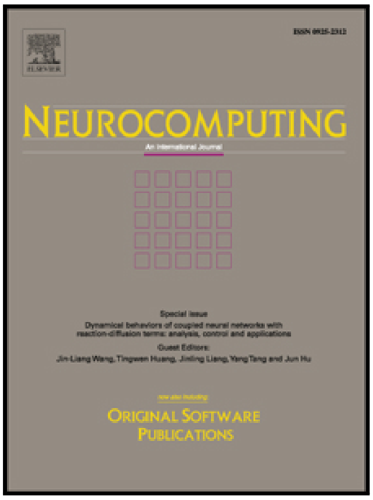
Communicated by Dr. Qi He



A Short Text Sentiment-Topic Model for Product Reviews

Shufeng Xiong, Kuiyi Wang, Donghong Ji, Bingkun Wang

PII:	S0925-2312(18)30169-3
DOI:	10.1016/j.neucom.2018.02.034
Reference:	NEUCOM 19324
To appear in:	*Neurocomputing*
Received date:	10 August 2017
Revised date:	16 December 2017
Accepted date:	8 February 2018

A Short Text Sentiment-Topic Model for Product Reviews

Shufeng Xiong[a], Kuiyi Wang[a], Donghong Ji[b], Bingkun Wang[a]

xsf@whu.edu.cn

[a]*Pingdingshan University*
[b]*Computer School, Wuhan University*

Abstract

Topic and sentiment joint modelling has been successfully used in sentiment analysis for product reviews. However, the problem of text sparse is universal with the widespread smart devices and the shorter product reviews. In this paper, we propose a joint sentiment-topic model WSTM (Word-pair Sentiment-Topic Model) for the short text reviews, detecting sentiments and topics simultaneously from the text, especially considering the text sparse problem. Unlike other topic models modelling the generative process of each document, our directly models the generation of the word-pair set from the whole global corpus. In the generative process of WSTM, all of the words in a sentence have the same sentiment polarity, and two words in a word-pair have the same topic. We apply WSTM to two real-life Chinese product review datasets to verify its performance. In three experiments, compared with the existing approaches, the results demonstrate WSTM is quantitatively effective on both topic discovery and document level sentiment.

1. Introduction

Online product reviews mining aims to determine the attitude of the reviewers about some topic or the polarity (positive, negative and neutral) at the document, sentence, or aspect level. In recent year, it has attracted more and more attentions [19, 18, 4, 10, 17, 28]. Especially in online shopping, one cannot touch nor check the quality by hand, the reviews from other consumers are more essential for decision making.

For product reviews, the descriptions of specific aspects of one product and the polarity of each aspect are quite informative. However, the same sentiment expression will bring different polarity in different topics. As shown in Figure 1,

$R1$: *外观不错，速度可以，键盘舒服，发热**小**，电池时间长,送货快,性价比高。*

Nice appearance, better running speed, keyboard is comfortable, **small** heat radiation, long battery life, fast delivery, high cost performance.

$R2$: 内存**小**！驱动不全！

The memory is **small**! The driver is incomplete!

Figure 1: Two reviews from notebook reviews corpus. For the same expression "small", its polarity is positive in $R1$ but negative in $R2$. In short review $R1$, there are 7 topics (underlined words) and traditional sampling method in common topic model will suffer from the data sparsity problem.

the word "小(small)" has opposite polarity when used to describe different topics "发热(heat)" and "内存(memory)".

Taking advantage of the relation between topic and sentiment, some work exploited unsupervised (weakly supervised) topic modelling to address this issue [11, 8, 15]. Although most of the previous studies [31, 27, 20, 35] had effective results by using supervised learning methods. Yet they relied on labelled corpus, demanding an expensive manual labour. Therefore, our work concentrates on unsupervised topic learning and polarity classification of document level (although our method could apply in fine-grain level with a small expand).

However, we have to face the text sparse problem in product reviews when jointly modelling sentiment and topic. Actually, most of the product reviews have a distinct viewpoint and expressed shortly, especially those online shopping site in Chinese. [48] has pointed out that the length of product reviews is getting shorter with the rapid spread of smart devices such as smart-phones and tablets. As shown in Figure 1, in review $R1$, the reviewer only used 27 Chinese characters to express his/her attitudes for 7 aspects (topics).

For the text sparse problem in topic model, some studies generated lengthy pseudo-documents by aggregating short texts before training [40, 9]. Another way of modelling is based on the assumption that a short text only covers a single topic [49, 7]. Recently, [43] proposed to direct model word-pair co-occurrence process. All these methods only model the latent topic in short text. In our model, we jointly detect sentiment and topic simultaneously in a global word-pairs generative process.

In this paper, we focus on document-level sentiment classification and topic modelling based on the proposed weakly supervised[1] Word-pair Sentiment-Topic Model (WSTM). Our model is a probabilistic mixture model, adapting the latent Dirichlet allocation (LDA) model [1] to learn sentiment and topic over short review text by directly modelling the generation of word-pairs in global scale. A word-pair is two unordered co-occurrence words in a context.

Specifically, the whole corpus is regarded as a bag of co-occurred word-pairs. And the corpus is generated by sampling each word-pair from a mixture model, which involving a set of topic language models and sentiment language models. By learning WSTM, we obtain the sentiment-topic components and a global sentiment-topic distribution of the corpus. And then, we derive the sentiment and topic distribution of each document based on the learned model. We evaluate our approach on two product review datasets. Experimental results show that WSTM accurately discover topics and further identify the polarity of reviewers than baseline methods.

