IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 21, 2019, accepted August 11, 2019, date of publication August 20, 2019, date of current version September 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936457

Bi-Level Attention Model for Sentiment Analysis of Short Texts

WEI LIU, GUOXI CAO, AND JIANQIN YIN
School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Wei Liu (twlw@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673192, in part by the Fund for Outstanding Youth of Shandong Provincial High School under Grant ZR2016JL023, and in part by the Basic Scientific Research Project of Beijing University of Posts and Telecommunications under Grant 2018RC31.

ABSTRACT Short text is an important form of information dissemination and opinion expression in various social media platforms. Sentiment analysis of short texts is beneficial for the understanding of customers' emotional state, obtaining customers' opinions and attitudes toward events, information and products, however, is difficult because the sparsity of the short-text data. Unlike the traditional methods using the external knowledge, this paper proposes a bi-level attention model for sentiment analysis of short texts, which does not rely on external knowledge to deal with the data sparsity. Specifically, at word level, our model improves the effect of word representation by introducing latent topic information into word-level semantic representation. Neural topic model is used to discover the latent topic of the text. A new topic-word attention mechanism is presented to explore the semantics of words from the perspective of topic–word association; At the sequence level, a secondary attention mechanism is used to capture the relationship between local and global sentiment expression. Experiments on the ChnSentiCorp-Htl-ba-10000 and NLPCC-ECGC datasets validate the effectiveness of the BAM model.

INDEX TERMS Attention mechanism, sentiment analysis, text analysis topic model, word embedding.

I. INTRODUCTION

Information dissemination, opinion expression, and other behaviors are increasingly presented in the form of short texts in various social media platforms, emerging news media, e-commerce, and other fields [1]. The construction of short-text sentiment analysis method is beneficial for the timely understanding of customers' emotional state, thereby obtaining customers' opinions and attitudes toward events, information, and products, which are of immense significance for personalized search, public opinion analysis, customer services, and product sales [2], [46].

Short-text data are sparse [1], [3]. On the one hand, the lack of co-occurrence of words affects word embedding methods, which depend on such a co-occurrence [3]–[5]. On the other hand, minimally valid features of short text restrict the ability of a model to extract text features and accurately grasp the relationship between text features and sentiment labels [5].

External knowledge is used to assist sentiment analysis in the majority of the existing methods to cope with the difficulties in sentiment analysis caused by data sparsity [5]–[7]. These methods are limited in the scope of application scope because numerous manual features are required or depend on high-quality external knowledge base in specific fields [5], [9], [10]. Moreover, some methods aim to solve sparsity issue by applying dimensionality reduction, such as topic models. Although some topic models (i.e., labeled topic model) can model the correlation of given label and latent topic directly, many studies see topic models as feature extractor that represent texts as mixture of latent topics [49], [50]. Topic models are used to explore the association between words and latent topics, while latent topic components are used to reduce the dimensionality of the text [11]–[14]. Furthermore, the topic model is used as a bridge to mine the semantic association among different words under the same latent topic and mitigate the impact of data sparsity on the comprehension of word meaning [1]. However, the majority of the related studies have focused on the role of topic information on text classification, without considering the association between topic and sentiment.

The current study was inspired by Zeng *et al.* [1] and presents a short-text sentiment analysis method of the

The associate editor coordinating the review of this article and approving it for publication was Biju Issac.

bi-level attention model on the bases of topic and sequence. However, this method does not introduce external knowledge to assist the comprehension of word meaning. Instead, topic model is combined with attention mechanism to complete the comprehension of word meaning. The main contributions of this study are presented as follows.

- A topic-word attention mechanism (T-WAM) is presented to adapt the characteristics of the neural topic model, and to explore the semantics of words from the perspective of topic – word meaning association, and establish semantic representation at the word level to improve the effects of word-level semantic representation.
- An end-to-end short-text sentiment analysis method based on bi-level attention model (BAM) is presented.
- Experiments on the ChnSentiCorp-Htl-ba-10000 and NLPCC-ECGC datasets validate the effectiveness of the BAM model.

The remainder of this paper is organized as follows. Section 2 reviews short-text sentiment analysis and the topic model-related research. Section 3 describes the BAM model proposed in this study. Section 4 introduces the experimental design in detail and analyzes the experimental result. Section 5 summarizes the work of this research.

II. RELATED WORKS

The short-text sentiment analysis has attracted wide attention [15], [16]. The sparsity issue of short text data makes it difficult to learn good word representation for researches in this field. We attempt to alleviate the impact of sparsity issue to word representation learning by introducing topic information (obtain by neural topic model) into word-level semantic representation. Therefore, this study reflects on two aspects, namely, sparsity of short-text data and topic model.

A. SPARSITY OF SHORT-TEXT DATA

At present, two main methods address data sparsity in short-text sentiment analysis [1], [17]: introduction of external knowledge to assist sentiment analysis and dimensionality reduction [1], [5].

- Introduction of external knowledge to address sparsity

Some studies rely on sentiment knowledge and hard-coded reversal rules [39], which recognize sentiment-indicative words in text by sentiment ontology [2] or sentiment lexicon [18]–[20] and combine evidence with certain rules; Many sentiment analysis methods focus on document-level feature engineering [21]–[23]. Pang pioneer this field by combining document features (e.g., unigram, bigram, POS, adjectives) with machine learning models (e.g., SVM, naïve Bayesian model and max entropy model) [18], [23]. Emoticons, sentiment lexicon, and other sentiment-related manual features are added into feature representation, to improve the performance of the sentiment analysis model [16], [24]. Some studies use external knowledge to improve the quality of word representation [7], [23]. Kumara *et al.* [25] apply Knowledge Graph Embedding on WordNet to get word embedding for each word in the text, improved the sentiment prediction by utilizing background knowledge base to build good word level representation. Fu *et al.* [7] proposed a lexicon-enhanced LSTM model, which use prior sentiment information of words as the supplement information to improve the quality of word representation. Yangsen *et al.* [9], Jiang *et al.* [26] and Wang *et al.* [27] recognized the special role of emoticons in emotional expression, to construct sentiment symbol knowledge base based on existing semantic resources including sentiment lexicons, negative words, emoticons degree adverbs and other sentiment-expressive symbols, and map this sentiment symbols into the vector space. Introduce external knowledge into neural network by unsupervised pretraining is also suggested by some studies, Li proposed a semantically rich hybrid neural network (HNN) which leverages unsupervised teaching models to incorporate semantic domain knowledge into neural network to bootstrap it inference power and interpretability [52].

