Knowledge-Based Systems 147 (2018) 43–54

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Weakly supervised topic sentiment joint model with word embeddings

Xianghua Fu, Xudong Sun, Haiying Wu, Laizhong Cui*, Joshua Zhexue Huang

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, PR China

ARTICLE INFO

Article history:
Received 15 June 2017
Revised 29 January 2018
Accepted 6 February 2018
Available online 9 February 2018

Keywords:
Sentiment analysis
Topic model
Topic sentiment joint model
Word embeddings

ABSTRACT

Topic sentiment joint model aims to deal with the problem about the mixture of topics and sentiment simultaneously from online reviews. Most of existing topic sentiment modeling algorithms are mainly based on the state-of-art latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA), which infer sentiment and topic distributions from the co-occurrence of words. These methods have been proposed and successfully used for topic and sentiment analysis. However, when the training corpus is small or when the documents are short, the textual features become sparse, so that the results of the sentiment and topic distributions might be not very satisfied. In this paper, we propose a novel topic sentiment joint model called weakly supervised topic sentiment joint model with word embeddings (WS-TSWE), which incorporates word embeddings and HowNet lexicon simultaneously to improve the topic identification and sentiment recognition. The main contributions of WS-TSWE include the following two aspects. (1) Existing models generate the words only from the sentiment-topic-to-word Dirichlet multinomial component, but the WS-TSWE model replaces it with a mixture of two components, a Dirichlet multinomial component and a word embeddings component. Since the word embeddings are trained on a very large corpora and can be used to extend the semantic information of the words, they can provide a certain solution for the problem of the textual sparse. (2) Most of previous models incorporate sentiment knowledge in the β priors. And the priors are usually set from a dictionary and completely rely on previous domain knowledge to identify positive and negative words. In contrast, the WS-TSWE model calculates the sentiment orientation of each word with the HowNet lexicon and automatically infers sentiment-based β priors for sentiment analysis and opinion mining. Furthermore, we implement WS-TSWE with Gibbs sampling algorithms. The experimental results on Chinese and English data sets show that WS-TSWE achieved significant performance in the task of detecting sentiment and topics simultaneously.

1. Introduction

With the rapid development of e-commerce and social media, the public can conveniently express their opinion and comment on everything with short texts in various types, such as microblogs, instant messages, and online products reviews. For instance, the existing products' reviews can help new users to decide whether or not to buy the products, and also can help the producers to mine business opportunities through analyzing users' valuable feedbacks. Therefore, it is extremely urgent and valuable to efficiently detect what users are interested in. After reading these reviews, the key points that users usually want to obtain are: (1) what is the overall sentiment orientation of a review, e.g., good/bad, like/dislike (i.e. document level sentiment identification); (2) what are the latent topics of the product talked about in the reviews (i.e. document level topic detection). For example, the reviews on a computer often contain detail comments on its different topics such as "battery life", "heat dispersion", etc; (3) what are the associations among topics and sentiments, including the topics associated with a specific sentiment and the sentiments associated with a specific topic (i.e. topic-sentiment identification). For instance, what features of a computer have made it to obtain negative sentiments? And what opinions are expressed on the topic "battery life", such as fast, bad, or something else? It is extremely labor intensive and time consuming to get the above information by reading the reviews one-by-one by manually. As a result, it is a great value to automatically analyze the reviews to extract sentiments, topics and the associations among them. Great efforts on new methodologies for automated sentiment detection and prob-

* Corresponding author.
E-mail addresses: fuxh@szu.edu.cn (X. Fu), snowalex@foxmail.com (X. Sun), 2160230425@email.szu.edu.cn (H. Wu), cuilz@szu.edu.cn (L. Cui), zx.huang@szu.edu.cn (J.Z. Huang).

https://doi.org/10.1016/j.knosys.2018.02.012
0950-7051/

ing latent knowledge from diversified text data have flourished in the recent years [7,30,33].

The problem to detect topic and sentiment simultaneously can not be solved by the traditional sentiment classification methods [38] and topic discovery methods, which only care about the overall sentiment of a review or the topics of a review separately, and don't perform an in-depth analysis to discover the latent topics and the associated topic sentiments. Actually, the problem consists of two sub-tasks. First, sentiment analysis is used to judge that the document-level sentiment polarity of a review, i.e., whether the polarity of a review is positive or negative. Second, the associations of sentiments and latent topics from text are detected. Actually, there are strong associations among the topics and sentiments, but most works research them separately. Traditional topic models like latent Dirichlet allocation (LDA) [2] and probabilistic latent semantic analysis (PLSI), and recent Bayesian Nonparametric Relational Topic Model [39] have been efficiently applied to topic discovery from a text [40]. The success of this approach has motivated the creation of numerous other models that extend these models in order to capture other text aspects. Several works extending probabilistic topic models have been designed to tackle the problem of the joint extraction of sentiments and latent topics from a document in the recent years [4,19,20,25]. The joint sentiment topic models (JST) [9,19] extend LDA to be a four-layer model by adding an additional sentiment layer between the document and the topic layer. Topic sentiment mixture (TSM) [25] jointly models topics and sentiments in the corpus built on the basis of PLSI. These approaches infer sentiment and topic distributions from the co-occurrence of words within documents. However, when the training corpus of documents is small or when the documents are short, the sparse feature vectors of words cause that it is difficult to distinguish the synonyms/homonyms and the resulting topic distributions might be not very satisfactory [28]. Some previous works show that it can help to improve the topic representation problem with external information [31,32]. For example, Phan et al. [32] used the hidden topics discovered in the large Wikipedia corpus to help shape the topic representations in the small corpus. Petterson et al. [31] used external information about word similarity, such as thesauri and dictionaries, to smooth the topic-to-word distribution. However, most of these methods are based on the bag of word method. The larger corpus maybe have many irrelevant topics, and it is not easy to directly calculate the word similarity between two words. It is necessary to find better methods to add external information into the short text.

Most recently, word embeddings are gaining more and more attention, since they show very good performance in a broad range of natural language processing (NLP) tasks [5,22,28]. Word embeddings are trained with large external corpus such as Wikipedia, which can get the sematic information of the words efficiently, and can be used to calculate the word similarity directly and combine multi-word semantics [26, 27]. The combination of values permitted by word embeddings forms a high dimensional vector space, which is suitable to model topics [28,22]. To achieve better performance, some studies try to combine word embeddings with topic model. Nguyen et al. [28] incorporate latent feature vector representations of words trained on a very large corpora to improve the word-topic mapping learnt on a smaller corpus. Liu et al. [22] employ latent topic models to assign topics for each word in the text corpus, and learn topical word embeddings (TWE) based on both words and their topics. However these models only complete the task of mining topic and do not detect the association of sentiment and topic. Little attention has been devoted to topic sentiment model with word embeddings so far. In this paper, we propose a new topic sentiment model, which incorporates word embeddings to overcome the above challenges. To our knowledge, this is the first work to formulate topic sentiment model with word embeddings.

Additionally, most of the recent works of sentiment analysis are committed to unsupervised and weakly supervised methods [4,6,30]. Many current approaches [4,6,20] provide some polarity lexicons such as general polarity words ("good", "bad", etc.) for each model polarity, to shape the prior word distributions of a model polarity. They assume a predefined dictionary of sentiment words, typically incorporating this information into the β priors for the topic-word distributions when sampling a sentiment for a word at the model initialization. However, these approaches have limitations. In many cases, such dictionaries may be unavailable. If the polarity lexicons are not rich, the impact of the prior is very limited [16]. Because they cannot obtain enough help to identify the topic-specific sentiment words, it is necessary to seek for other approaches.

In contrast with other topic sentiment modeling frameworks that use word co-occurrence statistics and polarity lexicons, our model is distinguished from them as follow:

(1) We introduce word embeddings which are trained on very large corpora into the new model. The word embedding can add external semantic information of words to improve the problems of data sparseness and synonyms/homonyms, and get better topic representations. Specifically, we replace the sentiment-topic-to-word Dirichlet multinomial component with a mixture of the Dirichlet multinomial component and the word embeddings component. The new model can significantly improve the word-sentiment-topic mapping, and extend semantic and syntactic information of words.
(2) Most of the existing methods only incorporate a predefined dictionary of sentiment words to help initialize the sentiment orientation of a word at the model initialization. Unlike the existing methods, our model utilizes HowNet lexicon to compute every word's sentiment orientation at first, and then initialize it with its sentiment value. It can significantly capture the sentiment information of each word in the corpus. Furthermore, the new model allows automatic updates of sentiment orientation for each word in a semi-supervised fashion, which can avoid relying on predefined domain knowledge to identify positive and negative words.

We implement our model (WS-TSWE) with Gibbs sampling algorithm and experiment on four real online review data sets (book, hotel, computer, and movie) for two kinds of language (English and Chinese). We compare our model with different sentiment initialization modes, such as without predefined sentiment polarity dictionary (TSWE−P), with predefined sentiment polarity dictionary (TSWE+P), and with HowNet lexicon (WS-TSWE), and reserved the words that not found in Google vector representations trained from Chinese Wikipedia corpus and English Wikipedia corpus (WS-TSWE'). We also compared with several other recent models, such as joint topic sentiment model (JST), Hidden Topic Sentiment Model (HTSM), Dependency-Sentiment-LDA (DSLDA), LDA and the latent feature LDA model (LFLDA). The experiments show that our models can get good performance to detect sentiments and topics simultaneously.