The major contribution of this paper can be summarized as follows:

- We propose a short-text sentiment-topic model to identify topics and sentiments simultaneously for product reviews (Section 3.2).
- We adopt Gibbs sampling to resolve the parameters inference (Section 3.3).
- We provide an effective way to estimate the topic and sentiment distribution in a document which cannot be directly obtained during the learning process (Section 3.4).
- Through experiments on two datasets, we demonstrate the effectiveness of our model in simultaneously detecting topics and sentiments of short reviews (Section 5).

2. Related work

In the previous section, we have discussed that topic (aspect) and sentiment are the main information cared by the customers. Therefore, it is a natural solution to joint model topic and sentiment in a probability mixture model.

After [1] proposing the topic model, Latent Dirichlet Allocation(LDA), there are many studies to address specific problems based on it. [14, 21, 22, 39] improved the performance of topic model for webblogs by leveraging the characters of webblogs. [16, 30] proposed specific topic model for the short text. [36, 37, 50]

[1] In our model, we use sentiment lexicon as prior knowledge, so we call it “weakly supervised”.

modelled the aspects of product reviews by exploiting topic model. [45, 47, 51, 41] combined topic model with other methods / properties of tasks to tackle different problems. [32, 52, 5, 44] exploited improved topic model to solve some special tasks, such as consensus topic identification, expert finding and topic tracking.

For the joint model for topic-sentiment modelling, there are some representative researches [34, 33, 3, 23, 26, 13, 25, 46]. In [24], the authors had listed some distinctive features used to distinguish these methods. These features are as follows:

- Using latent parameter modelling words / using different parameters modelling words and stars separately.
- Model all of the words / only model opinion expressions.
- Model the dependency between aspect and stars / not consider the dependency.
- Only use review corpus / use additional data.

Because of the first two features belong to intrinsic property, and the other two are about external knowledge and data requiring human intervention. According to the latter features, our method belongs to the category of modelling all words independently and not using additional input data. Under this classification standard, some representative researches are closely related to our method, as briefly introduced as follows.

Joint Sentiment-Topic Model (JST) It is a model based on the three hierarchical layers of LDA. In JST, it added an additional sentiment layer between the document and the topic layers [15]. In its four-layer model, sentiment labels are associated with documents and topics are associated with sentiment labels, and words are associated with both sentiment labels and topics.

Aspect and Sentiment Unification Model (ASUM) This model was proposed by [11], it had four layers like JST. The difference between JST and ASUM is that ASUM constrained the words in a single sentence to the same language model, while JST allowed individual word from different language models. In our model, we only constrain two words in a word-pair to the same language model.

Senti-Topic model with Decomposed Prior (STDP) In this model [12], it decomposed the generative process of sentiment labels into a hierarchy of two levels. They firstly determined a word as sentiment word or topic word, if it is a sentiment word, then identify its polarity.. In our model, we argued that sentiment label is determined by both sentiment and topic words (as discussed in Section 3.2). Another difference is that STDP required a prior rule by manually generating to discriminate sentiment words and topic words. At the same time, the prior discriminate rule will not fit all domains and cross-language scene (e.g. English and Chinese). In our model, we were trying to minimize the supervised behaviour. Therefore our model did not use any rules except a public available sentiment lexicon.

All of the three models mentioned above focus on lengthy reviews such as movie and restaurant reviews (please check original paper for details)[2]. Without specially considering the sparse problem in short reviews, a model has not enough word counts to effectively find topic relevance among words. This issue further harms the identification of sentiment labels. We overcame the drawback in this paper by modelling the corpus level word-pair set generative process, which is similar to BTM model [43]. The difference is we jointly detect sentiments and topics in a mixture model, while BTM only considers topic.

Recently, some studies on topic modelling have been aware of the spare problem in the short text [38, 29, 9, 40]. But them only modeled the topic information without considering the sentiment. In addition, [38] and [29] were designed for modeling topical trends over time and text classification, respectively. [9] and [40] were used for modeling Twitter message.

3. Our approach

In this section, we propose the Word-pair Sentiment Topic Model (WSTM), which learns topics and sentiment labels simultaneously over the short text by directly modelling the generation of the whole corpus. In WSTM, we adopt Gibbs sampling to infer the parameter. Thereafter, we estimate the sentiment label and topic in a document which are not directly inferred in the generative process.

3.1. Representation of the opinionated reviews

Intuitively, sentiment label of a word is determined by sentiment word and its context. As shown in Figure 1, the sentiment label of "小(small)" is *positive* when it appears in a word-pair ⟨发热(heat), 小(small)⟩ or *negative* when appears in ⟨内存(memory), 小(small)⟩. While reviews are not lengthy, the topic-sentiment model will face the text sparse problem. As pointed out in [43], it is an effective method to learn topic by directly using global word-pair co-occurrence pattern. Since word-pair can be used to identify sentiment label, we can learn both sentiment

[2] Although, [12] performed experiments on two short review sets, the STDP model itself did not take into account text sparse problem.

Table 1: Meanings of the notations.