- Solving sparsity issue with dimensionality reduction

Principal component analysis (PCA), latent Dirichlet allocation (LDA), latent semantic indexing (LSI), and other dimensionality reduction technologies have attracted considerable attention in short-text classification field [5]. Vinodhini and Chandrasekaran [28] studied the effect of feature reduction in sentiment analysis of online reviews with PCA method. Phan *et al.* [3] proposed a method that deals with short and sparse text by making the most of latent topic discovered from large scale data collections by topic models (e.g. LDA, LSI). Rao et al. proposed a topic-level maximum entropy (TME) model, which generate topic-level features by modeling latent topics, labels, and combine it with maximum entropy model, TME shows effectiveness on emotion classification over sparse words [51].

The method that introduces external knowledge to assist sentiment analysis is limited in the scope because of its dependence on high-quality external emotional knowledge base, and need for numerous manual features [5]. The method that aims to solve the sparsity issue by dimensionality reduction has high computational complexity because such a technique involves large matrix decomposition or extracting eigenvalues from large matrices.

B. TOPIC MODELS

Topic models are one of the most popular methods for learning representations of text [14]. The LSA method proposed by Deerwester *et al*. [30] apply singular value decomposition over matrix of terms by documents, to map documents to a vector space of reduced dimensionality [31]. Hofmann *et al.* [31] proposed probabilistic latent semantic analysis (PLSA) on the basis of LSA. The disadvantage of the PLSA method is that the model parameters increase linearly with the size of the training set, which is prone to overfitting. Many shortcomings of PLSA can be mitigated by using Latent Dirichlet Allocation (the LDA) proposed by

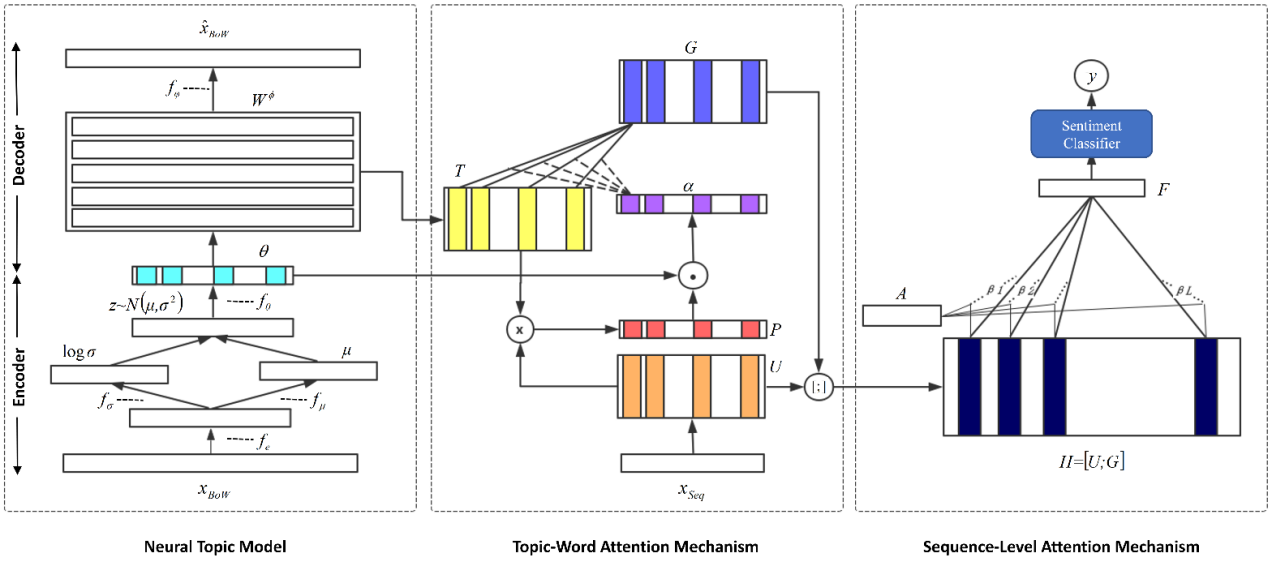

FIGURE 1. Bi-level attention model architecture.

Blei *et al*. [32]. Its parameters did not increase with the size of the training set and exhibited good generalization performance. It has been widely used in the field of natural language processing [3], [32], [33]. Cheng and Yan et al. argue that document-level topic models (i.e., LDA and PLSA) also suffer from sparsity problem because of sparse word co-occurrence in short text documents, and proposed biterm topic model (BTM), which learns topics by directly modeling the generation of word co-occurrence patterns (i.e., biterms) in the corpus [17], [47]. In recent years labeled topic models (i.e. ETM [48], UAM [49]) have attracted much attention. Many of those labeled topic models extend LDA (or BTM) model by introducing intermediate layers of label modeling, to model the correlation between given label (i.e., sentiment label) and latent topic [49], [50]. Classical topic models require substantial computing resources because of the need to decompose large matrices or work out eigenvalues. Moreover, combining these models with neural networks for end-to-end training is difficult.

The neural topic models (NTM), proposed by Miao *et al.* [13] and Srivastava and Sutton [14], are topic models parameterized with deep neural networks. Neural topic models use excellent ability of deep neural networks as function approximators to learn the complicated semantic association between latent topic and word meaning, and achieved good result. Zeng *et al*. [1] proposed that the combination of neutral topic model with memory network can alleviate the sparsity issue of short-text data. However, these researchers have mainly focused on the role of topic models in text categorization without considering the relationship between latent topic and sentiment.

III. BI-LEVEL ATTENTION MODEL

The structure of the bi-level attention model proposed in this study is shown in Fig.1, which is composed of four modules: (A) neural topic model (NTM), (B) topic-word attention mechanism (T-WAM), (C) attention mechanism at the sequence level, and (D) sentiment classifier.

At word level, our model combines NTM and T-WAM to improve the effect of word level representation, thereby alleviating the difficulty of word-level semantic understanding caused by the sparsity of short text. Neutral topic model extracts topic information on the basis of text diction. The diction of short text is expressed as $x_{BoW} \in R^{|V|}$ in the form of bag-of-words, where $|V|$ is the vocabulary size. The neutral topic model analyses latent topics, and generates embeddings for latent topics, which form a topic vector matrix. Thereafter, the topic information is transferred to the T-WAM module.

