster, BC V3M 5Z5, Canada.

Q. Li is with the Department of Computing, Hong Kong Polytechnic University, Hong Kong.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2940520

I. Introduction

With the development and popularization of social media services, users are increasingly inclined to communicate and share emotions on social network platforms, such as Twitter, Facebook, Sina Weibo, and WeChat. By using mobile devices, it is convenient for users to express comments on news or personal events, which generates large-scale short messages that are limited in length, usually spanning several sentences or less. Emotion detection on short messages is therefore quite valuable to capture the emotional tendency of social media users, for example, happy, sad, or surprise, toward entities, brands, or events. However, the feature sparsity of short texts brings huge challenges to traditional word-level algorithms [1], [2]. This is because two short documents may semantically related to each other without sharing any common words. Furthermore, a word can have multiple meanings depending on its context [3]. Thus, another solution to emotion detection attempts to extract topics first [4], [5], in which, a topic can represent a real-world event and the topic-level feature space is coherent by grouping semantically related words. Then, the emotions are associated with the topics for the emotion detection of unlabeled documents. Although the aforementioned issue of word-level algorithms can be alleviated by mapping the sparse word space to a coherent topic space, a traditional topic model, such as the latent Dirichlet allocation (LDA) [6], fails to generate accurate topics over short messages. This is because a short document lacks enough word occurrence patterns to draw statistical conclusions for such kind of topic models [7]. Recently, Cheng *et al.* [8] proposed the biterm topic model (BTM) to extract high-quality topics from short messages. BTM assumed that two words that co-occurred in a context (e.g., in the same document) are likely to belong to the same topic. However, the generated topic features of BTM may be unsuitable to predict emotions without any guidance from labels in the training corpus. Furthermore, BTM is too time consuming to model such large-scale word pairs.

To address the aforementioned issues, we here develop a weighted labeled topic model (WLTM) and an X-term emotion-topic model (XETM) to detect emotions toward certain topics. In the generative process of WLTM, we first define a one-to-many mapping among each emotion and multiple topics, by assuming that a single emotion may be evoked by several topics. Second, we use the emotion distributions of labeled documents to constrain the topic probability for each feature during the training process. Finally, we employ the support vector regression (SVR) [9] to predict emotion distributions of unlabeled documents given the estimated topic probability for each feature. In the generative process of XETM, we draw the emotion-topic probability which exploits abundant user scores over multiple emotions. Then, the topic-feature probability is derived for estimating the emotion probabilities of unlabeled documents. The main characteristics of WLTM and XETM are summarized as follows. First, both WLTM and XETM are supervised topic models which align the generated topics to emotions using the emotion distributions of training documents for guidance. Second, the abundant features are generated by jointly modeling emotion labels and term groups. Particularly, a term group with X words co-occurring in the same context is called X-term. With abundant features, the proposed models allow us to draw statistical conclusions for short documents. Although the sparse feature issue of short messages can be alleviated by WLTM and XETM, the time cost of estimating parameters is high due to the large-scale term groups and the sampling algorithm [10]. To improve the efficiency, we further propose the accelerated models dubbed fWLTM and fXETM for WLTM and XETM by combining the Alias method [11] and the Metropolis–Hastings (MH) sampling [12]. Experiments using a sensibly small and unbalanced news headlines with six emotions, a larger and balanced sentences annotated with seven emotions, and a Chinese corpus with eight emotions validate the effectiveness of the proposed methods.

The remainder of this article is organized as follows. In Section II, we summarize the related works on emotion detection and short text analysis. In Section III, we detail the basic WLTM and XETM methods, and corresponding accelerated models called fWLTM and fXETM for short text emotion detection. The experimental evaluations are shown in Section IV, and we draw the conclusions in Section V.

II. Related Work

As one of the basic tasks of affective computing and sentiment analysis [13], emotion detection aims to identify and extract the attitudes of a subject (i.e., an opinion holder, a commentator, and so forth) toward either a topic, an aspect, or the overall tone of a document [14]. Methods of emotion detection are mainly based on the lexicons, supervised learning, and unsupervised learning algorithms. The lexicon-based methods [5], [15]–[19] construct the word-level, concept-level, or topic-level emotional/sentimental dictionaries to detect emotions. For example, the emotion-term method [4] associated words with emotions and used the word-emotion dictionary for prediction. The contextual sentiment topic model (CSTM) [20] mined connections between topics and emotions by distilling context-independent information, which were further applied to social emotion classification. The models based on supervised learning used traditional classification algorithms (e.g., naïve Bayes [21], maximum entropy [22], and support vector machines [23]) or deep learning models (e.g., sentiment embedding-based method [24], deep memory network [25], hybrid neural network [26], and Sentic LSTM and H-Sentic-LSTM [27]) to detect emotions or sentiments from documents. The unsupervised learning methods detected the sentimental or emotional orientation by counting the co-occurrence frequency between words and positive/negative terms [28]. However, the aforementioned methods were mainly suitable to long articles which typically contain abundant features.

With the prevalence of tweets, questions, instant-messages, and news headlines, several strategies have been proposed to tackle the feature sparse issue of short messages. One solution expanded the content of short documents by transferring topical knowledge from large-scale data collections or auxiliary long texts [29], [30], but it only achieved a good topical distribution when the auxiliary data are closely related to the original corpus. Furthermore, it is difficult to determine the suitable size of external data collections. Another solution to short text analysis exploited the aggregated word co-occurrence patterns in the entire corpus for topic learning [8], [31]. For a short document with N words, C_N^2 unordered word pairs, namely, biterms, can be extracted by assuming that two words from the same document share a single topic. Unlike most existing document-level topic models, the above method learns topic components for a corpus using the generated rich biterms. However, it was unsuitable to model labeled documents due to the lack of supervision during the training process. Furthermore, Gibbs sampling was employed by the above model and many other topic models to estimate parameters [8], [32], which is quite time consuming with the increase of the number of documents, features/biterms, or topics. Therefore, we detect emotions of short text by two supervised topic models and further develop an MH sampling in conjunction with the Alias method for accelerating parameter estimation.

III. Fast Supervised Topic Models

Here, we first present the basic supervised topic models, namely, WLTM and XETM for detecting emotions over short messages. To make the topic sampling more efficient without reducing much topic quality, we further develop accelerated algorithms for both WLTM and XETM.

A. Problem Definition

Before illustrating our supervised topic models for short text emotion detection, we summarize notations, variables, and terms in Table I. Taking a collection of N_D short documents $\{d_1, d_2, \ldots, d_{N_D}\}$ as an example, the issue of emotion detection is defined as predicting the emotion distribution of unlabeled documents conditioned to labeled data. For each labeled document d, there are N_d words and scores/ratings over N_E emotions, which are denoted as $\omega_d = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_{N_d}\}$ and $E_d = \{E_{d,1}, E_{d,2}, \ldots, E_{d,N_E}\}$, respectively. Using each text as a context, we can generate N_G unordered term groups that are represented by $\mathbf{G} = \{g_i\}_{i=1}^{N_G}$. For instance, a short document with four words will get six term groups when X is 2: $(\omega_1, \omega_2, \omega_3, \omega_4) \Rightarrow \{(\omega_1, \omega_2), (\omega_1, \omega_3), (\omega_1, \omega_4), (\omega_2, \omega_3), (\omega_2, \omega_4), (\omega_3, \omega_4)\}$. We represent the emotion annotation information by a real-valued matrix γ with the size of

TABLE I
NOTATIONS

Symbol	Descriptions
τ	Multiplier between topic and emotion numbers
N_E	Number of emotion labels
N_z	Number of topics
N_D	Number of documents
N_ω	Number of words
X-term	A term group of X words
N_G	Number of term groups
g_i	The i-th X-term
z_i	The topic of the i-th X-terms
ε_i	The emotion label of the i-th X-term
$\mathbf{\Lambda}_{g_i}$	The binary indicator of X-term g_i over topics
$\mathbf{\Psi}$	$N_G \times N_E$ emotion label prior for term groups
λ_{g_i}	The vector of topics relative to X-term g_i
θ	$N_D \times N_z$ multinomial distributions of documents to topics
ϕ	$N_\omega \times N_z$ multinomial distributions of topics to words
φ	$N_E \times N_z$ multinomial distributions of topics to emotions
γ	$N_D \times N_E$ prior emotion frequencies in the corpus
α	Dirichlet prior of θ and φ
β	Dirichlet prior of ϕ

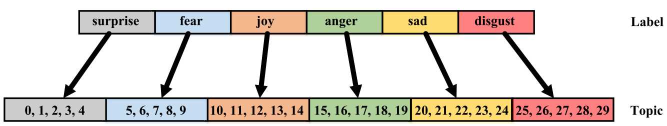


Fig. 1. Label-topic projection with $\tau = 5$.

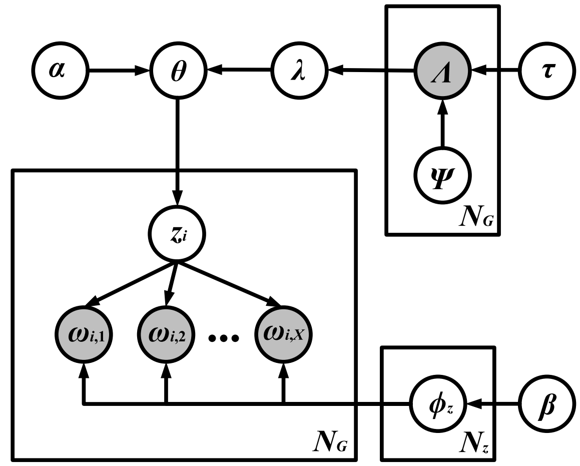


Fig. 2. Graphical representation of WLTM. Ψ is the emotion label prior for X-terms. τ indicates the number of topics associated with each emotion. $\mathbf{\Lambda}$ represents the topic binary (presence/absence) indictor. i means the ith X-term.

$N_D \times N_E$. Each row of γ is a document's real-valued vector over N_E emotion labels, for example, $\{1, 0\}$ means that the document is associated with the first emotion, and $\{3, 1\}$ indicates that the document is tagged to both emotions with strengths of 3 and 1, respectively.

In the first model called WLTM, we assume that each emotion can be associated with multiple topics. Take the following two short messages as an example: "I feel surprised about my Christmas gift" and "The examination results surprised me." Although both messages trigger the emotion of "surprise," we can observe that the distinct topics of "Christmas gift" and "examination" are embedded. To this end, we define a multiplier τ to represent how many topics per emotion involves. Specifically, Fig. 1 presents the projection of emotion labels and topics when τ equals 5, in which constant mapping of an emotion to τ topics is adopted. This is consistent to LDA's assumption that a document can be mapped to a given number of topics [6]. We leave the infinite mapping method, for example, in hierarchical Dirichlet processes to further research, because the parameter estimation is quite time consuming [33]. Through the above mapping of emotions to topics, we can conveniently develop the supervised mechanism in WLTM. In the second model called XETM, we use an $N_E \times N_z$ matrix φ to denote the multinomial distributions of emotions to topics.

