World Wide Web (2020) 23:1553–1587
https://doi.org/10.1007/s11280-019-00760-3

User group based emotion detection and topic discovery over short text

Jiachun Feng[1] · Yanghui Rao[1] · Haoran Xie[2] · Fu Lee Wang[3] · Qing Li[4]

Received: 25 April 2019 / Revised: 26 September 2019 / Accepted: 4 November 2019 /
Published online: 12 December 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
In recent years, with the development of social media platforms, more and more people express their emotions online through short messages. It is quite valuable to detect emotions and relevant topics from such data. However, the feature sparsity of short texts brings challenges to joint topic-emotion models. In many cases, it is necessary to know not only what people think of specific topics, but also which individuals have similar feedback, and what characteristics of these users have. In this paper, we propose a user group based topic-emotion model named UGTE for emotions detection and topic discovery, which can alleviate the above feature sparsity problem of short texts. Specifically, the characteristics of each user are used to discover groups of individuals who share similar emotions, and UGTE aggregates short texts within a group into long pseudo-documents effectively. Experiments conducted on a real-world short text dataset validate the effectiveness of our proposed model.

Keywords Joint topic-emotion model · Short text modeling · User characteristics · User group based mining

✉ Yanghui Rao
raoyangh@mail.sysu.edu.cn

Jiachun Feng
fengjch5@mail2.sysu.edu.cn

Haoran Xie
hrxie2@gmail.com

Fu Lee Wang
pwang@ouhk.edu.hk

Qing Li
csqli@comp.polyu.edu.hk

1 School of Data and Computer science, Sun Yat-sen University, Guangdong, China

2 Department of Computing and Decision Sciences, Lingnan University, New Territories, Hong Kong

3 School of Science and Technology, The Open University of Hong Kong, Kowloon, Hong Kong

4 Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

1 Introduction

The rapid growth of social media platforms results in the increasing number of people who express their emotions through short messages [20]. To extract the great value from this type of data for social emotion mining and monitoring, it is necessary to perform topic discovery and emotion detection to identify topics and emotions embedded in short texts [11].

Topic discovery aims to model topics from documents based on their content, and emotion detection identifies emotions from documents at the word, sentence or document level. In the scenario of emotion mining, public emotions always vary from one topic to another topic, and topics trigger public emotions. Therefore, topic discovery and emotion detection are closely related such that jointly modeling topics and emotions is an appropriate way to conduct these tasks [34]. Furthermore, we may also want to learn not only the emotions of a single document or user, but also the statistical results based on groups of individuals sharing similar interests. For example, editors of magazines often want to identify common interests among readers to ensure that all of the major interests are covered in each issue. The editors are also interested in their readers' characteristics (e.g., sex, age, and education level) to maintain the magazine's content appropriately. Thus, there are practical reasons for jointly modeling topics, emotions and user groups [33].

However, there are challenges to effectively detect topics and emotions in short texts. First, each short message includes only a few words, resulting the lack of significant context [36]. Models directly applied to these types of text often suffer from the feature sparsity problem leading to undesirable results. Second, there is the question of which information should be used when discovering representative groups of users. Third, there is the question of how to model content, emotions, and user information within groups to capture the relationships between topics, emotions, and users.

Conventional emotion-aware topic models only present results at the word, sentence, and document-level. Emotion Topic Model (ETM) [2] assumes every word is selected according to specific topics and emotions. Multi-label Supervised Topic Model (MSTM) and Sentiment Latent Topic Model (SLTM) [24] first discover topics within each document and then analyze emotions towards those topics at the document level. These methods can not produce high-level results for groups of users, which we denote as user group level in this paper. They also suffer from the sparsity problem in short texts. Time-User Sentiment/Topic Latent Dirichlet Allocation (TUS-LDA) [32] aggregates short texts from a single user or a single time interval into lengthy pseudo-documents to tackle the above problem when detecting burst topics and social sentiment feedback. TUS-LDA can work at the user level when topics belong to a user's static interest, or the global level when topics relate to current social issues. However, TUS-LDA can not discover groups of users, or the differences in topical interests and emotions between groups either. Besides, TUS-LDA needs pre-developed sentiment lexicons, which may be limited when dealing with a new emotion label.

Regarding the issue of how to divide users into groups, we observe that the more similar people are, the more likely they share similar interests. For example, in the 45th US presidential election, the Washington Post, an authoritative newspaper, used several sets of data to illustrate the characteristics of supporters of Donald Trump: the proportion of male supporters was 19% more than women; and 50% of those with annual incomes below $50,000 supported Trump versus 32% for those with higher income. Data broadly support the theory of homophily that relates similarity of interests to similarity of emotions [16]. Based on this phenomenon, this paper exploits user characteristics, content and emotions, to carry out topic discovery and emotion detection at the user group level.

We propose a method of emotion detection and topic discovery with the help of user characteristics, which is called User Group based Topic Emotion (UGTE) model. Our main contributions are summarized as follows. Firstly, UGTE models user characteristics, emotions, and content jointly to improve the effectiveness of both emotion detection and topic discovery. By influencing the process of topic generation, user characteristics help to identify semantic group structures. As mentioned above, individuals with similar characteristics are more likely to generate similar emotions. Therefore, when analyzing user group level results, UGTE considers not only topics and emotions information of an individual, but also the effects of user groups, as characterized by gender, social income, education, and others. Secondly, UGTE aggregates short texts into lengthy pseudo-documents and jointly models topics and emotions within each group to address the feature sparsity problem. Finally, different from existing methods, UGTE not only captures the relationship between topics and emotions for every group, presented as distributions over words, but also releases portraits for these groups presented as distributions over characteristics.

The rest of this paper is organized as follows. Section 2 introduces related work concerning topic models on short texts, joint topic emotion modeling, and community based sentiment/emotion detection. Section 3 demonstrates the proposed model and the inference of model parameters. Section 4 presents our experiments and discussions. In Section 5, conclusion is drawn.

2 Related work

2.1 Short text topic models

The topic model provides a solution for implicit semantic mining and understanding. Probabilistic Latent Semantic Analysis (PLSA) [10] is one of the first latent semantic models, which uses expectation-maximization (EM) algorithm for parameter inference. Given the fact that PLSA suffers from the overfitting problem, Latent Dirichlet Allocation (LDA) [3] introduces the Dirichlet distribution as the conjugate prior of topics. In recent years, LDA has achieved great success in information retrieval [21] and topic modeling [6, 15]. However, both LDA and PLSA perform well when mining topics from lengthy documents only. Nowadays, texts from the Internet are typically short and lacking context. The feature sparsity problem arises for LDA and PLSA when applied to short texts [36].

To overcome this limitation, the external document embedding method was first introduced to enrich contextual information in short texts [14, 19, 28]. This method is effective, but the enriched documents are not always consistent with the original messages. Thus, the method may have no effect or even a negative effect on the results. In addition, finding the auxiliary data is expensive and time-consuming. Besides the external document embedding, the Biterm Topic Model (BTM) [7] is an alternative method. It is proposed based on the idea that two words are more likely to belong to a same topic if they co-occurred more frequently. Such a kind of methods lengthen short texts by converting documents into biterm sets. However, the problem of biterm based methods lies in that they bring in little additional word co-occurrence information and therefore still face the feature sparsity problem [38]. Another alternative approach is integrating short texts into lengthy pseudo-documents, which solves the feature sparsity problem without carefully selecting external documents [39]. Twitter LDA [37] aggregates posts from a single user into a pseudo-document to identify topics from the words. TimeUserLDA [8], aggregates posts by user or timestamp to detect “breakout” topics. Such topics fall into two categories: personal static topics and

temporal dynamic topics. Similar to TimeUserLDA, the model that incorporates temporal, personal and extraction factor (TUK-TTM) [35] aggregates posts by time slices or users to produce personalized time-aware tag recommendations. However, these models can not be applied to emotion detection. To mine burst topics on social media, TUS-LDA [32] introduces a sentiment variable to every post aggregated in pseudo-documents. Taking advantage of the aggregation method, our proposed model also uses this idea to address the feature sparsity problem. UGTE differs, however, by aggregating short messages from a user group into a pseudo-document to give group level results.

2.2 Jointly modeling topics and emotions

Data from the Internet contains users' opinions and emotions. In recent years, to jointly model topics and emotions, several researchers have extended topic models to perform emotion detection of user-generated text, such as product and movie reviews [34]. ETM [2] uses emotion labels to implement a supervised emotion topic model for social emotion mining. Different from ETM which was developed from the writer's perspective, MSTM and SLTM [24] model topics and sentiment labels from the perspective of readers. Experiments show that they are more suitable for public voting articles when mining social emotions. The Contextual Sentiment Topic Model (CSTM) [23] proposes to classify reader emotions by explicitly distinguishing context-independent topics from nondiscriminative information such as some very common words, and a contextual theme which characterizes context-dependent information across different collections. However, models mentioned above are applied on regular documents rather than shor texts. Weighted Labeled Topic Model (WLTM) [25] based on BTM models multiple emotion labels and biterms for short text emotion detection jointly. Except LDA-based methods, neural based topic models arise for topic discovery and supervise learning recently. Supervised Neural Topic Model (sNTM) [4] extracts topics based on neural network by following the document-topic distribution in topic models. However, observable labels have a little effect on the process of topic discovery. Neural Siamese Labeled Topic Model (nSLTM) [12] incorporates the supervision of labels into topic modeling, which can be applied to both classification and regression.

Previous joint topic emotion models only model topics and emotions at the word, sentence, or document level. They do not capture emotions and topics within groups of users, nor do they identify which user would be interested in specific topics. Our UGTE approach integrates user characteristics with topics and emotions so that the model performance can be enhanced by exploiting the relationships among topics, emotions, and users.

2.3 Community based sentiment/emotion detection

LDA-based topic models are applied widely to community detection. Community detection has been studied from the perspective of network structural communities and semantic communities. Since most methods for detecting network structural communities use graph partitioning algorithms considering only users' relationships or interactions [18], we do not discuss them here due to their lack of relevance. Different from network structural community detection, semantic community detection takes both network structure and user semantic attributes into consideration. For example, the Group-Topic (GT) model uses entity relationships and textual attributes to simultaneously discover topics for events and communities among the entities [31]. The Topic User Community Mode (TUCM) uses social links, interaction types and context information to detect communities [27]. However, these methods do not take sentiments or emotions into consideration. To conduct sentiment analysis,

the Sentiment-Topic model for Community discovery (STC) aggregates topics, sentiments and interactions among users to detect sentiment-topic level communities [33]. Work in [30] detects sentiment communities with social relationships between users, context and sentiment labels. However, it is unable to discover topics or opinions across communities. The People Opinion Topic (POT) model introduces opinion based community detection to discover hot topics and analyze sentiment along with detecting social communities [5].

The methods mentioned thus far integrate sentiment analysis and community detection to improve model performance on both tasks. However, there are several differences between our work and these studies. First, existing models of community detection mostly work on discovering the best structural community by examining users with more interactions. They take sentiments or user context into consideration and fail to extract topics or sentiments of different communities. Instead, our model attempts to discover topics and emotions at the group level, which not only models topics and emotions but also identifies people sharing similar interests in the same group. Second, most community detection models depend on information from users' social relationships or online interactions. None of the existing models of community detection employ users' characteristics tags to analyze relationships between users interests and their profiles. However, in many cases, a decision maker may want to know what different groups of users think of an event, and which characteristics have the largest influence. Our model achieves such high level results when other models fail.

3 User group based topic emotion model

In this section, we introduce our UGTE model and present its structure. After defining the problem, relevant general terms and notations, we will describe our model in detail. We also present our method of learning parameters.

3.1 Problem definition

Given a set of documents $D = \{d_1, d_2, ..., d_{|D|}\}$ with $|D|$ elements, the vocabulary of D is $W = \{w_1, w_2, ..., w_{|W|}\}$ with size of $|W|$, the set of globally distinct emotion labels is $E = \{e_1, e_2, ..., e_{|E|}\}$ with $|E|$ elements, and the set of users is $U = \{u_1, u_2, ..., u_{|U|}\}$ with size of $|U|$. Suppose every document of D is generated by one user and labeled with one of the above emotions. Each document d_i can be further denoted as $d_i^{r,k}$, which means document d_i is generated by user u_r and labeled with emotion e_k. The words in document d_i are denoted as $W_{d_i} = \{w_{i,1}, w_{i,2}, ..., w_{i,N_i}\}$, where N_i is the total number of words in document d_i.

To discover user group based topics and emotions, we need exploit user characteristics such as age, gender, and country. For total J types of characteristics collected, we denote the set of characteristics tags of the jth type as $F_j = \{f_{j,1}, f_{j,2}, ..., f_{j,|F_j|}\}$ with $|F_j|$ elements. We denote the characteristics tags for each user u_r as $F^{u_r} = \{f_1^r, f_2^r, ..., f_J^r\}$ where f_j^r is the element of the jth type characteristic tag of u_r, that belong to F_j. For example, assume that three users u_1, u_2, and u_3 have characteristics tags $F^{u_1} =$ { 'Male', '22', 'America' }, $F^{u_2} =$ { 'Female', '23', 'America' } and $F^{u_3} =$ { 'Male', '24', 'America' }, respectively. There are totally $J = 3$ different types of characteristics tags: gender, age, and country. From the tag values, we determine that $F_1 =$ { 'Male', 'Female' } with $|F_1| = 2$, $F_2 =$ { '22', '23', '24' } with $|F_2| = 3$ and $F_3 =$ { 'America' } with $|F_3| = 1$.