Symbol	Description
D	number of reviews
M	number of word-pairs
T	number of topics
S	number of sentiment labels
V	the vocabulary size
b	word-pair, $b = (w_i, w_j)$
w	word
z	topic
l	sentiment label
$\pi_{k,l}$	distribution over topic k and sentiment l
Π	multinomial distribution over sentiment
$\phi_{k,l,w}$	distribution of word w over topic k and sentiment l
Φ	multinomial distribution over words
θ_k	distribution of topic k
Θ	multinomial distribution over topics
α	Dirichlet priors vector for θ
β	asymmetric Dirichlet priors for ϕ, $\beta = \{\{\{\beta_{z,l,i}\}_{k=1}^{T}\}_{l=1}^{S}\}_{i=1}^{V}$
γ	Dirichlet priors vector for π
z_t	the t-th word's topic
l_t	the t-th word's sentiment label
$\mathbf{B}$	the word-pair set
$\mathbf{z_{-t}}$	the topic assignment for all words except the t-th word
$\mathbf{l_{-t}}$	the sentiment label assignment for all words except the t-th word
$N_{k,l,i}$	number of word w_i in topic k and sentiment l
$N_{k,l}$	number of words in topic k and sentiment l
N_k	number of words in topic k
$N_{.,-t}^{(.)}$	counts without the t-th word

and topic from word-pair in an unified frame. In other word, WSTM models the generative process of all the word-pairs in the whole corpus instead of each document.

In WSTM, the first step is to represent D reviews as a bag of word-pairs $\mathbf{B} = \{b_m\}_{m=1}^{M}$ with $b_m = (w_i^m, w_j^m)$. A word pair b is unordered and composed by two co-occurred words w_i and w_j in a review R, where $|i - j| < t$, w_i denotes the ith words in R, and, t is a defined window size. In our experiments, the window size is 10. For example, in review $R2$ "内存小！驱动不全！(The memory is small! The driver is incomplete!)", there are six word-pairs, $\langle$内存(memory), 小(small)$\rangle$, $\langle$内存(memory), 驱动(driver)$\rangle$, $\langle$内存(memory), 不全(incomplete)$\rangle$, $\langle$小(small), 驱动(driver)$\rangle$, $\langle$小(small), 不全(incomplete)$\rangle$ and $\langle$驱动(driver), 不全(incomplete)$\rangle$. Each word-pair will be assigned the same sentiment and topic in generative process. We extract all word-pairs from the corpus to form a bag of word-pairs used to represent the opinionated reviews.

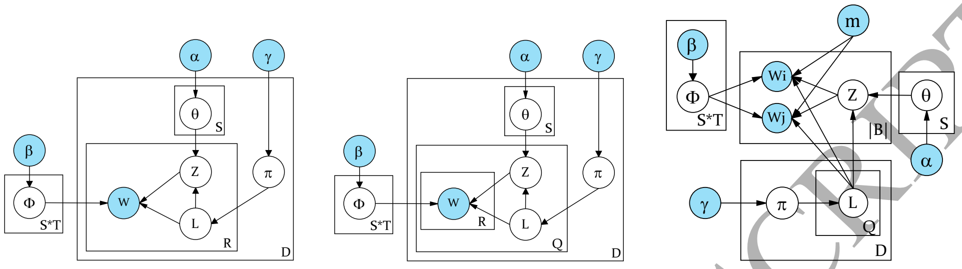
The same as other window-based methods, it may bring dirty data. Obviously, some word-pairs, such as $\langle$小(small), 驱动(driver)$\rangle$, $\langle$小(small) and 不全(incomplete)$\rangle$, should not be in the same sentiment and topic. In this case, there is a reasonable explanation that the frequency of these word-pairs is smaller than that of the other word-pairs when counting word-pair co-occurrence in the whole corpus. And that means counting global co-occurrence can reduce the negative impact from dirty data. Actually, except for [43], ASUM model [11] also used a similar window-based approach. Both of the two methods have achieved good performance. Especially in ASUM, it assumed that all of the words in the range of one sentence come from the same sentiment and topic. In our method, we relax the assumption in [11] by narrowing the range from whole sentence down to a word-pair.

3.2. Word-pair sentiment topic model

WSTM simulates the generative process of whole reviews written by all users. The process for one review is shown in the following scenario:

1. For each sentence, a reviewer firstly decides the distribution of sentiments. For example, suppose that the sentiments are 80% positive and 20% negative.
2. After determining the sentiment, she writes a subjective review about a notebook under a distribution of topic (aspect), for example, 30% about the memory, 30% about the appearance, 20% about the running speed and 20% about the battery.

3. Then she chooses some word-pairs to express her opinions under the identified topic and sentiment.



(a) JST (b) ASUM (c) WSTM

Figure 2: Graphical representation of (a) JST, (b) ASUM, (c) WSTM. Different from JST and ASUM: 1) WSTM does not model document generative process; 2) WSTM generates word-pair (w_i, w_j) under the same topic and sentiment, while ASUM generates all words (w) in the same sentence under the same topic and sentiment and JST generates word under separate topic and sentiment; 3) In WSTM, topic is dependent on sentiment. Q denotes the number of sentence in a document, R denotes the number of words in a sentence.