B. Weighted Labeled Topic Model

The graphical representation of WLTM is shown in Fig. 2, where observed and unobservable data are represented by shaded and blank nodes, respectively.

After mapping each emotion to multiple topics via multiplier τ, we could incorporate the supervision of emotion labels of training documents into topic generation. Furthermore, the number of topics N_z can be tuned by setting different values of τ, that is, $N_z = N_E \times \tau$. Although there were supervised topic models, such as labeled LDA (LLDA) [32] being proposed, existing methods mainly exploited the one-to-one correspondence between labels and topics, which renders the number of topics must equal the size of the label set. On the other hand, the label-topic projection in our WLTM is one-to-many. Thus, different aspects can be discovered for each emotion label as mentioned earlier.

To explore document labels in generating topics effectively, we propose to extract an $N_G \times N_z$ indicator matrix $\mathbf{\Lambda}$ for all X-terms. For the above matrix, each row $\mathbf{\Lambda}_{g_i}$ is a list of binary topic indictors (i.e., presence/absence) related to the emotion labels of the document that contains X-term g_i. In particular, the generation of $\mathbf{\Lambda}_{g_i}$ is as follows. Given τ and emotion label prior information Ψ, for each X-term g_i, Ψ_{g_i} is the prior emotion label with size of $1 \times N_E$. Then, we construct an $N_E \times \tau$ matrix L_{g_i} which means each emotion label is linked to τ topics, as follows:

$$\left\{L_{g_i}^j\right\}_{j=1}^{N_E} = \begin{cases} \{1\}^\tau & \text{if } \Psi_{g_i}^j != 0 \\ \{0\}^\tau & \text{if } \Psi_{g_i}^j = 0 \end{cases} \tag{1}$$

where $L_{g_i}^j$ is the jth row of L_{g_i} and $\Psi_{g_i}^j$ is the jth element of Ψ_{g_i}. $\{1\}^\tau$ and $\{0\}^\tau$ are τ-dimensional vectors with 1 and 0, respectively. Then, we transform L_{g_i} to a 1-D vector $\mathbf{\Lambda}_{g_i}$ with $N_E \times \tau$ (i.e., N_z) elements by appending the vector of following rows to the first row in turn. Take $\tau = 2$, $N_E = 2$, $N_z = 4$, and a labeled document that contains g_i with emotion ratings $\Psi_{g_i} = \{2, 0\}$ as an example, we obtain $L_{g_i} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ according to (1) and $\mathbf{\Lambda}_{g_i} = \{1, 1, 0, 0\}$. The generative process of WLTM is presented as follows:

1) For each topic $z \in [1, N_z]$, draw $\phi_z \sim$ Dirichlet (β);
2) For each X-term $g_i \in \mathbf{G}$:
3) For each topic $z \in [1, N_z]$:
4) Generate $\mathbf{\Lambda}_{g_i}^z \in \{0, 1\} \sim$ Bernoulli$(\cdot|\mathbf{\Psi}_{g_i}, \tau)$;
5) Generate $\lambda_{g_i} = \{z|\mathbf{\Lambda}_{g_i}^z = 1\}$;

6) Draw $\theta \sim$ Dirichlet $(\cdot|\alpha, \lambda)$;
7) For each X-term $g_i \in \mathbf{G}$:
8) Generate $z_i \in \lambda_{g_i} \sim$ Multinomial$(\cdot|\theta, \alpha)$;
9) Generate $\omega_{i,1}, \ldots, \omega_{i,X} \in g_i \sim$ Multinomial(ϕ_{z_i}).

After generating $\boldsymbol{\Lambda}_{g_i}$ in step 4, we obtain the related topic distribution for g_i in step 5. Then, the topic assignment z_i is drawn from the above distribution according to step 8, which indicates that this restriction limits all topic assignments to the labels of X-term. In the above key steps, note that we explore the topics of each X-term g_i in the range of label relative topics and discard the topics not assigned to λ_{g_i}.

To estimate model parameters, an approximate inference method based on Gibbs sampling [34]–[36] can be used. First, the conditional probability of X-term g_i is estimated as follows:

$$P\left(g_i|\hat{\theta}, \hat{\phi}, \lambda_{g_i}\right) = \sum_{z=1}^{N_z} P\left(\omega_{i,1}, \ldots, \omega_{i,X}, z_i = z|\theta, \phi, \lambda_{g_i}\right) = \sum_{z=1}^{N_z} \theta_z \prod_{x=1}^{X} \phi_{z,\omega_{i,x}}. \tag{2}$$

Second, the likelihood function of all X-terms that should be maximized is given as follows:

$$P\left(\mathbf{G}|\hat{\theta}, \hat{\phi}, \Lambda\right) = \prod_{i=1}^{N_G} \sum_{z=1}^{N_z} \theta_z \prod_{x=1}^{X} \phi_{z,\omega_{i,x}}. \tag{3}$$

Finally, the topic of each X-term g_i is sampled by the following conditional probability:

$$P\left(z_i = z, z \in \lambda_{g_i}|\hat{z}_{-i}, \mathbf{G}\right) \propto \left(n_{-i,z} + \alpha\right) \times \frac{\gamma_{d_i,\left|\frac{z}{\tau}\right|}}{\sum_{z'} \gamma_{d_i,\left|\frac{z'}{\tau}\right|}} \prod_{x=1}^{X} \frac{\left(n_{-i,\omega_{i,x}|z} + \beta\right)}{\left(n_{-i,\cdot|z} + N_\omega \beta\right)} \tag{4}$$

where $\hat{z}_{-i}$ denotes the assigned topics for the group of X-terms, $n_{-i,z}$ represents the number of X-terms that are assigned to topic z, $n_{-i,\omega|z}$ is the number of times that word ω is assigned to topic z, $n_{-i,\cdot|z}$ is the number of times for all words that are assigned to topic z, and the notation $-i$ indicates that the number does not include the current assignment of X-term g_i. We use d_i to represent the document from which g_i is sampled, and the absolute value of z divides by τ (i.e., $|(z/\tau)|$) to achieve the emotion index. Since the sampling of topics for the ith X-term is restricted according to the emotion labels of documents containing g_i (i.e., $z \in \lambda_{g_i}$), the label information is injected into the probability distribution to supervise the topic generation through a weighted mechanism.

After a given number of iterations, we record the number of X-terms that are assigned to topic z, that is, n_z, and the number of times word ω being assigned to topic z, that is, $n_{\omega|z}$. Then, the probabilities of words conditioned to topics ϕ and the probabilities of topics over the corpus θ are, respectively, calculated as follows:

$$\phi_{z,\omega} = \frac{n_{\omega|z} + \beta}{n_{\cdot|z} + N_\omega \beta}, \quad \theta_z = \frac{n_z + \alpha}{N_G + N_z \alpha}. \tag{5}$$

Based on the generated topic of each X-term, WLTM calculates the topic proportion via computing each document's posterior topic probability. For each document d, the topic of X-term $g_i^{(d)} = (\omega_{i,1}^{(d)}, \ldots, \omega_{i,X}^{(d)})$ is assumed to be conditionally independent with each other. After the generation of X-terms, we have $P(z|d) = \sum_i P(z|g_i^{(d)})P(g_i^{(d)}|d)$, where $P(g_i^{(d)}|d)$ is the frequency of X-term g_i in document d, and $P(z|g_i^{(d)})$ can be calculated by the following Bayes rule:

$$P\left(z_i = z|g_i^{(d)}\right) = \frac{\theta_z}{\sum_{z'} \theta_{z'}} \prod_{x=1}^{X} \frac{\phi_{z,\omega_{i,x}^{(d)}}}{\sum_{z'} \phi_{z',\omega_{i,x}^{(d)}}}. \tag{6}$$

Algorithm 1 Gibbs Sampling Algorithm for WLTM

Input:
τ: Multiplier between topic and emotion numbers;
N_E: Number of emotion labels;
α: Hyperparameter of θ;
β: Hyperparameter of ϕ_z;
$\mathbf{G}$: The X-term groups in the training set;
Output:
ϕ: Multinomial distributions of words for topics;
θ: Multinomial distributions of topics for the corpus;
procedure BUILD WLTM
  Calculate topic numbers N_z by $\tau \times N_E$;
  Randomly initialize topic assignments for all X-terms;
  repeat
    for all $g_i = (\omega_{i,1}, ..., \omega_{i,X}) \in \mathbf{G}$ **do**
      Draw topic z according to Equation (4);
      Update n_z, $n_{\omega_{i,1}|z}$, ..., and $n_{\omega_{i,X}|z}$;
    end for
  until N_{iter} times
  Compute ϕ and θ by Equation (5).
end procedure

We present the Gibbs sampling algorithm that is used for WLTM in Algorithm 1. After computing the topic probability of each document $P(z|d)$ as mentioned earlier, we employ the SVR [9] to predict the emotion distributions of unlabeled documents using $P(z|d)$ as the input.

C. X-Term Emotion-Topic Model

Fig. 3 presents the graphical model of XETM, in which, an emotion label ε is first generated under the constraint of prior emotion frequencies. Second, a topic related to emotion label ε is sampled. Finally, we generate an X-term (i.e., a group of X words) for each document.

The generative process of XETM is shown as follows:
1) For emotion $\varepsilon \in [1, N_E]$, draw $\varphi_\varepsilon \sim$ Dirichlet (α);
2) For each topic $z \in [1, N_z]$, draw $\phi_z \sim$ Dirichlet (β);
3) For each document $d \in D$:
3) For each X-term $g_i \in d$:
4) Generate $\varepsilon_i \sim$ Multinomial(γ_d);
5) Generate $z_i \sim$ Multinomial(δ_{φ_i});
6) Generate $\omega_{i,1}, \ldots, \omega_{i,X} \in g_i \sim$ Multinomial(ϕ_{z_i}).

In the above, $\varepsilon_i \in E$ and $z_i \in Z$ are the assigned emotion and topic for X-term g_i, respectively. Specifically, ε, which is normalized and summed up to 1, is sampled from a multinomial distribution with emotion ratings that are parameterized by γ. Accordingly, we can estimate the joint probability of all

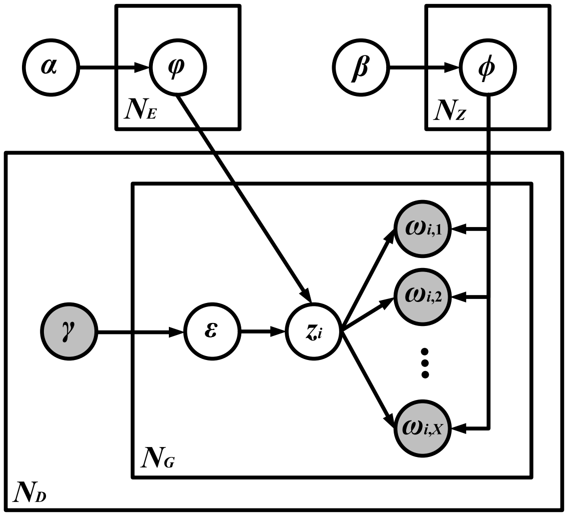


Fig. 3. Graphical representation of XETM.