Our primary task is to jointly discover the topics $Z = \{z_1, z_2, ..., z_{|Z|}\}$ with size of $|Z|$ and the emotions of given documents at the user group level. In other words, we should detect different user groups $G = \{g_1, g_2, ..., g_{|G|}\}$ with $|G|$ elements, infer the topic distributions of different groups θ_g, and analyze the emotion distributions $\phi_{g,z}$ of each topic within groups simultaneously. In our UGTE model, the user groups are latent variables as topics. The number of user groups is a predefined parameter, whose impact will be detailed in Section 4.2. Table 1 provides a summary of the notations used in our presentation.

3.2 Generative process

Conventional joint topic-emotion models focus on the association between emotions and topics at the level of documents or users. To model topics and emotions jointly at the user group level, we propose the UGTE model by adding a user group layer to the generation module of topics and emotions. Figure 1 shows the structure of UGTE. In UGTE, every user u_r is related to a global group distribution π. Every group g is associated with its own characteristics tags distributions $\psi_{g,j}$, topic distribution θ_g, and emotion distribution $\phi_{g,z}$. For a

Table 1 Notations used in UGTE

Notations	Description
W	Vocabulary set
G	Set of user groups
U	user set
Z	Set of topics as latent variables
E	Set of emotions as observable variables
D_{train}	Set of training documents
D_{test}	Set of testing documents
J	Number of types of user characteristics
g, z, e, w	Specific user group g, topic z, emotion label e, and word w
g_{d_i}	Group g that document d_i belongs
$d_i^{r,k}$	Document d_i, which is generated by user u_r and labeled with emotion e_k
F^{u_r}	Characteristics tags of user u_r
F_j	Set of the jth type characteristics tags
N_i	Length of document d_i
f_j^r	The jth type of characteristics tag of user u_r
$z_{i,n}$	Topic assignment of nth word of document d_i
$d_{i,e}$	Emotion label e of document d_i
$e_{i,n}$	Emotion assignment of nth word of document d_i
$w_{i,n}$	The nth word in document d_i
π	The multinomial distribution of groups
$\psi_{g,j}$	The multinomial distribution of the jth type characteristics tags specific to group g
θ_g	The multinomial distribution of topics specific to group g
$\phi_{g,z}$	The multinomial distribution of emotions specific to topic z of group g
$\varphi_{z,e}$	The multinomial distribution of words specific to topic z and emotion e
$\alpha, \beta, \gamma, \lambda$	Hyperparameters of Dirichlet distributions

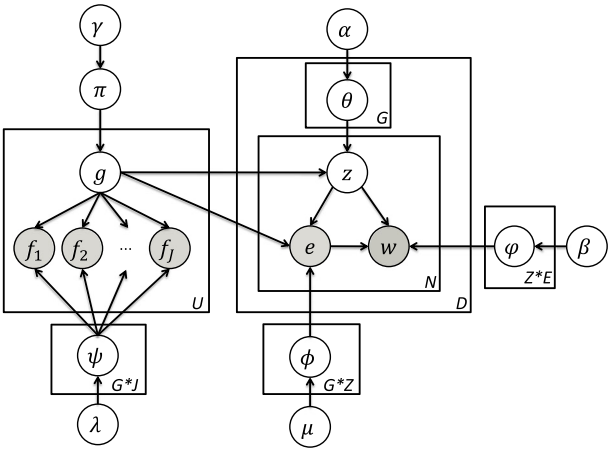


Figure 1 The graphical model of UGTE

document $d_i^{r,k}$, user group of this document g_{d_i} will be sampled according to π. After determining the group assignment, we generate each of the user's characteristics tags f_j^r from the characteristic distribution $\psi_{g,j}$. UGTE identifies topics and emotions according to each group's parameters by assuming that documents of each group follow the same topic distributions, and introducing emotions to topics in each group separately. When a user writes each word $w_{i,n}$ in document $d_i^{r,k}$, s/he first chooses a topic $z_{i,n}$ from a group's topic distribution θ_g. Then, emotion $e_{i,n}$ is determined from the emotion distribution $\phi_{g,z}$. According to the specific topic and emotion, the user draws word $w_{i,n}$ from the word distribution $\varphi_{z,e}$.

With respect to the group membership, it is reasonable and natural for UGTE to assume that one document belongs to one group while one user of several documents can belong to multiple groups with different probabilities. Although a short message often expresses one central idea, the user may write several messages on different topics with different attitudes. For example, if a person is fond of comics but has little interest in political news, s/he may regularly post about comics but far less often about political news. Such a person would be strongly related to a group whose users are fond of comics and weakly related to another group with heated political discussion. Besides, different from conventional joint topic-emotion models, one of the contributions of UGTE is its use of each user's characteristics tags for group discovery. UGTE identifies groups according to the documents' topics, emotion labels, and characteristics tags of corresponding users. The basis for this idea is that people sharing similar characteristics are more likely to share similar emotions on specific topics so that can be treated as a group. For example, individuals from different regions or social classes and with different ages are often interested in different topics. Even for a given topic, different groups of people may hold different emotions. The size of group set G is a predetermined parameter as that of topic set Z, enabling UGTE to mine different levels of group based emotions.

Formally, the generative process for each document is as follows:

1. For every type of characteristic tag $f_j \in F$, draw $\psi_j \sim Dirichlet(\lambda)$;
2. Draw the distribution over groups $\pi \sim Dirichlet(\gamma)$;
3. For each group g, draw the distribution over topics $\theta_g \sim Dirichlet(\alpha)$;

4. For each topic z of each group g, draw the distribution over emotions $\phi_{g,z} \sim Dirichlet(\mu)$;
5. For each topic z of specific emotion e, draw the distribution over words $\varphi_{z,e} \sim Dirichlet(\beta)$;
6. For each document $d_i^{r,k}$:
   (a) Draw group $g_r \sim Multinomial(\pi)$;
   (b) Draw each characteristics tag $f_j^r \sim Multinomial(\psi_j)$;
   (c) For each word $w_{i,n}$ document $d_i^{r,k}$:
      (i) Draw topic $z_{i,n} \sim Multinomial(\theta_{g_r})$;
      (ii) Draw emotion $e_{i,n} \sim Multinomial(\phi_{g_r,z_{i,n}})$;
      (iii) Draw word $w_{i,n} \sim Multinomial(\varphi_{e_{i,n},z_{i,n}})$.

3.3 Parameter inference

As a variant of the joint topic-emotion model, the inference of latent variables in our model are intractable. To address this, Gibbs sampling [9] or variational inference [3] is often employed. Gibbs sampling is a special case of Markov Chain Monte Carlo [13], which could achieve an accurate posterior distribution for parameter inference. On the other hand, variational inference can only provide an analytic approximation. Furthermore, it is mathematically arduous for variational inference to derive the approximation when the model structure is complex. Thus, following the previous works [2, 7, 32], we use Gibbs sampling when discovering groups and modeling the topics and emotions. According to the generative process, the joint probability of all the random variables for a document collection is shown as follows:

$$\begin{aligned} & p(z, w, e, g, f, \psi, \pi, \theta, \phi, \varphi, \alpha, \beta, \mu, \lambda, \gamma) \\ = \; & p(\pi; \gamma) p(\psi; \lambda) p(\theta; \alpha) p(\varphi; \beta) p(\phi; \mu) \\ & p(g|\pi) p(f|g, \psi) p(z|g, \theta) p(e|g, z, \phi) p(w|z, e, \varphi). \end{aligned} \tag{1}$$

During group discovery, a posterior probability for inferring the group g_r of a user u_r can be derived by marginalizing the above joint probability. The posterior probability is related to the user characteristics tags F^{u_r}, topics and the emotion label of document d_i, as follows:

$$\begin{aligned} & p(g_r = g \mid F^{u_r}, d_i^{r,k}, g_{-r}, \alpha, \beta, \mu, \lambda, \gamma) \\ \propto \; & p(g_r = g \mid g_{-r}, \gamma) p(F^{u_r} \mid g_{-r}, \lambda) p(d_i \mid d_{i,e} = e_k, g_{-r}, \alpha, \beta, \mu) \\ \propto \; & p(g_r = g \mid g_{-r}, \gamma) \prod_{j=1}^{J} p(f_j^r = f \mid g_{-r}, \lambda) \prod_{n=1}^{N_i} p(w_{i,n} \mid d_{i,e} = e_k, g_{-r}, \alpha, \beta, \mu). \end{aligned} \tag{2}$$

Specially, emotion label $d_{i,e}$ is an observable variable, so that the posterior probability of a group generates a word with a specific emotion can be derived by marginalizing the topic variable according to (1). The formulas is shown as follows:

$$\begin{aligned} & p(w_{i,n}, \mid d_{i,e} = e_k, g_{-r}, \alpha, \beta, \mu) \\ = \; & \sum_{z=1}^{|z|} p(z_{i,n} = z \mid g_{-r}, \alpha) p(e_{i,n} = e_k \mid z, g_{-r}, \mu) p(w_{i,n} \mid z, e_{i,n} = e_k, g_{-r}, \beta). \end{aligned} \tag{3}$$

According to the detailed derivation, we can estimate the posterior probability by (4):

$$
\begin{aligned}
& p(g_r = g \mid F^{u_r}, d_i, d_{i,e} = e_k, g_{-r}, \alpha, \beta, \mu, \lambda, \gamma) \\
\propto\ & \frac{N_{-r}^{g} + \gamma}{\sum_{g'=1}^{|G|}(N_{-r}^{g'} + \gamma)} \times \prod_{j=1}^{J} \frac{N_{f,j}^{g,-r} + \lambda}{\sum_{f'=1}^{|F_j|} N_{f',j}^{g,-r} + |F_j|\lambda} \\
& \times \prod_{n=1}^{N_i} \sum_{z=1}^{|z|} \frac{N_{z}^{-i,g} + \alpha}{\sum_{z'=1}^{|Z|}(N_{z'}^{-i,g} + \alpha)} \times \frac{N_{e_k}^{-i,g,z} + \mu}{\sum_{e'=1}^{|E|}(N_{e'}^{-i,g,z} + \mu)} \times \frac{N_{w_{i,n}}^{-i,z,e} + \beta}{\sum_{w'=1}^{|W|}(N_{w'}^{-i,z,e} + \beta)},
\end{aligned} \tag{4}
$$

where N_{-r}^{g} is the number of users assigned to group g excluding user u_r, and $N_{f,j}^{g,-r}$ is the number of characteristics tags f_j assigned to group g excluding tag f_j^r. Furthermore, $N_z^{-i,g}$ is the number of words assigned to topic z in group g excluding words of document d_i, $N_e^{-i,g,z}$ is the number of words assigned to emotion e of topic z in group g excluding words of document d_i, and $N_w^{-i,z,e}$ is the number of words w assigned to emotion e of topic z excluding words of document d_i.

After sampling the group of document d_i, the assignment of topics and emotions of words can be inferred by parameters characterizing the group. Differing from conventional unsupervised LDA-based model, UGTE is a supervised joint topic-emotion model, which utilizes the emotion labels of documents when performing Gibbs sampling. Inspired by Labeled LDA [22], we incorporate supervision by simply constraining the emotion assignment of words same as the emotion labels of corresponding documents. We formulate this process as follows:

$$
\begin{aligned}
& p(z_{i,n} = z, e_{i,n} = e_k \mid d_{i,e} = e_k, z_{-i,n}, e_{-i,n}, w_{i,n}, g, \theta, \phi, \varphi, \alpha, \beta, \mu) \\
\propto\ & p(w_{i,n} = w \mid z_{i,n} = z, e_{i,n} = e_k, d_{i,e} = e_k, z_{-i,n}, e_{-i,n}, \theta, \phi, \varphi, \alpha, \beta, \mu) \\
\propto\ & \frac{N_{z}^{g,-w_{i,n}} + \alpha}{\sum_{z'=1}^{|Z|}(N_{z'}^{g,-w_{i,n}} + \alpha)} \cdot \frac{N_{e_k}^{g,z,-w_{i,n}} + \mu}{\sum_{e'=1}^{|E|}(N_{e'}^{g,z,-w_{i,n}} + \mu)} \cdot \frac{N_{w}^{z,e_k,-w_{i,n}} + \beta}{\sum_{w'=1}^{|W|}(N_{w'}^{z,e,-w_{i,n}} + \beta)}.
\end{aligned} \tag{5}
$$

After the sampling process converging according to (4) and (5), the distribution of π, θ_g, $\phi_{g,z}$ and $\varphi_{z,e}$ is convenient to be estimated according to (8)- (10), as follows:

$$
\pi_g = \frac{N_g + \gamma}{\sum_{g'=1}^{|G|}(N_{g'} + \gamma)}, \tag{6}
$$

$$
\psi_{g,j,f} = \frac{N_{f,j}^{g} + \lambda}{\sum_{f'=1}^{|F_j|}(N_{f',j}^{g} + \lambda)}, \tag{7}
$$

$$
\theta_{g,z} = \frac{N_{z}^{g} + \alpha}{\sum_{z'=1}^{|Z|}(N_{z'}^{g} + \alpha)}, \tag{8}
$$

$$
\phi_{g,z,e} = \frac{N_{e}^{g,z} + \mu}{\sum_{e'=1}^{|E|}(N_{e'}^{g,z} + \mu)}, \tag{9}
$$

$$
\varphi_{z,e,w} = \frac{N_{w}^{z,e} + \beta}{\sum_{w'=1}^{|W|}(N_{w'}^{z,e} + \beta)}. \tag{10}
$$

With all the parameters derived above, we can further infer the emotions of unlabeled document d_{test} as follows:

$$p(e_{test} \mid d_{test}) = p(e_{test} = e \mid d_{test}, F^{u_r}) = \prod_{n=1}^{N_{test}} \sum_{g=1}^{|G|} \sum_{z=1}^{|Z|} \begin{matrix} \tilde{\pi}_g \\ \cdot \prod_{j=1}^{J} \tilde{\psi}_{g,j,f^{u_{test}}} \\ \cdot \tilde{\theta}_{g,z} \\ \cdot \tilde{\phi}_{g,z,e} \\ \cdot \tilde{\varphi}_{z,e,w} \end{matrix}, \tag{11}$$

where $\tilde{\pi}_g, \tilde{\psi}_{g,j,f^{u_{test}}}, \tilde{\theta}_{g,z}, \tilde{\phi}_{g,z,e}, \tilde{\varphi}_{z,e,w}$ can be inferred according to (6) - (10) with the number of corresponding instances including the union of documents in D_{train} and document d_{test}.