The topic information is used to extend word-level semantic representation by T-WAM. Text is initially expressed as $x_{Seq} \in R^{L}$ in the form of word sequence, where L refers to the text sequence length. Thereafter x_{Seq} is represented as a word-embedding matrix U through word embedding. T-WAM produce word topic embedding G on the basis of topic vector matrix T and relationship α between topic information and word embedding, thereby construction word representation from the perspective of topic. We concatenate word embedding U and word topic embedding G to obtain word-level semantic representation H.

The sequence-level attention mechanism is applied to perform semantic fusion at the sequence level based on the semantic representation H at the word level, thereby enabling the construction of the semantic representation of the text. Lastly, the sentiment classification model is used to infer sentiment label y of the text based on the semantic representation of the text.

A. NEURAL TOPIC MODEL

The topic information of the text is extracted on the basis of text diction x_{BoW}, while short text is represented as a mixture of latent topics by the neutral topic model [1]. In particular, $\theta \in R^{K}$ is used to represent the components of latent

topics, with each dimension corresponding to a latent topic. Meanwhile, K refers to the number of latent topics, and each latent topic corresponds to a probability distribution over the vocabulary [1], [17].

The topic model of this study is established based on studies on neutral topic models, which are proposed by Zeng *et al.* [1], Miao *et al.* [13], and Srivastava and Sutton [14]. The NTM is based on VAE [42] model, which is divided into two parts, namely, encoder and decoder. In the encoder network, the input variables are mapped to intermediate variables $z \sim \mathrm{N}(\mu, \sigma^2)$. Unlike the VAE model, NTM model further maps the intermediate variable z in the encoder to obtain the topic component θ:

$$\theta = soft\max(f_\theta(z)) \tag{1}$$

In the decoder network, the topic component θ is used as input to obtain $\hat{x}_{BoW}$:

$$w_n \sim sigmoid(f_\phi(\theta)) \tag{2}$$

where $f_\theta(\bullet)$ refers to a multi-layer perceptron with three full-connected layers, $f_\phi(\bullet)$ refers to a perceptron, and w_n corresponds to the n th dimension of the topic model output, which represents the probability of the occurrence of the n th word in the vocabulary, given topic component θ. In the NTM, $f_\phi(\bullet)$ map topic component θ into a probability distribution $\hat{x}_{BoW}$ over the vocabulary. $W^\phi \in R^{K\times|V|}$ is the kernel weight matrix of $f_\theta(\bullet)$, where K is the number of latent topics, and $|V|$ is the vocabulary size. The rows of W^ϕ correspond to the topics, while the columns correspond to the vocabulary.

In addition to the latent topic component θ, the NTM produce embeddings for latent topics, which form a topic vector matrix $T \in R^{K\times E}$. T is obtained through the matrix mapping:

$$T = f_T(W^\phi) \tag{3}$$

where $f_T(\bullet)$ is a perceptron, whose activation function is *relu*. Each row vector of T, $t^k \in R^E$, corresponds to a latent topic.

B. TOPIC-WORD ATTENTION MECHENISM

In this study, we propose a topic-word attention mechanism (T-WAM) to introduce topic information into word representation, thereby alleviate the impact of sparsity issue at word level. The topic information (including topic component θ and topic vector matrix T), provided by the neutral topic model, and word semantic information (represent as word embedding U) are used to compute the relationship α between topics and words. Thereafter, the topic vector matrix T is reformed according to α, to obtain the word topic embedding G. Word topic embeddings $G \in R^{L\times E}$ is a word-level representation of text from the perspective of topics. Thereafter we concatenate word embedding U and word topic embedding G to obtain word-level semantic representation H. The relationship between latent topic and word is calculated as follows:

$$U' = f_U(U) \tag{4}$$

$$P_{l,k} = u'_l t_k^T \tag{5}$$

$$\alpha_{l,k} = \frac{\exp(P_{l,k}\theta_k)}{\sum_{j=1}^{K}\exp(P_{l,j}\theta_j)} \tag{6}$$

where $U' \in R^{L\times E}$ is the result of the dimensionality reduction of the single layer perceptron $f_U(\bullet)$ against matrix U; $P \in R^{L\times K}$ is an intermediate variable matrix for calculating α. Intuitively, $P_{l,k}$ measures how well word embedding u'_l can match topic vector t_k, thereby in a certain degree reflect the correlation between word and topics, however, we argue that more topic information could be add as topic component θ_k involved; $\alpha_{l,k}$ refers to the relationship between the sequence word at position l and k th latent topics; u'_l refers to the l st row vector of matrix U; And t_k refers to the k th row vector of the topic vector matrix T. Thereafter T-WAM construct word topic embedding vector g_l with topic-word correspondence $\alpha_{l,k}$ and topic embedding vector t_k.

$$g_l = \sum_{k=1}^{K}\alpha_{l,k}t_k \tag{7}$$

Concatenation of Word topic embedding vector g_l and word embedding vector u_l is the word semantic representation h_l:

$$h_l = [u_l; g_l] \tag{8}$$

where [;] refers to the concatenation function, and u_l is the row vector of word embedding matrix U. Lastly, h_l is spliced into word-level semantic representation H.

C. SEQUENCE-LEVEL ATTENTION MECHANISM

Sequence-level attention mechanism based on the word-level semantic representation H, further integrates the text semantics in the sequence dimension, thereby construct text sentiment semantic representation F. According to the contribution β_l of the input state h_l to the sentiment expression, the sequence-level attention mechanism constructs the sentiment semantic representation of text as follows:

$$F = \sum_{l=1}^{L}\beta_l h_l \tag{9}$$

Similar to by Zhang *et al.* [9] and Zhou *et al.* [37], the current study designs a sentiment prior vector A. Sequence-level attention mechanism aims to compute the contribution β_l of the input state h_l to sentiment expression based on the matching between the input state h_l and sentiment prior vector A. Sentiment prior vector A, which is a model parameter, is obtained through training. The sentiment prior vector A can be regarded as a vectorized representation of sentiment recognition knowledge. In particular, α_l is calculated as follows:

$$\beta_l = \frac{\exp(e_l A)}{\sum_{j=1}^{L}\exp(e_j A)} \tag{10}$$

$$e_l = \tanh(W_e h_l + b_e) \tag{11}$$

where W_e and b_e are model parameters, and are obtained through training respectively. Sequence-level attention mechanism integrates the semantic information into a fixed-length sentiment semantic vector F.