Algorithm 2 Gibbs Sampling Algorithm for XETM

Input:
N_z: Number of topics;
N_E: Number of emotion labels;
α: Hyperparameter of φ;
β: Hyperparameter of ϕ;
G: The X-term groups in the training set;
Output:
$P(\varepsilon|d)$: The emotion proportion of document d;
procedure BUILD XETM
    Randomly initialize topic assignments for all X-terms;
    Randomly initialize emotion assignments for X-terms;
    repeat
        for all $g_i = (\omega_{i,1}, ..., \omega_{i,X}) \in \mathbf{G}$ **do**
            Draw emotion ε according to Equation (8);
            Draw topic z according to Equation (9);
            Update $n_{z|\varepsilon}$, $n_{\omega_{i,1}|z}$, ..., and $n_{\omega_{i,X}|z}$;
        end for
    until N_{iter} times
    Estimate $P(\varepsilon|d)$ by Equation (12).
end procedure

variables for each document as follows:

$$P(\gamma, \varepsilon, \mathbf{z}, \mathbf{G}, \phi, \varphi; \alpha, \beta) = P(\varphi; \alpha)P(\phi; \beta)P(\gamma) \times P(\varepsilon|\gamma)P(\mathbf{z}|\varepsilon, \phi)P(\mathbf{G}|\mathbf{z}, \varphi). \quad (7)$$

Particularly, the posterior probability of emotion ε for term g_i conditioned to topics is given as follows:

$$P\big(\varepsilon_i = \varepsilon|\hat{\varepsilon}_{-i}, \hat{\mathbf{z}}, \gamma, \mathbf{G}; \alpha, \beta\big) \propto \frac{\alpha + n_{-i,z_i|\varepsilon}}{N_z\alpha + \sum_z n_{-i,z|\varepsilon}} \times \frac{\gamma_{d_i,\varepsilon}}{\sum_{\varepsilon'} \gamma_{d_i,\varepsilon'}}. \quad (8)$$

Then, we sample a new topic conditioned to the set of X-terms **G** as follows:

$$P\big(z_i = z|\hat{\mathbf{z}}_{-i}, \hat{\varepsilon}, \gamma, \mathbf{G}; \alpha, \beta\big) \propto \frac{\alpha + n_{-i,z|\varepsilon_i}}{N_z\alpha + \sum_{z'} n_{-i,z'|\varepsilon_i}} \times \prod_{x=1}^{X} \frac{\beta + n_{-i,\omega_{i,x}|z}}{N_\omega\beta + \sum_{\omega'} n_{-i,\omega'|z}} \quad (9)$$

where the candidate topic and emotion for sampling are, respectively, denoted as z and ε, the number of times that topic z assigned to emotion ε is represented by $n_{z|\varepsilon}$, the number of times that word ω assigned to topic z is denoted as $n_{\omega|z}$, and each X-term g_i in **G** contains X words (i.e., $\omega_{i,1}, \ldots, \omega_{i,X}$). The subscript $-i$ is used for $n_{z|\varepsilon}$ and $n_{\omega|z}$ to indicate that the count does not include the current ith assignment of emotions or topics.

After the sampling of topics and emotions, the posterior probabilities of φ and ϕ can be calculated as follows:

$$\varphi_{\varepsilon,z} = \frac{\alpha + n_{z|\varepsilon}}{N_z\alpha + \sum_{z'} n_{z'|\varepsilon}} \quad (10)$$

and

$$\phi_{z,\omega} = \frac{\beta + n_{\omega|z}}{N_\omega\beta + \sum_{\omega'} n_{\omega'|z}}. \quad (11)$$

Finally, the predicted emotion distribution for a testing document d can be estimated by

$$P(\varepsilon|d) = \frac{P(\varepsilon)\prod_{\omega,\omega\in d} P(\omega|\varepsilon)}{\sum_\varepsilon P(\varepsilon)\prod_{\omega,\omega\in d} P(\omega|\varepsilon)} \quad (12)$$

where $P(\varepsilon)$ is the emotion probability distribution for the entire training set, and the probability of word ω conditioned to emotion ε can be estimated by integrating the latent topic z: $P(\omega|\varepsilon) = \sum_z \varphi_{\varepsilon,z}\phi_{z,\omega}$. To detail the estimation of parameters, we present the Gibbs sampling algorithm in Algorithm 2.

D. Accelerated Algorithm

Due to the high complexity of Gibbs sampling, we propose an accelerated algorithm for WLTM and XETM via a supervised MH sampling [12] in conjunction with the Alias method [11].

1) Alias Method: The number of topics N_z is one of the factors that determine the time complexity in topic modeling. The sampling procedure is very time consuming when N_z is large. Particularly, a general discrete probability distribution $P = \{p_1, p_2, \ldots, p_{N_z}\}$ will take $O(N_z)$ operations to generate a sample. On the other hand, it will take just $O(1)$ operations if the discrete probability distribution is a uniform distribution. Inspired by the above property, the Alias method simulates the characteristics of a uniform distribution by building up an Alias table and a probability table [11]. The generation processes of these two tables are shown in Fig. 4.

Take the sample generation from a discrete probability distribution $P = \{0.1, 0.2, 0.3, 0.4\}$ as an example, the objective is to make each entry in P to be equal to 1. We first multiply each entry in P by 4, thus the third and the fourth entries of P are larger than 1, while the first and the second entries of P are less than 1. Then, we use the third and the fourth entries to supplement the first and the second entries. During the process, the values of the probability table (ProbTable) are from the value of each relative entry. Furthermore, the Alias table (AliasTable) is the index number of the supplement entry. After the above process, we can sample an entry from these two tables with $O(1)$ operations as shown in Algorithm 3.

2) Metropolis–Hastings Sampling: As mentioned earlier, we implement the Gibbs sampling algorithm for our basic models WLTM and XETM, but with a high time cost (the

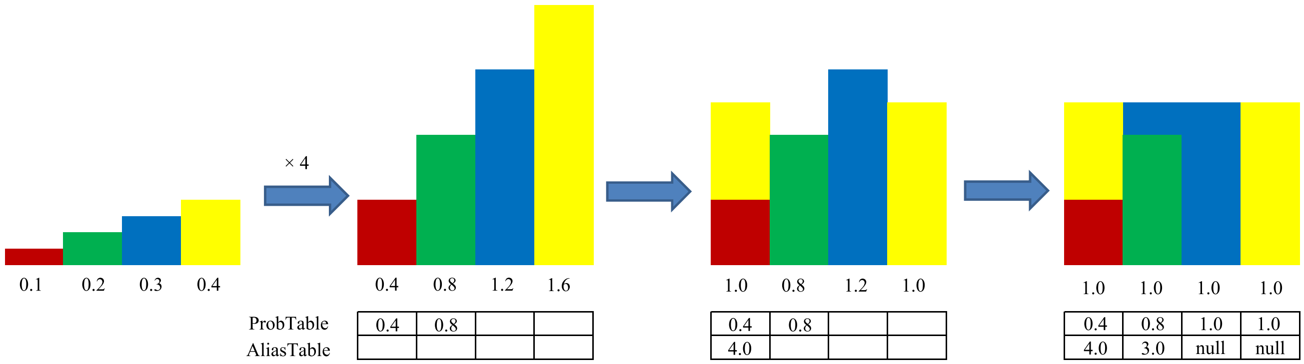


Fig. 4. Illustration of generating the alias table and the probability table.

Algorithm 3 Sampling Process of the Alias Method

Input:
AliasTable and *ProbTable* with N_z iterms
procedure SAMPLING
    $x = randint(N_z)$
    $y = random(0, 1)$
    if $y < ProbTable[x]$ **then**
        return x
    else
        return $AliasTable[x]$
    end if
end procedure

running time will be shown in Table IX). In the Gibbs sampling algorithm for WLTM and XETM, we need to draw a topic for each X-term in each iteration. This process will be very time consuming if the number of X-terms N_G is too large. Moreover, if we only build up an Alias table for Gibbs sampling, we have to save two matrices in the Alias and probability table for all X-terms with a total size of $N_G \times N_z$. Since N_G is quite large, the above operation not only costs time but also wastes storage space. In light of this consideration, we use the MH sampling [12], [37] in conjunction with the Alias method to estimate model parameters, in which, we only need to build up an Alias table and a probability table for each word. The accelerated models are, respectively, called fWLTM and fXETM, and the parameter derivation is shown in the next part.

3) Parameter Derivation: As for the parameter derivation of fWLTM, we decompose (4) into separated parts: $[(n_{-i,\omega_{i,1}|z}+\beta)/(n_{-i,\cdot|z}+N_\omega\beta)],\ldots,[(n_{-i,\omega_{i,X}|z}+\beta)/(n_{i,\cdot|z}+N_\omega\beta)]$ and $n_{-i,z}+\alpha$ for each X-term g_i in the conditional distribution. According to the MH sampling method, these parts are called proposal distributions. Specifically, we denote $[(n_{-i,z}+\alpha)/(N_G+N_z\alpha)]$ as the corpus proposal $p_{z,c}$ and $[(n_{-i,\omega|z}+\beta)/(n_{-i,\cdot|z}+N_\omega\beta)]$ as the word proposal $p_{\omega|z}$. The MH sampling algorithm draws a topic from $p_{z,c}$, and $p_{\omega_{i,1}|z},\ldots,p_{\omega_{i,X}|z}$ in turns, thus $p_{g_i}(z) \propto p_{z,c}\prod_{x=1}^{X} p_{\omega_{i,x}|z}$, where X is the number of words in a term group g_i, and it is called "cycle proposal" [38].

For the corpus proposal distribution, we have

$$p_{z,c} \propto (n_z+\alpha) \tag{13}$$

where the acceptance probability is $\min(1,\pi_c^{s\to t})$ for topic translation $s \to t$, and $\pi_c^{s\to t}$ is given as follows:

$$\pi_c^{s\to t} = \frac{(n_{-i,t}+\alpha)}{(n_{-i,s}+\alpha)}\frac{(n_{-i,\cdot|s}+N_\omega\beta)}{(n_{-i,\cdot|t}+N_\omega\beta)}\frac{(n_s+\alpha)}{(n_t+\alpha)} \times \prod_{x=1}^{X}\frac{(n_{-i,\omega_{i,x}|t}+\beta)}{(n_{-i,\omega_{i,x}|s}+\beta)} \tag{14}$$

where n_s is the number of X-terms assigned to topic s.