4 Experiments

To evaluate our proposed method, we perform topic discovery and emotion classification, and compare our method with other state-of-the-art models.

4.1 Experimental setup

4.1.1 Dataset

We use a real-world dataset to verify the effectiveness of our model. ISEAR[1] is a typical dataset for emotion detection, which contains 7,666 sentences/short texts annotated by 1,096 users with different cultural backgrounds. It is completed in the form of a questionnaire, which includes their personal information, experiences, and their expressions over seven emotions, i.e., anger, disgust, fear, joy, sadness, shame and guilt. Each sample contains 42 attributes,including discrete types and some discriptions. We use 11 discrete attributes and contents for experiments: ID, CITY, COUNTRY, SEX, AGE, RELI, PRAC, FOCC, MOCC, FIEL, EMOT and SIT. After pre-processing by removing stop words and filtering punctuation marks, there are totally 7,652 samples left for experiments. By default, we use 80% data (6,122 samples) as the training set and the remaining 20% data (1,530 samples) as the testing set. To further explore how effective the model could address the feature sparsity problem brought by extremely short text, we divide examples into two different groups based on the length of its content. Texts longer than 10 words are grouped as the "short text" subset, while those shorter than 10 words are classified into the "extremely short text" subset. For each subset, 80% samples are used as the training set and the remaining 20% samples are used as the testing set. Particularly, there are 891 training samples and 223 testing samples in the "short text" subset, and 5,230 training samples and 1,308 testing samples in the "extremely short text" subset. Details of attributes of ISEAR are shown in Table 2.

4.1.2 Baselines

To evaluate the effectiveness of UGTE, we employ several representative algorithms that jointly model topics and emotions/sentiments as baselines: Author-Topic model (AT) [26],

[1] http://www.affective-sciences.org/researchmaterial

Table 2 Selected attributes of ISEAR

Attribute	Description
ID	User ID
CITY	User's city
COUNTRY	User's country
SEX	User's gender
AGE	User's age
RELI	User's religion
PRAC	User's practising religion
FOCC	User father's occupation
MOCC	User mother's occupation
FIEL	User's field of study
SIT	Free description of an event or a situation
EMOT	Emotion category

Multi-label Supervised Topic Model (MSTM) and Sentiment Latent Topic Model (SLTM) [24], Contextual Sentiment Topic Model (CSTM) [23], supervised Neural Topic Model (sNTM) [4], and neural Siamese Labeled Topic Model (nSLTM) [12]. AT extends LDA to include authorship information by jointly modeling users and topics. MSTM and SLTM are topic models for social emotion mining from the perspective of readers. CSTM classifies reader emotions across different contexts by distinguishing context-independent topics from both a background theme and a contextual theme. sNTM is in essence a neural network by following the document-topic distribution in topic models. nSLTM is a supervised topic model based on the Siamese network, which can trade off label-specific word distributions with document-specific label distributions in a uniform framework.

4.1.3 Metrics

Topic coherence [17] is an effective measure for the quality of topic discovered by the models. Between any two words in top-n words for each topic, the more the words co-occurred within a document, the better the generated topic is. $Coherence@n$ denotes models' performance on topic discovery, as measured by the average of coherence values for each topic in the model. The calculation can be formulated as follows:

$$Coherence@n = \frac{1}{Z}\sum_{z=1}^{Z} C(z,n), \tag{12}$$

$$C(z,n) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} \log \frac{D(w_{z,j}, w_{z,i}) + 1}{D(w_{z,i})}, \tag{13}$$

where $C_{z,n}$ is the coherence value of topic z according to top-n words, $w_{z,i}$ is the ith most probable word of topic z, $D(w_{z,i})$ is the frequency of word $w_{z,i}$ appeared in the dataset, $D(w_{z,j}, w_{z,i})$ is the co-occurrence frequency of $w_{z,j}$ and $w_{z,i}$ within documents in the dataset. For the task of emotion classification, the accuracy and the Cohen's kappa score [1] are used as the evaluation metrics.

4.1.4 Parameter setting

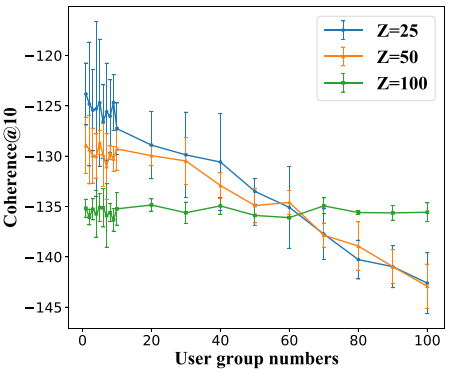
We verify the effectiveness of our proposed model by conducting topic discovery and emotion classification. Experiments of comparing UGTE and baselines are set up. For topic discovery, we run all models with different topic numbers $|Z| \in \{25, 30, 35, 40, 45, 50, 100, 150, 200, 250, 300\}$. We selecte hyper-parameters for the Dirichlet priors as symmetric Dirichlet prior vectors according to other studies [2, 23, 24, 26], where $\alpha = 50/|Z|$, $\beta = 0.1$, $\gamma = 0.1$, $\lambda = 0.1$, $\mu = 0.1$. We set the number of user groups $|G|$ to 10 based on a preliminary study. For completeness, we also evaluate the influence of user group numbers on our model in Section 4.2, by setting $|Z| \in \{25, 50, 100\}$ and $|G| \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. Tasks of topics discovery and emotion classification are conducted on UGTE_ID (UGTE that exploits only ID, EMOT and SIT of users) in this part. Since LDA-based model is insensitive to values of the hyper-parameters for the Dirichlet priors [29], we set parameters for the baselines according to the corresponding papers. Except baselines of AT, MSTM, SLTM, CSTM, sNTM and nSLTM, the proposed UGTE_ALL (UGTE that use all attributes of users) is also compared with UGTE_ID. We use the training set to estimate model parameters. Then we infer parameters and evaluated $Coherence@10$, $Coherence@20$ and $Coherence@30$ on the testing set. For emotion classification, we use a similar process to set parameters and evaluated accuracy and Cohen's kappa on the testing set. MSTM, SLTM, CSTM, sNTM, nSLTM, UGTE_ALL, UGTE_ID are adopted for comparison since AT can not be applied to emotion classification directly.

Then, to explore the impact of different user characteristics tags on UGTE, topic discovery and emotion classification are conducted on 12 variant models of UGTE. These variant models include UGTE_NULL, UGTE_ID, UGTE_CITY, UGTE_COUN, UGTE_SEX, UGTE_AGE, UGTE_RELI, UGTE_PRAC, UGTE_FOCC, UGTE_MO-CC, UGTE_FIEL, UGTE_ALL, which refer models that use no characteristics tags, ID, CITY, COUN, SEX, AGE, RELI, PRAC, FOCC, MOCC, FIEL and all characteristics tags, respectively. Experiments are run in a similar process to the above.

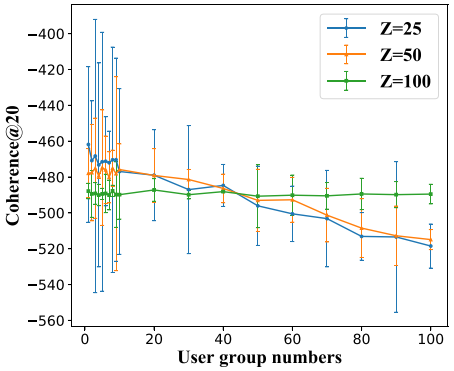
Finally, a case study is conducted to demonstrate how does the user characteristic help improve the performance of topics discovery and emotions detection. User portraits are illustrated to show how dose the UGTE discover the relationship between topics and emotions at the group level. The number of iterations is set to 3,000 for all experiments. We run each model 10 times to reduce noise and randomness, and both the mean and the variance are presented.

4.2 Influence of user group numbers

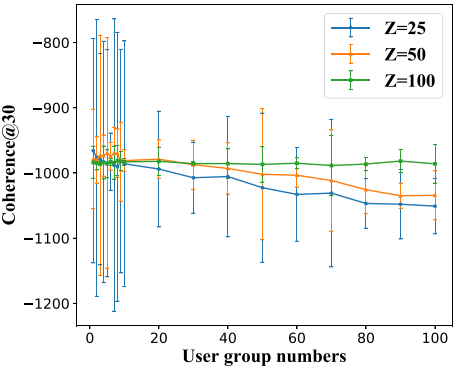
To investigate the relationship between the number of user groups and model performance, we conduct topic discovery and emotion classification tasks with different numbers of user groups under a fixed number of topics. Results are shown in Figures 2 and 3. From Figure 2 we can observe that when $|Z| = 25$ and $|Z| = 50$, the coherence score fluctuates firstly and then decreases as the number of user groups increased. It indicates that UGTE_ID performs better when the number of user groups is small for this dataset. The optimal size of user groups is $G \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ according to that UGTE_ID performs well conditioned on $1 \leq G \leq 10$. However, when $|Z| = 100$, the performance of UGTE_ID is more stable under three coherence metrics with smaller variances. It indicates that the number of user groups has little influence on UGTE when the number of topics is large. Although the variances of UGTE_ID with small numbers of topics and user groups are bigger, it can



(a) *Coherence*@10



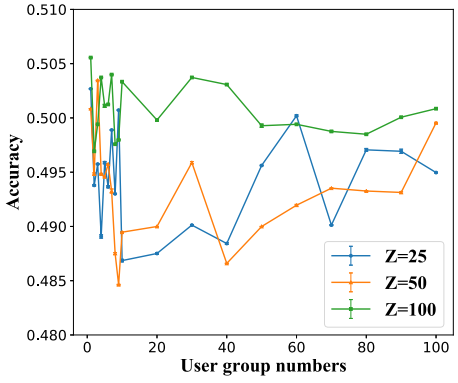
(b) *Coherence*@20



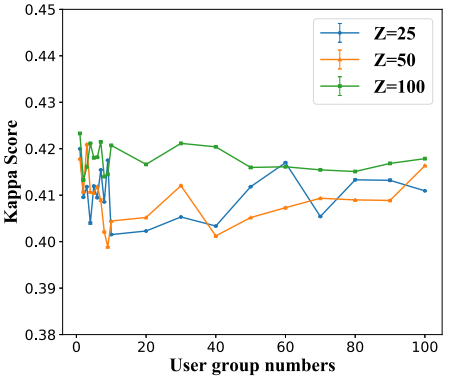
(c) *Coherence*@30

Figure 2 Topic coherence of UGTE_ID with different numbers of user groups

achieve more competitive results in average. For the task of emotion classification, we can see that as the number of user groups increased, UGTE_ID performs stabler under $|Z| = 100$ than that under $|Z| = 50$ and $|Z| = 25$, as shown in Figure 3. Furthermore, UGTE_ID



(a) Accuracy



(b) Kappa Score

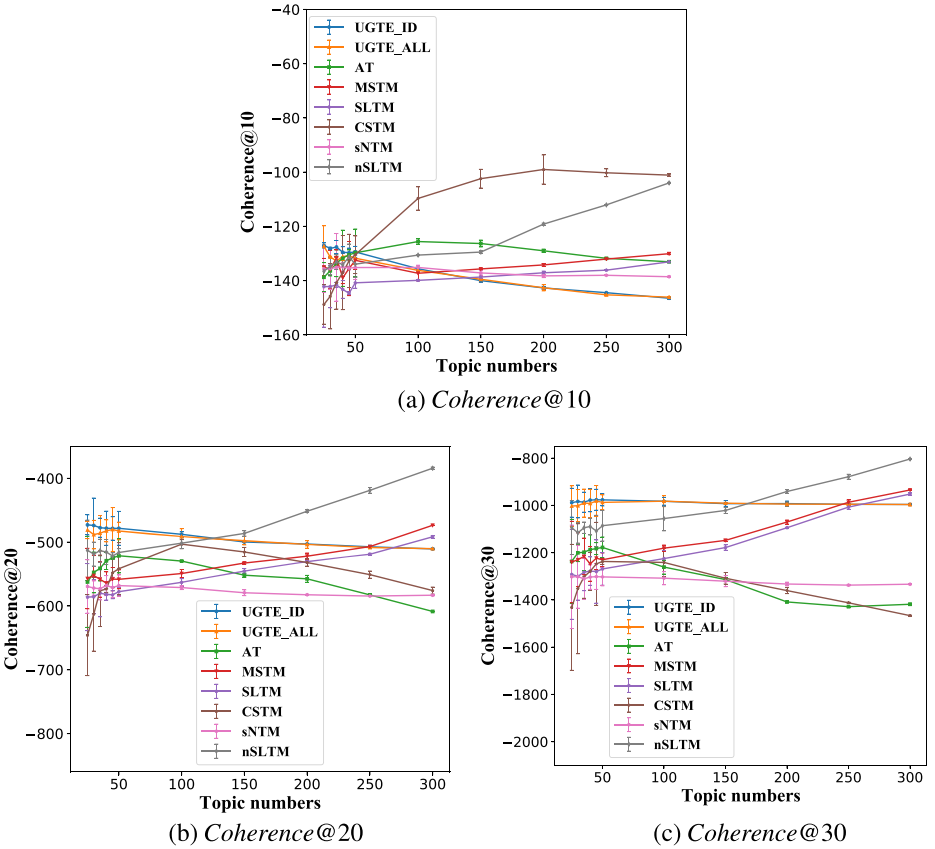
Figure 3 Emotion classification of UGTE_ID with different numbers of user groups

achieves higher accuracy and Kappa score when $|Z| = 100$. It indicates that UGTE_ID with a large number of topics performs better in the task of emotion classification.