D. SENTIMENT CLASSIFIER

Sentiment classifier infer the sentiment labels y on the basis of the sentiment semantic vector F. Sentiment Classifier is a multi-layer perceptron composed of two fully connected layers. Thereafter *soft* max (or *sigmoid*) function is applied to map sentiment analysis result into the form of probability distribution.

E. LOSS FUNCTION

The global loss function of BAM model is composed of the loss function of the NTM and the loss function of sentiment classifier.

$$L = L_{NTM} + L_{CLF} \tag{12}$$

where L_{NTM} refers to the loss function of the NTM. In the training process, the loss function used in the NTM is as follows:

$$L_{NTM} = D_{KL}(q(z)||p(z|x)) - E_{q(z)}[p(x|z)] \tag{13}$$

$D_{KL}(q(z)||p(z|x))$ refers to the KL divergence between the posterior distribution and prior distribution of the hidden variable z and $E_{q(z)}[p(x|z)]$ is used to measure how well NTM can match the training data [1], [14]. We leave out derivation details and refer readers to [1], [14], due to space limitation. The loss function of sentiment classifier L_{CLF} is a cross-entropy loss function:

$$L_{CLF} = \frac{1}{N}\sum_{i=1}^{N} y_i \log p(y_i|x) \tag{14}$$

where N refers to the number of train examples, $p(y|x)$ refers to the output of the sentiment classifier, and y_i refers to true sentiment label for the i th sample.

IV. EXPERIMENTS

A. DATASET

To verify the proposed model[1] (sourse code is presented below), the current study conducts experiments on the ChnSentiCorp-Htl-ba-10000[2] and NLPCC-ECGC[3] datasets. Detailed information can be seen from TABLE 1. Both of the datasets are Chinese datasets.

- **ChnSentiCorp-Htl-ba-10000** created by [43], includes 10000 piece of hotel reviews, labeled as positive and negative. It is an unbalanced corpus in which 7000 positive samples exist.
- **NLPCC-ECGC** dataset [44] come from online comments on Weibo, which includes 1 million dialogues in the post-response form. Each post and response dialogue had its corresponding label.

TABLE 1. Dataset partition.

Corpus	Classes	Instance	Avg length	Total words	Train/test
ChnSentiCorp-Htl-ba-10000	2	10000	80	12645	9000/1000
NLPCC-ECGC	6	1100000	22	96580	1095000/5000

B. EXPERIMENTAL SETUP

To provide a point of reference for the bi-level attention model (BAM) result, we used SVM + BoW [39], MCNN [34], RCNN [35], VDCNN [36], Bi-GRU, AttBiLSTM [37], TMN [1], BERT [45], and TAM-CNN as baselines. All of the above are effective methods for sentiment classification. SVM + BoW is a simple and effective text classification model, it uses SVM to classify texts upon Bag of Word features. MCNN is a classic convolutional neural network for text classification tasks, which is also baseline model for many studies. RCNN, VDCNN, AttBiLSTM, TMN and BERT are advanced neural network models for text classification and sentiment analysis. The current study is inspired by the work of Zeng *et al.* [1] (namely Topic Memory Network, TMN), but our model is different from TMN from in the following way:

- At word level, TMN use memory network to build word level representation, while in this article we present T-WAM to improve word representation.
- At sequence level, TMN [1] use CNN to encode short texts, while we argue that attention mechanism is a better choice to capture sequence level local features for short text sentiment analysis tasks. Experiment results shows that our model outperforms TMN on ChnSentiCorp-Htl-ba-10000 and NLPCC-ECGC datasets.

A total of 1000 texts are randomly extracted from ChnSentiCorp-Htl-ba-10000 to construct the testing set, and the rest of the texts are used to construct the training set. A total of 5000 texts are randomly extracted from the NLPCC-ECGC dataset to construct the testing set, while the remainder is used to construct the training set. First, a series of pre-processing of text data in dataset is carried out (apply the source code), we tokenize each text with jieba and remove some special tokens. Second, a Word2Vec training corpus is constructed on the basis of two datasets, and word vector is trained with the help of the gensim tool, and the word vector is trained, to obtain a word embedding matrix that contains 96580 words, whose word vector dimensionality is 50, word vectors for GloVe and Fasttext are obtained from the same corpus. We did not use stop words in the experiments. For BERT model, we apply fine-tuning over pre-trained model 'BERT-Base, Chinese'.[4]

[1] https://github.com/Alex-sanda/bam/
[2] http://www.searchforum.org.cn/tansongbo/corpus/
[3] http://coai.cs.tsinghua.edu.cn/hml/challenge2017/
[4] https://github.com/google-research/bert

TABLE 2. Experimental results on ChnSentiCorp-Htl-ba-10000 dataset.

Models	ChnSentiCorp-Htl-ba-10000	
	Acc	F1-score
SVM + BoW [39]	0.711	0.775
MCNN [34]	0.845	0.862
RCNN [35]	0.861	0.844
VDCNN [36]	0.824	0.821
Bi-GRU	0.822	0.82
AttBiLSTM [37]	0.853	0.86
TMN [1]	0.884	0.883
BERT [45]	0.894	-
T-WAM + MCNN	0.869	0.866
BAM （our model）	**0.900**	**0.892**

TABLE 3. Experimental results on NLPCC-ECGC dataset.

Models	NLPCC-ECGC	
	Acc	F1-score
SVM + BoW [39]	0.506	0.459
MCNN\ [34]	0.901	0.924
RCNN [35]	0.925	0.901
VDCNN [36]	0.935	0.953
Bi-GRU	0.886	0.901
AttBiLSTM [37]	0.920	0.952
TMN [1]	0.931	0.931
BERT [45]	**0.948**	-
T-WAM + MCNN	0.932	0.945
BAM (our model)	0.945	0.956

C. RESULTS AND DISSCUSSION

- ChnSeniCorp-Htl-ba-10000 dataset

On ChnSentiCorp-Htl-ba-10000 date set, the experimental results of the BAM and baseline models are shown in TABLE 2.