Given a corpus C containing D documents $d \in \{1, 2, 3, ..., D\}$; we have extracted word-pair set $B = \{b_1, b_2, ..., b_M\}$ from C using the method described in Section 3.1. WSTM models the generative process of B, then provides a subsequent approach to estimate the sentiment-topic distribution of each document d. Assume that there are S sentiment labels indexed by $S = \{1, 2, ..., S\}$. For each sentiment label l, there are T topics associated with it. Suppose α, β and γ are the Dirichlet priors. The formal definition of the generative process corresponding to the graphical model shown in Figure 2 (c) is as follows:

- For each document d,
  - * Draw a distribution $\pi_d \sim Dir(\gamma_k)$
  - * For each sentiment label l, draw a topic distribution $\theta \sim Dir(\alpha)$
  - * For each sentence M,
    - - Draw a sentiment label $l \sim Mult(\pi_d)$
    - - For the sentiment label l, draw a topic $z \sim Mult(\theta_l)$
- For each word-pair $b \in B$
  - * Assign its sentiment label l according to the sentence M where $b \in M$

* Draw a topic $z \sim Mult(\theta_l)$

* Draw two words: $w_i, w_j \sim Mult(\phi_{l,z})$

In WSTM, it must provide proper prior to tell the model what is the “true” human sentiments. In other words, it should set up correspondence relationship between the modelled sentiments and the human sentiments. We use sentiment lexicon to inject sentiment information. Specifically, it is implemented by asymmetric prior parameter β. If a word w_i is found in sentiment lexicon, then the Dirichlet prior about its sentiment distribution is

$$\beta_{w_i,l} = P_{wi}(l) * \beta_0, \tag{1}$$

where $P(l)$ is a predefined probability for a word in sentiment lexicon over sentiment label l, and β_0 is the basic factor (e.g. 0.05). Assume there are three polarities: *neutral, positive and negative*, and w_i has a positive polarity in sentiment lexicon, we define the probability as $P_{wi}(neutral) = 0.009$, $P_{wi}(positive) = 0.99$ and $P_{wi}(negative) = 0.001$. It means w_i has 99% chance as a positive polarity in corpus, 0.9% chance as neutral and 0.1% chance as negative.

3.3. Model inference

In order to estimate the parameters Π, Θ and Φ in WSTM, it needs to evaluate the posterior distribution $P(z, l|B)$, i.e., the probability of topic z and sentiment label l over word-pair set B. This distribution is difficult to evaluate directly, so we adopt Gibbs sampling to perform approximate inference. The full conditional distribution for Gibbs sampler is:

$$P(l_t = l, z_t = k|\mathbf{B}, \mathbf{z_{-t}}, \mathbf{l_{-t}}) \propto \frac{P(\mathbf{B}|\mathbf{z},\mathbf{l})}{P(\mathbf{B_{-t}}|\mathbf{z_{-t}},\mathbf{l_{-t}})} \cdot \frac{P(\mathbf{z}|\mathbf{l})}{P(\mathbf{z_{-t}}|\mathbf{l_{-t}})} \cdot \frac{P(\mathbf{l})}{P(\mathbf{l_{-t}})}, \tag{2}$$

For the numerator in the first term, by integrating out ϕ, we obtain:

$$P(\mathbf{B}|\mathbf{l},\mathbf{z}) = \prod_{l=1}^{S}\prod_{k=1}^{T} \frac{\Gamma(V\beta_{k,l})}{\Gamma(\beta_{k,l})^V} \frac{\prod_{i=1}^{V}\Gamma(N_{k,l,i} + \beta_{k,l,i})}{\Gamma(\sum_{i=1}^{V} N_{k,l,i} + V\beta_{k,l})}, \tag{3}$$

For the numerator in the second term, we can integrate out π:

$$P(\mathbf{z}|\mathbf{l}) = \prod_{l=1}^{S} \frac{\Gamma(T\alpha_l)}{\Gamma(\alpha_l)^T} \frac{\prod_{k=1}^{T}\Gamma(N_{k,l} + \alpha_{k,l})}{\Gamma(\sum_{k=1}^{T} N_{k,l} + T\alpha_l)}, \tag{4}$$

For the numerator in the third term, by integrating out θ, we obtain:

$$P(\mathbf{l}) = \prod_{d=1}^{D} \frac{\Gamma(S\gamma)}{\Gamma(\gamma_l)^S} \frac{\prod_{l=1}^{S} \Gamma(N'_{d,l} + \gamma_l)}{\Gamma(\sum_{l=1}^{S} N'_{d,l} + S\gamma)}, \tag{5}$$

And the three denominator can be worked out in the same way:

$$P(\mathbf{B_{-t}}|\mathbf{l_{-t}}, \mathbf{z_{-t}}) = \prod_{l=1}^{S} \prod_{k=1}^{T} \frac{\Gamma(V\beta_{k,l})}{\Gamma(\beta_{k,l})^V} \frac{\prod_{i=1}^{V} \Gamma(\{N_{k,l,i}\}_{-t} + \beta_{k,l,i})}{\Gamma(\sum_{i=1}^{V} \{N_{k,l,i}\}_{-t} + V\beta_{k,l})}, \tag{6}$$

$$P(\mathbf{z_{-t}}|\mathbf{l_{-t}}) = \prod_{l=1}^{S} \frac{\Gamma(T\alpha_l)}{\Gamma(\alpha_l)^T} \frac{\prod_{k=1}^{T} \Gamma(\{N_{k,l}\}_{-t} + \alpha_{k,l})}{\Gamma(\sum_{k=1}^{T} \{N_{k,l}\}_{-t} + T\alpha_l)}, \tag{7}$$

$$P(\mathbf{l_{-t}}) = \prod_{d=1}^{D} \frac{\Gamma(S\gamma)}{\Gamma(\gamma_l)^S} \frac{\prod_{l=1}^{S} \Gamma(\{N'_{d,l}\}_{-t} + \gamma_l)}{\Gamma(\sum_{l=1}^{S} \{N'_{d,l}\}_{-t} + S\gamma)}, \tag{8}$$