The rest of the paper is organized as follow. Some related works are presented in Section 2. The topic sentiment joint model with word embeddings is introduced in Section 3. The experimental setting and results on the multi-domain sentiment data sets are described in Section 4. Finally, the conclusion and the future work are drawn in Section 5.

2. Related works

2.1. Topic sentiment models

The traditional topic models like LDA [2] and PLSI [12], which have been effectively applied to the latent topic discovery from a text. The LDA topic model has three hierarchical layers, where topics are associated with documents, and words are generated according to the topics. LDA model is based on the assumption that a document is the mixture of topics, where each topic is a probability distribution over words in a fixed vocabulary. Generally, to generate a word in a document with LDA model can be divided into two procedures. Firstly, a distribution over a mixture of K topics for the document is generated. Then, a topic is selected randomly from the topic distribution, and a word is drawn from that topic based on the corresponding topic-word distribution.

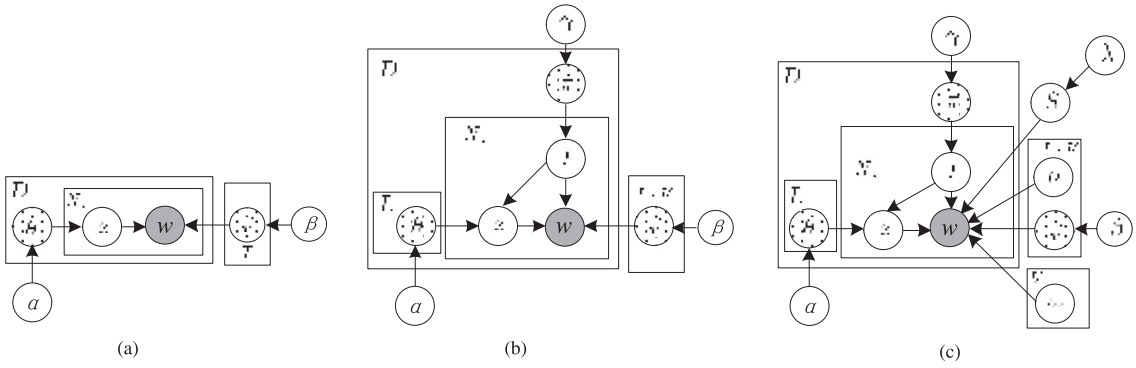
There are some works to extend topic models to formulate joint topic sentiment models, and achieved good performance. For example, The joint topic sentiment model (JST) and reverse joint sentiment-topic model (Reverse-JST) [19,20] are LDA extension models that detect both sentiments and latent topics simultaneously from a text. They transform LDA model to a four-layer model by adding an additional sentiment layer between the document and the topic layer. In JST model, sentiment labels are generated according to documents, and then the topics are generated based on sentiment labels and words generated based on both sentiment labels and topics [19,20]. Lin et al. also incorporate two combined lexicons as prior information into the JST model and put forward the Reverse-JST model [20]. While in Reverse-JST the association direction is reversed: sentiments are associated with a topic in the modeling process. It means that Reverse-JST consider to model the topic level sentiments. They also observed that its performance on document level sentiment analysis is consistently worse than that of JST when the sentiment prior information is encoded. Topic sentiment mixture (TSM) [25] jointly models topics and sentiments in the corpus based on the PLSI with an extra background component and two additional sentiment sub-topics. The topic-sentiment correlation is derived through a post-processing to calculate the topic-word and the word-sentiment distributions. Topic-sentiment modeling (TS) [6] is another topic model for topic-specific sentiment modeling from a text, which adopts a “bottom-up” approach, and it is similar with Reverse-JST. Sentiment topic model with decomposed prior (STDP) [4] is another variant of JST, where the sentiment polarity of a word depends on its part-of-speech category. The model decomposes the generative process of a word’s sentiment polarity to be a two-level hierarchy. The first level determines whether a word is used as a sentiment word or just an ordinary topic word. The second level (if the word is used as a sentiment word) determines the polarity of this word. With this decomposition, STDP provides separate priority for the first level to encourage the discrimination between sentiment words and ordinary topic words. Pavitra et al. [30] proposed a weakly supervised document level sentiment classification in conjunction with topic detection and topic sentiment analysis of bigrams simultaneously based on the weakly supervised joint sentiment topic (JST) model. Jo et al. [13] present the aspect and sentiment unification model (ASUM), and intend to analyze user reviews of goods and services by incorporating both aspect-based analysis and sentiment analysis. The model is similar to JST, but it breaks a review down into sentences, assuming that all words in a single sentence are generated from the same topic (aspect), i.e., a single sentence is assumed to speak about only one aspect of the item under review. Li et al. [17] put forward the Dependency-Sentiment-LDA (DSLDA), where they consider the relationship between sentiment and topic, and the relationship between sentiment depends on the context. Moreover, Rahman et al. [33] present Hidden Topic Sentiment Model (HTSM) to explicitly capture topic coherence and sentiment consistency to extract latent aspects and corresponding sentiment polarities, where they enforce words in the same sentence to share the same topic assignment and guide both topic transition and sentiment transition by a parameterized logistic function. Although there are many works to analyze topics and sentiments at the same time, all of these aforementioned models share some similar limitations:

(1) They infer sentiment and topic distributions from the co-occurrence of words within the documents. However, the research result shows that when the training corpus is small or when the documents are short, the resulting distributions might be not very satisfactory [28].
(2) In order to model sentiments efficiently in accordance with human sentiments and improve the performance of sentiment analysis, most previous works need to extract phrases that contain polarity words in predefined lexicons as prior information. With the lexicons, these works usually distinguish whether a word is affective to express feelings, and evaluate whether a word can express some sentiment about a specific aspect. However, the polarity words in the lexicons are usually constructed by manual. If the polarity lexicons are not rich, the impact of the prior is very limited. Because they are a little help to identify the topic-specific sentiment words. In this paper, we improve these drawbacks by using word embeddings that trained on a large external corpus to supply a multinomial sentiment-topic model that estimated from the training corpus.

2.2. Word embeddings models

With the recent surge of interests in deep neural networks, many works have concentrated on learning a real-valued word embeddings in a continuous space, where similar words are likely to have similar vectors. This technique is called word embeddings [36], which can capture distributional similarity (lexical, semantic) between words.

Word embeddings have been widely used for Natural Language Processing (NLP) tasks and have demonstrated improvements in generalization accuracy on a variety of tasks [3,5,8,15,18,22,23,28,34]. Topic models have also been constructed by using word embeddings [5,22,28]. Nguyen et al. [28] proposed the latent feature LDA model (LFLDA) which incorporate latent feature vector representations of words trained on very large corpora to improve the word-topic mapping learnt on a smaller corpus. Liu et al. [22] employ latent topic models to assign topics for each word in the text corpus and learn topical word embeddings (TWE) based on both words and their topics. The word2vec toolkit [26] is an open source tool from Google, which can represent words as vectors effectively and can be trained directly on a new corpus. Therefore, it has the ability to represent the context words well and calculate similarity between words accurately. Many researches focus on introducing it to the topic model for learning semantic meaning of words [3,23,34]. But they rely solely on a multinomial or latent feature model. In Ma et al. and Sridhar et al. [23,34], the word embeddings have also only used for short text classification. Ma et al. [23] assume that a short text document is a specific sample of one distribution in a Bayesian framework. These mentioned models only complete the task of mining topic and don’t detect the association of sentiment and topic. However, in our new models, we introduce the word embeddings into the topic sentiment model. Specifically, since the sentiment-topic-to-word Dirichlet multinomial component generates the words from sentiment-topics in each Dirichlet multinomial topic mode, in order to incorporate the rich information from external large data set, we replace the sentiment-topic-to-word Dirichlet multinomial



(a) (b) (c)

Fig. 1. (a) LDA model. (b) JST model. (c) WS-TSWE model. The shaded nodes w are the observations, which are inputs. And the dotted nodes θ, π and φ need to be inferenced, which are outputs.

component with a mixture of the Dirichlet multinomial component and the word embeddings component. These unsupervised models can take advantages of the mixture of Dirichlet multinomial component and word embeddings component, thus achieving a good effect.

3. Weakly supervised topic and sentiment model with word embeddings

As depicted in the previous section, in order to capture the sentiment-topic correlation, number of topic models have been built on the basis of the well-known LDA, as shown in Fig. 1(a). In this section, we propose a novel topic sentiment model with word embeddings and HowNet Lexicon called WS-TSWE in Fig. 1(c), which combines word embeddings and Hownet lexicon with JST model in Fig. 1(b). We use the word embeddings to expand the semantic information of words, and the HowNet Lexicon to capture the sentiment prior information of words. First, we need to get sentiment orientation of each word in the corpus during the model initialization and update the β priors in the process of model learning.