1) **Attention mechanism plays an important role in short-text sentiment analysis.** We compare models that (not) introduce attention mechanism, the experimental result shows that the accuracy of the AttBiLSTM model that introduces such a mechanism at the sequence level reaches 0.853, which is higher than those of Bi-GRU with similar structure (accuracy 0.822) and advanced VDCNN model (accuracy 0.824) and is comparable to the performance of RCNN (accuracy 0.861), and BERT model (accuracy 0.894), which use multiple self-attention layers, outperforms all above CNN models. Moreover, the accuracy of the BAM model using bi-level attention mechanism reaches 0.900, which is the highest among all the models. Evidently, attention mechanism plays an important role in short-text sentiment analysis.
2) **T-WAM can improve the effect of word-level semantic representation.** We compare the models (not) introducing T-WAM to construct a word-level semantic representation. Experimental result shows that the accuracy of TAM-CNN and BAM is high (i.e., up to 0.869 and 0.900 respectively) and higher than those of MCNN and AttBiLSTM (i.e., 0.845 and 0.853, respectively). This result indicates that the extension of word-level semantic representation by means of T-WAM can improve word-level representation, and thereby alleviate the impact of sparsity issue at word-level representation.
3) **The bi-level attention model is effective for short-text sentiment analysis**. A comparison between the BAM model and other baseline models indicates that the accuracy of BAM is 0.900, while that of F1-score is 0.892. The two indicators are the highest, which shows the validity of the proposed model in short-text sentiment analysis.

- Evaluation of the NLPCC-ECGC dataset

The experimental results on NLPCC-ECGC dataset are shown in TABLE 3. BAM (accuracy 0.945) achieved comparable result with BERT (accuracy 0.948) and outperforms all other baselines, which indicates that our model is effective in sentiment analysis tasks. The other observations in TABLE 3. is consistent with that in TABLE 2, thereby further supports the previous discussion. The following observations are further obtained based on the experimental results in TABLE 2 and TABLE 3.

Topic-word attention mechanism (T-WAM) is more effective in small datasets. The comparison of the differences in the performance of models, that (not) introducing T-WAM on ChnSentiCorp-Htl-ba-10000 dataset (small scale) and NLPCC-ECGC dataset (large scale) shows that the performance of the model is improved after introducing T-WAM. However, the improvement effect on small-scale dataset is more evident. This phenomenon can be explained from the following perspectives. Recall that, at word level, sparsity issue affects word-level representations by cutting down the word co-occurrence frequency. On small scale dataset, words appear less frequently, while the co-occurrence of vocabulary is even scarce because the corpus on small-scale dataset is minimal, thereby resulting in inadequate understanding on word-level semantics. Therefore, the introduction of topic information into word-level representation results in an evident strengthening effect on word representation.

D. HYPER PARAMETERS

To further explore the impact of the hyper parameters on the proposed model, experiments are carried out against three key hyper parameters, namely, sentiment priori vector A, numbers of latent topics K, and word-level vector representation method.

- Evaluation on ChnSentiCorp-Htl-ba-10000 dataset

The experimental result of adjusting the sentiment priori vector A on ChnSentiCorp-Htl-ba-10000 dataset is shown in Fig 2. When the dimension of A ranges in the interval of

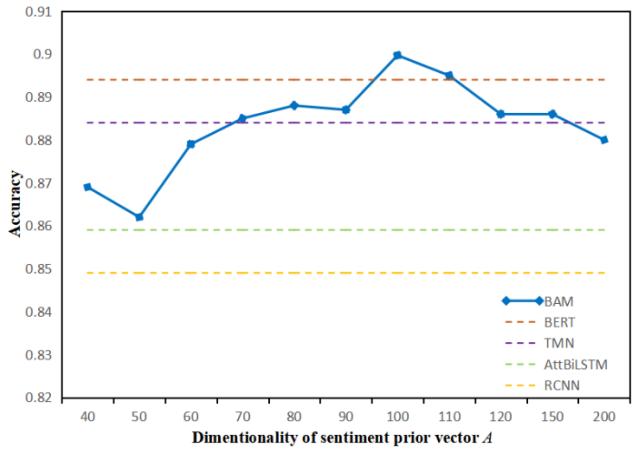

FIGURE 2. The impact of dimensionality sentiment prior vector *A* on accuracy.

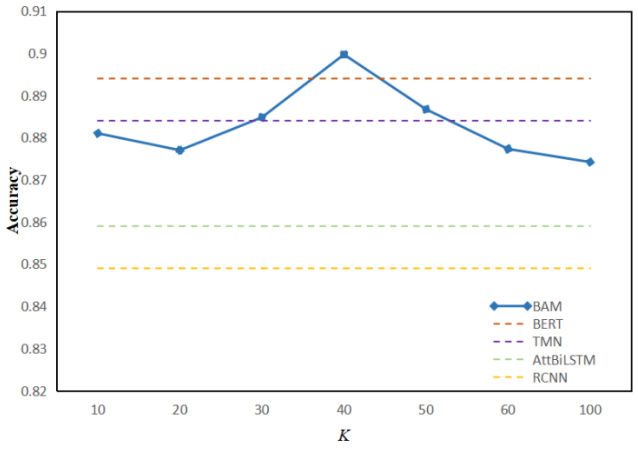

FIGURE 3. The impact of topic number *K* on accuracy.

[100, 110], the proposed model outperforms BERT, TMN, AttBiLSTM and RCNN models. When adjusting other hyper parameters, we use Fasttext [38] to initialize word representation before training as default option.

The impact of topic number K is shown in Fig 3. Moreover, the performance of the model stably exceeds those of the other contrast models. If $K = 40$, then the optimal result is obtained.

To explore the impact of word representation method, this study used several different word representation methods to obtain word-level representations, thereafter word-level representations are sent to MCNN [34] model and sequence-level attention model respectively. TABLE 4. shows the experimental results:

From TABLE 4, we obtain the following observations:

1) **T-WAM can be used as supplement to classical word representation methods.** The contrast between (not) introducing the T-WAM mechanism to extend word representation shows that, introducing T-WAM into Fasttext [38], GloVe [39], and word2vec [40], improves the accuracy from 0.879, 0.848, and 0.854 to 0.900, 0.869, and 0.883., respectively. However, T-WAM (only), which uses word topic embedding as the word-level semantic representation, does not exhibit an ideal effect (i.e., 0.700 and 0.699). Therefore, T-WAM, can be used as the supplement of classical word representation methods (e.g. Fasttext, GloVe, skip-gram Word2Vec) to improve word-level semantic representations.