By replacing terms in Eq.(2) with those in Eqs.(3-8), we have the conditional distribution probability in each iteration of Gibbs sampling:

$$\begin{aligned} &P(l_t = l, z_t = k|\mathbf{B}, \mathbf{l_{-t}}, \mathbf{z_{-t}}) \propto \\ &\frac{(\{N_{k,l,w_{i,1}}\}_{-t} + \beta)(\{N_{k,l,w_{i,2}}\}_{-t} + \beta)}{(\{N_{k,l}\}_{-t} + V\beta + 1)(\{N_{k,l}\}_{-t} + V\beta)} \frac{(\{N_{k,l}\}_{-t} + \alpha_{k,l})}{(\{N_k\}_{-t} + T\alpha_l)} \frac{(\{N'_{d,l}\}_{-t} + \gamma_l)}{(\{N'_d\}_{-t} + S\gamma)}, \end{aligned} \tag{9}$$

The number of Gibbs sampling iterations is 1000 in our experiments.

Given the hyperparameters α, β and γ, word-pair set B and their topic assignments z, sentiment label l, we can derive the probability of the parameters Φ, θ and π by utilizing the Bayes's rule and Dirichlet-multinomial conjugate property:

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

$$\phi_{k,l,i} = \frac{N_{k,l,i} + \beta}{N_{k,l} + V\beta}, \tag{11}$$

$$\pi_{d,l} = \frac{N'_{d,l} + \gamma_l}{N'_d + S\gamma_l}, \tag{12}$$

3.4. Inferring sentiment and topic in a document

To address the text sparse problem, WSTM does not model the document generation process. Therefore, we must provide an approximate estimation of the sentiment and topic in a document as an extension step. We define the document-level sentiment as:

$$L_d = \arg\max_{l \in L} \pi_{d,l}, \tag{13}$$

Since the generative process is based on word-pair, we estimate the topic z of word w_i in document d based on the learnt parameters:

$$P(z|w_i) = \frac{\sum_B P(z|b)P(w_i|b)}{\sum_B P(w_i|b)}, \tag{14}$$

where:

$$P(z|b) = \frac{\sum_l P_d(l)P(z|l)P(w_i|l,z)P(w_j|l,z)}{\sum_z(\sum_l P_d(l)P(z|l)P(w_i|l,z)P(w_j|l,z))}, \tag{15}$$

where $P_d(l) = \pi_{d,l}$, $P(z|l) = \theta_{k,l}$ and $P(w_i|l,z) = \phi_{k,l,i}$. Similarly, the probability of word w with sentiment label l can be calculated using:

$$P(l|w_i) = \frac{\sum_M P_d(l)P(w_i|m)}{\sum_M P(w_i|m)}, \tag{16}$$

where $P(w_i|m)$ is the probability of word w_i appeared in sentence m.

Although it is a straightforward way, we find the experimental results are always better. More sophisticated ways need to be further studied.

4. Experimental setup

4.1. Datasets

We used two sets of reviews from two different online shopping sites. One dataset is a collection of laptop reviews from JINGDONG [3] and another one is mobile phone reviews from IT168 [4]. All the product reviews were tagged by their authors as either positive or negative in the two e-commence websites. After performing Chinese word segmentation, we remove punctuation, numbers and stop words as the pre-process. The statistics of pre-processed datasets are described in Table 2. In our experiments, we used a 50% random split for tuning parameters and 50% for testing. The datasets are available on `https://github.com/pdsujnow/WSTM`.

[3] http://www.jd.com

[4] http://product.it168.com

Table 2: Statistics of the review corpus. # donotes the size.

	LAPTOP	**MOBILE**
Avg.# words/review	20	32
# reviews	3988	2289
Vocab.size	7964	8787
# positive review	1993	1146
# negative review	1995	1143

4.2. Sentiment lexicon

Since sentiment lexicon is a necessary knowledge for discriminating sentiment, we incorporated HowNet[5] sentiment lexicon to our model as sentiment prior. HowNet sentiment lexicon includes approximately 5000 positive and 5000 negative segmented Chinese words. As introduced in Section 3.2, we used sentiment lexicon to affect the prior β.

4.3. Comparison partners and parameter settings

In order to evaluate the performance of WSTM model, we designed three tasks: topic discovery, sentiment-specific topic discovery and document level sentiment identification. We provided two methods for two topic discovery experiments: **BTM** and **LDA**. The former is a topic model for the short text and the later for the common text. Meanwhile, we provided three methods for sentiment identification task: The **baseline** method classifies each document according to Eq. (13), where the word sentiment label is directly obtained from sentiment lexicon. Both **JST** and **ASUM** have been introduced in section 2.