During the corpus proposal sampling, we do not need to build the Alias table and the probability table. Particularly, we store the topic that is assigned to the ith X-term g_i as ZG_i, which can be considered as an N_G length vector. After randomly sampling a topic ZG_j of an X-term g_j from ZG, the current assigned topic ZG_j of g_j can be considered as the translation state. Because the probability of sampling entry from vector ZG is equal with each other, ZG is a uniform distribution and the time complexity is $O(1)$. Considering the hyperparameter α in corpus proposal, we randomly set a float number f in the range of $(0, N_G+N_z\alpha)$. If f is less than N_G, we set an integer $f_{\text{int}} = \lfloor f \rfloor$, else $f_{\text{int}} = \lfloor f - N_G \rfloor$. Then, the translation state/topic is $ZG_{f_{\text{int}}}$.

For the word proposal distribution (e.g., $\omega_{i,x}$), we have

$$p_{\omega_{i,x}|z} \propto \frac{(n_{\omega_{i,x}|z}+\beta)}{(n_{\cdot|z}+N_\omega\beta)} \tag{15}$$

where the acceptance probability is $\min(1,\pi_{\omega_{i,1}}^{s\to t})$ when topic s translates to topic t, and $\pi_{\omega_{i,1}}^{s\to t}$ is estimated as follows:

$$\pi_{\omega_{i,x}}^{s\to t} = \frac{(n_{-i,t}+\alpha)}{(n_{-i,s}+\alpha)}\frac{(n_{-i,\cdot|s}+N_\omega\beta)^2}{(n_{-i,\cdot|t}+N_\omega\beta)^2}\frac{(n_{\omega_{i,x}|s}+\beta)}{(n_{\omega_{i,x}|t}+\beta)} \times \frac{(n_{\cdot|t}+N_\omega\beta)}{(n_{\cdot|s}+N_\omega\beta)}\prod_{x=1}^{X}\frac{(n_{-i,\omega_{i,x}|t}+\beta)}{(n_{-i,\omega_{i,x}|s}+\beta)}. \tag{16}$$

During the word proposal topic sampling, we restrict the states/topics of the g_i to its relative topics via only sampling from its label-related topic set λ_{g_i} as mentioned earlier.

As for the parameter deviation of fXETM, we decompose the conditional distribution (9) into separated parts: $[(\alpha+n_{-i,z|\varepsilon_i})/(N_z\alpha+\sum_{z'} n_{-i,z'|\varepsilon_i})]$ and $[(\beta+n_{-i,\omega_{i,1}|z})/(N_\omega\beta+\sum_{\omega'} n_{-i,\omega'|z})],\ldots,[(\beta+n_{-i,\omega_{i,X}|z})/(N_\omega\beta+\sum_{\omega'} n_{-i,\omega'|z})]$. Similarly, the first part is the topic-emotion proportion which is called emotion proposal $p_{z|\varepsilon_i}$, the remaining parts are word proposal $p_{\omega_{i,1}|z}, \ldots, p_{\omega_{i,X}|z}$, respectively. Specifically, the MH sampling for XETM draws a topic from these three proposal in turns, thus $p_{g_i}(z) \propto p_{z|\varepsilon_i} \prod_{x=1}^{X} p_{\omega_{i,x}|z}$.

For the emotion proposal distribution, we have

$$p_{z|\varepsilon_i} \propto \frac{\alpha + n_{-i,z|\varepsilon_i}}{N_z\alpha + \sum_{z'} n_{-i,z'|\varepsilon_i}} \tag{17}$$

where the acceptance probability is $\min(1, \pi_{\varepsilon_i}^{s\to t})$ when topic s translates to topic t, and $\pi_{\varepsilon_i}^{s\to t}$ is estimated as follows:

$$\pi_{\varepsilon_i}^{s\to t} = \frac{(\alpha + n_{s|\varepsilon_i})}{(\alpha + n_{t|\varepsilon_i})} \frac{(\alpha + n_{-i,t|\varepsilon_i})}{(\alpha + n_{-i,s|\varepsilon_i})} \frac{\left(N_\omega\beta + \sum_{\omega'} n_{-i,\omega'|s}\right)^2}{\left(N_\omega\beta + \sum_{\omega'} n_{-i,\omega'|t}\right)^2} \times \prod_{x=1}^{X} \frac{(\beta + n_{-i,\omega_{i,x}|t})}{(\beta + n_{-i,\omega_{i,x}|s})}. \tag{18}$$

For the word proposal distribution (e.g., $\omega_{i,x}$), we have

$$p_{\omega_{i,x}|z} \propto \frac{\beta + n_{-i,\omega_{i,x}|z}}{N_\omega\beta + \sum_{\omega'} n_{-i,\omega'|z}} \tag{19}$$

where the acceptance probability is $\min(1, \pi_{\omega_{i,1}}^{s\to t})$ when topic s translates to topic t, and $\pi_{\omega_{i,1}}^{s\to t}$ for the proposed fXETM is estimated as follows:

$$\pi_{\omega_{i,x}}^{s\to t} = \frac{(\beta + n_{\omega_{i,x}|s})}{(\beta + n_{\omega_{i,x}|t})} \frac{\left(N_\omega\beta + \sum_{\omega'} n_{\omega'|t}\right)}{\left(N_\omega\beta + \sum_{\omega'} n_{\omega'|s}\right)} \frac{(\alpha + n_{-i,t|\varepsilon_i})}{(\alpha + n_{-i,s|\varepsilon_i})} \times \frac{\left(N_\omega\beta + \sum_{\omega'} n_{-i,\omega'|s}\right)^2}{\left(N_\omega\beta + \sum_{\omega'} n_{-i,\omega'|t}\right)^2} \prod_{x=1}^{X} \frac{(\beta + n_{-i,\omega_{i,x}|t})}{(\beta + n_{-i,\omega_{i,x}|s})}. \tag{20}$$

During each iteration of topic generation, we first sample an emotion ε_i according to (8), which is less time consuming because there are generally a few emotion labels in the datasets (e.g., 6–8 for our employed datasets in the experiment). We apply the above MH sampling method for the topic generation based on emotion ε_i, which alleviates the time-consuming problem under a large number of topics.

E. Complexity Analysis

In the sampling process of each term group g_i, $n_{\omega|z}$ in WLTM or XETM changes slow, that is, there are only two counters reduced and two counters added for old and new topics, respectively. Therefore, it is unnecessary to update the Alias table and the probability table for each sample, which will reduce much running time. Especially, the Alias method keeps the MH proposal (i.e., corpus proposal, emotion proposal, and word proposal) over one iteration, rather than modify it after every sampling. For the MH sampling, the acceptance probability can be computed in $O(1)$ time. To achieve a better mixing rate, we combine the proposals into a cycle proposal, such as $p_{g_i}(z) \propto p_{z,c} \prod_{x=1}^{X} p_{\omega_{i,x}|z}$ for the fWLTM and $p_{g_i}(z) \propto p_{z|\varepsilon_i} \prod_{x=1}^{X} p_{\omega_{i,x}|z}$ for the fXETM, where a sequence is constructed for each token by alternating between corpus proposal and word proposal. Such cycle proposals are theoretically guaranteed to converge as shown in [38].

According to the above formulas, we summarize different models' time complexity in Table II. For the accelerated models (i.e., fWLTM and fXETM), we update the Alias tables over each iteration rather than each sampling. In this table, N_{iter} is the number of iteration, N_G is the number of generated term groups, N_ω is the number of distinct words in the corpus, and N_E and N_z are the numbers of emotion labels and topics, respectively. During each iteration of the proposed WLTM, we have to compute the topic probability distribution for each term group using (4) and sample one topic, so its time complexity is $O(N_{\text{iter}} \times N_G \times N_z)$. As for that of fWLTM, after the initialization of topic assignment for each term group, we build up the Alias table and the probability table for each word, which takes $O(N_\omega \times N_z)$ time, then we update these two tables over each iteration. Thus, the time complexity of fWLTM is $O(N_{\text{iter}} \times (N_\omega \times N_z + N_G))$. In each iteration of XETM, (8) computes the emotion probability distribution for each topic to sample one emotion, and (9) computes the topic probability distribution for each term group to sample one topic. So the time complexity of XETM is $O(N_{\text{iter}} \times N_G \times (N_E + N_z))$. As for fXETM, the MH sampling is applied in the topic sampling process according to (9). In each iteration of the topic sampling step, we update the Alias and probability tables for emotion and word proposal distributions, respectively. Thus, the time complexity of fXETM is $O(N_{\text{iter}} \times (N_G \times N_E + N_\omega \times N_z + N_E \times N_z + N_G))$.

Specifically, we can observe that N_G is always larger than N_ω when the value of X is larger than 1, and the running time of WLTM and XETM will increase when the number of

TABLE II
TIME COMPLEXITY OF DIFFERENT MODELS

Model	Time complexity
WLTM	$O(N_{iter} \times N_G \times N_z)$
XETM	$O(N_{iter} \times N_G \times (N_E + N_z))$
fWLTM	$O(N_{iter} \times (N_\omega \times N_z + N_G))$
fXETM	$O(N_{iter} \times (N_G \times N_E + N_\omega \times N_z + N_E \times N_z + N_G))$

TABLE III
STATISTICS OF DATASETS

Datasets	Emotion labels	# of documents	Mean words
SemEval	anger	87	7
	disgust	42	7
	fear	194	7
	joy	441	6
	sad	265	7
	surprise	217	7
	all	1,246	7
ISEAR	anger	1,096	24
	disgust	1,096	21
	fear	1,095	24
	joy	1,094	20
	sadness	1,096	20
	shame	1,096	22
	guilt	1,093	24
	all	7,666	22
RenCECps	joy	3,819	12
	hate	1,070	13
	love	4,756	13
	sorrow	3,392	13
	anxiety	3,770	13
	surprise	288	12
	anger	575	13
	expect	1,668	13
	all	19,338	13

TABLE IV
PERFORMANCE OF WLTM WITH VARIOUS X. (a) AP ON *SemEval*. (b) AP ON *ISEAR*. (c) AP ON *RenCECps*. (d) HD ON *SemEval*. (e) HD ON *ISEAR*. (f) HD ON *RenCECps*. (g) *Accuracy* ON *SemEval*. (h) *Accuracy* ON *ISEAR*. (i) *Accuracy* ON *RenCECps*

(a)

X	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.1585	1.28E-06	0.2252	1.59E-07
2	**0.1952**	0.0006	**0.2411**	0.0002
3	0.1865	0.0030	0.2378	0.0012
4	0.1812	0.0002	0.2297	0.0003

(b)

X	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.4061	1.84E-10	0.4061	1.84E-10
2	**0.4299**	4.12E-05	**0.4496**	9.31E-05
3	0.4136	1.78E-05	0.4252	2.11E-05
4	0.4065	1.05E-05	0.4101	3.11E-05

(c)

X	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.4061	1.84E-10	0.4061	1.84E-10
2	**0.4299**	4.12E-05	**0.4496**	9.31E-05
3	0.4136	1.78E-05	0.4252	2.11E-05
4	0.4065	1.05E-05	0.4101	3.11E-05

(d)

X	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
1	**0.4619**	1.26E-05	0.4799	3.67E-06
2	0.4655	3.19E-06	0.4816	8.88E-07
3	0.4782	0.0013	0.4789	0.0010
4	0.4823	0.0042	**0.4720**	0.0002