4.3 Comparison with baselines

4.3.1 Topic discovery

The coherence of topics for our models and baselines over ISEAR are illustrated in Figure 4. Under $Coherence@10$, UGTE_ID performs the best when $|Z| \leq 50$, but achieves worse results when the number of topics increased. It indicates that UGTE_ID is more suitable to discover a small number of topics. On the other hand, the baseline model of CSTM performs better than other models when $|Z| \geq 100$. Though UGTE_ALL and UGTE_ID do not achieves competitive results under $Coherence@10$, they both perform better and more steadily under $Coherence@20$ and $Coherence@30$. Under $Coherence@30$, when $|Z| \leq 150$, UGTE_ID achieves the best performance. Neural network based model nSLTM achieves coherence values as the number of topics increased, which indicates that nSLTM is suitable for mining a large number of topics. sNTM does not perform as well as nSLTM, which achieves lower coherence values than nSLTM as $|Z|$ increased.



(a) $Coherence@10$

(b) $Coherence@20$

(c) $Coherence@30$

Figure 4 Topic coherence of UGTE_ID and baselines with different topic numbers when $|G| = 10$

To evaluate the differences of these models statistically, we also perform two kinds of statistical test on paired models. The first one is conducted to evaluate the stability of performance in terms of variances, and the second one is to evaluate the averaged performance in terms of means. The p-values are estimated for both kinds of statistical test. The conventional significance level (i.e., p-value) is 0.05, which means the null hypothesis can be rejected with a probability of 95%. The difference between paired models is statistically significant if the p-value is lower than 0.05. Firstly, the analysis of variance in terms of F-test is employed to test the underlying assumption of homoscedasticity. The F-tests are conducted on UGTE_ID, UGTE_ALL, AT, CSTM, MSTM, SLTM, sNTM and nSLTM. Results are shown in Table 3 where the significant values are highlighted in boldface. UGTE_ID is statistically significantly different from AT, CSTM, MSTM, SLTM, sNTM under $Coherence@10$, $Coherence@20$ and $Coherence@30$. It indicates that UGTE_ID is statistically stabler than AT, CSTM, MSTM, SLTM, sNTM over different topic numbers. UGTE_ID differs from nSLTM significantly under $Coherence@20$ and $Coherence@30$, indicating that UGTE_ID performs stabler than nSLTM under large numbers of top words in the topic coherence metric.

Secondly, t-tests are conducted to test the underlying assumption that the difference of performance between paired models has a mean value of zero (i.e., the null hypothesis implies identical performance). The results are shown in Table 4 where the significant values are highlighted in boldface. We can observe that UGTE_ID outperforms the baselines of CSTM and SLTM significantly under three coherence metrics. UGTE_ID is statistically significantly different from AT, MSTM, sNTM under $Coherence@20$ and $Coherence@30$.

4.3.2 Emotion detection

For the task of emotion classification, it needs to estimate parameters firstly on the training set, and then make prediction of emotions of unlabeled documents in the testing set. Figure 5a and b present accuracy and Cohen's Kappa score of emotion classification on ISEAR. nSLTM achieves highest accuracy and Kappa score when $|Z| \geq 100$, which show the effectiveness of neural based algorithms when mining large topics. UGTE_ALL, UGTE_ID and nSLTM perform better than CSTM, MSTM, SLTM and sNTM. UGTE_ID performs better than UGTE_ALL when $|Z| \leq 100$ and $|Z| = 200$, indicating that multi-characteristics may have negative effects on emotion detection. In this task, UGTE_ID do not outperform nSLTM. However, nSLTM can not neither detect groups of individuals sharing similar interest nor give group-based topic and emotion analysis.

Statistical tests are performed on the results, as shown in Tables 3 and 4. The results show that the variance of UGTE_ID is statistically different from those of MSTM, SLTM

Table 3 P-values of F-test between UGTE_ID and other models

Models	$Coherence@10$	$Coherence@20$	$Coherence@30$	Accuracy	Kappa score
UGTE_ALL	3.29E-01	2.28E-01	4.09E-01	3.84E-01	3.87E-01
AT	**2.79E-02**	**3.20E-02**	**5.65E-10**	–	–
MSTM	**1.45E-03**	**2.20E-02**	**2.54E-10**	**1.53E-04**	**1.79E-04**
SLTM	**1.09E-02**	**8.54E-03**	**7.56E-11**	**8.95E-05**	**9.63E-05**
CSTM	**2.22E-03**	**1.18E-03**	**5.60E-09**	3.37E-01	3.22E-01
sNTM	**5.95E-06**	**8.04E-03**	**2.48E-02**	6.28E-02	**8.92E-05**
nSLTM	1.39E-01	**4.43E-04**	**3.74E-10**	**4.73E-04**	**7.93E-04**

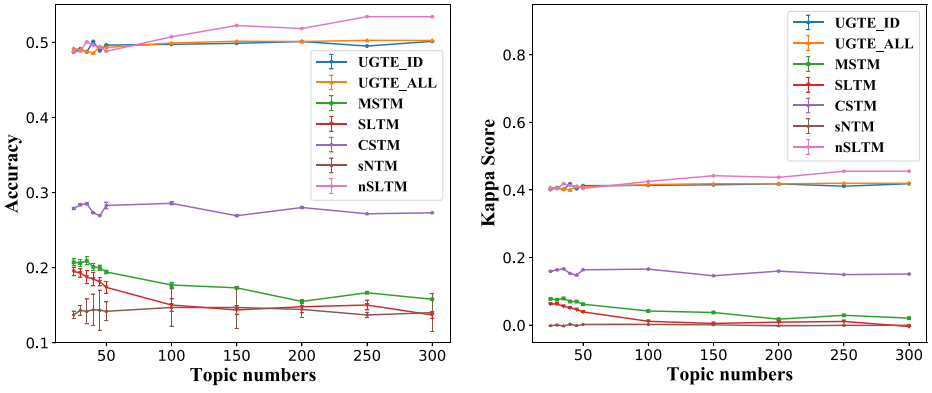
Table 4 P-values of T-test between UGTE_ID and other models

Models	*Coherence*@10	*Coherence*@20	*Coherence*@30	Accuracy	Kappa score
UGTE_ALL	3.36E-01	2.38E-01	**3.34E-02**	4.28E-01	4.30E-01
AT	1.21E-01	**2.72E-06**	**1.37E-06**	–	–
MSTM	4.72E-01	**3.85E-05**	**2.30E-04**	**1.98E-14**	**2.10E-15**
SLTM	**2.42E-02**	**5.84E-06**	**1.11E-04**	**1.94E-14**	**2.07E-14**
CSTM	**4.21E-02**	**6.58E-05**	**2.96E-08**	**2.44E-27**	**3.18E-27**
sNTM	2.41E-01	**1.74E-11**	**2.55E-20**	**7.99E-34**	**2.93E-21**
nSLTM	**3.64E-02**	4.36E-01	1.18E-01	**3.55E-02**	**2.75E-02**

and nSLTM on both accuracy and Kappa score. It indicates that UGTE_ID performs stabler than MSTM, SLTM and nSLTM. The mean of UGTE_ID is statistically different from MSTM, SLTM, CSTM, sNTM and nSLTM. The results show that UGTE_ID outperforms than MSTM, SLTM, CSTM and sNTM.

4.4 Impact of user characteristics

Our proposed model aggregates characteristics tags to conduct group-based topic and emotion analysis. To explore the impact of different characteristics tags on UGTE, topic discovery and emotion classification are performed over 12 variant models of UGTE. Topic discovery and emotion analysis are conducted on each variant model with different numbers of topics are above experiments. Average results are taken as illustrated in Table 5. Not all the characteristics are helpful for UGTE to get a good result. We can observe that when compared with UGTE_NULL, some characteristics have positive effective on average like ID, CITY, COUNT, SEX, AGE, RELI, PRAC and FOCC under *Coherence*@10. Under *Coherence*@20, only ID, CITY, SEX, AGE, PRAC, FOCC and MOCC have positive effective on average while others have negative effects with lower coherence values. However, in the task of emotion classification, show that AGE and SEX can be helpful. UGTE_ALL performs worse than other variant models of UGTE on both topic discovery and emotion classification, which verifies that not all the characteristics tags can help to discover user



(a) Accuracy over topics (b) Kappa Score over topics

Figure 5 Emotion classification results of UGTE_ID and baselines with different topic numbers when $|G| = 10$

Table 5 The mean and variance over different characteristics tags where the best results are highlighted in boldface

Model	Coherence@10	Coherence@20	Coherence@30	Accuracy	Kappa score
GTSM_NULL	−134.7500 | 3.3102	−488.3700 | 14.0521	−986.3800 | 49.5280	0.5014 | 0.0001	0.4184 | 0.0001
GTSM_ID	−134.2900 | 2.0679	**−487.2700** | 15.6484	−986.0400 | 77.1833	0.4973 | 0.0001	0.4137 | 0.0000
GTSM_CITY	−134.4300 | 3.4606	−487.8000 | 11.6895	−985.4600 | 62.4715	0.5009 | 0.0001	0.4179 | 0.0001
GTSM_COUN	−134.2800 | 2.9503	−488.5500 | 18.2365	−988.0600 | 39.4006	0.5000 | 0.0000	0.4169 | 0.0000
GTSM_SEX	**−134.2500** | 2.8470	−487.7000 | 14.8564	−985.8800 | 58.8609	**0.5057** | 0.0001	**0.4234** | 0.0001
GTSM_AGE	−134.5400 | 2.0708	−487.9000 | 17.0055	−986.4100 | 55.8719	0.5020 | 0.0001	0.4191 | 0.0001
GTSM_RELI	−134.3200 | 2.0995	−488.5700 | 23.5460	−987.4100 | 50.3316	0.5007 | 0.0001	0.4177 | 0.0000
GTSM_PRAC	−134.4400 | 2.0312	−488.1100 | 17.3123	−986.4800 | 91.1240	0.5011 | 0.0001	0.4181 | 0.0001
GTSM_FOCC	−134.5200 | 2.0380	−487.6500 | 18.1535	**−984.9100** | 62.9523	0.4990 | 0.0001	0.4156 | 0.0000
GTSM_MOCC	−134.8500 | 1.0696	−487.6600 | 16.0361	−986.1200 | 43.2076	0.5014 | 0.0000	0.4185 | 0.0000
GTSM_FIEL	−135.0600 | 1.5845	−488.4600 | 12.3045	−986.7500 | 68.3423	0.4995 | 0.0001	0.4163 | 0.0001
GTSM_ALL	−135.9000 | 2.2920	−491.8900 | 14.0651	−992.0200 | 41.9525	0.4955 | 0.0001	0.4116 | 0.0001

Table 6 P-values of F-test over GTSM_NULL and different characteristics tags

Models	Coherence@10	Coherence@20	Coherence@30	Accuracy	Kappa score
UGTE_ID	2.74E-01	4.12E-01	3.90E-01	3.15E-01	2.76E-01
UGTE_CITY	**1.36E-02**	4.33E-01	4.10E-01	2.33E-01	**1.35E-02**
UGTE_COUN	**4.65E-02**	4.26E-01	4.93E-01	8.94E-02	**4.92E-02**
UGTE_SEX	**3.22E-02**	4.15E-01	4.51E-01	2.14E-01	**3.36E-02**
UGTE_AGE	**9.80E-03**	4.44E-01	4.04E-01	3.42E-01	**9.96E-03**
UGTE_RELI	**1.41E-03**	3.91E-01	4.55E-01	2.23E-01	**1.50E-03**
UGTE_PRAC	**1.22E-02**	4.62E-01	4.70E-01	1.46E-01	**1.23E-02**
UGTE_FOCC	**7.04E-04**	4.54E-01	4.42E-01	4.02E-01	**7.12E-04**
UGTE_MOCC	**1.65E-05**	4.96E-01	4.29E-01	3.15E-01	**1.48E-05**
UGTE_FIEL	**4.44E-03**	4.64E-01	4.96E-01	4.16E-01	**4.82E-03**
UGTE_ALL	2.82E-01	3.63E-01	2.84E-01	1.50E-01	2.81E-01

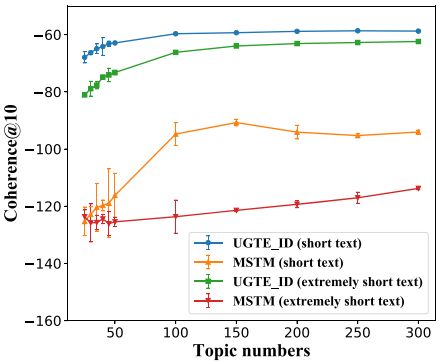
groups. Statistics tests are performed on results between UGTE_NULL and other variant models, and the values are shown in Tables 6 and 7. It is obvious that different characteristics tags do not have statistically significant differences in terms of variances and means under $Coherence@20$, $Coherence@30$ and $Accuracy$.