In our experiments, we used the same hyper-parameter settings for comparison partners as given in their original papers. For WSTM model, we used symmetric α of 0.03 and symmetric γ of 0.02. The parameter β is asymmetric, as discussed in Section 3.2, we set a basic value β_0 as 0.05.

4.4. Evaluation metrics

For topic discovery, we conduct a qualitative evaluation and a quantitative one. The former is to check the founded topic words by topic model. The latter is to compare the topic coherence by using coherence measure (CM) [42], defined

[5] http://www.keenage.com/html/e_index.html

as the ratio between the numbers of relevant words and candidate words. For sentiment-specific topics discovery, we checked the founded sentiment-specific topic words as qualitative evaluation. For document level sentiment identification, we compared the accuracy of sentiment identification of different models.

5. Results and analysis

5.1. Topic discovery

For topic model, extracting topic words is one of the main tasks. Since our model is designed for product reviews analysis, aspect category about the product is naturally treated as a topic. We found that 15 is a better topic number setting for discovering topics and identifying sentiments. Therefore all of our experiment results in this section are based on this setting. Table 3 and Table 4 list the topic words found by topic model for LAPTOP and MOBILE.

For easy understanding, we had manually assigned a label for each topic. We only listed the top 10 (descending order by probability) words in 4 example topics for each dataset. We can see: 1) The listed words of each topic are corresponding well to the aspect category of each product; 2) These words have better internal coherence within the topic. For example, in Table 3, column 2 is possibly related to "battery", the words may be explained like this: the battery can be used for a "long" "time", "endurance" is "good".

However, it seems that there are a few "noise" words in some topic. For example, in column 1 of Table 3, "piano" and "fingerprint" are likely to be irrelevant words, but by doing corpus analysis, we found the reasons. 1) The Chinese word segementer incorrectly divides the word "钢琴烤漆(mirage-black)" into "钢琴(piano)" and "烤漆(painting)". 2) Some reviewers mention that *it is easy to leave fingerprint on the mirage black surface of laptop.* Moreover, there are some adverbs or mood words, such as "hehe" (a common used Chinese mood word that indicates someone is speechless for something) in column 4. It is not difficult to understand, in reviews, some mood words and adverbs are usually used by someone to emphasis. And some words appear in more than one topic, such as "phone" in column 2, 4, 6 of Table 4. These words coming from some global topics, which may be used as a common subtopic of other topics. When using WSTM for fine-grained aspect discovery, it is necessary to filter common topics, we leave it to the future work.

We also performed quantitative evaluation for this task. The measure method for topic model is still an open research problem[2]. Recently, [42] proposed a reasonable method CM(coherence measure) based on human judgement. We used

Table 3: Examples of topics discovered from LAPTOP dataset.

WSTM			**BTM**			**LDA**		
appearance	**battery**	**cooling**	**appearance**	**battery**	**cooling**	**appearance**	**battery**	**cooling**
钢琴	电池	散热	太	电池	散热	容易	电池	好
漂亮	小时	热	容易	时间	好	指纹	小时	散热
指纹	时间	好	指纹	小时	不错	外壳	时间	声音
烤漆	长	温度	键盘	键盘	电池	钢琴	长	风扇
键盘	比较	烫	烤漆	比较	度	烤漆	续航	小
好	续航	CPU	比较	长	热	表面	比较	温度
屏幕	使用	硬盘	不错	好	温度	亮点	使用	热
外壳	好	机器	外壳	不错	声音	感觉	键盘	运行
容易	不错	比较	钢琴	使用	使用	说	小巧	轻
呵呵	小	风扇	屏幕	续航	CPU	屏幕	芯	时
piano	battery	cooling	too	battery	cooling	easy	battery	good
beautiful	hour	hot	easy	time	good	fingerprint	hour	cooling
fingerprint	time	good	fingerprint	hour	good	shell	time	voice
painting	long	temperature	keypad	keypad	battery	piano	long	fans
keypad	comparison	hot	painting	comparison	degree	painting	endurance	small
good	endurance	CPU	comparison	long	hot	surface	comparison	temperature
screen	usage	harddisk	good	fine	temperature	highlights	usage	hot
shell	good	machine	shell	good	voice	felling	keypad	running
easy	good	comparison	piano	usage	usage	talk	smarty	light
hehe	small	fans	screen	endurance	CPU	screen	core	hour

CM method to evaluate our experiments. We invited 4 annotators to judge the first 10 candidate words in each topic. The annotator determined whether a topic was able to abstract out an understandable subject. If not, then the 10 words would be labelled as irrelevant. Or, the annotator judged if each word was relevant to the topic. CM is defined as a ratio between the number of relevant words and the total candidate. For each dataset, we randomly chose 10 topics for evaluation. The results are shown in Table 5 and Table 6. The performance of WSTM is comparable with BTM and ASUM, all of them outperform LDA and JST. The effectiveness on MOBILE dataset is better than on LAPTOP dataset on the whole. It shows the word-pair sampling method in WSTM achieves the same level as BTM, and it tackles the text sparse problem in sentiment-topic model for the short text to some extend. At the same time, WSTM outperforms ASUM, attributed to the word-pair sampling method. By using word-pair sampling method, WSTM constrains the two words in a word-pair to the same topic and sentiment. And, ASUM constrains the words in a single sentence to the same topic and sentiment. The assumption of WSTM is more suitable for product reviews. Take review $R1$ in Figure 1 as an example, it describes seven topics (aspects) about the mobile phone. In ASUM model, all of the words in $R1$ from the same topic, not in accordance with the fact.