(e)

X	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.7379	0.0007	0.7408	0.0005
2	**0.6796**	2.54E-05	**0.6956**	6.13E-06
3	0.6841	2.71E-05	0.7012	3.01E-06
4	0.6896	1.94E-05	0.7098	2.87-06

(f)

X	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.7379	0.0007	0.7408	0.0005
2	**0.6796**	2.54E-05	**0.6956**	6.13E-06
3	0.6841	2.71E-05	0.7012	3.01E-06
4	0.6896	1.94E-05	0.7098	2.87-06

(g)

X	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
1	0.3140	0.5310	**0.7590**
2	**0.3567**	**0.5799**	0.7411
3	0.3230	0.5410	0.7480
4	0.3130	0.5380	0.756

(h)

X	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
1	**0.3650**	0.5244	0.6204
2	0.3567	**0.5799**	**0.7411**
3	0.3301	0.5231	0.6301
4	0.3602	0.5430	0.6412

(i)

X	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
1	0.2829	0.4897	0.6735
2	**0.4103**	**0.6298**	**0.7603**
3	0.2606	0.4757	0.6792
4	0.2708	0.4870	0.6859

TABLE V
PERFORMANCE OF XETM WITH VARIOUS X. (a) AP ON *SemEval*. (b) AP ON *ISEAR*. (c) AP ON *RenCECps*. (d) HD ON *SemEval*. (e) HD ON *ISEAR*. (f) HD ON *RenCECps*. (g) *Accuracy* ON *SemEval*. (h) *Accuracy* ON *ISEAR*. (i) *Accuracy* ON *RenCECps*

(a)

X	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.2572	0.0007	0.0950	0.0003
2	**0.3121**	0.0007	**0.1995**	0.0004
3	0.3015	0.0041	0.1921	0.0040
4	0.2989	0.0090	0.1876	0.0012

(b)

X	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.2535	6.65E-05	0.1896	0.5509
2	**0.2977**	1.93E-05	**0.3424**	0.0001
3	0.2901	3.32E-05	0.3389	7.32E-05
4	0.2882	4.23E-05	0.3312	4.73E-05

(c)

X	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.4061	1.84E-10	0.4061	1.84E-10
2	**0.4299**	4.12E-05	**0.4496**	9.31E-05
3	0.4136	1.78E-05	0.4252	2.11E-05
4	0.4065	1.05E-05	0.4101	3.11E-05

(d)

X	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.4839	9.02E-09	0.4898	4.55E-09
2	**0.4837**	1.97E-07	**0.4890**	3.15E-08
3	0.4894	1.58E-05	0.4941	1.40E-05
4	0.4931	1.05E-05	0.4989	1.59E-05

(e)

X	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.7879	2.42E-09	0.7879	2.42E-09
2	**0.7874**	2.97E-09	**0.7874**	2.97E-09
3	0.7943	1.14E-05	0.7884	1.75E-05
4	0.7987	7.63E-05	0.7997	2.01E-05

(f)

X	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
1	0.7379	0.0007	0.7408	0.0005
2	**0.6796**	2.54E-05	**0.6956**	6.13E-06
3	0.6841	2.71E-05	0.7012	3.01E-06
4	0.6896	1.94E-05	0.7098	2.87-06

(g)

X	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
1	**0.3620**	0.5720	0.7490
2	0.3567	**0.5799**	**0.7411**
3	0.3420	0.5510	0.7280
4	0.2670	0.5350	0.7070

(h)

X	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
1	0.3321	0.5642	0.7109
2	**0.3567**	**0.5799**	**0.7411**
3	0.3471	0.5610	0.7312
4	0.3392	0.5578	0.7201

(i)

X	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
1	0.3590	0.5710	0.7340
2	**0.3784**	**0.5868**	**0.7550**
3	0.3620	0.5720	0.7490
4	0.3510	0.5660	0.7280

topics N_z getting larger. On the other hand, we first update the Alias table over each iteration for fWLTM and fXETM, then we can use the MH sampling method to sample a topic in $O(1)$ time for each term group. Furthermore, the accelerated models, fWLTM and fXETM, only sample the topics from the related topics of each term group. Thus, the actual value of N_z in fWLTM is smaller than other topic models which compute each topic probability for training.

IV. EXPERIMENTS

This section presents the experimental results on the proposed models and baselines in terms of effectiveness and efficiency.

A. Datasets

SemEval: This dataset contains 1246 news headlines with the total score larger than 0, which is used in the 14th task of the 4th International Workshop on Semantic Evaluations (SemEval-2007) [1] and officially divided into a training set with 246 documents and a testing set with 1000 documents. The emotion labels include anger, disgust, fear, joy, sad, and surprise, which are posited to be basic [39].

ISEAR: This dataset contains 7666 sentences annotated by 1096 participants manually according to seven emotions [40]. The emotion categories are anger, disgust, fear, joy, sadness, shame, and guilt. For this dataset, 60%, 20%, and 20% of sentences are selected randomly as the training set, the validation set, and the testing set, respectively.

RenCECps: This corpus contains 1487 Chinese blogs with a total of 35 096 sentences [41]. The emotion categories are joy, hate, love, sorrow, anxiety, surprise, anger, and expect. We limit the number of words in a sentence from 5 to 20, so as to generate a labeled short text dataset with 19 338 sentences. For this dataset, 60% and 40% of sentences are selected randomly as the training set and the testing set, respectively.

TABLE VI
PERFORMANCE STATISTICS OF DIFFERENT MODELS. (a) AP OVER *SemEval*. (b) AP OVER *ISEAR*. (c) AP OVER *RenCECps*. (d) HD OVER *SemEval*. (e) HD OVER *ISEAR*. (f) HD OVER *RenCECps*. (g) *Accuracy* OVER *SemEval*. (h) *Accuracy* OVER *ISEAR*. (i) *Accuracy* OVER *RenCECps*

(a)

Models	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
WLTM	0.1952	0.0006	**0.2411**	0.0002
XETM	**0.3121**	0.0007	0.1995	0.0004
LLDA [32]	0.0032	0.0032	0.00827	6.77E-05
BTM [8]	0.1895	0.0011	0.2262	0.0008
ETM [4]	0.2268	0.0009	0.0666	0.0001
CSTM [20]	0.3001	0.0001	0.1114	0.0009
SLTM [15]	0.1746	0.0044	0.0205	0.0005
SNSTM [43]	0.2468	0.0011	0.1495	0.0001

(b)

Models	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
WLTM	**0.4299**	4.12E-05	**0.4496**	9.31E-05
XETM	0.2977	1.93E-05	0.3424	0.0001
LLDA [32]	0.0142	3.41E-05	0.0239	7.90E-05
BTM [8]	0.3327	0.0014	0.3590	0.0015
ETM [4]	0.3470	6.87E-05	0.4149	0.0002
CSTM [20]	0.2111	0.0004	0.2269	0.0007
SLTM [15]	0.0957	0.0012	0.0896	0.0010
SNSTM [43]	0.3323	0.0001	0.3716	0.0002

(c)

Models	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
WLTM	**0.4005**	4.12E-05	**0.3112**	9.31E-05
XETM	0.3436	4.27E-06	0.2350	6.54E-05
LLDA [32]	0.1442	3.41E-05	0.0239	7.90E-05
BTM [8]	0.2704	0.0005	0.1164	0.0025
ETM [4]	0.3470	6.87E-05	0.4149	0.0002
CSTM [20]	0.2111	0.0004	0.2269	0.0007
SLTM [15]	0.0957	0.0012	0.0896	0.0010
SNSTM [43]	0.3710	0.0001	0.2509	0.0002

(d)

Models	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
WLTM	0.4655	3.19E-06	**0.4816**	8.88E-07
XETM	0.4837	1.97E-07	0.4890	3.15E-08
LLDA [32]	0.4833	1.79E-11	0.4899	3.58E-12
BTM [8]	0.4713	1.11E-05	0.4817	4.65E-10
ETM [4]	0.4852	8.97E-09	0.4999	7.60E-10
CSTM [20]	**0.4598**	1.09E-05	0.4867	3.39E-06
SLTM [15]	0.4758	5.60E-05	0.4901	3.44E-07
SNSTM [43]	0.7623	0.0001	0.7793	0.0001

(e)

Models	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
WLTM	**0.6796**	2.54E-05	**0.6956**	6.13E-06
XETM	0.7874	2.97E-09	0.7874	2.97E-09
LLDA [32]	0.7887	8.75E-13	0.7887	8.78E-13
BTM [8]	0.7289	0.0002	0.7335	0.0002
ETM [4]	0.7102	9.61E-06	0.7122	0.0002
CSTM [20]	0.7637	4.91E-05	0.7669	2.96E-05
SLTM [15]	0.7878	9.19E-10	0.7856	2.37E-06
SNSTM [43]	0.7835	0.0001	0.7835	0.0001

(f)

Models	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
WLTM	**0.6671**	2.54E-05	**0.7170**	6.13E-06
XETM	0.7002	1.32E-06	0.7525	1.57E-09
LLDA [32]	0.7413	8.75E-13	0.7649	8.78E-13
BTM [8]	0.7519	0.0003	0.7646	0.0004
ETM [4]	0.7884	8.75E-13	0.7884	6.48E-11
CSTM [20]	0.7637	4.91E-05	0.7669	2.96E-05
SLTM [15]	0.7878	9.19E-10	0.7856	2.37E-06
SNSTM [43]	0.7194	0.0001	0.7765	0.0001

(g)

Models	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
WLTM	0.3643	**0.5909**	**0.7761**
XETM	0.3567	0.5799	0.7411
LLDA [32]	0.2020	0.2680	0.4520
BTM [8]	0.3160	0.5330	0.7000
ETM [4]	0.2541	0.5019	0.6684
CSTM [20]	0.2977	0.5384	0.7225
SLTM [15]	0.2084	0.4253	0.6161
SNSTM [43]	**0.3890**	0.5630	0.7380

(h)

Models	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
WLTM	0.4012	0.5675	0.6608
XETM	0.3867	0.5891	0.7175
LLDA [32]	0.1429	0.2857	0.4286
BTM [8]	0.3235	0.5577	0.7065
ETM [4]	**0.4850**	**0.6622**	**0.7786**
CSTM [20]	0.2898	0.4661	0.6176
SLTM [15]	0.2040	0.3648	0.5112
SNSTM [43]	0.4540	0.6119	0.7221

(i)

Models	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
WLTM	0.4103	**0.6298**	**0.7603**
XETM	0.3784	0.5868	0.7550
LLDA [32]	0.2280	0.4993	0.6002
BTM [8]	0.3255	0.5389	0.6581
ETM [4]	0.3671	0.5728	0.7510
CSTM [20]	0.3012	0.5419	0.6401
SLTM [15]	0.2310	0.4291	0.5930
SNSTM [43]	**0.4129**	0.6164	0.7584