4.5 Performance on Extremely Short Text

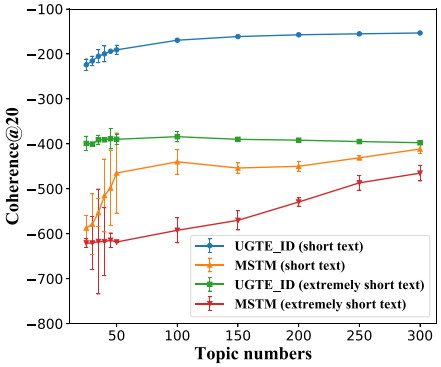
To further explore the performance of joint topic-emotion models on extremely short texts, we run UGTE_ID and MSTM on "short text" and "extremely short text" subsets and present their results in Figures 6 and 7. The results indicate that both UGTE_ID and MSTM achieve higher coherence values on "short text" and perform a little unstably on "extremely short text". It suggests that extremely short text brings more serious feature sparsity problem to joint topic-emotion models. However, by exploiting user characteristics, UGTE_ID achieves higher coherence values than MSTM consistently in the task of topic coherence. According to Figures 6a and b, we can observe that UGTE_ID achieves the best results on both "short text" and "extremely short text". In Figure 6c, UGTE_ID performs better than MSTM when $|Z| \leq 200$, since UGTE may be more suitable to discovery a small number of topics. But UGTE_ID always performs much more stable than MSTM. In the task of emotion detection, UGTE_ID achieves much higher accuracy and Kappa score than MTSM. On the

Table 7 P-values of T-test over UGTE_NULL and different characteristics tags

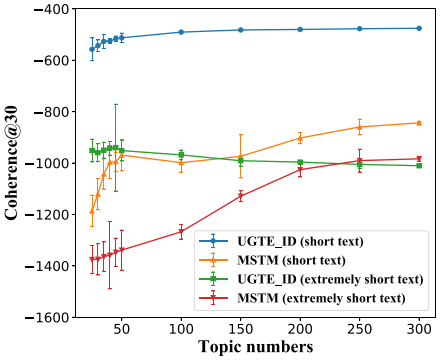
Models	Coherence@10	Coherence@20	Coherence@30	Accuracy	Kappa score
UGTE_ID	**4.93E-03**	4.44E-01	4.29E-01	4.74E-01	**4.79E-03**
UGTE_CITY	**4.00E-02**	4.61E-01	4.63E-01	4.04E-01	**4.96E-03**
UGTE_COUN	**2.28E-02**	4.43E-01	4.88E-01	3.17E-01	**2.24E-02**
UGTE_SEX	5.83E-02	4.39E-01	4.55E-01	4.47E-01	5.78E-02
UGTE_AGE	8.07E-02	4.74E-01	4.69E-01	4.97E-01	7.92E-02
UGTE_RELI	**3.10E-02**	4.48E-01	4.87E-01	3.92E-01	**3.03E-02**
UGTE_PRAC	**4.44E-02**	4.61E-01	4.83E-01	4.89E-01	**4.30E-02**
UGTE_FOCC	**7.78E-03**	4.71E-01	4.52E-01	3.58E-01	**7.51E-03**
UGTE_MOCC	**4.65E-02**	4.88E-01	4.54E-01	4.74E-01	**4.53E-02**
UGTE_FIEL	**1.28E-02**	4.61E-01	4.93E-01	4.66E-01	**1.23E-02**
UGTE_ALL	**1.22E-03**	3.50E-01	2.57E-01	6.58E-02	**1.19E-03**



(a) Coherence@10



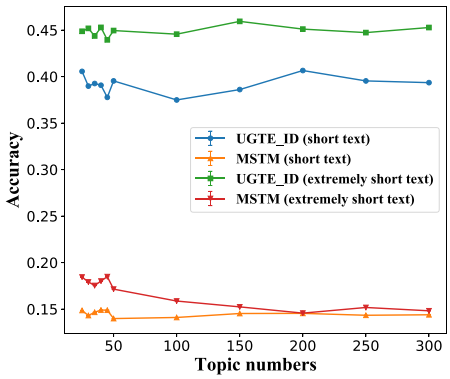
(b) Coherence@20



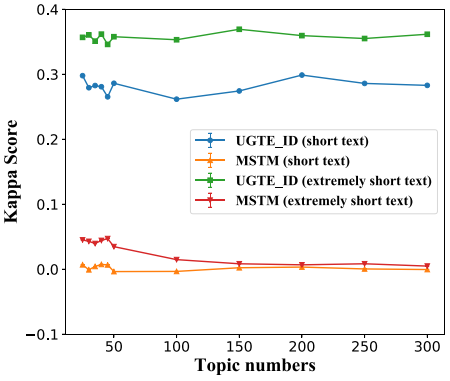
(c) Coherence@30

Figure 6 Topic coherence of UGTE_ID and MSTM over "short text" and "extremely short text" subsets

"extremely short text" subset, UGTE_ID achieves the best accuracy of 0.4594 and Kppa Score of 0.3693, while MSTM only achieves the best accuracy of 0.1849 and Kappa Score of 0.0452. Results show that user characteristics can improve the model perfomance by



(a) Accuracy



(b) Kappa Score

Figure 7 Emotion detection results of UGTE_ID and MSTM over "short text" and "extremely short text" subsets

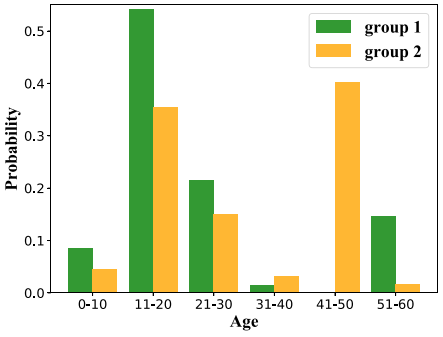
Table 8 Top 10 words of selected topic "Intimate relationship"

Emotion	Model	Top words
Joy	UGTE_AGE	entrance music engaged remember kiss falling love expensive dinner continously
	UGTE_NULL	period falling love hearing involved met job smoking dinner helped
Sadness	UGTE_AGE	died friend grandmother close father passed hospital sad left accident
	UGTE_NULL	difficulties hanging announced failed love uninteresting parents person beating accident

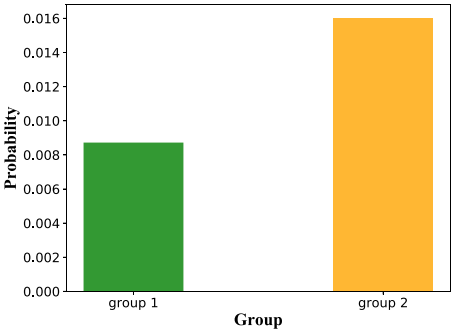
addressing the feature sparsity problem of short texts for both topic discovery and emotion detection.

4.6 Case study

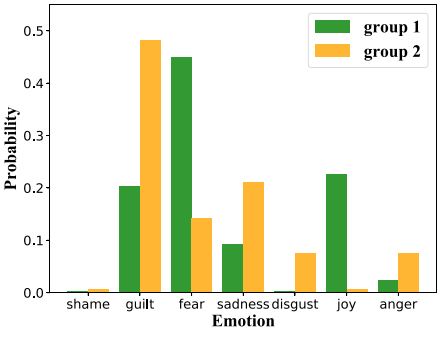
To verify the effect of user characteristics (e.g.,"Age") on topic discovery and emotion detection, we compare UGTE_AGE and UGTE_NULL under $|Z| = 100$ and $|G| = 10$.



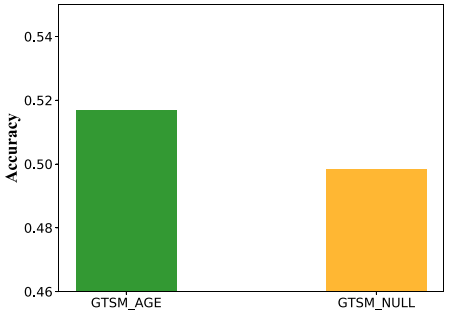
(a) Age Distribution



(b) Probability of "Intimate relationship" topic



(c) Emotion Distribution



(d) Accuracy of Emotion detection

Figure 8 Portraits of selected groups generated by UGTE_AGE

Table 8 presents 10 representative words of topics with joy and sadness emotions, respectively. By checking these top words manually, we can conclude that the topics are related to "Intimate relationship". The results indicate that words under the emotion of "joy" generated by UGTE_AGE are more about "Love" and "Wedding". On the other hand, "period", "helped" , "smoking", and "job" discovered by UGTE_NULL seem to be less associated with "Love" or "Wedding", making the topic less coherent. Similarly, the topic under the emotion of "sadness" discovered by UGTE_AGE is more about "injury" and "death". By contrast, many words of UGTE_NULL are incoherent, such as "announced" and "handing".

Different from conventional joint topic-emotion models, our method can identify portraits for these groups presented as distributions over characteristics. User portraits in Figure 8 show that the distribution over ages of group 1 achieves the maximum value within 11 to 20, while users in group 2 are between 41 to 50 mostly. Besides, group 1 concerns a little less about the topic than group 2. As we can see in Figure 8c, emotions of users in group 1 are more about "fear" and "joy", which coincides with the mentality of youth. In contrast, users in group 2 feel more about "sadness" and "guilt" on the "Intimate relationship" topic, which could be understood according to the top words of topic with "sadness" emotion in Table 8. As shown in Figure 8d, UGTE_AGE achieves a higher accuracy than UGTE_NULL, which validates that the user characteristics of "Age" can help to improve the performance of emotion detection.

5 Conclusion and future work

To address the issue of feature sparsity in short text, we proposed a method named UGTE by modeling topics, emotions and user characteristics jointly. UGTE can explore the relationships of topics, emotions and users characteristics among different groups. In addition, short messages popular online bring challenges of feature sparsity problems to traditional joint topic-emotion models. So, introducing a user group layer to the topic-based emotion detection model, UGTE can efficiently aggregate short text into long pseudo-documents to address the feature sparsity problem of short text. Experiments conducted on a real-world dataset ISEAR showed that UGTE is not only effective in emotion detection, but also can mine significant topics concerned by each user group. With the development of neural network technologies, we plan to combine our proposed model with neural networks to improve its capacity for modeling topics and emotions at the user group level. Besides, considering the generality of the model, we also plan to propose a general framework for topic discovery by integrating other information, such as word position, context relevance, and so forth.

Acknowledgment This work has been supported by Top-Up Fund (TFG-04) and Seed Fund (SFG-10) for General Research Fund / Early Career Scheme and Interdisciplinary Research Scheme of the Dean's Research Fund 2018-19 (FLASS/DRF/IDS-3), Departmental Collaborative Research Fund 2019 (MIT/DCRF-R2/18-19), Funding Support to General Research Fund Proposal (RG 39/2019-2020R) and the Internal Research Grant (RG 90/2018-2019R) of The Education University of Hong Kong, and LEO Dr David P. Chan Institute of Data Science, Lingnan University, Hong Kong. The work has also been supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Collaborative Research Fund, project number C1031-18G).

Appendix

For clarity, numerical results of Figures 2–7 are provided as follows.