3.1. The WS-TSWE model

Since the sentiment-topic-to-word Dirichlet multinomial component generates words from sentiment-topics, we build our new model by utilizing the original Dirichlet multinomial topic sentiment model JST, and replacing the sentiment-topic-toward Dirichlet multinomial component with a two-component mixture of the sentiment-topic-to-word Dirichlet multinomial component and the word embeddings component. The new model has the structure of the original JST model, with two additional matrices υ and ω of embeddings weight, where υ_k and ω_i are the embeddings representations associated with sentiment-topic k and word i respectively. We also add a binary indicator variable s_i obeying Bernoulli distribution, to determine whether the word w_i is to be generated by the Dirichlet multinomial or the word embeddings component. Our model defines the probability that a word is generated from embeddings component as the multinomial distribution *MulT* with:

$$MulT(w_i|\upsilon_k\omega^{\mathrm{T}}) = \frac{\exp(\upsilon_k \cdot \omega_{w_i})}{\sum_{w_i' \in W} \exp(\upsilon_k \cdot \omega_{w_i'})} \tag{1}$$

MulT is a multinomial distribution with log-space parameters. ω is pre-trained word embeddings learned from an external big corpus, which is fixed. w_i and υ_k are word and topic embeddings weights. To approximate the topic embeddings υ_k, we apply to the regularized maximum likelihood estimation. Learning log-linear models for topic models with MAP estimation is also used in model [10,29]. Nevertheless, it is worth noting that the application of MAP in WS-TSWE is different from those models. Because they do not use word embeddings to characterize sentiment-topic-word distributions. The negative log likelihood L in our model factorizes topic-wise into factors L_k for each topic associated with sentiment. With L_2 regularization for each topic z_k under sentiment, we set L_2 to 0.01. N^{k,w_i} denotes the number of occurrences of word w_i assigned with topic k and sentiment label l, we derive:

$$L_k = \mu||\upsilon_k||_2^2 - \sum_{w_i \in W} N^{k,w_i}\left(\upsilon_k\omega_{w_i} - \log\left(\sum_{w_i' \in W}\exp(\upsilon_k\omega_{w_i'})\right)\right) \tag{2}$$

We obtain the MAP estimation of topic vectors υ_k by minimizing the regularized negative log likelihood. The derivative with respect to the mth element of the embeddings for topic k is:

$$\frac{\partial L_k}{\partial \upsilon_{k,m}} = 2\mu\upsilon_{k,m} - \sum_{w_i \in W} N^{k,w_i}\left(\omega_{w_i,m} - \sum_{w_i' \in W}\omega_{w_i',m} MulT(w_i'|\upsilon_k^{\mathrm{T}})\right) \tag{3}$$

Then, we applied L-BFGS implementation [21] from the Mallet toolkit[1] to discover the topic vector υ_k that minimizes L_k.

3.2. Generative process for the WS-TSWE model

The WS-TSWE model generates a word w_i in document d as follow. First, a sentiment multinomial distribution π_d is sampled from the Dirichlet distribution γ. In other words, sentiment multinomial distribution π_d is generated by the hyper-parameter γ for the Dirichlet distribution. Second, a sentiment label l is chosen from the per-document sentiment multinomial distribution π_d. Third, a topic multinomial distribution $\theta_{d,l}$ corresponding to the sentiment label l (see Table 1 for notation) is generated from Dirichlet distribution α. That is, topic multinomial distribution $\theta_{d,l}$ is generated by the hyper-parameter α for the Dirichlet distribution. Fourth, a topic k is selected from the topic multinomial distribution $\theta_{d,l}$, where $\theta_{d,l}$ is based on the sampled sentiment label l. Fifth, for each word w_i in document d, the WS-TSWE model chooses a binary indicator variable s_i that is a Bernoulli distribution, to determine whether the word w_i is to be generated by the Dirichlet multinomial or word embeddings component. Finally, the word is generated from the selected topic by the determined sentiment-topic-word model on both sentiment and topic label. This is different from JST, since in JST a word is sampled from the Dirichlet multinomial according to the sentiment topic.

The formal definition of the generative process of WS-TSWE model is as follow:

[1] http://mallet.cs.umass.edu/

Table 1
Notations.

Symbol	Description
D	Number of documents in the corpus
V	Number of words in the vocabulary
K	Number of topics in each document
L	Number of sentiment labels in each document
R	Number of iterations
W	Word sequence set in the vocabulary
l	Sentiment labels
N_d	Number of words in document d
$N_{d,\,l}$	Number of words in document d assigned with sentiment l
$N_{d,\,l,\,k}$	Number of occurrences of word i in document d being associated with sentiment label l and topic k
$N_{l,\,k,\,i}$	Number of occurrences of word i appeared in topic k and sentiment label l by the embedding component and Dirichlet multinomial component
$N_{l,\,k}$	Number of occurrences of word i assigned to sentiment label l and topic k

1. For each of sentiment-topic pair (l, z) generate the word distribution of the sentiment-topic pair $\varphi_{l,\,k} \sim Dir(\beta)$
2. For each document d draw a multinomial distribution $\pi_d \sim Dir(\gamma)$
3. For each sentiment label l under document d draw a multinomial distribution $\theta_{d,\,l} \sim Dir(\alpha)$
4. For each word w_i in document d
   - draw a sentiment label $l_i \sim Mul(\pi_d)$
   - draw a topic $z_i \sim Mul(\theta_{d,\,l})$
   - draw a binary indicator variable $s_i \sim Ber(\lambda)$
   - draw a word $w_i \sim (1 - s_i)mul(\varphi_{z_i}) + s_i MulT(\upsilon_{z_i}\omega^{\mathrm{T}})$

Dir and *Mul* represent a Dirichlet distribution and a multinomial distribution, respectively. *MulT* is a multinomial distribution with a log-space parameter. $Ber(\lambda)$ is a Bernoulli distribution with success probability λ. In our implementation, we use asymmetric prior α and symmetric prior β and γ. There are three sets of latent variables that we need to infer in our new model WS-TSWE, the per-document sentiment multinomial distribution π, the per-document sentiment label specific topic multinomial distribution θ, and the joint sentiment-topic word distribution φ. We will present Gibbs sampling procedures for our novel model in the following subsection.

3.3. Learning algorithm

There are two common used approximation methods to infer the graph models. One is the Gibbs sampling method, and the other is the variational inference method. Variational inference methods essentially decouple all the nodes, and introduce a new parameter, called a variational parameter. For each node, they iteratively update these parameters so as to minimize the cross-entropy (KL distance) between the approximate and true probability distributions [1,11,14,37]. Gibbs sampling algorithm is also a popular approach for parameter estimation and inference in many topic models [4,6,19]. The advantage of the Gibbs sampling method includes that it is simple and easy to implement, and it has theoretical guarantees of convergence. On the other side, there are some drawbacks of Gibbs sampling algorithms such as they often have slow convergence and they are not sufficient to deal with massive corpus [1,11,14,37]. However, we do not consider large data sets to evaluate the effectiveness and performance of our WS-TSWE model, so it is also feasible for us to implement our WS-TSWE model with Gibbs sampling algorithm. We also adopt Gibbs sampling algorithm to calculate the conditional joint sentiment topic assignment probabilities for each word as following. The pseudocode of the Gibbs sampling algorithm for the WS-TSWE model is given in Algorithm 1, and the meanings of all variables are shown in Table 1.

Algorithm 1 Learning algorithm for the WS-TSWE model.

Data: Document in the input, word embeddings ω
Begin
Initialize $\pi_{d,\,l}$, $\theta_{d,\,l,\,k}$, $\varphi_{l,\,k,\,i}$ using the JST sampling algorithm in [19]
for each document $d \in [1, D]$ **do**
  for each word $i \in [1, N_d]$ in document d **do**
    Sample a sentiment label and topic for word i
    Increment counts:
    $N_{d,l} + 1, N_d + 1, N_{d,l,k} + 1, N_{l,k,i} + 1, N_{l,k} + 1$
  end for
end for
for iteration $n = 1$ to max Gibbs sampling iterations **do**
  Optimize topic vectors
  for each document $d \in [1, D]$ **do**
    for each word $i \in [1, N_d]$ in document d **do**
      for the current assignment for word i
      Decrement counts:
      $N_{d,l} - 1, N_d - 1, N_{d,l} - 1, N_{l,k,i} - 1, N_{l,k} + 1$
      Sample a new sentiment and topic label using Eq. (8)
      Increment counts:
      $N_{d,l} + 1, N_d + 1, N_{d,l,k} + 1, N_{l,k,i} + 1, N_{l,k} + 1$
  end for
end for
  Update β_{lkw} priors with Eq. (12)
end for
  Update $\pi_{d,\,l}$, $\theta_{d,\,l,\,k}$, $\varphi_{l,\,k,\,i}$ with Eqs. (9), (10), (11) respectively

3.3.1. Gibbs sampling for the WS-TSWE model

In order to integrate π, θ and φ, we first estimate the posterior distribution over l and z, and the assignment of word tokens to sentiment labels and topics. The hyper-parameters in our WS-TSWE model contain α, β, γ, which are given according to the experience. The latent parameters include l, z, π, θ, φ, which need to be estimated through the observed variables.

According to the Fig. 1(c) of the WS-TSWE model, the joint probability distribution for a word given the remaining sentiment labels and topics can be factored as follow:

$$p(w, z, l|\alpha, \beta, \gamma, \upsilon, \omega) = p(w|z, l, \beta, \lambda, \upsilon, \omega) \cdot p(z|l, \alpha) \cdot p(l|\gamma) \tag{4}$$

where the $p(w|z, l, \beta, \lambda, \upsilon, \omega)$ is the process of sampling the words according to the joint sentiment topic, and it can be generated by Dirichlet multinomial word embeddings component. $p(z|l, \alpha)$ obtains the topic based on the certain sentiment l and prior distribution parameter α. $p(l|\gamma)$ estimates the sentiment of the document according to hyper-parameter γ. These three terms are the unknown parameters we need to compute as following.