TABLE VII
PERFORMANCE OF THE WORD-LEVEL BASELINE MODELS. (a) *Semeval*. (b) *ISEAR*. (c) *RenCECps*

(a)

Models	$AP_{document}$	$AP_{emotion}$	$HD_{document}$	$HD_{emotion}$	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
SWAT [1]	0.2603	0.2204	0.8201	0.4927	0.3080	0.5220	0.6650
ET [4]	0.2391	0.2304	0.8170	0.5013	0.3060	0.5180	0.6840
SVR [9]	0.0192	0.0013	0.0521	0.5573	0.2610	0.4640	0.6170

(b)

Models	$AP_{document}$	$AP_{emotion}$	$HD_{document}$	$HD_{emotion}$	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
SWAT [1]	0.2112	0.2173	0.7809	0.7201	0.3069	0.4870	0.6350
ET [4]	0.3791	0.4325	0.7913	0.7870	0.3123	0.4912	0.7021
SVR [9]	0.0501	0.0710	0.7204	0.7248	0.1949	0.3356	0.4628

(c)

Models	$AP_{document}$	$AP_{emotion}$	$HD_{document}$	$HD_{emotion}$	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
SWAT [1]	0.2092	0.1641	0.7852	0.7853	0.2975	0.4781	0.6230
ET [4]	0.1790	0.0030	0.7471	0.7953	0.2013	0.3768	0.5375
SVR [9]	0.1021	0.0983	0.7291	0.7381	0.2151	0.3639	0.5921

Table III summarizes the statistics of these three datasets, where the number of documents and mean words of each emotion label are calculated based on the sum of documents having the largest score over that emotion. Note that an emotion can be assessed for both categories and the strength in SemEval and RecCECps. For example, the users annotated four categories (i.e., "Joy," "Fear," "Surprise," and "Sad") for a single news headline—"Test to predict breast cancer relapse is approved" in SemEval, and scores of these categories are 38, 15, 11, and 9, respectively. Therefore, it is suggested to take all emotion scores into account for evaluation [42], rather than only concern about the emotion with the largest score.

B. Experimental Design

We denote WLTM and XETM that incorporate our accelerated algorithm as fWLTM and fXETM, respectively. The term groups are generated for the above three datasets. For instance, when X is equal to 2, there are 5123 and 1 571 829 2-terms in SemEval and ISEAR, respectively. Since the scale of SemEval is too limited, we employ ISEAR to evaluate the efficiency of fWLTM, fXETM, and other models. Some classical approaches that do not exploit topics [1], [4], [9], and topic-level baselines, including LLDA [32], BTM [8], emotion-topic model (ETM) [4], CSTM [20], sentiment latent-topic model (SLTM) [15], and siamese network-based supervised topic model (SNSTM) [43] are implemented as baselines.

For BTM, WLTM, fWLTM, XETM, and fXETM, all term groups are generated by taking each short text as an individual context unit. We employ SVR [9] with radial basis function (RBF) as the kernel function to predict emotion distributions of unlabeled documents for WLTM, fWLTM, LLDA, and BTM. To tune the parameters of SVR, five-fold cross-validation is performed on the training set for SemEval and RenCECps and on the validation set for ISEAR. For XETM and fXETM, the emotion distribution of each testing document is estimated by (12). Similar to the previous studies [4], [32], [34], the hyperparameters α and β are, respectively, set to symmetric Dirichlet priors with values of 0.1 and 0.01, and the number of Gibbs sampling iteration is set to 500. The running time is recorded on a 24 core high-performance computational node with 64G memory. To ensure the effectiveness of MH sampling, we set MH sampling times to 2, which means that the topic of an X-term is sampled twice at each iteration.

To take emotion scores into account, two fine-grained metrics, the averaged Pearson's correlation coefficients (AP) and the averaged Hellinger distance (HD), are used for evaluation [1], [20], [44]. Given two vectors p and q with element x, AP and HD are estimated as follows:

$$\mathrm{AP}(p,q) = \frac{\sum_x (p(x)-\overline{p})(q(x)-\overline{q})}{\sqrt{\sum_x (p(x)-\overline{p})^2}\sqrt{\sum_x (q(x)-\overline{q})^2}}$$

$$\mathrm{HD}(p,q) = \sqrt{\frac{1}{2}\sum_x \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2}.$$

In the above, $\overline{p}$ and $\overline{q}$ are the mean values of p and q, respectively. For completeness, both AP and HD are measured between the predicted probabilities and the actual votes over the document level ($\mathrm{AP}_{\mathrm{document}}$ and $\mathrm{HD}_{\mathrm{document}}$), and over the emotion level ($\mathrm{AP}_{\mathrm{emotion}}$ and $\mathrm{HD}_{\mathrm{emotion}}$), respectively. The value of AP ranges from -1 to 1, where 1 indicates a perfect prediction with the maximum correlation coefficient, and the value of HD ranges from 0 to 1, where 0 indicates a perfect prediction with the minimum Hellinger distance.

We also compare the performance of different models by a coarse-grained metric, that is, Accuracy@N ($N = 1, 2, 3$) [3]. Specifically, given a document d, an actual emotion set $E_{\mathrm{top}N@d}$ which includes N top-ranked emotions, and the top-ranked predicted emotion ε_p, $\mathrm{Accuracy}_d@N$ is first calculated as

$$\mathrm{Accuracy}_d@N = \begin{cases} 1 & \text{if } \varepsilon_p \in E_{\mathrm{top}}N@d \\ 0 & \text{else.} \end{cases}$$

Then, Accuracy@N for the testing set D is

$$\mathrm{Accuracy}@N = \sum_{d\in D} \frac{\mathrm{Accuracy}_d@N}{|D|}.$$

As mentioned earlier, the topic number of WLTM that indicates documents' latent aspects depends on the multiplier (i.e., τ) between topic and emotion numbers. To evaluate the performance of our models with different numbers of topics, we vary τ from 1 to 15 for three datasets in our experiments, thus the topic numbers of SemEval, ISEAR, and RenCECps range from $|N_{E_{\mathrm{SemEval}}} * 1| = 6$ to $|N_{E_{\mathrm{SemEval}}} * 15| = 90$, from $|N_{E_{\mathrm{ISEAR}}} * 1| = 7$ to $|N_{E_{\mathrm{ISEAR}}} * 15| = 105$, and from $|N_{E_{\mathrm{RenCECps}}} * 1| = 8$ to $|N_{E_{\mathrm{RenCECps}}} * 15| = 120$, respectively.

C. Influence of X

In the first part of experiments, we evaluate the influence of X (i.e., the number of words for each term group) on the

TABLE VIII
PERFORMANCE OF THE ACCELERATED MODELS. (a) AP OVER *ISEAR*. (b) HD OVER *ISEAR*. (c) *Accuracy* OVER *ISEAR*

(a)

Models	$AP_{document}$		$AP_{emotion}$	
	Mean	Variance	Mean	Variance
fWLTM	**0.3468**	0.0271	0.3519	0.0436
fXETM	0.2744	0.0048	**0.3805**	0.0109

(b)

Models	$HD_{document}$		$HD_{emotion}$	
	Mean	Variance	Mean	Variance
fWLTM	**0.7138**	0.0210	**0.7272**	0.0163
fXETM	0.7878	2.93E-05	0.7878	2.95E-05

(c)

Models	$Accuracy@1$	$Accuracy@2$	$Accuracy@3$
fWLTM	**0.3821**	**0.5412**	**0.6431**
fXETM	0.3672	0.5307	0.6214

model performance by setting X to 1, 2, 3, and 4. The experimental results in Tables IV and V indicate that the proposed models perform the worst when X is set to 1 in most cases. It is reasonable because of the number of words in a short message is limited. We can also observe that the proposed models perform the best when X is set to 2 mostly, which indicates that two words are more likely to form a phrase (i.e., a semantically related term) than others for these three datasets. Unless otherwise specified, we set X to 2 in the following experiments.

D. Comparison With Baselines

Table VI presents the mean and variance of model performance in terms of AP, HD, and Accuracy, where the top values of each metric are highlighted in boldface.

According to the AP results, the proposed WLTM achieves better performance than baselines on these three datasets in most cases, except for a sightly worse performance than some other models in terms of $\mathrm{AP}_{\mathrm{document}}$ over SemEval. A possible reason is that there are 28 words appearing in the 1000 testing documents but not in the 246 training documents. Since the lack of samples in tuning parameters, WLTM, LLDA, and BTM which employ SVR for prediction may underfit emotional distributions at the document level. By generating emotion-topic and topic-word probabilities without parameter tuning, the proposed XETM yields competitive performance on $\mathrm{AP}_{\mathrm{document}}$. In terms of $\mathrm{AP}_{\mathrm{emotion}}$, WLTM achieves the best mean value of 0.2411 and XETM ranks top 3 with a value of 0.1995. Particularly, the variances of WLTM and XETM indicate the performance stability of our two models. According to the results over ISEAR, WLTM yields competitive performance on both evaluation metrics and the corresponding variances rank top 3. On the other hand, XETM cannot achieve the best results on AP, but its variances with different multiplier values also rank top 3.

Note that the Hellinger distance measures the similarity between two probability distributions. Table VI shows that WLTM achieves the best performance except for $\mathrm{HD}_{\mathrm{document}}$ on *SemEval*, in which CSTM is slightly better. These results indicate that the predicted emotion distribution for WLTM is quite close to the prior emotion label distribution. This is because the generation of topics is constrained by one-to-many

TABLE IX
RUNNING TIME WITH DIFFERENT VALUES OF τ OVER *ISEAR* ($N_{\text{iter}} = 500$), UNIT: SECOND

τ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
fWLTM	477	526	536	548	578	578	591	608	617	627	646	655	664	674	676
WLTM	448	618	786	913	1031	1153	1297	1424	1559	1679	1816	1958	2067	2202	2337
fXETM	2921	3079	3156	3176	3173	3226	3241	3264	3277	3294	3309	3324	3324	3341	3327
XETM	985	1245	1514	1780	2032	2285	2551	2818	3041	3330	3607	3846	4109	4354	4628
BTM [8]	402	660	980	1238	1497	1751	2011	2266	2534	2824	3126	3349	3737	3893	4205

projection between emotions and topics for WLTM, which renders the extracted topics corresponding to relative emotions. On the other hand, XETM achieves modest performance among these three datasets. The reason may be that XETM first samples one emotion label ε and then generates a topic conditioned to ε. However, there are more than one emotion label for most sentences in both *SemEval* and *RenCECps*.

As for the metrics of Accuracy@1, Accuracy@2, and Accuracy@3, the proposed two models also perform competitively. WLTM outperforms other models on both *SemEval* and *RenCECps*. However, ETM performs better on *ISEAR*, in which there are only one label for each document. As mentioned earlier, the sampling of topics is constrained by one emotion for ETM, thus ETM mostly samples an emotion which is the actual label of the document.