Table 9 Coherence@10 of UGTE_ID and baselines with different topic numbers when $|G| = 10$, where the best results are highlighted in boldface

(a)

Model	25	30	35	40	45
UGTE_ID	**−127.0237** | 0.9396	**−128.1418** | 0.4671	**−127.6175** | 2.4091	**−129.5531** | 2.0617	**−129.6932** | 1.6861
UGTE_ALL	−127.3680 | 7.5889	−131.3043 | 0.3031	−132.7116 | 1.2696	−131.1469 | 1.5027	−130.8406 | 0.3758
AT	−138.7526 | 5.3924	−136.3771 | 1.8968	−133.6252 | 4.9163	−131.7716 | 10.3600	−130.5205 | 1.8186
MSTM	−135.0000 | 3.1754	−135.8164 | 6.9620	−131.6409 | 2.9654	−138.9608 | 2.1493	−135.5555 | 9.9791
SLTM	−142.2412 | 14.9133	−142.1818 | 7.8352	−141.6045 | 0.9472	−143.2605 | 3.2688	−144.6565 | 0.5000
CSTM	−148.8208 | 7.2536	−145.8034 | 12.0064	−140.7302 | 9.7631	−137.0355 | 13.6627	−132.7447 | 9.8129
sNTM	−135.6900 | 8.6965	−135.5800 | 7.5662	−135.1000 | 12.4666	−134.9600 | 1.5577	−135.3900 | 0.6939
nSLTM	−136.5600 | 1.9029	−134.7200 | 3.1548	−133.0900 | 2.6930	−134.1400 | 4.1879	−131.2000 | 4.6858

(b)

Model	50	100	150	200	250	300
UGTE_ID	**−129.4202** | 2.0221	−135.6328 | 0.5888	−140.0148 | 0.6207	−142.6985 | 0.5662	−144.4550 | 0.0764	−146.4922 | 0.3277
UGTE_ALL	−131.7624 | 2.1850	−136.2210 | 0.7076	−139.5120 | 0.8791	−142.6429 | 1.1468	−145.2449 | 0.3464	−146.0954 | 0.1239
AT	−129.7481 | 8.7731	−125.5992 | 0.9994	−126.3202 | 1.2694	−129.0001 | 0.5902	−131.7856 | 0.2282	−133.0581 | 0.1674
MSTM	−132.5110 | 3.3665	−137.2911 | 2.8133	−135.6930 | 0.2177	−134.2135 | 0.4396	−132.0621 | 0.0607	−130.0798 | 0.3245
SLTM	−140.8132 | 2.0079	−139.9092 | 0.2329	−138.6894 | 0.6887	−137.1148 | 0.6131	−136.1514 | 0.1311	−133.0861 | 0.1790
CSTM	−130.4588 | 6.9379	**−109.7005** | 4.2767	**−102.4071** | 3.4400	**−99.0179** | 5.3653	**−100.2382** | 1.4324	**−101.0950** | 0.4457
sNTM	−135.1800 | 4.4410	−135.1500 | 0.8348	−137.1700 | 0.8195	−138.2600 | 0.6314	−138.0400 | 0.1131	−138.5700 | 0.2458
nSLTM	−133.9400 | 4.8281	−130.5900 | 0.2056	−129.4800 | 0.4201	−119.1700 | 0.3844	−112.0800 | 0.1056	−103.9700 | 0.2104

(a) Coherence@10 with $|Z| \in \{25, 30, 35, 40, 45\}$. (b) Coherence@10 with $|Z| \in \{50, 100, 150, 200, 250, 300\}$

Table 10 Coherence@20 of UGTE_ID and baselines with different topic numbers when $|G| = 10$, where the best results are highlighted in boldface

(a)

Model	25	30	35	40	45
UGTE_ID	**−472.8675** | 15.7981	**−473.7831** | 42.9177	**−477.3060** | 14.8741	**−478.1656** | 26.7178	**−478.4086** | 18.3734
UGTE_ALL	−480.8306 | 13.6080	−488.1540 | 22.3221	−485.9116 | 27.6891	−482.2417 | 17.2646	−480.1103 | 34.3768
AT	−562.2700 | 71.2536	−547.4200 | 32.0064	−541.0400 | 19.7631	−529.1500 | 13.6627	−524.5200 | 1.8129
MSTM	−557.0061 | 47.3639	−552.2588 | 31.9858	−558.2339 | 27.9613	−564.0604 | 17.5203	−557.8820 | 18.1813
SLTM	−586.0753 | 52.5419	−585.6113 | 26.8187	−579.9124 | 36.9702	−581.7932 | 8.8410	−582.4983 | 6.1192
CSTM	−645.6893 | 63.3989	−612.7027 | 58.0097	−576.1643 | 56.0269	−573.3450 | 5.3249	−547.8586 | 23.4765
sNTM	−569.4500 | 41.8099	−571.8500 | 11.0838	−572.9800 | 17.4909	−568.0000 | 15.6657	−570.9300 | 15.9171
nSLTM	−512.7800 | 47.4189	−518.5900 | 18.7456	−513.2900 | 17.8056	−515.4300 | 19.6309	−521.7100 | 5.0765

(b)

Model	50	100	150	200	250	300
UGTE_ID	**−478.2550** | 26.4989	**−487.6648** | 3.7684	−499.7023 | 0.8025	−502.8695 | 1.0815	−507.2252 | 2.4839	−510.5402 | 1.2564
UGTE_ALL	−482.4778 | 13.8110	−491.2869 | 12.4179	−497.6480 | 6.5973	−503.6150 | 5.8417	−508.3999 | 0.3619	−510.1580 | 0.4259
AT	−521.3000 | 26.9379	−529.5800 | 0.2767	−551.7500 | 3.4400	−557.4300 | 5.3653	−582.6200 | 1.4324	−608.5500 | 0.4457
MSTM	−558.1949 | 14.7818	−549.1293 | 6.3246	−532.6921 | 1.5809	−521.9660 | 4.9757	−506.8969 | 1.2361	−473.5094 | 0.2212
SLTM	−577.5587 | 6.7715	−562.9262 | 5.8307	−545.3276 | 3.9174	−530.4544 | 1.9730	−518.9439 | 1.5796	−491.5813 | 2.0568
CSTM	−541.5672 | 40.8956	−503.1077 | 7.0876	−515.3516 | 7.0819	−531.8953 | 5.6414	−550.8603 | 5.6205	−575.8195 | 5.4421
sNTM	−568.0800 | 16.6584	−571.0800 | 3.1587	−579.2800 | 4.5099	−582.4100 | 0.6796	−584.4300 | 0.6874	−583.1400 | 0.9229
nSLTM	−516.1500 | 6.0068	−501.5300 | 8.4042	**−486.2800** | 4.0286	**−451.4800** | 1.9489	**−418.7700** | 4.4428	**−384.0500** | 1.8403

(a) Coherence@20 with $|Z| \in \{25, 30, 35, 40, 45\}$. (b) Coherence@20 with $|Z| \in \{50, 100, 150, 200, 250, 300\}$

Table 11 Coherence@30 of UGTE_ID and baselines with different topic numbers when $|G| = 10$, where the best results are highlighted in boldface

(a)

Model	**25**	**30**	**35**	**40**	**45**
UGTE_ID	**−988.4709** | 61.0751	**−983.1960** | 69.5678	**−986.4559** | 42.3027	**−977.5956** | 49.1749	**−976.1411** | 44.2658
UGTE_ALL	−1002.5797 | 85.5218	−1001.2881 | 68.6300	−989.9886 | 59.0986	−991.7126 | 59.7629	−981.0827 | 65.2007
AT	−1236.9500 | 177.9269	−1200.7100 | 37.0210	−1198.0500 | 79.3390	−1187.1300 | 62.7150	−1182.6700 | 25.2279
MSTM	−1239.7780 | 172.4073	−1229.8218 | 114.0211	−1215.9444 | 77.6506	−1247.1960 | 73.3765	−1223.4319 | 49.4679
SLTM	−1291.8175 | 190.3193	−1300.0810 | 100.5746	−1282.8348 | 78.0375	−1277.9796 | 60.0175	−1279.9932 | 133.7248
CSTM	−1431.5930 | 266.5040	−1352.6545 | 274.9734	−1302.2942 | 90.7964	−1278.8657 | 79.4288	−1247.5063 | 175.5945
sNTM	−1302.1700 | 219.0994	−1301.8000 | 133.1085	−1310.6200 | 87.6051	−1303.2200 | 51.8289	−1301.3900 | 54.2377
nSLTM	−1093.6600 | 114.1605	−1116.1300 | 43.1291	−1094.4900 | 24.2220	−1089.2900 | 35.2711	−1108.3700 | 66.4202

(b)

Model	**50**	**100**	**150**	**200**	**250**	**300**
UGTE_ID	**−976.4773** | 25.7348	−982.4643 | 16.2643	−992.5024 | 12.8972	−992.2731 | 10.9598	−995.3995 | 0.9553	−995.9672 | 1.6108
UGTE_ALL	−987.1877 | 31.6687	**−982.1087** | 23.0584	**−990.6200** | 1.7070	−994.4214 | 7.9142	−995.0471 | 3.5257	−996.2114 | 5.3899
AT	−1177.9100 | 44.2355	−1260.9600 | 29.4490	−1313.8700 | 8.2983	−1408.6000 | 3.8219	−1428.5200 | 2.1515	−1418.4100 | 5.0981
MSTM	−1229.9652 | 28.7204	−1180.8602 | 13.1571	−1147.9769 | 4.9726	−1070.2926 | 8.5150	−986.6450 | 10.6102	−934.6358 | 0.7483
SLTM	−1268.9048 | 68.6943	−1224.9951 | 34.2277	−1178.3816 | 11.0170	−1094.3976 | 1.7230	−1007.1394 | 9.1355	−951.9653 | 5.0043
CSTM	−1237.2050 | 12.3721	−1241.2599 | 60.8073	−1308.2944 | 24.3348	−1360.4824 | 12.7288	−1412.2278 | 3.0130	−1466.4400 | 1.8114
sNTM	−1302.6100 | 37.8062	−1307.9600 | 27.8940	−1320.7400 | 23.1953	−1332.6600 | 7.8509	−1337.4000 | 3.0403	−1333.5500 | 1.8040
nSLTM	−1085.8700 | 64.8624	−1055.6700 | 51.8895	−1021.2700 | 12.9144	**−941.0300** | 8.0439	**−878.5300** | 10.2779	**−803.5600** | 2.2154

(a) Coherence@30 with $|Z| \in \{25, 30, 35, 40, 45\}$. (b) Coherence@30 with $|Z| \in \{50, 100, 150, 200, 250, 300\}$

Table 12 Accuracy of UGTE_ID and baselines with different topic numbers when $|G| = 10$, where the best results are highlighted in boldface

(a)

Model	25	30	35	40	45
UGTE_ID	0.4870 | 0.0000	**0.4914** | 0.0001	0.4872 | 0.0001	**0.5012** | 0.0000	0.4888 | 0.0001
UGTE_ALL	**0.4921** | 0.0001	0.4900 | 0.0000	0.4882 | 0.0001	0.4857 | 0.0000	**0.4947** | 0.0001
MSTM	0.2072 | 0.0046	0.2062 | 0.0046	0.2091 | 0.0053	0.2011 | 0.0046	0.1998 | 0.0036
SLTM	0.1949 | 0.0056	0.1929 | 0.0054	0.1874 | 0.0085	0.1848 | 0.0089	0.1813 | 0.0056
CSTM	0.2788 | 0.0014	0.2838 | 0.0016	0.2852 | 0.0014	0.2731 | 0.0003	0.2693 | 0.0003
sNTM	0.1372 | 0.0050	0.1432 | 0.0065	0.1418 | 0.0165	0.1440 | 0.0209	0.1434 | 0.0267
nSLTM	0.4870 | 0.0001	0.4883 | 0.0001	**0.5005** | 0.0001	0.4963 | 0.0001	0.4938 | 0.0000

(b)

Model	50	100	150	200	250	300
UGTE_ID	**0.4964** | 0.0001	0.4973 | 0.0002	0.4988 | 0.0000	0.5011 | 0.0000	0.4951 | 0.0000	0.5014 | 0.0001
UGTE_ALL	0.4933 | 0.0001	0.4991 | 0.0000	0.5015 | 0.0002	0.5011 | 0.0000	0.5027 | 0.0001	0.5026 | 0.0001
MSTM	0.1942 | 0.0024	0.1769 | 0.0035	0.1732 | 0.0012	0.1550 | 0.0023	0.1666 | 0.0012	0.1579 | 0.0003
SLTM	0.1738 | 0.0079	0.1503 | 0.0082	0.1437 | 0.0057	0.1480 | 0.0079	0.1503 | 0.0066	0.1370 | 0.0043
CSTM	0.2829 | 0.0045	0.2856 | 0.0020	0.2691 | 0.0011	0.2801 | 0.0015	0.2717 | 0.0002	0.2731 | 0.0002
sNTM	0.1419 | 0.0124	0.1470 | 0.0249	0.1470 | 0.0278	0.1444 | 0.0107	0.1370 | 0.0036	0.1401 | 0.0251
nSLTM	0.4880 | 0.0000	**0.5074** | 0.0001	**0.5224** | 0.0000	**0.5184** | 0.0000	**0.5342** | 0.0001	**0.5341** | 0.0001

(a) Accuracy with $|Z| \in \{25, 30, 35, 40, 45\}$. (b) Accuracy with $|Z| \in \{50, 100, 150, 200, 250, 300\}$

Table 13 Kappa Score of UGTE_ID and baselines with different topic numbers when $|G| = 10$, where the best results are highlighted in boldface

(a)