Since these three processes are independent, we can deal with them separately. $p(w|z, l, \beta, \lambda, \upsilon, \omega)$ is obtained to compute φ.

$$p(w|z, l, \beta, \lambda, \upsilon, \omega) = (1-\lambda)\left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{L\cdot K} \prod_l \prod_k \frac{\Pi_i \Gamma(N_{l,k,i}+\beta)}{\Gamma(N_{l,k}+V\beta)} + \lambda \cdot MulT(w|\upsilon_k\omega^{\mathrm{T}}) \tag{5}$$

The remaining terms of Eq. (4) are obtained in the same way by integrating θ and π respectively, and we derive

$$p(z|l,\alpha)=\left(\frac{\Gamma(\sum_{k=1}^{T}\alpha_{l,k})}{\prod_{k=1}^{T}\Gamma\alpha_{l,k}}\right)^{D\cdot L}\prod_{d}\prod_{l}\frac{\prod_{k}\Gamma(N_{d,l,k}+\alpha)}{\Gamma(N_{d,l}+T\alpha)} \tag{6}$$

$$p(l|\gamma)=\left(\frac{\Gamma(L\gamma)}{\Gamma(\gamma)^{L}}\right)^{D}\prod_{d}\frac{\prod_{l}\Gamma(N_{d,l}+\gamma)}{\Gamma(N_{d}+L\gamma)} \tag{7}$$

Given the current values of all other variables and data for Bayesian analysis with the joint distribution, we can calculate conditional distribution (posterior distribution) of the latent variables l and z. We make the superscript $\neg i$ denote a quantity that excludes data from ith position. For instance, $N_{d,l}^{\neg i}$ is the number of words appearing under a sentiment label l and disregarding the ith word of document d. Therefore, the Posterior probability can be obtained from the joint probability as follow:

$$P(z_i=k,l_i=l|w,z^{\neg i},l^{\neg i},\alpha,\beta,\gamma,\lambda,\upsilon,\omega)$$
$$\propto\left((1-\lambda\cdot\frac{N_{l,k,w_i}^{\neg i}+\beta}{N_{l,k}^{\neg i}+V\beta}+\lambda\cdot MulT(w_i|\upsilon_k\omega^{\mathrm{T}})\right)\cdot\frac{N_{d,l,k}^{\neg i}+\alpha}{N_{d,l}^{\neg i}+T\alpha}\cdot\frac{N_{d,l}^{\neg i}+\gamma}{N_{d}^{\neg i}+L\gamma} \tag{8}$$

Samples derived from the Markov chain are then used to estimate the per-document sentiment multinomial distribution π, per-document sentiment topic multinomial distribution θ and the per-corpus sentiment-topic word distribution, as depicted in ((9)–(11)).

$$\pi_{d,l}=\frac{N_{d,l}+\gamma}{N_{d}+L\gamma} \tag{9}$$

$$\theta_{d,l,k}=\frac{N_{d,l,k}+\alpha}{N_{d,l}+T\alpha} \tag{10}$$

$$\varphi_{l,k,i}=(1-\lambda)\cdot\frac{N_{l,k,i}+\beta}{N_{l,k}+V\beta}+\lambda MulT(w_i|\upsilon_k\omega^{\mathrm{T}}) \tag{11}$$

3.3.2. Learning sentiment prior information

We can get the prior sentiment polarity information based on an external lexicon in the model initialization. In JST and Reverse-JST, word sentiment priors λ are drawn from an external dictionary and incorporated into β priors; $\beta_{lkw}=\beta$ if word w can have sentiment label k and $\beta_{lkw}=0$ otherwise. In our model, we treat β_{lkw} as latent variables in the model and they are trained with the Gibbs sampling algorithm. We can get every word sentiment prior information β_{lkw} based on the external lexicon in the initial approximation and we update $\beta_{lkw}\propto N_{l,\,k,\,i}$ in the Gibbs sampling algorithm process.

The Dirichlet prior parameters do not have to sum up to 1, and their sum affects the variance of the Dirichlet distribution. Here, it makes sense to start with a relatively large variance at the stage where we are unsure of the sentiments and then gradually refine the β_{lkw} estimates. So the β_{lkw} can be represented as $\beta_{lkw}=\frac{N_{l,k,w}}{\eta}$, where η is a regularization coefficient that should start to be large and then decrease. For the iteration n, β_{lkw} is updated as:

$$\beta_{lkw}=\frac{1}{\eta_n}N_{l,k,w},\ \text{where}\ \eta_n=\max(1,\frac{100}{n}) \tag{12}$$

3.4. Sentiment orientation calculation based on HowNet lexicon

The sentiment prior information can be calculated with different sentiment lexicon. We use HowNet lexicon in this paper. HowNet lexicon is a general knowledge base of Chinese and English words and it is very suitable for our work, since our data sets contain both Chinese and English data set. There are two basic terminologies in HowNet, which are "concept" and "sememe". Concept describes the semantic of words and every word can express one concept or a few concepts. Sememe denotes the minimum semantic unit to describe the concept. Each concept in HowNet lexicon is described by several sememes. We can use the similarity between two words to get the sentiment polarity. The similarity between two words in HowNet is calculated by the similarity between their concepts. Furthermore, the similarity of two concepts is calculated by the similarity of their sememes. Given two sememes o_i and o_j, their sememes' similarity is calculated by their path distance in their sememe tree [41]. The calculation formula is as follow:

$$Sim_o(o_i,o_j)=\frac{\alpha}{dis_{ij}+\alpha} \tag{12}$$

where α is an adjustable parameter, $dis_{ij}>0$ is a positive integer that denotes the path length between o_i and o_j in the sememes tree.

The similarity of two concepts U_i and U_j is determined by their maximal sememe similarity.

$$Sim(U_i,U_j)=\max_{m,n}Sim(o_{im},o_{in}) \tag{13}$$

Given two Chinese words w_i and w_j, if w_i contains g concepts, denoted by U_{i1}, U_{i2}, ..., U_{ig}, and likewise, w_j contains h concepts, denoted by U_{j1}, U_{j2}, ..., U_{jh}, the similarity between w_i and w_j equals to the maximum similarity of all the concepts of the words:

$$Sim_w(w_i,w_j)=\max_{e,r}Sim(U_{ie},U_{jr}), \tag{14}$$

where $e=1,2,...,g$, $r=1,2,...,h$.

In order to calculate the sentiment polarity of each word in our training data, first of all, we choose a pairs of words with strong positive and negative polarity as benchmark words. Afterward, we calculate the similarity between the training word and seed word. The similarity value is used to measure the word's sentiment orientation. The sentiment orientation $G_{label}(w)$ will be calculated as:

$$G_{label}(w)=\sum_{i=1}^{a}(wp_i\cdot Sim_w(p_i,w)-wn_i\cdot Sim_w(n_i,w)) \tag{15}$$

where p_i denotes the positive benchmark words and n_i denotes the negative seed words. Obviously, if $G_{label}(w)>0$, the sentiment polarity of w is positive; if $G_{label}(w)<0$, it means the sentiment polarity of w is negative; otherwise, if $G_{label}(w)=0$, it implies a neutral sentiment polarity of w. Therefore, every word prior sentiment polarity in the corpus can be captured through this process.

To analyze our algorithm, obviously it has a similar running process with the JST algorithm [19,20]. The time complexity of single iteration of our approach is O(MNKL), where M is the number of documents, N is the average length of the document, K is the number of topics, and L is the number of sentiment categories.

4. Experiments

In this section, we evaluate the performance of our new WS-TSWE model on document-level sentiment polarity identification and topic-sentiment identification with different data sets for English and Chinese. The performance of the document-level sentiment identification is measured based on the probability of a sentiment label given a document. We only take into account the probability of positive and negative labels for a given document in our experiments, and the probability of the neutral labels are ignored. Therefore, we define that a document d will be identified to be a positive-sentiment document if its probability of a positive sentiment label is bigger than its probability of negative sentiment

label, and vice versa. On the other side, topic-sentiment identification determines the word distribution conditioned on both topics and sentiment labels in the whole corpus, which performance can be evaluated by perplexity and normalized mutual information (NMI) [24]. In addition, we also compare the performance with different sentiment initialization modes, such as mode incorporating the prior sentiment polarity dictionary, mode without introducing prior sentiment polarity dictionary, and mode combining HowNet lexicon. We also explore the influence of the super-parameter λ on the performance.

4.1. Experimental setup

4.1.1. Training word embeddings

We train 300 dimensional word vectors on two corpora of different sizes by using the Google word2vec toolkit [26]: a Chinese Wikipedia and an English Wikipedia. Wikipedia is the largest encyclopedia, which is open, web-oriented and multi-lingual so far. Since documents in Wikipedia are clearly organized by topics, it's very suitable for training our word embeddings.

The Chinese Wikipedia corpus[2] has 777,961 articles, which contains several languages, such as Continental Simplified, Taiwan traditional, Hong Kong and Macao traditional, and so on. Before training, we first use wiki extractor to get the content of each text, and then convert them to Simplified Chinese. Meanwhile all useless tags in each document are removed. Then the whole corpus is expressed as a big document of 777,961 lines, and each line represents a text. Another preprocessing tasks we should do include Chinese word segment, filtering out low/high frequency words, stop words, improper characters, and single words. In terms of Chinese word segment, we used the NLPIR 2014 system[3], which supports the user dictionary and new words mining. And the threshold value of low/high frequency words is set according to the experience. In our experiment, we assume that word occurring less than 5 times in the training data is a low frequency word, and if its frequency is larger than 15%, it is regarded as a high frequency. Finally, we get a corpus with about a total of 160,373 tokens.