TABLE 4. Impact of word representation method on accuracy, experimental results on ChnSentiCorp-Htl-ba-10000 dataset.

word representation method	ChnSentiCorp-Htl-ba-10000	
	MCNN [34]	Seq-Attention
T-WAM (only)	0.700	0.699
Fasttext [38]	0.840	0.879
GloVe [39]	0.836	0.849
word2vec [40]	0.843	0.854
T-WAM + Fasttext	**0.869**	**0.900**
T-WAM + GloVe	0.854	0.869
T-WAM + word2vec	0.845	0.883

2) **Attention mechanism is better than CNN in the semantic integration of short text at the sequence level.** The contrast between Seq-Attention models and CNN models shows that when the word-level semantic representation methods are Fasttext + T-WAM, GloVe + T-WAM, and word2vec + T-WAM, the accuracy of the Seq-Attention model are 0.900, 0.869, and 0.883, which are higher than those of the corresponding CNN models, namely, 0.869, 0.854, and 0.845. Evidently, Seq-Attention can accurately capture the association between local and global sentiment expression at sentence level.

- Evaluation of the NLPCC-ECGC dataset

The experimental result of the hyper parameter A on NLPCC-ECGC dataset is shown in Fig.4. Fig 2 and Fig 4 show that, 100 to 120 is the reasonable interval of the hyper parameter A.

The regulation result of the hyper parameter K on the NLPCC-ECGC dataset is shown in Fig 5. Figs. 3 and Fig 5 indicate that the performance of BAM exceeds those of the RCNN [35], VDCNN [36], AttBiLSTM [37], and TMN [1]. Moreover, BAM achieve at least comparable result with the advanced BERT model on both datasets. Hence, BAM shows an excellent performance.

We explore the impact of word-level semantic representation method on NLPCC-ECGC dataset. TABLE 5. shows the experimental result. The observations from TABLE 5 is consistent with that from TABLE 4, thereby further prove the previous discussion. The comparison of TABLE 5 with TABLE 4 shows that when the scale of training set corpus is large, the introduction of the T-WAM mechanism has minimal influence on the model performance.

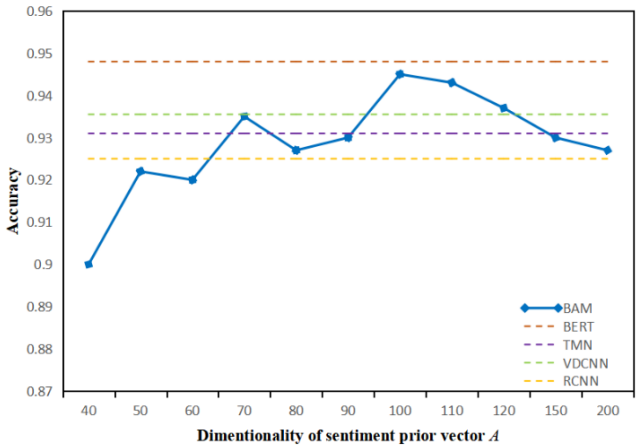

FIGURE 4. The impact of dimensionality of sentiment prior vector *A* on accuracy.

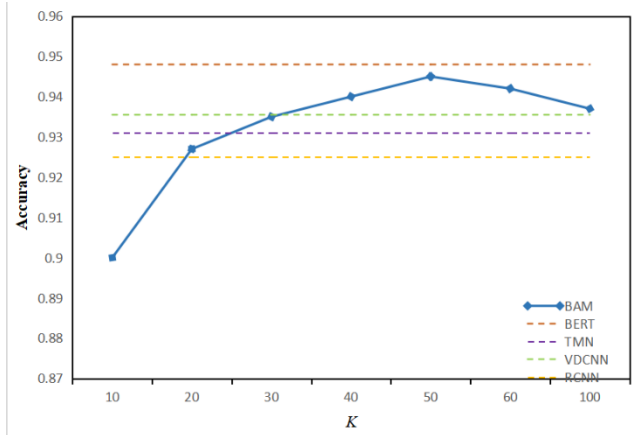

FIGURE 5. The impact of topic number *K* on accuracy.

E. CASE STUDY

We present a case study to further understand how T-WAM and sequence-level attention mechanism assist the model in semantic sentiment analysis. We chose hotel review instance "*前台服务不错, 询问旅游事宜都详细解答* (*The reception provides a good service, and answers to inquire about travel in detailed*)" as the case. Fig 6 shows a visualization of topic component analysis result, T-WAM, and sequence-level attention for the given instance.

Word "*前台(reception)*" is a low-frequency word in the corpus, which means its word representation could easily be influenced by the sparsity issue. At word level, T-WAM found that "*前台(reception)*" is associated with several latent topics, whose top key words are "*服务(service), 服务员(waitress),*" and "*入住(check in).*" These key words can represent the real meaning of "*前台(reception)*" from a certain perspective. The T-WAM integrates the topic vector of those latent topic according to association between topic and word, to obtain word topic embedding for "*前台(reception)*". T-WAM combines word embedding and word topic embedding to improve word-level representation of "*前台(reception)*". The semantic representation of other words is the same as that of "*reception*". At the sentence level, the attention mechanism accurately grasps two positive local feature of the given instance, namely "*不错(good)* and *详细(detailed)*," which indicates positive sentiment. Model induce correct label for the given instance with the help of T-WAM and sequence-level attention.

As is shown above, T-WAM introduces topic information into word representation, to improve word-level semantic understanding and alleviate the sparsity issue. Moreover, sequence-level attention captures local feature for sentiment expression, thereby infer sentiment label. The performance of the proposed model in the case shows that BAM is effective in short text sentiment analysis.

TABLE 5. Impact of word representation method on accuracy, experimental results on NLPCC-ECGC dataset.

word representation method	NLPCC-ECGC	
	MCNN [34]	Seq-Attention
T-WAM (only)	0.504	0.510
Fasttext [38]	0.804	0.916
GloVe [39]	0.905	0.906
word2vec [40]	0.909	0.902
T-WAM + Fasttext	0.905	0.914
T-WAM + GloVe	0.922	0.935
T-WAM + word2vec	**0.937**	**0.945**

TABLE 6. Top-4 words of topics indicated by T-WAM.