Model	25	30	35	40	45
UGTE_ID	0.4016 | 0.0000	**0.4069** | 0.0000	0.4018 | 0.0001	**0.4183** | 0.0000	0.4038 | 0.0001
UGTE_ALL	**0.4075** | 0.0000	0.4051 | 0.0000	0.4030 | 0.0001	0.4000 | 0.0000	**0.4107** | 0.0001
MSTM	0.0777 | 0.0004	0.0753 | 0.0007	0.0792 | 0.0005	0.0700 | 0.0006	0.0698 | 0.0004
SLTM	0.0629 | 0.0006	0.0625 | 0.0004	0.0558 | 0.0008	0.0512 | 0.0005	0.0463 | 0.0002
CSTM	0.1594 | 0.0000	0.1636 | 0.0002	0.1668 | 0.0000	0.1528 | 0.0002	0.1480 | 0.0000
sNTM	−0.0015 | 0.0000	0.0004 | 0.0000	−0.0023 | 0.0000	0.0025 | 0.0000	−0.0012 | 0.0000
nSLTM	0.4032 | 0.0000	0.4047 | 0.0001	**0.4183** | 0.0001	0.4136 | 0.0000	0.4104 | 0.0000

(b)

Model	50	100	150	200	250	300
UGTE_ID	**0.4127** | 0.0000	0.4137 | 0.0001	0.4153 | 0.0000	0.4182 | 0.0000	0.4111 | 0.0000	0.4185 | 0.0001
UGTE_ALL	0.4089 | 0.0001	0.4159 | 0.0000	0.4186 | 0.0000	0.4181 | 0.0001	0.4199 | 0.0000	0.4198 | 0.0000
MSTM	0.0619 | 0.0006	0.0419 | 0.0007	0.0375 | 0.0006	0.0178 | 0.0004	0.0292 | 0.0003	0.0207 | 0.0008
SLTM	0.0392 | 0.0003	0.0114 | 0.0004	0.0051 | 0.0009	0.0089 | 0.0005	0.0109 | 0.0001	−0.0040 | 0.0003
CSTM	0.1638 | 0.0000	0.1659 | 0.0001	0.1461 | 0.0001	0.1599 | 0.0001	0.1495 | 0.0000	0.1512 | 0.0000
sNTM	0.0019 | 0.0000	0.0020 | 0.0000	0.0011 | 0.0000	−0.0019 | 0.0000	−0.0006 | 0.0000	−0.0005 | 0.0000
nSLTM	0.4044 | 0.0000	**0.4252** | 0.0001	**0.4422** | 0.0000	**0.4373** | 0.0000	**0.4553** | 0.0001	**0.4556** | 0.0001

(a) Kappa Score with $|Z| \in \{25, 30, 35, 40, 45\}$. (b) Kappa Score with $|Z| \in \{50, 100, 150, 200, 250, 300\}$

Table 14 The mean and variance of topic discovery and emotion discovery of UGTE_ID over different numbers of user groups, where the best results are highlighted in boldface

Group Number	Z=25	Z=50	Z=100
(a)			
1	**−123.8282** | 3.0761	−128.9259 | 2.7922	−135.1777 | 0.8902
2	**−124.8229** | 6.1371	−129.3543 | 3.3956	−136.0454 | 0.7786
3	**−125.4330** | 4.0319	−129.9477 | 2.7302	−135.2791 | 1.0459
4	**−125.3082** | 8.6745	−129.9907 | 2.1933	−135.7210 | 2.3208
5	**−124.6989** | 6.3020	−128.6609 | 1.1947	−135.1003 | 1.3900
6	**−126.5772** | 3.6613	−129.8699 | 3.1354	−135.1326 | 1.8942
7	**−125.5934** | 4.8152	−131.0385 | 3.2536	−135.8892 | 3.1443
8	**−126.0309** | 3.6204	−129.6428 | 0.6667	−135.6102 | 0.8877
9	**−124.6777** | 2.7678	−130.2807 | 1.2375	−136.3205 | 1.1138
10	**−127.2744** | 2.5617	−129.2806 | 2.1171	−135.2322 | 1.6146
20	**−128.8997** | 3.3402	−129.9516 | 1.0019	−134.8486 | 0.6064
30	**−129.8742** | 4.2374	−130.4765 | 2.3407	−135.6203 | 1.0488
40	**−130.5837** | 4.8134	−132.9117 | 1.2479	−134.9361 | 0.8259
50	**−133.4985** | 1.2831	−134.9123 | 1.7117	−135.8846 | 0.9894
60	−135.0881 | 4.0638	**−134.5896** | 1.1725	−136.1052 | 0.0960
70	−137.7421 | 2.5095	−137.8466 | 1.2079	**−134.9371** | 0.7920
80	−140.2717 | 1.9243	−138.9284 | 2.3987	**−135.5992** | 0.1904
90	−140.9768 | 2.0798	−140.9756 | 1.6865	**−135.6350** | 0.7278
100	−142.6105 | 3.0235	−142.9163 | 2.2073	**−135.5691** | 0.9332
(b)			
1	**−461.8391** | 43.3912	−477.9031 | 13.7765	−487.8411 | 4.3157
2	**−470.8320** | 33.3073	−477.5539 | 26.8731	−489.4252 | 13.0675
3	**−468.2822** | 76.2225	−475.0971 | 27.9649	−489.1515 | 6.0255
4	**−473.2406** | 56.8750	−479.3433 | 1.4839	−490.0553 | 5.9319

Table 14 (continued)

5	**−471.4755** | 72.2202	−474.7813 | 32.2194	−489.4264 | 3.0962
6	**−471.2181** | 24.8386	−475.9743 | 18.8449	−489.1645 | 10.7042
7	**−472.0221** | 17.6211	−480.1584 | 14.2602	−490.0278 | 8.2974
8	**−470.3862** | 62.8673	−475.0753 | 10.0474	−487.7244 | 3.1128
9	**−470.3284** | 56.6883	−478.1293 | 54.1056	−489.8041 | 18.2835
10	−476.9194 | 46.2270	**−475.8162** | 14.3190	−489.8370 | 13.6409
20	**−478.9845** | 25.3742	−479.1241 | 14.9650	−487.2175 | 6.3009
30	−487.0167 | 35.7185	**−481.3901** | 5.5966	−489.8622 | 2.3853
40	**−484.6942** | 11.7033	−486.3666 | 7.9941	−488.1429 | 0.4411
50	−496.0669 | 22.1057	−493.0257 | 17.3397	**−490.6975** | 17.5568
60	−500.5333 | 15.4620	−492.6950 | 12.5078	**−490.1357** | 11.2360
70	−503.2043 | 26.9036	−501.2808 | 14.9594	**−490.4923** | 7.4278
80	−513.1288 | 13.2617	−508.4942 | 16.3863	**−489.4492** | 8.7502
90	−513.4376 | 41.9970	−512.7702 | 16.5883	**−489.8309** | 7.2871
100	−518.4933 | 12.2850	−514.9007 | 5.5863	**−489.5230** | 5.4761
(c)			
1	**−965.8252** | 171.9881	−978.5236 | 76.1347	−983.7645 | 24.6176
2	**−976.9719** | 212.1470	−979.9890 | 35.4531	−984.7415 | 10.4838
3	−978.8473 | 161.9569	**−972.9784** | 184.0852	−985.3553 | 3.2876
4	−983.1173 | 184.9980	**−973.3633** | 31.5091	−983.9500 | 25.2068
5	−985.0156 | 174.0750	**−969.2186** | 176.6199	−985.8712 | 21.4486
6	−982.6703 | 43.8376	**−974.5103** | 21.2742	−983.1227 | 5.0223
7	−987.8648 | 224.0784	**−969.9208** | 39.9077	−984.2788 | 25.2481
8	−990.4951 | 206.0546	**−969.8918** | 37.3688	−981.2195 | 23.2299
9	−981.8750 | 171.2419	−982.7219 | 60.5603	**−981.8180** | 14.7489

Table 14 (continued)

10	−985.8702 | 188.2467	**−980.9669** | 16.5535	−983.0013 | 4.9354
20	−993.9860 | 88.6885	**−978.9197** | 29.5323	−982.1190 | 21.5686
30	−1007.3011 | 54.2997	−987.3268 | 37.5943	**−985.6545** | 2.7158
40	−1005.5088 | 92.2993	−993.0100 | 39.3091	**−985.5782** | 22.4872
50	−1022.5639 | 114.3346	−1002.0531 | 100.4693	**−986.7750** | 27.1315
60	−1032.8760 | 71.8156	−1003.4838 | 18.2392	**−985.0538** | 8.5261
70	−1030.8940 | 113.0801	−1011.5551 | 77.5889	**−988.5810** | 46.3901
80	−1046.7456 | 38.0490	−1025.6196 | 36.6561	**−986.3313** | 10.0813
90	−1047.8796 | 53.2804	−1034.9163 | 19.4256	**−981.7170** | 17.8968
100	−1050.8502 | 42.4042	−1034.2433 | 37.4420	**−985.9698** | 29.6594

Table 14 (continued)

Group Number	Accuracy			Kappa Score		
	Z=25	Z=50	Z=100	Z=25	Z=50	Z=100
(d)						
1	0.5027 | 0.0000	0.5009 | 0.0000	**0.5056** | 0.0000	0.4200 | 0.0000	0.4179 | 0.0000	**0.4233** | 0.0000
2	0.4938 | 0.0001	0.4948 | 0.0001	**0.4969** | 0.0001	0.4096 | 0.0001	0.4108 | 0.0001	**0.4133** | 0.0001
3	0.4957 | 0.0000	**0.5035** | 0.0000	0.4994 | 0.0000	0.4118 | 0.0001	**0.4210** | 0.0000	0.4161 | 0.0000
4	0.4891 | 0.0002	0.4948 | 0.0001	**0.5037** | 0.0000	0.4040 | 0.0002	0.4107 | 0.0001	**0.4212** | 0.0000
5	0.4959 | 0.0001	0.4946 | 0.0000	**0.5011** | 0.0001	0.4120 | 0.0002	0.4105 | 0.0001	**0.4181** | 0.0002
6	0.4937 | 0.0001	0.4957 | 0.0000	**0.5012** | 0.0000	0.4095 | 0.0001	0.4119 | 0.0000	**0.4182** | 0.0000
7	0.4989 | 0.0000	0.4933 | 0.0002	**0.5040** | 0.0000	0.4154 | 0.0001	0.4089 | 0.0003	**0.4215** | 0.0000
8	0.4930 | 0.0000	0.4875 | 0.0001	**0.4976** | 0.0000	0.4085 | 0.0001	0.4021 | 0.0001	**0.4140** | 0.0001
9	**0.5007** | 0.0000	0.4846 | 0.0000	0.4980 | 0.0000	**0.4175** | 0.0001	0.3989 | 0.0000	0.4145 | 0.0000
10	0.4869 | 0.0001	0.4895 | 0.0000	**0.5033** | 0.0001	0.4015 | 0.0002	0.4044 | 0.0001	**0.4207** | 0.0001
20	0.4875 | 0.0001	0.4900 | 0.0000	**0.4998** | 0.0001	0.4023 | 0.0001	0.4052 | 0.0000	**0.4166** | 0.0001
30	0.4901 | 0.0001	0.4959 | 0.0001	**0.5037** | 0.0001	0.4053 | 0.0001	0.4121 | 0.0001	**0.4212** | 0.0001
40	0.4884 | 0.0001	0.4866 | 0.0000	**0.5031** | 0.0001	0.4034 | 0.0001	0.4012 | 0.0000	**0.4204** | 0.0001
50	0.4956 | 0.0000	0.4900 | 0.0000	**0.4993** | 0.0002	0.4118 | 0.0000	0.4052 | 0.0000	**0.4160** | 0.0002
60	**0.5002** | 0.0001	0.4920 | 0.0001	0.4994 | 0.0000	**0.4170** | 0.0001	0.4073 | 0.0001	0.4161 | 0.0000
70	0.4901 | 0.0000	0.4935 | 0.0000	**0.4988** | 0.0001	0.4054 | 0.0001	0.4093 | 0.0000	**0.4155** | 0.0001
80	0.4971 | 0.0001	0.4933 | 0.0000	**0.4985** | 0.0001	0.4133 | 0.0002	0.4090 | 0.0001	**0.4151** | 0.0001
90	0.4969 | 0.0002	0.4931 | 0.0001	**0.5001** | 0.0000	0.4132 | 0.0003	0.4089 | 0.0001	**0.4168** | 0.0000
100	0.4950 | 0.0000	0.4995 | 0.0001	**0.5009** | 0.0001	0.4109 | 0.0001	0.4164 | 0.0001	**0.4179** | 0.0001

(a) Coherence@10 of topic discovery. (b) Coherence@20 of topic discovery. (c) Coherence@30 of topic discovery. (d) Accuracy and Kappa Score of emotion detection

Table 15 The mean and variance values of impact of extremely short text on UGTE_ID and MSTM