The English Wikipedia corpus[4] we used is the April 2010 snapshot, which contains about 2 million articles and 990 million tokens. We remove the stop words, digits and the word with too low word frequency and too high word frequency, and convert all the characters to a lower case in the preprocessing. After this processing, we construct a vocabulary that contains the top 100,000 most frequent words in the corpus for model training.

4.1.2. Experimental data sets

WS-TSWE model is a general topic sentiment joint model and can be used for different language documents. To demonstrate the effectiveness of WS-TSWE model for different language, we perform experiments on two kinds of language sentiment mining data sets, Chinese and English.

4.1.2.1. Chinese data sets. The Chinese sentiment mining data sets consists of three categories of product reviews data sets[5], including book, hotel, and computer, with 2000 positive and 2000 negative examples for each domain.

4.1.2.2. English data set. The English sentiment mining corpora called MR04 data set is the polarity data set version 2.0[6], which was introduced by Pang and Lee in 2004, consisting of 1000 positive and 1000 negative movie reviews.

Table 2
The experimental data sets. #doc: number of documents; #p-doc: number of positive documents; #n-doc: number of negative documents; V: the number of word types.

Data set	#doc	#p-doc	#n-doc	#average doc length	V
Computer	4000	2000	2000	11	3776
Hotel	4000	2000	2000	27	8702
Book	3975	2000	1975	18	2473
MR04	1385	700	685	183	4365

4.1.2.3. Preprocessing. Preprocessing is conducted on all of these data sets. For Chinese product reviews data sets, we finish tasks such as Chinese word segment, filter out low/high frequency words, stop words, improper characters, and single words. For English movie review data set, we run the following preprocessing: (1) remove the repetitive comments; (2) convert letters into lowercase; (3) remove non-latin characters and stop words. For both Chinese and English data sets, the words with document frequency less than 2 or larger than 15 are also removed. In addition, the words that not be found in Google vector representations which trained from Chinese Wikipedia corpus and English Wikipedia corpus are also removed, respectively. After preprocessing, the basic statistics for each data set are listed in Table 2.

Additionally, we also perform experiments on the data sets that do not remove the words which not found in Google vector representations trained from Chinese Wikipedia corpus and English Wikipedia corpus. And the experimental results are presented in Section 4.3.2.

4.2. Parameter setting

We set the symmetric prior hyper-parameter β to be 0.01 in our WS-TSWE model, as this is a common setting in the literature [35]. The symmetric hyper-parameter γ is set to $\frac{0.05 \cdot A}{L}$, where A is the average document length, L is the total number of sentiment labels, and the value of 0.05 on average allocates 5% of probability mass for mixing, as noted by Lin et al. [20]. And α is set to a standard setting $\alpha = \frac{50}{T}$.

For our WS-TSWE model, we run the baseline JST model for 1000 iterations, and then use the outputs from the last sample of the 1000 iterations to initialize our new model, which we run for 500 further iterations.

4.3. Experimental results and analysis

In this section, we present and discuss the experimental results of both document-level sentiment identification and topic-sentiment identification. As WS-TSWE consider to model sentiment and topic mixtures simultaneously, it is worth exploring how the sentiment identification and topic detection tasks effect separately. In addition, how the model behaves with different topic number settings on different data sets is also evaluated. With this in mind, we conduct a set of experiments on WS-TSWE, with different topic numbers as (1,5,10,20,40,60,80,100).

4.3.1. Document level sentiment identification evaluation

This subsection evaluates the performance of sentiment identification of each document. The proposed method is weakly supervised, which only uses the sentiment lexicons and does not use the sentiment labels in the training process. After using the topic sentiment model to calculate the sentiment probabilities of a document, we assign the sentiment label that has the highest probability to each document. It means if the positive sentiment word

[2] http://download.wikipedia.com/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2
[3] http://ictclas.nlpir.org/downloads
[4] http://nlp.stanford.edu/data/WestburyLab.wikicorp.201004.txt.bz2
[5] http://www.datatang.com/data/11937
[6] www.cs.cornell.edu/people/pabo/movie-review-data/

Table 3
Forty pairs of sentiment polarity benchmark words.

Positive	健康/healthy	友善/friendly	美丽/beautiful	乖巧/cute	喜欢/love
	天使/angel	精英/elite	权威/authoritative	优秀/excellent	精选/choiceness
	欢喜/joyful	幸福/happiness	容易/easy	文明/civilization	积极/positive
	著名/notability	完美/perfect	和平/peace	开明/enlightened	亮点/lightspot
	真实/real	先进/advanced	便宜/inexpensive	正确/correct	坚定/firm
	不错/good	精神/spiritual	新/new	诚信/honesty	安静/quiet
	完整/full	保准/guarantee	宝贵/precious	聪明/clever	顶级/top-level
	风趣/humor	干净/clean	公平/fair	成熟/mature	方便/convenience
Negative	陈旧/obsolete	嘈杂/noisy	缺少/lack	讨厌/hate	脏/dirty
	烦恼/annoy	害怕/fear	失望/disappoint	气愤/angry	谴责/denounce
	伤心/sad	怀疑/suspect	失败/failure	错误/error	不良/bad
	浪费/waste	生病/sickness	虚假/sham	恶意/malevolence	不安/uneasy
	愚/foolish	谣言/rumor	变态/abnormal	脆弱/weak	不合格/unqualified
	失误/mistake	自负/overconfident	麻烦/troublesome	陷阱/trap	淫秽/bawdry
	暴力/violence	无聊/bored	责备/blame	犹豫/hesitate	羞愧/shame
	非法/illegality	鄙视/contempt	消沉/depressed	委屈/grievance	挑剔/nitpick

count is greater than that of the negative words, a document is identified as positive, and vice versa. So we use the common metric accuracy as the evaluation criteria of the whole sentiment identification ability, which is same way as [20,28]. The accuracy is in the range of 0.0 to 1.0, and a higher value reflects better performance. Furthermore, taking the deficiency of accuracy measurement into consideration, we also evaluate the sentiment identification performance of each sentiment polarity (positive and negative) with the measurement of precision, recall and F1 score. When to identify the sentiment polarity of a document, the identification result can be classified four types: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). As we all know, these evaluation criteria are defined as: $precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN}$, $F1 = 2\frac{precision*recall}{precision+recall}$ $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. In this paper, all the results are the mean and standard deviation of Gibbs sampling running 5 times under every number of a topic.

4.3.1.1. Setting. In the experiments, in order to verify the effectiveness of the word embeddings and HowNet Lexicon, we compare the sentiment identification results of the model with sentiment polarity benchmark words, the model without sentiment polarity benchmark words and the model combining with HowNet Lexicon.

In this paper, TSWE+P represents our model incorporating sentiment polarity benchmark words, i.e., if a word can be found in the polarity benchmark words, its sentiment polarity is determined according the polarity benchmark words. Otherwise, its sentiment polarity is initialized randomly. TSWE−P denotes our model without sentiment polarity benchmark words, i.e. the sentiment polarity of each word is randomly initialized. And WS-TSWE denotes the model combining with HowNet Lexicon. The sentiment polarity benchmark words include two subjectivity lexicons. The English benchmark words are the MPQA[7] and the Chinese benchmark words are HowNet emotional word set[8]. They are incorporated as prior information into the model initialization. These two lexicons contain lexical words whose polarity orientation has been fully specified. Finally, the prior information is produced by retaining all words in the MPQA and HowNet emotional word set that occur in the experimental data sets, respectively. The WS-TSWE also selects the sentiment benchmark words with strong emotional polarity from HowNet lexicon. These sentiment benchmark words are selected according to the ordered Hits value returned by Google search engine. Furthermore, the words that have similar semantics are merged out. The final forty pairs of sentiment polarity benchmark words are listed in Table 3.

[7] http://www.cs.pitt.edu/mpqa/.
[8] http://www.datatang.com/datares/go.aspx?dataid=603399

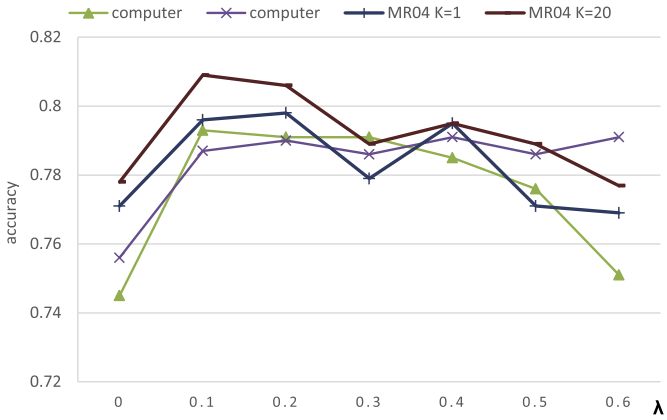


Fig. 2. Accuracy on the computer and MR04 with number of topics K = 1 and K = 20, varying the mixture weight λ from 0.0 to 0.6.