To compare the performance of our supervised topic models on short text emotion detection statistically, we conduct t-tests to test the assumption that the difference in performance between paired models has a mean value of zero. T-test is conducted on the proposed models (i.e., WLTM and XETM) and the baseline models. The results indicate that the proposed WLTM outperforms the baselines of LLDA, BTM, ETM, CSTM, SLTM, and SNSTM significantly with p-values much less than 0.05. The p-values between XETM and most of baselines, except BTM, are less than 0.05. The difference in performance between XETM and BTM is not statistically significant with a p-value equal to 0.2856 in terms of $\text{HD}_{\text{document}}$. Similar results are observed for other metrics.

We also implement some word-level baseline models for comparison, that is, ET [4], SVR [9], and SWAT [1]. The experimental results are shown in Table VII. ET computes $P(\varepsilon)$ and $P(w|\varepsilon)$ by extending naïve Bayes, and SWAT predicts emotions via aggregating the emotion of each word. Since the input is quite sparse for both datasets, SVR based on words achieves the worst performance in terms of AP. On the other hand, our WLTM that employs SVR using topic distributions as the input can outperform the conventional SVR by a large margin. These results indicate that our models effectively extract valuable features for short text emotion detection.

E. Evaluation on the Accelerated Algorithm

In this part, we evaluate the performance of fWLTM and fXETM on *ISEAR* in terms of AP, HD, and *Accuracy*, and compare the corresponding running time with topic models based on 2-terms, i.e., WLTM, XETM, and BTM. For all metrics, we vary τ from 1 to 15 and present the mean and variance values in Table VIII. Specifically, the $\text{AP}_{\text{document}}$ of fWLTM reaches the best value of 0.3943 with $\tau = 5$ and has an averaged value of 0.3519. Although the above mean value is less than the averaged $\text{AP}_{\text{document}}$ value of WLTM (i.e., 0.4299), it is better than BTM with an averaged value of 0.3327. Furthermore, the averaged $\text{AP}_{\text{emotion}}$ value of fWLTM is 0.3519, which is very close to that of BTM (i.e., 0.3590). Particularly, the best value of $\text{AP}_{\text{emotion}}$ is 0.4175 for fWLTM when τ equals to 4 or 5, which is higher than the best value of BTM. Although WLTM achieves the highest values of $\text{AP}_{\text{document}}$ and $\text{AP}_{\text{emotion}}$, fWLTM is much less time consuming and more efficient than WLTM as shown in the following evaluation. On the other hand, the results of fXETM indicate that although the averaged $\text{AP}_{\text{document}}$ value of fXETM (i.e., 0.2744) is less than that of XETM with a value of 0.2977, the averaged $\text{AP}_{\text{emotion}}$ reaches a value of 0.3806, which outperforms XETM and BTM. The performance variances of fWLTM and fXETM show that both of them have good stability. In terms of HD, fWLTM achieves a smaller value than those of baselines, and outperforms that of XETM, which means fWLTM generates better topic distribution conditioned to emotion labels during the supervised training process. As for *Accuracy*, the results of fWLTM are slightly lower than WLTM but still competitive for those of baselines above.

TABLE X
EMOTION LEXICON SAMPLES FROM WLTM AND FWLTM OVER *ISEAR*

Models	Topics	Representative words	Emotions
WLTM	1	back, angry, parent, bad, work	anger
	7	disgust, film, woman, felt, drunk	disgust
	23	win, team, competition, prize, game	joy
fWLTM	2	corrupt, degenerate, bureaucrat, decade, tapism	anger
	6	disgust, felt, cigarette, tax, sexuality	disgust
	31	ashamed, calm, confuse, chicken, toilet	shame

To evaluate the above results statistically, we conduct t-tests between the performance of fWLTM and those of WLTM and BTM. The p-values of BTM are almost larger than 0.05, and the mean values and variances of AP are closed to fWLTM. Therefore, the performance of fWLTM is as competitive as BTM. Although WLTM achieves the best performance of correlation coefficients, it is expensive when training on a lager scale of documents or features.

The t-tests between the performance of fXETM and those of XETM and BTM is also evaluated. Specifically, the p-values between the $\text{AP}_{\text{emotion}}$ of fXETM and that of XETM are less than 0.05, which means that the performance on $\text{AP}_{\text{emotion}}$ of fXETM is better than that of XETM statistically. Compared to BTM, the proposed fXETM also achieves better performance on the metric of $\text{AP}_{\text{emotion}}$.

Table IX presents the running time of those models with different values of τ. In the experiment, we set the iteration time N_{iter} to 500, vary τ from 1 to 15, and record how many seconds are used for these models. Although BTM performs

TABLE XI
EMOTION LEXICON SAMPLES FROM XETM AND fXETM OVER *ISEAR*

Models	Topics	Representative words	Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame
XETM	6	carnival bright organizer champagne fascinating	3.5962E-05	0.0002	0.0001	0.0004	**0.9991**	8.6301E-05	3.5962E-05
	17	divide freshmen dash seclude spate	7.5754E-05	0.0005	**0.9991**	7.5754E-05	7.5754E-05	7.5754E-05	7.5754E-05
	35	diagnosis pretentious fortuitous employee tablet	**0.9974**	0.0001	0.0002	0.0010	0.0004	0.0002	0.0004
fXETM	1	reserve festival impatient bureaucrat corruption	**0.9067**	0.0050	0.0111	0.0338	0.0034	0.0068	0.0332
	20	skid swerve steep aircraft precipice	0.0557	0.0336	**0.8305**	0.0295	0.0127	0.0136	0.0245
	32	locker supplementary trainer housemen guiltier	0.0751	0.0479	0.0169	**0.7491**	0.0154	0.0262	0.0694

competitively, it is quite time consuming. In general, the running time of fWLTM is always less than that of BTM and WLTM when τ is larger than 1. Specifically, in terms of a larger τ value, when τ increases by 1, the running time of fWLTM increases by 10 s while that of WLTM increases by more than 100 s. On the other hand, the running time of fXETM is almost 3300 s with different values of τ, but that of XETM is more than 4000 s finally. Moreover, with τ getting larger, the values of running time of fWLTM and fXETM increase slowly while those of the WLTM, XETM, and BTM increase fast and become expensive. This is because the accelerated models spend stable time to construct the Alias table with different τ values. As mentioned above, we construct the Alias table for each word in each iteration, so the running time relies on the number of different words instead of the number of topics N_z. After constructing the Alias table, we can spend $O(1)$ time to sample a topic for each term group. As for Gibbs sampling, we need compute the topic probability distribution for each topic, which has a time complexity of $O(N_z)$. Thus, Gibbs sampling is computationally prohibitive under a large N_z. In this article, by employing the Alias method and developing the supervised MH sampling, both fWLTM and fXETM are efficient with competitive performance.

F. Emotion Lexicon Samples

As stated earlier, both WLTM and XETM are supervised topic models using prior emotion scores to restrict the topic probability during each sampling process. Specifically, the proposed WLTM directly maps topics to emotion labels, while XETM employs a topic-emotion layer to connect words and topics, from which we can conduct a topic-emotion probability distribution using (11). Therefore, for the example as shown in Fig. 1, each topic of WLTM is mapped to the corresponding emotion label. Each topic of XETM performs a probability distribution for each emotion after sampling. In the following text, we show the emotion lexicon samples over *ISEAR* for WLTM, fWLTM, XETM, and fXETM, in which the value of τ is 5, so the number of topics is $\tau \times N_{E_{\text{ISEAR}}} = 42$. For XETM and fXETM, the distribution of an emotion specific to each topic can be estimated by (10). For these four proposed models, the probabilities of words conditioned to each topic are estimated according to (5) or (11).

Table X shows the emotion lexicon samples that are generated by WLTM and fWLTM. In the second and the third columns, we present sample topics and their representative words. The last column is the relative emotion label from our one-to-many mapping method. As shown in the sample results, it is convinced that both WLTM and fWLTM can effectively generate the emotion-related words for each topic. For example, in topic 23 from WLTM, the sample words are "win, team, competition, prize, game," which mostly means "a team win in a competition and win the price," and the emotion of that is exactly "joy." In topic 2 of fWTLM, the sample word "corrupt" means the phenomenon of corruption and the related emotion is "anger."

Table XI shows the emotion lexicon samples from XETM and fXETM, where the sample topics and their representative words with the largest conditional probabilities are presented in the second and the third columns. The distributions of seven emotions for each topic are listed in the other columns, and the largest values are boldfaced. First, the samples indicate that the topics are strongly relative to one emotion label, for example, the topic 6 from XETM has a probability of 99% relating to the emotion of "joy," the topic 1 from fXETM is almost 90% relating to the emotion of "anger." Second, the sample words have the exact emotional expression like the topic-related emotion. For example, the word "carnival" in topic 6 from XETM is mostly implied in a festival event, the word "skid" in topic 20 from fXETM means "stop the car," which is probably used for a traffic accident news with the emotion of "fear."

V. CONCLUSION

Emotion detection aims to predict emotional responses embedded in documents. This article proposed two models, WLTM and XETM, to address the issue of feature sparsity in detecting emotions over short messages. In this article, we evaluated the influence of the number of words in a term group and compare the performance with state-of-the-art baselines. To reduce the time cost of estimating parameters, we proposed the accelerated methods, fWLTM and fXETM to generate topics and detect emotions efficiently. The experimental results indicated that the accelerated models were quite less time consuming without reducing much quality, especially for the proposed fWLTM. Considering that users often use sarcasm for emphasizing their sentiment [45], our future work will focus on incorporating sarcasm detection into our method. Furthermore, we intend to evaluate the model performance on multimodal sentiment analysis [46]. We also plan to extend the fast parametric topic models to nonparametric ones [47]–[49], so as to handle text streams where the number of topics is hard to be specified manually.

ACKNOWLEDGMENT

Some contents from the conference version are reused in this journal article as this article is a follow-up work of the conference paper. The new contents of this article are more

than 70% according to the regulation of the published journal. The new contents can be summarized in the following aspects.

1) We extend the basic proposed models by setting the length of term groups as a flexible variable.
2) To reduce the time complexity of the generation process, we newly propose an accelerated algorithm for our basic models.
3) In the experiments, we evaluate and compare the performance of our models with different lengths of a term group.
4) To conduct in-depth analysis, we present the emotion lexicon samples that are generated by our models.
5) We add a Chinese corpus and two metrics (i.e., the Hellinger distance and accuracy) to evaluate the effectiveness of different models comprehensively.

REFERENCES

[1] P. Katz, M. Singleton, and R. H. Wicentowski, "SWAT-MP: The SemeVal-2007 systems for task 5 and task 14," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 308–313.

[2] C. Strapparava and R. Mihalcea, "SemeVal-2007 task 14: Affective text," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 70–74.

[3] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov./Dec. 2017.

[4] S. Bao *et al.*, "Mining social emotions from affective text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 9, pp. 1658–1670, Sep. 2012.

[5] Y. Rao, J. Lei, W. Liu, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web*, vol. 17, no. 4, pp. 723–742, 2014.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[7] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledgebase," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2330–2336.