(a)				
Coherence@10	UGTE_ID (short text)	MSTM (short text)	UGTE_ID (extremely short text)	MSTM (extremely short text)
25	−67.9337 | 1.9686	−125.2921 | 4.9813	−81.0953 | 0.3438	−123.6870 | 2.3916
30	−66.3398 | 0.5736	−122.8285 | 2.6585	−78.9038 | 2.5659	−125.8367 | 6.6384
35	−64.9278 | 1.6594	−120.3800 | 8.3506	−77.6888 | 1.2641	−125.7000 | 2.5135
40	−64.1684 | 3.0847	−119.7305 | 1.8957	−74.8293 | 0.5797	−124.4655 | 1.6141
45	−63.1678 | 0.8834	−118.9447 | 11.9979	−74.1155 | 2.3352	−126.1198 | 4.1895
50	−62.9171 | 0.0757	−116.1973 | 7.6968	−73.2623 | 0.8009	−125.4669 | 1.6027
100	−59.6979 | 0.1562	−94.7710 | 3.9766	−66.1968 | 0.3717	−123.6672 | 5.7860
150	−59.3381 | 0.0886	−90.7548 | 1.1470	−63.9791 | 0.1949	−121.4538 | 0.2576
200	−58.8641 | 0.0188	−94.0969 | 2.3978	−63.1348 | 0.0713	−119.3321 | 1.1826
250	−58.6697 | 0.0901	−95.2936 | 0.6904	−62.7651 | 0.0472	−117.0551 | 1.9442
300	−58.7751 | 0.0488	−94.0578 | 0.6874	−62.4116 | 0.0319	−113.7766 | 0.4559
(b)				
Coherence@20	UGTE_ID (short text)	MSTM (short text)	UGTE_ID (extremely short text)	MSTM (extremely short text)
25	−224.0668 | 12.7039	−587.2539 | 28.0626	−399.1152 | 15.8461	−620.9208 | 9.4251
30	−215.5474 | 10.4016	−578.8302 | 67.4005	−400.5570 | 2.9390	−620.6153 | 59.0241
35	−205.0287 | 14.3585	−552.9686 | 30.8117	−390.9206 | 9.7951	−617.3729 | 116.5202
40	−199.8369 | 17.6423	−514.9208 | 81.0895	−390.8898 | 4.3081	−617.6413 | 75.4847
45	−194.0536 | 0.3153	−498.1536 | 83.3026	−388.5121 | 22.7391	−614.5683 | 16.1756
50	−190.8648 | 10.1737	−464.9679 | 89.0391	−389.8786 | 11.5109	−618.3922 | 0.5068
100	−169.4500 | 1.1323	−439.9617 | 27.6227	−384.0398 | 11.4235	−592.3600 | 27.7738
150	−161.3123 | 0.1950	−453.8012 | 11.3582	−390.0078 | 3.5649	−570.1901 | 21.3781
200	−157.3825 | 0.1013	−449.9891 | 10.9651	−391.8322 | 2.0831	−529.4262 | 9.1291
250	−155.2101 | 0.3482	−431.0458 | 4.6867	−394.9236 | 1.9617	−486.7823 | 16.2153
300	−153.3774 | 0.0807	−411.5242 | 9.8505	−397.4808 | 0.5136	−465.2004 | 17.3309

Table 15 (continued)

(c)				
Coherence@30	UGTE_ID (short text)	MSTM (short text)	UGTE_ID (extremely short text)	MSTM (extremely short text)
25	−557.5358 | 44.1923	−1186.3438 | 61.0624	−951.2371 | 44.1839	−1375.4125 | 54.8620
30	−543.2405 | 22.6519	−1120.8832 | 60.7645	−959.5559 | 34.8954	−1374.5276 | 60.8654
35	−526.7139 | 26.9596	−1042.2577 | 58.7206	−949.9567 | 30.5716	−1364.1326 | 57.2947
40	−525.2440 | 8.3395	−996.4760 | 63.9572	−942.2908 | 26.4009	−1358.4812 | 130.8195
45	−516.5046 | 10.3561	−994.4593 | 36.9967	−939.9171 | 168.8056	−1347.1129 | 54.7513
50	−512.6368 | 17.9373	−968.9523 | 61.2160	−951.5307 | 40.3218	−1338.8333 | 77.8130
100	−490.4139 | 3.7409	−998.5308 | 36.7486	−968.3013 | 18.8755	−1267.3512 | 28.4627
150	−482.1151 | 0.7218	−973.7064 | 83.5750	−990.9090 | 20.5843	−1129.0113 | 21.0184
200	−480.1174 | 0.6773	−902.4509 | 22.0120	−996.2109 | 5.3002	−1025.5737 | 28.2384
250	−477.2773 | 0.3147	−859.5349 | 30.3163	−1005.1819 | 14.5636	−990.4725 | 44.3866
300	−475.5740 | 0.2716	−843.5639 | 4.0196	−1010.3193 | 4.8989	−983.4701 | 9.2100
(d)				
Accuracy	UGTE_ID (short text)	MSTM (short text)	UGTE_ID (extremely short text)	MSTM (extremely short text)
25	0.4056 | 0.0003	0.1485 | 0.0000	0.4487 | 0.0001	0.1844 | 0.0002
30	0.3898 | 0.0003	0.1431 | 0.0000	0.4519 | 0.0001	0.1791 | 0.0000
35	0.3926 | 0.0003	0.1466 | 0.0001	0.4437 | 0.0001	0.1753 | 0.0002
40	0.3907 | 0.0004	0.1490 | 0.0000	0.4529 | 0.0001	0.1801 | 0.0002
45	0.3777 | 0.0005	0.1488 | 0.0000	0.4395 | 0.0000	0.1849 | 0.0001
50	0.3953 | 0.0005	0.1398 | 0.0000	0.4495 | 0.0000	0.1714 | 0.0001

Table 15 (continued)

100	0.3749 | 0.0003	0.1410 | 0.0000	0.4456 | 0.0001	0.1586 | 0.0001
150	0.3860 | 0.0003	0.1452 | 0.0000	0.4594 | 0.0001	0.1524 | 0.0001
200	0.4065 | 0.0007	0.1454 | 0.0001	0.4511 | 0.0002	0.1456 | 0.0002
250	0.3953 | 0.0004	0.1433 | 0.0000	0.4473 | 0.0000	0.1517 | 0.0000
300	0.3935 | 0.0005	0.1439 | 0.0000	0.4528 | 0.0001	0.1481 | 0.0002
(e)				
Kappa Score	UGTE_ID (short text)	MSTM (short text)	UGTE_ID (extremely short text)	MSTM (extremely short text)
25	0.2979 | 0.0004	0.0069 | 0.0000	0.3568 | 0.0001	0.0452 | 0.0004
30	0.2793 | 0.0004	−0.0008 | 0.0000	0.3606 | 0.0001	0.0429 | 0.0000
35	0.2827 | 0.0004	0.0043 | 0.0000	0.3510 | 0.0001	0.0396 | 0.0004
40	0.2809 | 0.0006	0.0075 | 0.0000	0.3618 | 0.0001	0.0443 | 0.0002
45	0.2653 | 0.0007	0.0067 | 0.0000	0.3460 | 0.0000	0.0472 | 0.0001
50	0.2862 | 0.0008	−0.0036 | 0.0000	0.3578 | 0.0001	0.0348 | 0.0001
100	0.2617 | 0.0005	−0.0033 | 0.0000	0.3531 | 0.0001	0.0148 | 0.0001
150	0.2744 | 0.0004	0.0023 | 0.0000	0.3693 | 0.0002	0.0084 | 0.0001
200	0.2988 | 0.0011	0.0033 | 0.0000	0.3595 | 0.0002	0.0068 | 0.0002
250	0.2860 | 0.0006	0.0005 | 0.0000	0.3550 | 0.0000	0.0084 | 0.0000
300	0.2830 | 0.0007	−0.0004 | 0.0000	0.3616 | 0.0001	0.0049 | 0.0003

(a) Coherence@10 of topic discovery. (a) Coherence@20 of topic discovery. (a) Coherence@30 of topic discovery. (a) Accuracy of emotion detection. (a) Kappa Score of emotion detection

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
2. Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., Yu, Y.: Mining social emotions from affective text. IEEE Trans. Knowl. Data Eng. **24**(9), 1658–1670 (2012)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Cao, Z., Li, S., Liu, Y., Li, W., Ji, H.: A novel neural topic model and its supervised extension. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, pp. 2210–2216 (2015)
5. Chen, H., Yin, H., Li, X., Wang, M., Chen, W., Chen, T.: People opinion topic model: Opinion based user clustering in social networks. In: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, pp. 1353–1359 (2017)
6. Chen, Z., Liu, B.: Mining Topics in Documents: Standing on the Shoulders of Big Data. In: The 20Th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, pp. 1116–1125 (2014)
7. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)
8. Diao, Q., Jiang, J., Zhu, F., Lim, E.: Finding bursty topics from microblogs. In: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Jeju Island, Korea - Volume 1: Long Papers, pp. 536–544 (2012)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
10. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1/2), 177–196 (2001)
11. Huang, F., Zhang, S., Zhang, J., Yu, G.: Multimodal learning for topic sentiment analysis in microblogging. Neurocomputing **253**, 144–153 (2017)
12. Huang, M., Rao, Y., Liu, Y., Xie, H., Wang, F.L.: Siamese network-based supervised topic modeling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, pp. 4652–4662 (2018)
13. Huang, T., Nevmyvaka, Y.: A practical markov chain monte carlo approach to decision problems. In: Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference, Key West, pp. 520–524 (2001)
14. Jin, O., Liu, N.N., Zhao, K., Yu, Y., Yang, Q.: Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, pp. 775–784 (2011)
15. Lin, T., Tian, W., Mei, Q., Cheng, H.: The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text. In: 23Rd International World Wide Web Conference, WWW '14, Seoul, pp. 539–550 (2014)
16. Mcpherson, M., Smithlovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annu. Rev. Sociol. **27**(1), 415–444 (2001)
17. Mimno, D.M., Wallach, H.M., Talley, E.M., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 262–272 (2011)
18. Parthasarathy, S., Ruan, Y., Satuluri, V.: Community Discovery in Social Networks: Applications, methods and emerging trends. In: Social Network Data Analytics, pp. 79–113 (2011)
19. Phan, X.H., Nguyen, M.L., Horiguchi, S.: Learning to classify short and sparse text & Web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, pp. 91–100 (2008)
20. Poria, S., Gelbukh, A.F., Hussain, A., Howard, N., Das, D., Bandyopadhyay, S.: Enhanced senticnet with affective labels for concept-based opinion mining. IEEE Intell. Syst. **28**(2), 31–38 (2013)
21. Pu, X., Jin, R., Wu, G., Han, D., Xue, G.: Topic modeling in semantic space with keywords. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, pp. 1141–1150 (2015)
22. Ramage, D., Hall, D.L.W., Nallapati, R., Manning, C.D.: 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, EMNLP 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 248–256 (2009)

23. Rao, Y.: Contextual sentiment topic model for adaptive social emotion classification. IEEE Intell. Syst. **31**(1), 41–47 (2016)
24. Rao, Y., Li, Q., Mao, X., Wenyin, L.: Sentiment topic models for social emotion mining. Inf. Sci. **266**, 90–100 (2014)
25. Rao, Y., Pang, J., Xie, H., Liu, A., Wong, T., Li, Q., Wang, F.L.: Supervised Intensive Topic Models for Emotion Detection over Short Text. In: Database Systems for Advanced Applications - 22Nd International Conference, DASFAA 2017, Suzhou, Proceedings, Part I, pp. 408–422 (2017)
26. Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, pp. 487–494 (2004)
27. Sachan, M., Contractor, D., Faruquie, T.A., Subramaniam, L.V.: Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, pp. 331–340 (2012)
28. Sahami, M., Heilman, T.D.: A Web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, pp. 377–386 (2006)
29. Wallach, H.M., Mimno, D.M., McCallum, A.: Rethinking LDA: why priors matter. In: Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, pp. 1973–1981 (2009)
30. Wang, D., Li, J., Xu, K., Wu, Y.: Sentiment community detection: exploring sentiments and relationships in social networks. Electron. Commer. Res. **17**(1), 103–132 (2017)
31. Wang, X., Mohanty, N., McCallum, A.: Group and topic discovery from relations and text. In: Proceedings of the 3rd international workshop on Link discovery, LinkKDD 2005, Chicago, pp. 28–35 (2005)
32. Xu, K., Qi, G., Huang, J., Wu, T., Fu, X.: Detecting bursts in sentiment-aware topics from social media. Knowl.-Based Syst. **141**, 44–54 (2018)
33. Yang, B., Manandhar, S.: STC: A Joint Sentiment-Topic Model for Community Identification. In: Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2014 International Workshops: DANTH, BDM, MobiSocial, BigEC, CloudSD, MSMV-MBI, SDA, DMDA-Health, ALSIP, SocNet, DMBIH, BigPMA, Tainan, 2014. Revised Selected Papers, pp. 535–548 (2014)
34. Zhang, L., Liu, B.: Sentiment Analysis and Opinion Mining. In: Encyclopedia of Machine Learning and Data Mining, pp. 1152–1161 (2017)
35. Zhang, Q., Gong, Y., Sun, X., Huang, X.: Time-aware personalized hashtag recommendation on social media. In: COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, Dublin, pp. 203–212 (2014)
36. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., Li, X.: Comparing Twitter and Traditional Media Using Topic Models. In: Advances in Information Retrieval - 33Rd European Conference on IR Research, ECIR 2011, Dublin, 2011. Proceedings, pp. 338–349 (2011)
37. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., Li, X.: Comparing Twitter and Traditional Media Using Topic Models. In: Advances in Information Retrieval - 33Rd European Conference on IR Research, ECIR 2011, Dublin, 2011. Proceedings, pp. 338–349 (2011)
38. Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., Xiong, H.: Topic modeling of short texts: A pseudo-document view. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, pp. 2105–2114 (2016)
39. Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., Xiong, H.: Topic modeling of short texts: A pseudo-document view. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, pp. 2105–2114 (2016)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.