We do some experiments to see the effect of the mixture weight λ. Fig. 2 presents the identification accuracy results obtained by WS-TSWE model on the computer and MR04 data sets with the numbers of topics *K* set to either 1 or 20. The results are conducted on WS-TSWE which is combined with HowNet Lexicon, and the value of the mixture weight λ varies from 0.0 to 0.6. By varying λ, as shown in Fig. 2, the WS-TSWE model obtains its best result at λ = 0.1, and we can see the result is better than λ = 0.0 when the λ is set from 0.1 to 0.6. It can improve the performance when λ is small, such as 0.1 in our experiments. That shows the word embeddings are effective in capturing positive and negative sentiments. So we fix the mixture weight λ to be 0.1, and discuss the experimental results based on this value for the rest experiments.

4.3.1.2. TSWE+P vs TSWE−P and WS-TSWE. In Fig. 3, we can see that the identification results of the model incorporating lexicon scores (TSWE+P) are almost similar with the identification results of the model without lexicon (TSWE−P) on the same topic number on most tests. TSWE+P and TSWE−P achieved 80.5% and 79.7% respectively on MR04 data set at $K = 20$. Also, the yielding result on computer data set (TSWE+P 78.7% and TSWE−P 76.3%) can actually get better performance by incorporating any aforementioned prior information, which demonstrates that the sentiment priors have already been captured by the embeddings. Furthermore, we can also see the accuracy of the model with HowNet lexicon (WS-TSWE) performs better than TSWE+P and TSWE−P on all of the data sets. So the HowNet Lexicon indeed makes contributions to the sentiment accuracy. These results show that our weakly supervised topic sentiment model with word embeddings is effective in capturing positive and negative sentiments.

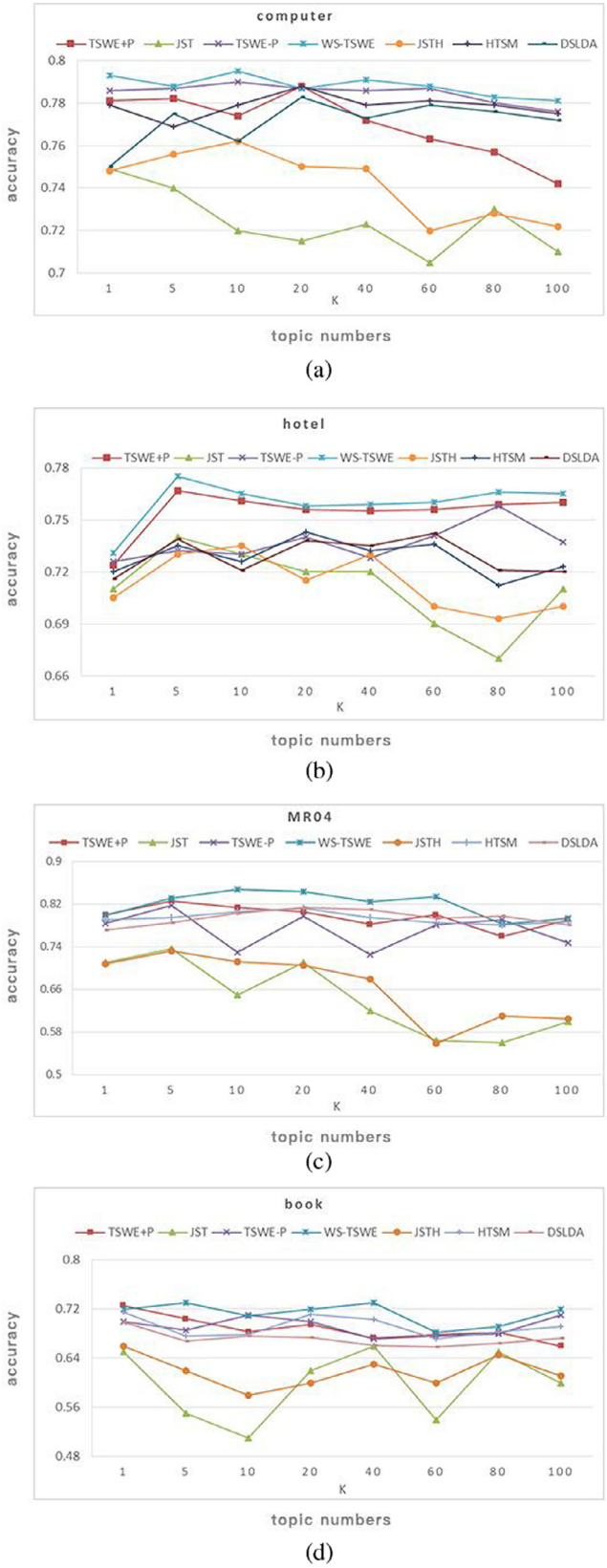


Fig. 3. Accuracy with different topic number settings on the four data sets. (a) Accuracy on computer for TSWE+P, TSWE−P, JST, JSTH, WS-TSWE, HTSM and DSLDA. (b) Accuracy on hotel for TSWE+P, TSWE−P, JST, JSTH, WS-TSWE, HTSM and DSLDA. (c) Accuracy on MR04 for TSWE+P, TSWE−P, JST, JSTH, WS-TSWE, HTSM and DSLDA. (d) Accuracy on book for TSWE+P, TSWE−P, JST, JSTH, WS-TSWE, HTSM and DSLDA.

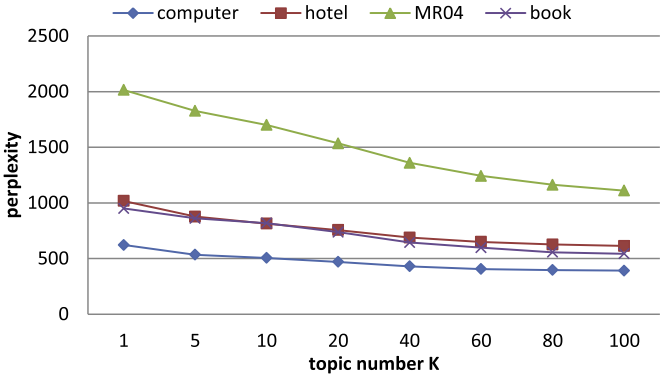


Fig. 4. Perplexity with different topic number settings on computer, hotel, MR04 and book data set.

4.3.1.3. Our models vs other models with different numbers of topics. Fig. 3 also shows identification results produced by our models (WS-TSWE, TSWE+P, TSWE−P) and other models (JST [19,20], HTSM [33], DSLDA (F. [17]), and JSTH) on the four data sets with different numbers of topics, where JSTH is the JST model with HowNet lexicon. We can see that our new model WS-TSWE significantly outperforms the JST and JSTH on all of the data sets, the JSTH is better than JST at most topic numbers. To be specific, on the MR04 data set, we get 30.0% improvement on accuracy at $K = 80$, and on the book data set we obtain 17% higher accuracy at $K = 10$. Those results show that an improved model of sentiment-topic-word mappings also improves the document-sentiment assignments. When the topic number is set to 1, both our models and JST essentially become the standard LDA model with only two sentiment topics, and hence they leave out the correlation between sentiment labels and topics. Fig. 3 also shows that our model performs better with multiple topic settings in the hotel domains. Because it is longer than other data sets on the size of each document. For the cases where a single topic performs the best on computer, book and MR04 data set, it is observed that the change in sentiment identification accuracy by additionally modeling mixtures of topics is only marginal, but the new model is able to extract sentiment-oriented topics in addition to document-level sentiment detection. The above results show that the word embeddings can help extend the semantic information of words, and the HowNet Lexicon can capture the sentiment information of words.

4.3.2. Topic-sentiment identification evaluation

Another task is to identify topics and sentiment from the data sets, and evaluate the effectiveness of sentiment topic captured by these models. Unlike the sentiment identification task, the topic-sentiment identification evaluates the word distribution over topics under positive and negative sentiment label for each document. So we need to evaluate the topic clustering performance under the corresponding sentiment polarity. We used two common metrics to evaluate the performance: perplexity and normalized mutual information (NMI) [24].

4.3.2.1. Perplexity. The perplexity used by convention in language modeling is monotonically decreasing in the likelihood of the test data sets. A lower perplexity scores reflect better generalization performance. More formally, for a test set of D documents, the perplexity is:

$$perplexity = \exp\left(-\frac{\sum_{d=1}^{D} \log p(w_d)}{\sum_{d=1}^{D} N_d}\right)$$

To determine an appropriate topic number K, we run TWSE on different data sets. Fig. 4 shows the perplexity values with the different number of topics on the four data sets. We can see the per-

Table 4
NMI results on computer, hotel, MR04 and book data set.