1st Topic	2nd Topic
服务 (service)	服务员 (waitress)
入住 (check in)	去 (go)
很 (very)	价格 (price)
所以 (so)	不错 (good)

F. ERROR ANALYSIS

This section performs error analysis on the basis of the experimental results involving the ChnSentiCorp-Htl-ba-10000 dataset. We observe that one major error type come from the failure to understand the priority of local features. When positive and negative feature appear in the same text, the model can be confused. Taking "very good position, complete chaotic service, and depressed hotel." as an example. This example includes both positive expression ("Very good"), and negative expression ("chaotic" and "depressed"). Visualization of sequence-level attention for the given example is shown at Fig 7. Fig 7 shows that, without understanding the priority between local features, our model

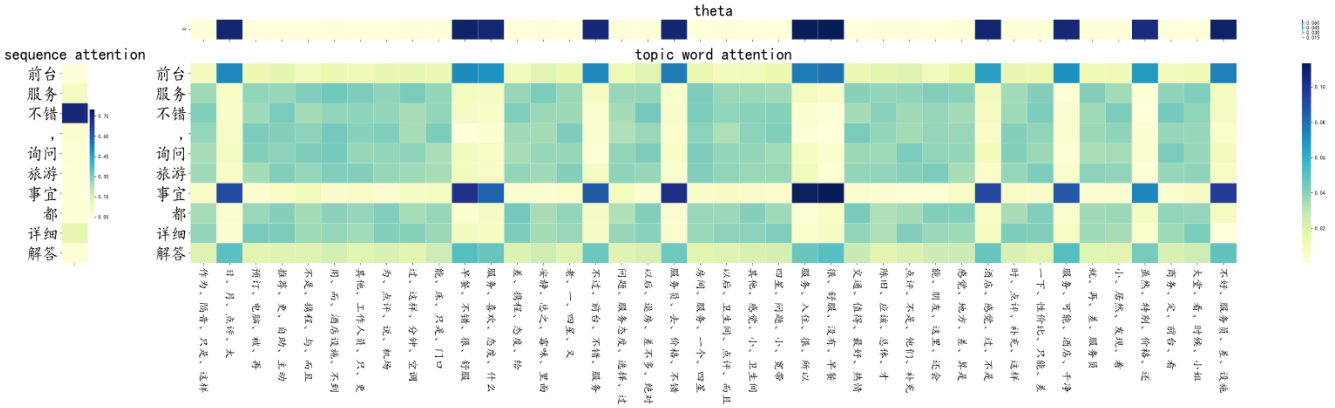

FIGURE 6. Visualization of topic components, T-WAM, sequence-level attention for the given case, sequence attention block is the visualization of vector β. The theta block is the visualization of topic component vector θ, and topic word attention block is the visualization of topic-word attention matrix α.

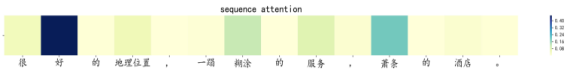

FIGURE 7. Visualization of sequence-level attention for instance in error analysis.

pay too much attention on the positive expression (''Very good'') and thus predicted wrong sentiment label. Such errors can be reduced by fine-grained sentiment analysis, which explore the relationship between sentiment expression and the targeted objects. Priority between local features are associated with the priority of targeted objects. In the succeeding research, we will explore fine-grained sentiment analysis methods, which investigates priority between features.

V. CONCLUSION

This paper proposes a short text sentiment analysis method based on bi-level attention model, which does not rely on manual features or external knowledge. Our model combines neural topic model and attention mechanism to mitigate the impact of data sparsity. At word level, our model improves the effect of word representation by introducing latent topic information into word-level semantic representation. A new topic-word attention mechanism is proposed to explore the semantics of words from the perspective of topic. At sequence level, attention mechanism is used to capture the relationship between local and global sentiment expression. Experiments on the ChnSentiCorp-Htl-ba-10000 and NLPCC-ECGC datasets validate the effectiveness of the BAM model.

Error analysis shows that the model remains prone to confusion when positive and negative features appear in the same text. This type of error came from the failure to understand the priority of local features. Therefore, the next step of our research, we will involve building a fine-grained sentiment analysis model, which explore the relationship between sentiment expression and targeted objects, thereby understand the priority between local features.

REFERENCES

[1] J. Zeng, J. Li, Y. Song, C. Gao, M. R. Lyu, and I. King, "Topic memory networks for short text classification," in *Proc. Conf. EMNLP*, Brussels, Belgium, 2018, pp. 3120–3131.

[2] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017. doi: 10.1016/j.imavis.2017.08.003.

[3] 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*, Beijing, China, 2008, pp. 91–100.

[4] H.-Y. Lu, L.-Y. Xie, N. Kang, C.-J. Wang, and J.-Y. Xie, "Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery," in *Proc. AAAI*, San Francisco, CA, USA, 2017, pp. 1192–1198.

[5] Z. Yang, Y.-X. Lai, L.-J. Duan, and Y.-J. Li, "Short text sentiment classification based on context reconstruction," *Acta Automatica Sinica*, vol. 38, no. 1, pp. 55–67, 2012.

[6] F. H. Khan, U. Qamar, and S. Bashir, "A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet," *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 851–872, 2017. doi: 10.1007/s10115-016-0993-1.

[7] X. Fu, J. Yang, J. Li, M. Fang, and H. Wang, "Lexicon-enhanced LSTM with attention for general sentiment analysis," *IEEE Access*, vol. 6, pp. 71884–71891, 2018.

[8] J. Wang, Z. Wang, D. Zhang, and J. Yan, "Combining knowledge with deep convolutional neural networks for short text classification," in *Proc. IJCAI*, Melbourne, Vic, Australia, 2017, pp. 2915–2921.

[9] Z. Yangsen, Z. Jia, H. Gaijuan, and J. Yuru, "Microblog sentiment analysis method based on a double attention model," *J. Tsinghua Univ. Sci. Technol.*, vol. 58, no. 2, pp. 122–130, Feb. 2018. doi: 10.16511/j.cnki.qhdxxb.2018.22.015.