[8] X. Cheng, Y. Lan, J. Guo, and X. Yan, "BTM: Topic modeling over short texts," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2928–2941, Dec. 2014.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 389–396, 2011.

[10] X. He, H. Xu, J. Li, L. He, and L. L. Yu, "FastBTM: Reducing the sampling time for biterm topic model," *Knowl. Based Syst.*, vol. 132, pp. 11–20, Sep. 2017.

[11] A. J. Walker, "New fast method for generating discrete random numbers with arbitrary frequency distributions," *Electron. Lett.*, vol. 10, no. 8, pp. 127–128, Apr. 1974.

[12] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[13] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[14] A. Gangemi, V. Presutti, and D. R. Recupero, "Frame-based detection of opinion holders and topics: A model and a tool," *IEEE Comput. Intell. Mag.*, vol. 9, no. 1, pp. 20–30, Feb. 2014.

[15] Y. Rao, Q. Li, X. Mao, and W. Liu, "Sentiment topic models for social emotion mining," *Inf. Sci.*, vol. 266, pp. 90–100, May 2014.

[16] Y. Rao, Q. Li, W. Liu, Q. Wu, and X. Quan, "Affective topic model for social emotion detection," *Neural Netw.*, vol. 58, pp. 29–37, Oct. 2014.

[17] Q. Yang, Y. Rao, H. Xie, J. Wang, F. L. Wang, and W. H. Chan, "Segment-level joint topic-sentiment model for online review analysis," *IEEE Intell. Syst.*, vol. 34, no. 1, pp. 43–50, Jan./Feb. 2019.

[18] M. Dragoni, S. Poria, and E. Cambria, "OntoSenticNet: A commonsense ontology for sentiment analysis," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 77–85, May/Jun. 2018.

[19] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1795–1802.

[20] Y. Rao, "Contextual sentiment topic model for adaptive social emotion classification," *IEEE Intell. Syst.*, vol. 31, no. 1, pp. 41–47, Jan./Feb. 2016.

[21] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive Bayes text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006.

[22] J. Li, Y. Rao, F. Jin, H. Chen, and X. Xiang, "Multi-label maximum entropy model for social emotion classification over short text," *Neurocomputing*, vol. 210, pp. 247–256, Oct. 2016.

[23] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Nat. Lang. Process.*, 2002, pp. 79–86.

[24] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment embeddings with applications to sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 496–509, Feb. 2016.

[25] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2016, pp. 214–224.

[26] X. Li, Y. Rao, H. Xie, R. Y. K. Lau, J. Yin, and F. L. Wang, "Bootstrapping social emotion classification with semantically rich hybrid neural networks," *IEEE Trans. Affective Comput.*, vol. 8, no. 4, pp. 428–442, Oct./Dec. 2017.

[27] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis," *Cogn. Comput.*, vol. 10, no. 4, pp. 639–650, 2018.

[28] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 417–424.

[29] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & Web with hidden topics from large-scale data collections," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 91–100.

[30] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, 2011, pp. 775–784.

[31] Y. Rao, H. Xie, J. Li, F. Jin, F. L. Wang, and Q. Li, "Social emotion classification of short text via topic-level maximum entropy model," *Inf. Manag.*, vol. 53, no. 8, pp. 978–986, 2016.

[32] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2009, pp. 248–256.

[33] M. A. Taddy, "On estimation and selection for topic models," in *Proc. 15th Int. Conf. Artif. Intell. Stat.*, 2012, pp. 1184–1193.

[34] S. Bao *et al.*, "Joint emotion-topic modeling for social affective text mining," in *Proc. 9th IEEE Int. Conf. Data Min.*, 2009, pp. 699–704.

[35] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. s1, pp. 5228–5235, 2004.

[36] R. Y. K. Lau, Y. Xia, and Y. Ye, "A probabilistic generative model for mining cybercriminal networks from online social media," *IEEE Comput. Intell. Mag.*, vol. 9, no. 1, pp. 31–43, Feb. 2014.

[37] J. Geweke and H. Tanizaki, "Bayesian estimation of state-space models using the Metropolis–Hastings algorithm within Gibbs sampling," *Comput. Stat. Data Anal.*, vol. 37, no. 2, pp. 151–170, 2001.

[38] L. Tierney, "Markov chains for exploring posterior distributions," *Ann. Stat.*, vol. 22, no. 4, pp. 1701–1728, 1994.

[39] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 384–392, 1993.

[40] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *J. Pers. Soc. Psychol.*, vol. 66, no. 2, pp. 310–328, 1994.

[41] C. Quan and F. Ren, "Sentence emotion analysis and recognition based on emotion words using REN-CECPS," *Int. J. Adv. Intell. Paradigms*, vol. 2, no. 1, pp. 105–117, 2010.

[42] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, "Emotion distribution learning from texts," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2016, pp. 638–647.

[43] M. Huang, Y. Rao, Y. Liu, H. Xie, and F. L. Wang, "Siamese network-based supervised topic modeling," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2018, pp. 4652–4662.

[44] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*. New York, NY, USA: Springer, 2012.

[45] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, Jan. 2019.

[46] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.

[47] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proc. 17th Adv. Neural Inf. Process. Syst.*, 2004, pp. 1385–1392.

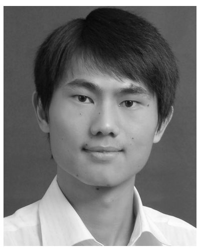
[48] S. J. Gershman and D. M. Blei, “A tutorial on Bayesian nonparametric models,” *J. Math. Psychol.*, vol. 56, no. 1, pp. 1–12, 2012.

[49] J. Xuan, J. Lu, and G. Zhang, “A survey on Bayesian nonparametric learning,” *ACM Comput. Surveys*, vol. 52, no. 1, pp. 1–36, 2019.

[50] Y. Rao *et al.*, “Supervised intensive topic models for emotion detection over short text,” in *Proc. 22nd Int. Conf. Database Syst. Adv. Appl.*, 2017, pp. 408–422.

Jianhui Pang received the Bachelor of Engineering degree in computer science from Sun Yat-sen University, Guangzhou, China.

His current research interests include topic modeling and emotion detection.



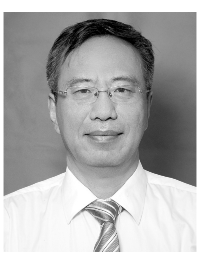
Yanghui Rao (M’18) received the master’s degree from the Graduate University of the Chinese Academy of Science, Beijing, China, in 2010, and the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2014.

He is an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has published over 20 refereed journals and conference papers, including the *ACM Transactions on Information Systems*, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE INTELLIGENT SYSTEMS, ACL, EMNLP, CIKM, and DASFAA. His current research interests include topic modeling, emotion detection, and natural language processing.

Haoran Xie (M’15) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong.

He is an Associate Professor with Lingnan University, Hong Kong. He has totally published 170 research publications, including 71 journal articles. His current research interests include artificial intelligence, big data, and educational technology.

Dr. Xie was a recipient of ten research awards, including the Golden Medal and British Innovation Award from International Invention Innovation Competition in Canada and the Second Prize Winner from Multimedia Grand Challenges of ACM Multimedia in 2019. His proposed LSGAN published in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and ICCV, with over 700 citations in two years, has been included in a course in Stanford University and implemented by Google TensorFlow.



Xizhao Wang (M’03–SM’04–F’12) received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1998.

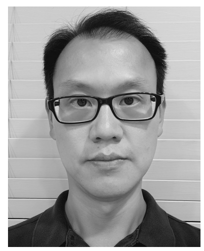
He was a Research Fellow with Hong Kong Polytechnic University, Hong Kong, from 1998 to 2001, and served with Hebei University, Baoding, China, as a Professor and the Dean of the School of the Mathematics and Computer Sciences before from 2001 to 2014. After 2014, he was a Professor with the Big Data Institute of ShenZhen University, Shenzhen, China. He has edited over 10 special issues and published 3 monographs, 2 textbooks, and over 200 peer-reviewed research papers. As a Principle Investigator (PI) or Co-PI, he has completed over 30 research projects. He has supervised over 150 M.Phil. and Ph.D. students. His current research interests include uncertainty modeling and machine learning for big data.

Prof. Wang was a recipient of the IEEE SMCS Outstanding Contribution Award in 2004 and the IEEE SMCS Best Associate Editor Award in 2006. He is the General Co-Chair of the 2002–2018 International Conferences on Machine Learning and Cybernetics, cosponsored by IEEE SMCS. He was a Distinguished Lecturer of the IEEE SMCS. He is the previous BoG Member of IEEE SMC Society, the Chair of the IEEE SMC Technical Committee on Computational Intelligence, the Chief Editor of the *Machine Learning and Cybernetics*, and an associate editor for a couple of journals in the related areas.

Fu Lee Wang (SM’15) received the B.Eng. degree in computer engineering and the M.Phil. degree in computer science and information systems from the University of Hong Kong, Hong Kong, and the Ph.D. degree in systems engineering and engineering management from the Chinese University of Hong Kong, Hong Kong.

He is with the School of Science and Technology, Open University of Hong Kong, Hong Kong. He has published over 150 academic articles in refereed journals and conference proceedings. His current research interests include educational technology, information retrieval, computer graphics, and bioinformatics.

Prof. Wang is a fellow of BCS and a Senior Member of ACM. He was the Chair of the IEEE Hong Kong Section Computer Chapter and ACM Hong Kong Chapter.



Tak-Lam Wong (M’09) received the Bachelor of Engineering and M.Phil. degrees in systems engineering and engineering management, the Postgraduate Diploma of Education degree in mathematics, and the Ph.D. degree in systems engineering and engineering management from the Chinese University of Hong Kong, Hong Kong.

He is a Professor with the Department of Computing Studies and Information Systems, Douglas College, New Westminster, BC, Canada. He has published papers in different journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *ACM Transactions on Information Systems*, and conferences, including SIGIR, SIGKDD, AAAI, and WWW. His current research interests include Web mining, data mining, information extraction, machine learning, e-learning, programming education, and knowledge management.

Prof. Wong also served as the Chair of the IEEE Hong Kong Section Computer Chapter in 2016 and 2017.

Qing Li (SM’07) received the B.Eng. degree in computer science from Hunan University, Changsha, China, and the M.Sc. and Ph.D. degrees in computer science from the University of Southern California, Los Angeles, CA, USA.

He is a Chair Professor with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. His current research interests include multimodal data management, conceptual data modeling, social media, Web services, and e-learning systems. He has authored/coauthored over 400 publications in the above areas.

Dr. Li is actively involved in the research community and has served as an Associate Editor of a number of major technical journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the *ACM Transactions on Internet Technology*, *Data Science and Engineering*, World Wide Web, and the *Journal of Web Engineering*, in addition to being a Conference and Program Chair/Co-Chair of numerous major international conferences. He also sits in the Steering Committees of DASFAA, ER, ACM RecSys, IEEE U-MEDIA, and ICWL. He is a fellow of IEE/IET, U.K., and a Distinguished Member of CCF, China.