Data set	Model	NMI							
		$K = 1$	$K = 5$	$K = 10$	$K = 20$	$K = 40$	$K = 60$	$K = 80$	$K = 100$
Computer	**WS-TSWE**	0.569	0.523	**0.571**	0.521	0.556	0.522	0.548	0.546
	LDA	0.211	0.231	0.241	0.245	0.211	0.265	0.234	0.312
	LFLDA	0.311	0.341	0.356	0.362	0.411	0.368	0.425	0.396
	JST	0.441	0.442	0.390	0.377	0.391	0.360	0.421	0.380
Hotel	**WS-TSWE**	0.395	**0.479**	0.465	0.453	0.451	0.452	0.459	0.461
	LDA	0.256	0.231	0.354	0.369	0.344	0.322	0.302	0.296
	LFLDA	0.386	0.478	0.423	0.431	0.424	0.446	0.356	0.431
	JST	0.370	0.420	0.400	0.38	0.395	0.330	0.310	0.377
Book	**WS-TSWE**	**0.398**	0.356	0.314	0.342	0.343	0.352	0.311	0.231
	LDA	0.211	0.301	0.295	0.196	0.255	0.195	0.183	0.146
	LFLDA	0.365	0.321	0.301	0.295	0.301	0.312	0.301	0.246
	JST	0.260	0.083	0.070	0.195	0.270	0.062	0.24	0.168
MR04	**WS-TSWE**	0.536	**0.565**	0.535	0.516	0.527	0.536	0.482	0.525
	LDA	0.301	0.295	0.225	0.305	0.206	0.195	0.206	0.211
	LFLDA	0.305	0.396	0.306	0.398	0.256	0.231	0.411	0.397
	JST	0.358	0.420	0.248	0.370	0.195	0.101	0.100	0.164

Table 5
NMI, accuracy, precision, recall and F1 results on computer, hotel, MR04 and book data set.

Data sets	Models	NMI	Accuracy	Positive			Negative		
				Precision	Recall	F1	Precision	Recall	F1
Computer	**WS-TSWE'**	**0.593**	**0.789**	0.724	**0.935**	**0.816**	**0.908**	0.644	0.753
	WS-TSWE	**0.556**	**0.791**	**0.772**	0.827	**0.798**	0.813	**0.756**	**0.783**
	TSWE−P	0.493	0.786	0.761	**0.833**	0.796	**0.816**	0.739	0.775
	TSWE+P	0.487	0.772	0.762	0.792	0.776	0.783	0.753	0.767
	JST	0.391	0.723	0.685	0.827	0.749	0.781	0.620	0.691
	JSTH	0.436	0.749	0.718	0.820	0.766	0.790	0.679	0.730
	DSLDA	0.549	0.773	**0.800**	0.728	0.762	0.750	**0.818**	**0.783**
	HTSM	0.543	0.779	0.757	0.823	0.788	0.806	0.736	0.769
Hotel	**WS-TSWE'**	**0.456**	**0.765**	0.718	**0.873**	**0.788**	**0.838**	0.657	0.737
	WS-TSWE	0.451	**0.759**	**0.802**	0.689	0.741	0.727	**0.830**	**0.775**
	TSWE−P	0.401	0.728	0.759	0.669	0.711	0.704	0.787	0.743
	TSWE+P	**0.453**	0.755	0.701	**0.889**	**0.784**	**0.848**	0.621	0.717
	JST	0.395	0.721	0.707	0.756	0.730	0.738	0.686	0.711
	JSTH	0.398	0.730	**0.763**	0.667	0.712	0.704	**0.793**	**0.746**
	DSLDA	0.413	0.735	0.728	0.750	0.739	0.742	0.720	0.731
	HTSM	0.426	0.732	0.738	0.720	0.729	0.727	0.744	0.735
Book	**WS-TSWE'**	**0.391**	**0.728**	**0.731**	0.727	**0.729**	**0.725**	**0.729**	**0.727**
	WS-TSWE	**0.343**	**0.731**	**0.712**	**0.780**	**0.745**	**0.754**	0.681	**0.715**
	TSWE−P	0.295	0.672	0.683	0.649	0.666	0.662	**0.695**	0.678
	TSWE+P	0.296	0.673	0.676	0.673	0.674	0.670	0.673	0.672
	JST	0.273	0.665	0.652	0.717	0.683	0.681	0.613	0.645
	JSTH	0.269	0.630	0.608	**0.743**	0.669	0.664	0.515	0.581
	DSLDA	0.296	0.661	0.652	0.700	0.675	0.672	0.622	0.646
	HTSM	0.323	0.703	0.699	0.720	0.709	0.708	0.686	0.697
MR04	**WS-TSWE'**	**0.633**	**0.841**	**0.877**	0.797	**0.835**	0.810	**0.886**	**0.847**
	WS-TSWE	0.527	**0.824**	**0.839**	0.806	**0.822**	0.809	**0.842**	**0.825**
	TSWE−P	0.465	0.726	0.681	**0.859**	0.760	0.803	0.590	0.680
	TSWE+P	0.507	0.782	0.751	0.850	0.798	**0.823**	0.712	0.764
	JST	0.201	0.507	0.513	0.486	0.499	0.501	0.528	0.515
	JSTH	0.301	0.681	0.676	0.707	0.691	0.686	0.654	0.670
	DSLDA	**0.529**	0.809	0.784	0.857	0.819	0.839	0.759	0.797
	HTSM	0.503	0.796	0.742	**0.914**	0.819	**0.885**	0.674	0.766

plexity values decrease with larger topic number. The generalization performance increases when multiple topics are considered. However, the perplexity value begins to be stable after $K = 40$. Also we can see that the perplexity on the MR04 data set is higher than that on the other data sets. The reason is that the word number in the corpus is more than others. It also proves that it is appropriate to analyze the topic and sentiment in a unified way.

4.3.2.2. NMI. For a test set of D documents, and topic number K, the NMI is: $NMI = \sum_{k=1}^{K} \sum_{1 \le i < j \le D} \frac{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)}$

NMI scores always range from 0.0 to 1.0, and a higher score shows better clustering performance.

We also run WS-TSWE, LDA, LFLDA and JST on four data sets with different topic number. The NMI results are list in Table 4. From Table 4, we can see that the WS-TSWE model has better NMI value than JST, LDA and LFLDA. The NMI for the WS-TSWE model is around 0.231∼0.569, and the JST obtains NMI only around 0.101∼0.442, and the LFLDA is slightly better than JST on hotel, book and MR04 data sets in some topic numbers, which validates the effectiveness of the word embeddings on topic clustering performance.

4.3.2.3. Comprehensive evaluation. The results above are performed on the data sets that excluding the words that not in the trained word embeddings. Considering that the excluding words are in-

Table 6
Extracted topics under different sentiment labels by JST and WS-TSWE.

JST	Positive	T1	**漂亮/nice**	**散热/cooling**	**外观/appearance**	**喜欢/like**	设计/design
			比较/very	配置/configuration	硬盘/hard disk	**噪音/noise**	内存/memory
		T2	本本/machine	**完美/perfect**	**时尚/fashion**	钢琴/piano	键盘/keyboard
			京东/jingdong	价格/price	东西/stuff	**送货/delivery**	朋友/friends
			速度/speed	**便宜/cheap**	应该/should	鼠标/mouse	**收到/receive**
		T1	电脑/computer	本来/originally	原装/innate	**服务/service**	选择/select
			声音/voice	**风扇/fan**	**温度/temperature**	**发热量/calorific value**	**散热/cooling**
			硬盘/hard disk	接受/accept	开机/starting up	**噪音/ noise**	**发热/heat**
		T2	感觉/feeling	确实/indeed	**运行/operation**	控制/control	触摸/touch
	Negative		**京东/jingdong**	**服务/service**	问题/service	电话/phone	人员/staff
			快递/expressage	发现/find	**订单/order**	**发货/shipments**	检测/detection
		T1	希望/hope	根本/fundamental	时间/time	告诉/tell	**垃圾/rubbish**
WS-TSWE	Positive		**散热/cooling**	**风扇/fan**	**不错/good**	**声音/voice**	**安静/quietness**
			温度/temperature	**完美/perfect**	**散热器/radiator**	**做工/workmanship**	**喜欢/like**
		T2	**漂亮/nice**	**运行/operation**	游戏/game	**合适/suitable**	效果/effect
			京东/jingdong	**速度/speed**	沉稳/calm	**送货/delivery**	效果/effect
			服务/service	要求/demand	**满意/satisfaction**	**发货/shipment**	**很快/very fast**
		T1	**稳定/stabilize**	**态度/attitude**	特别/especially	**收到/receive**	**便宜/cheap**
			散热/cooling	**风扇/fan**	**声音/ voice**	**温度/ temperature**	**一般/general**
			不好/bad	**噪音/noise**	**散热器/ radiator**	**发热量/calorific value**	机器/machine
	Negative	T2	**运行/operation**	**发热/heat**	游戏/game	硬盘/hard disk	效果/effect
			京东/jingdong	**速度/speed**	**一般/general**	**发货/shipments**	**服务/service**
			物流/logistics	**订单/order**	**快递/expressage**	**送货/delivery**	有待/remain
			电话/phone	实在/indeed	**不好/bad**	**不足/insufficient**	**垃圾/rubbish**

teresting to capture the sentiment polarity and extract the topics, we also perform the experiments that not excluding the words that not in the trained word embeddings. We call this mode as WS-TSWE'. Because these words don't have word embeddings, the words are trained completely based on the Dirichlet multinomial component.

According Fig. 4, the perplexity begins to be stable when $K = 40$. So we list the NMI, accuracy, precision, recall, and F1 at $K = 40$ on four data sets in Table 5 to compare different models further. We can see that in most cases the WS-TSWE' and WS-TSWE can get the best two results in NMI, accuracy and F1. DSLDA also get the best result in F1 of negative sentiment on computer data set. JSTH get a second-best result in F1 of negative sentiment on hotel data set. Furthermore, we can find the WS-TSWE' perform better than all others' results in NMI and accuracy, which prove that the excluding words have an impact on the topic identification and classification accuracy.