[10] 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. doi: 10.1109/MIS.2018.033001419.

[11] L. Peng, L. Chen, X. Wu, H. Guo, and G. Chen, "Hierarchical complex activity representation and recognition using topic model and classifier level fusion," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 6, pp. 1369–1379, Jun. 2017. doi: 10.1109/TBME.2016.2604856.

[12] Q. Chen, L. Yao, and J. Yang, "Short text classification based on LDA topic model," in *Proc. ICALIP*, Shanghai, China, Jul. 2016, pp. 749–753.

[13] Y. Miao, E. Grefenstette, and P. Blunsom, "Discovering discrete latent topics with neural variational inference," in *Proc. ICML*, Sydney, NSW, Australia, 2017, pp. 2410–2419.

[14] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," 2017, *arXiv:1703.01488*. [Online]. Available: https://arxiv.org/abs/1703.01488

[15] D. Jiang, X. Luo, J. Xuan, and Z. Xu, "Sentiment computing for the news event based on the social media big data," *IEEE Access*, vol. 5, pp. 2373–2382, 2016. doi: 10.1109/ACCESS.2016.2607218.

[16] L. Rotim and J. Šnajder, "Comparison of short-text sentiment analysis methods for croatian," in *Proc. BSNLP*, Valencia, Spain, 2017, pp. 69–75.
[17] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 1445–1456.
[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2002, pp. 79–86.
[19] Z. Lei, Y. Yang, and M. Yang, "Sentiment lexicon enhanced attention-based LSTM for sentiment classification," in *Proc. AAAI*, New Orleans, LA, USA, 2018, pp. 8105–8106.
[20] L. Wang and R. Xia, "Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision," in *Proc. EMNLP*, Copenhagen, Denmark, 2017, pp. 502–510.
[21] K. Yang, Y. Cai, D. Huang, J. Li, Z. Zhou, and X. Lei, "An effective hybrid model for opinion mining and sentiment analysis," in *Proc. IEEE BigComp*, Jeju, South Korea, Feb. 2017, pp. 465–466.
[22] L. Zheng, H. Wang, and S. Gao, "Sentimental feature selection for sentiment analysis of chinese online reviews," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 1, pp. 75–84, 2018.
[23] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings," in *Proc. AAAI*, Phoenix, AZ, USA, 2016, pp. 3038–3044.
[24] D.-T. Vo and Y. Zhang, "Target-dependent twitter sentiment classification with rich automatic features," in *Proc. IJCAI*, Buenos Aires, Argentina, 2015, pp. 1347–1353.
[25] A. Kumar, D. Kawahara, and S. Kurohashi, "Knowledge-enriched two-layered attention network for sentiment analysis," in *Proc. NAACL*, New Orleans, LA, USA, 2018, pp. 253–258.
[26] F. Jiang, Y.-Q. Liu, H.-B. Luan, J.-S. Sun, X. Zhu, M. Zhang, and S.-P. Ma, "Microblog sentiment analysis with emoticon space model," *J. Comput. Sci. Technol.*, vol. 30, no. 5, pp. 1120–1129, Sep. 2015. doi: 10.1007/s11390-015-1587-1.
[27] J. H. Wang, T. W. Liu, X. Luo, and L. Wang, "An LSTM approach to short text sentiment classification with word embeddings," in *Proc. ROCLING*, Hsinchu, Taipei, 2018, pp. 214–223.
[28] G. Vinodhini and R. M. Chandrasekaran, "Effect of feature reduction in sentiment analysis of online reviews," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 6, pp. 2165–2172, 2013.
[29] F. Colace, M. De Santo, L. Greco, V. Moscato, and A. Picariello, "Probabilistic approaches for sentiment analysis: Latent Dirichlet allocation for ontology building and sentiment extraction," in *Sentiment Analysis and Ontology Engineering*. Cham, Switzerland: Springer, 2016, pp. 75–91.
[30] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[31] T. Hofmann, "Probabilistic LSA," in *Proc. ACM SIGIR Forum*, New York, NY, USA, 2017, pp. 211–218.
[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[33] V. K. R. Sridhar, "Unsupervised topic modeling for short texts using distributed representations of words," in *Proc. EMNLP*, Copenhagen, Denmark, 2017, pp. 502–510.
[34] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
[35] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. AAAI*, Austin, TX, USA, 2015, pp. 2167–2173.
[36] A. Conneau, H. Schwenk, Y. LeCun, and L. Barrault, "Very deep convolutional networks for text classification," in *Proc. EACL*, 2017, pp. 1107–1116. [Online]. Available: https://arxiv.org/abs/1606.01781
[37] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. ACL*, Berlin, Germany, 2016, pp. 207–212.
[38] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.
[39] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, Doha, Qatar, Oct. 2014, pp. 1532–1543.
[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[41] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev. Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
[42] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2014, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114
[43] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Syst. With Appl.*, vol. 34, no. 4, pp. 2622–2629, 2008.
[44] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. AAAI*, New Orleans, LA, USA, 2018, pp. 730–738.
[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805
[46] H. Xie, X. Li, T. Wang, R. Y. K. Lau, T.-L. Wong, L. Chen, F. L. Wang, and Q. Li, "Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 61–72, 2016.
[47] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic modeling over short texts," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2928–2941, Dec. 2014.
[48] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu, "Joint emotion-topic modeling for social affective text mining," in *Proc. 9th IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 699–704.
[49] W. Liang, H. Xie, Y. Rao, R. Y. K. Lau, and F. L. Wang, "Universal affective model for Readers' emotion classification over short texts," *Expert Syst. With Appl.*, vol. 114, pp. 322–333, Dec. 2018.
[50] 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.
[51] 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. Manage.*, vol. 53, no. 8, pp. 978–986, 2016.
[52] 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.

WEI LIU received the Ph.D. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2003. He is currently a Professor with the Automation School, Beijing University of Posts and Telecommunications, Beijing. His research interests include human factors, human cognitive process, and user's behavior analysis and prediction.

GUOXI CAO received the master's degree from the Automation School, Beijing University of Posts and Telecommunications, in 2019. His main research interests include natural language processing and pattern recognition.

JIANQIN YIN received the Ph.D. degree from Shandong University, Jinan, China, in 2013. She is currently a Professor with the Automation School, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include service robot, pattern recognition, machine learning, and image processing.