Table 4: Examples of topics discovered from MOBILE dataset.

WSTM			**BTM**			**LDA**		
photo	**media**	**screen**	**photo**	**media**	**screen**	**photo**	**media**	**screen**
拍摄	播放	屏幕	像素	MP3	屏幕	效果	支持	屏幕
功能	好	好	摄像头	播放	色	摄像头	MP3	显示
屏幕	速度	不错	拍摄	耳机	显示	像素	播放	比较
支持	不错	感觉	数码	效果	TFT	拍照	内存	色彩
像素	电脑	显示	*手机*	好	效果	照片	蓝牙	色
材质	*手机*	色	支持	音乐	色彩	拍摄	卡	清晰
效果	支持	效果	倍	听	*手机*	拍	格式	高
照片	影音	设计	效果	功能	好	数码	扩展	铃声
拍照	处理器	色彩	相机	不错	26万	相机	文件	方便
摄像头	格式	TFT	拍照	比较	像素	倍	视频	TFT
shoot	play	screen	pixel	MP3	screen	effect	support	screen
function	good	good	camera	play	color	camera	MP3	display
screen	speed	better	shoot	earphone	display	pixel	play	comparison
support	better	felling	digital	effect	TFT	shoot	memory	color
pixel	computer	display	*phone*	good	effect	photo	blue tooth	color
texture	*phone*	color	support	music	color	shoot	slow	clear
effect	support	effect	times	listen	*phone*	shoot	format	high
photo	video	design	effect	function	good	digital	extend	ring
shoot	CPU	color	camera	good	260M	camera	file	convenient
camera	format	TFT	shoot	comparison	pixel	times	video	TFT

WSTM relaxes the assumption in ASUM, allowing the word-pairs generated from one sentence have different topics and sentiments. We will further evaluate the effectiveness of sentiment identification of WSTM.

Table 5: CM(%) on LAPTOP dataset.

	LDA	**BTM**	**JST**	**ASUM**	**WSTM**
annotator1	58	70	57	67	71
annotator2	50	66	50	58	64
annotator3	60	75	64	66	73
annotator4	56	72	52	65	69
average	56	70.75	55.75	64	69.25

Table 6: CM(%) on MOBILE dataset.

	LDA	**BTM**	**JST**	**ASUM**	**WSTM**
annotator1	69	76	65	75	73
annotator2	65	74	61	73	72
annotator3	71	81	73	70	80
annotator4	74	81	71	78	78
average	69.75	78	67.5	74	75.75

5.2. Sentiment-specific topics discovery

The second experiment is to discover sentiment-specific topics. This task is to classify words into topics according to their co-relations and sentiment polarity. The identified topics can be used to find some useful information, i.e. which topics (aspect) are positive or negative? why a topic (aspect) is not optimistic?

As shown in Table 7, we listed four sentiment-specific (two positive and two negative) topics for each dataset. Take the laptop dataset as an example, for the question: why topic “screen” is positive? The listed words give the reason–it use mainstream screen, good colors, suitable for game and eye-protecting. A topic may have both positive and negative comments E.g. workmanship of laptop, it gets positive comments for its brand quality and better sturdy structure, and, negative comments for its keypad with twin adhesive and gaps among components.

5.3. Document level sentiment identification

In this section, we present the quantitative evaluation results of sentiment identification by WSTM. For each dataset, each review has a binary sentiment label (positive and negative). The identification results are presented in Table 8 in terms of accuracy. To ensure a fair comparison, we applied a unified setting of topic number 15 for each sentiment in Table 8. The influence of topic numbers on sentiment identification is reported in Figure 3. It is shown that, although WSTM lightly fluctuates, with the increasing number of topics, it performs better. JST model is relative stable while WSTM and ASUM have some movement, the main reasons are: 1) In JST model, each word may have different topics, the topic number only decides the granularity of topic partition but not affects the sentiment identification; 2) ASUM suppose each word in the same sentence have the same topic, WSTM supposes two words in a word-pair have the same topic. A topic corresponds to the common subject of several words (form word-pair or sentence), therefore both of ASUM and WSTM are sensitive to the topic number.

Table 7: Examples of sentiment-topics discovered by WSTM.

LAPTOP				**MOBILE**			
positive		**negative**		**positive**		**negative**	
screen	**workmanship**	**workmanship**	**service**	**function**	**keypad**	**service**	**speed**
LED	好	双面胶	问题	功能	按键	修	慢
效果	做工	发现	京东	游戏	手感	差	速度
屏幕	感觉	键盘	出现	部分	设计	坏	反应
好	不错	粘	坏	强大	感觉	月	太
屏	键盘	差	本本	内置	好	维修	手机
主流	品牌	明显	屏幕	MP3	操作	售后服务	操作
显示	强	装	退换	支持	容易	换	按键
色彩	质量	螺丝	现象	效果	不错	服务	卡
识别	手感	太	换	压缩	使用	时间	开机
游戏	设计	缝隙	更换	图片	太	使用	机子
LED	good	twin adhesive	problem	function	keypad	mend	slow
effect	workmanship	find	JINGDONG	game	handfeel	poor	speed
screen	feeling	keypad	appear	part	design	bad	response
good	nice	adhere	bad	strong	feeling	month	too
screen	keypad	bad	notebook	built-in	good	repair	phone
mainstream	brand	evident	screen	MP3	operation	service	operation
display	strong	install	exchange	support	easy	change	keypad
color	quality	bolt	phenomenon	effect	nice	service	slow
identify	handfeel	too	change	compression	usage	time	power on
game	design	gap	exchange	picture	too	usage	machine

Table 8: Sentiment identification results.