4.3.2.4. Extracted topics. A topic is multinomial distribution over words based on both topics and sentiments. The top words (most probable words) for each distribution could approximately reflect the meaning of the topic.Table 6 shows the selected examples of global topics extracted from computer data set with JST and WS-TSWE. Each row shows the top 15 words for corresponding two topics (T1 and T2). The purpose of the table is to demonstrate the model can well identify the topic words and sentiment words under different sentiment. We list the top words under two topics in Table 6. We can obtain that the two topics are about two different aspect topics. T1 is about the Heat-dissipation of the computer and T2 is about the transport logistics. Our model can recognize more sentiment words and topic words than JST. We can find that only a few words like "cooling, design, memory" are about the topics of the computer Heat-dissipation problem in the JST model. There are only four emotional words such as "nice, appearance, like, perfect", and some redundant words such as "appearance", "piano", "keyboard" in the JST model. But in our WS-TSWE, we can see that more words such as "cooling, fan, radiator, voice, temperature, workmanship, operation" are about the computer Heat-dissipation problem, and more words such as "good, quietness, perfect, like, nice, suitable" are the emotional tendencies of the computer Heat-dissipation problem. It shows that WS-TSWE can extract topic and sentiment simultaneously. Overall, the above analysis demonstrates the effectiveness of WS-TSWE in extracting opinionated topics under sentiment from a corpus.

5. Conclusions and future work

In this paper, we propose a novel weakly supervised generative model (WS-TSWE) for jointly mining sentiments, topics and their associations from online reviews. In the WS-TSWE model, we incorporates word embeddings, which can be trained on very large external corpora, and HowNet lexicon to train sentiment orientation of each word, and explored how to take advantage of both word embeddings component and Dirichlet multinomial component. We implement the WS-TSWE model with Gibbs sampling algorithm, and do experiments on four real online review data sets (book, hotel, computer, and movie) for two kinds of language (English and Chinese). We compare our model with different sentiment initialization modes, such as without predefined sentiment polarity dictionary (TSWE−P), with predefined sentiment polarity dictionary (TSWE+P), and with HowNet lexicon (WS-TSWE), and reserved the words that not found in Google vector representations trained from Chinese Wikipedia corpus and English Wikipedia corpus (WS-TSWE'). We also compare with several other recent models, such as JST, HTSM, DSLDA, LDA and LFLDA. The experimental results show that WS-TSWE is effective in discovering sentiment and extracting sentiment-topics.

In our experiments, we find DSLDA which consider the dependency of the words can get good results on some data sets. In the future work, we consider to incorporate the sequence information of words. For example, we can introduce the dependency among word embeddings by recurrent neural network. It can help to extend the semantic information and redefines the topic sentiment-word distribution. Also, we consider to provide fast variational inference algorithm to process larger corpus.

Acknowledgment

This research is supported by the National Nature Science Foundation of Chinaunder Grant nos. 61472258 and 61772345, the Major Fundamental Research Project in the Science and Technology Plan of Shenzhenunder Grant no. JCYJ20160310095523765.

References

[1] D. Blei, M.I. Jordan, Variational inference for Dirichlet process mixtures, Bayesian Anal. 1 (2006) 121–144.
[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] Z. Cao, S. Li, Y. Liu, W. Li, H. Ji, in: A Novel Neural Topic Model and Its Supervised Extension, AAAI. Publishing, 2015, pp. 2210–2216.
[4] Z. Chen, C. Li, J.-T. Sun, J. Zhang, C. Li, J. Zhang, J.-T. Sun, Z. Chen, in: Sentiment Topic Model With Decomposed Prior, SDM, 2013, pp. 767–775.
[5] R. Das, M. Zaheer, C. Dyer, Gaussian LDA for topic models with word embeddings, in: Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics, 2015, pp. 795–804.
[6] M. Dermouche, L. Kouas, J. Velcin, S. Loudcher, A joint model for topic-sentiment modeling from text, in: Proceedings of the 30th Annual ACM Symposium on Applied Computing, 2015, pp. 819–824.
[7] X. Fu, L. Guo, Y. Guo, Z. Wang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowl. Based Syst. 37 (2013) 186–195.
[8] X. Fu, W. Liu, Y. Xu, L. Cui, Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis, Neurocomputing 241 (2017) 18–27.
[9] X. Fu, K. Yang, J.Z. Huang, L. Cui, Dynamic non-parametric joint sentiment topic mixture model, Knowl. Based Syst. 82 (2015) 102–114.
[10] M. Goto, A predominant-F 0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models, in: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2001, pp. 3365–3368.
[11] M.D. Hoffman, D.M. Blei, C. Wang, J. Paisley, Stochastic variational inference, J. Mach. Learn. Res. 14 (2013) 1303–1347.
[12] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 50–57.
[13] Y. Jo, A.H. Oh, Aspect and sentiment unification model for online review analysis, in: Proceedings of the Forth International Conference on Web Search and Web Data Mining (WSDM), Hong Kong, China, 2011, pp. 815–824.
[14] M.I. Jordan, Graphical models, Stat. Sci. 19 (2004) 140–155.
[15] Z. Lai, W.K. Wong, Y. Xu, J. Yang, Approximate orthogonal sparse embedding for dimensionality reduction, IEEE Trans. Neural Netw. Learn. Syst. 27 (2015) 723–735.
[16] C. Li, J. Zhang, J. Sun, Z. Chen, Sentiment topic model with decomposed prior, Soc. Ind. Appl. Math. (2013) 767–775.
[17] F. Li, M. Huang, X. Zhu, Sentiment analysis with global topics and local dependency, in: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010, pp. 1371–1376.
[18] J. Li, J. Li, X. Fu, M.A. Masud, J.Z. Huang, Learning distributed word representation with multi-contextual mixed embedding, Knowl. Based Syst. 106 (2016) 220–230.
[19] C. Lin, Y. He, Joint sentiment/topic model for sentiment analysis, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 375–384.
[20] C. Lin, Y. He, R. Everson, S. Rüger, Weakly supervised joint sentiment-topic detection from text, Knowl. Data Eng. IEEE Trans. 24 (2012) 1134–1145.
[21] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Math. Program. 45 (1989) 503–528.
[22] Y. Liu, Z. Liu, T.-S. Chua, M. Sun, in: Topical Word Embeddings, AAAI, 2015, pp. 2418–2424.
[23] C. Ma, W. Xu, P. Li, Y. Yan, Distributional representations of words for short text classification, in: Proceedings of NAACL-HLT, 2015, pp. 33–38.
[24] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, 1, Cambridge university press, Cambridge, 2008.
[25] Q. Mei, X. Ling, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 171–180.
[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of Workshop at (ICLR), 2013, pp. 1–12.
[27] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in neural information processing systems, 2013, pp. 3111–3119.
[28] D.Q. Nguyen, R. Billingsley, L. Du, M. Johnson, Improving topic models with latent feature word representations, Trans. Assoc. Comput. Linguist. 3 (2015) 299–313.
[29] M.J. Paul, M. Dredze, SPRITE: Generalizing topic models with structured priors, Trans. Assoc. Comput. Linguist. 3 (2015) 43–57.
[30] R. Pavitra, P. Kalaivaani, Weakly supervised sentiment analysis using joint sentiment topic detection with bigrams, in: Proceedings of the 2nd International Conference on Electronics and Communication Systems (ICECS), 2015, pp. 889–893.
[31] J. Petterson, W. Buntine, S.M. Narayanamurthy, T.S. Caetano, A.J. Smola, Word features for latent Dirichlet allocation, Adv.Neural Inf. Process. Syst. 23 (2010) 1921–1929.
[32] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Neuyen, S. Horiguchi, Q.-T. Ha, A hidden topic-based framework toward building applications with short web documents, IEEE Trans. Knowl. Data Eng. 23 (2011) 961–976.
[33] M.M. Rahman, H. Wang, Hidden topic sentiment model, in: Proceedings of the International Conference on World Wide Web, 2016, pp. 155–165.
[34] V.K.R. Sridhar, Unsupervised topic modeling for short texts using distributed representations of words, in: Proceedings of NAACL-HLT, 2015, pp. 192–200.
[35] M. Steyvers, T. Griffiths, Probabilistic topic models, Handb. Latent Semant. Anal. 427 (2007) 424–440.
[36] J. Turian, L. Ratinov, Y. Bengio, Word representations: a simple and general method for semi-supervised learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 384–394.
[37] M. Wainwright, M. Jordan, Graphical models, exponential families, and variational inference, Foundations and Trends in Machine Learning 1 (1-2) (2008) 1–305.
[38] Y. Wilks, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2016) 102–107.
[39] J. Xuan, J. Lu, G. Zhang, R.Y.D. Xu, X. Luo, Bayesian nonparametric relational topic model through dependent gamma processes, IEEE Trans. Knowl. Data Eng. 29 (2017) 1357–1369.
[40] Y. Zhang, G. Zhang, H. Chen, J. Lu, Topical analysis and forecasting for science, technology and innovation: methodology and a case study focusing on big data research, Technol. Forecast. Social Change 105 (2016) 179–191.
[41] Y.L. Zhu, J. Min, Y.Q. Zhou, X.J. Huang, W.U. Li-De, Semantic orientation computing based on HowNet, J. Chin. Inf. Process. 1 (2006) 14–24.