	LAPTOP	**MOBILE**
baseline	0.637645	0.602188
JST	0.50677	0.53698
ASUM	0.57754	0.43694
WSTM	**0.65503**	**0.65731**

In both two datasets, WSTM outperforms the other methods. We believe that the better sentiment identification result of WSTM is benefited from the better performance of topic discovery. As discussed in [11], JST performs better on movie reviews in the original paper [15] than in the other datasets, but is not doing well on our datasets as well as the datasets of [11]. ASUM is based on an assumption that all of the words in a single sentence come from the same sentiment-specific topic. That is a strong assumption. For short review, it does not sample enough sentences to estimate the parameters. We relax the assumption, WSTM allows the words in a sentence have different topics. In our experiments, the performance of ASUM is even below baseline in MOBILE dataset. It is worth noting that JST and ASUM have opposite effectiveness in two datasets, we argue that this is caused by the text sparse. WSTM outperforms the other methods in both two datasets, demonstrating the effectiveness of our method.

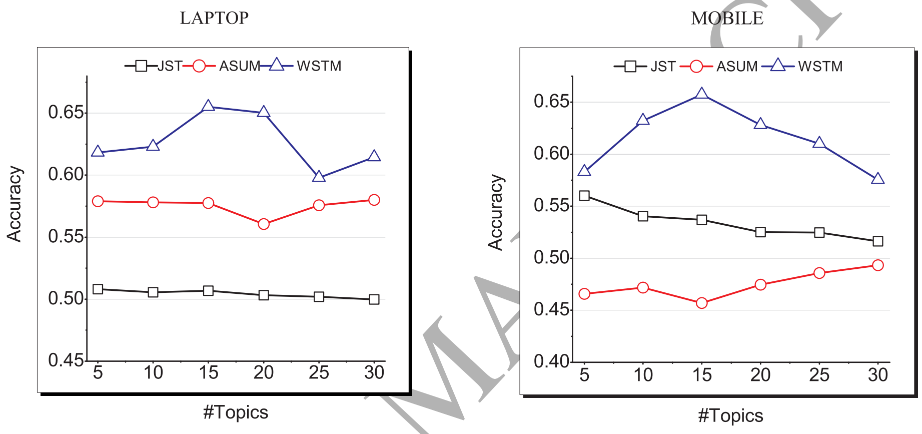


Figure 3: The impact of topic numbers in three topic models.

6. Conclusion

In this paper, we proposed a weakly supervised sentiment-topic model for short text reviews. In this model, we provided a novel method for jointly modelling sentiments and topics, reducing the negative impact of the text sparse problem in product reviews. In our WSTM model, we represented the whole reviews corpus as a bag of word-pairs. By modelling the generative process of the word-pair set, WSTM effectively captures sentiment and topic information that are implicit in words co-occurrence pattern. The experimental results, on two real-world Chinese product reviews datasets, demonstrates that WSTM not only learns higher quality topics, but also accurately identifies the document-level sentiment.

There are still some efforts to be made in the future. For example, we can filter common topics by adding a global topic model. As a generative model, Generative Adversarial Networks [6] provides a novel way to train model. Hence, it is a future direction to use adversarial training in topic model. In addition, a sophisticated approach may be found to estimate the document-level sentiment, while we currently apply a straightforward and effective method. Moreover, we can apply the model to analyse other domain data such as twitter, which is another typical short text.

Acknowledgement

This work is supported by the Scientific Research Starting Foundation for High-level Talents of Pingdingshan University (No. PXY-BSQD2017001), the Foundation for Fostering the National Foundation of Pingdingshan University (No. PXY-PYJJ-2018003), the Educational Commission of Henan Province, China (No. 17A520050), the National Natural Science Foundation of China (No. 61772378, 61373108), the major program of the National Social Science Foundation of China (No. 11&ZD189) and the High Performance Computing Center of Computer School, Wuhan University.

References

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of NIPS*, pages 288–296. MIT Press, 2009.

[3] M. H. F. Li and X. Sentiment analysis with global topics and local dependency. In *Proceedings of AAAI*, pages 1371–1376. AAAI, 2010.

[4] L. Fang, M. Huang, and X. Zhu. Exploring weakly supervised latent sentiment explanations for aspect-level review analysis. In *Proceedings of CIKM*, pages 1057–1066. ACM, 2013.

[5] X. Fu, K. Yang, J. Z. Huang, and L. Cui. Dynamic non-parametric joint sentiment topic mixture model. *Knowledge-Based Systems*, 82(""):102 – 114, 2015.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *Proceedings of NIPS*, pages 2672–2680. ACM, 2014.

[7] A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden topic markov models. In *AISTATS*, pages 163–170. JMLR, 2007.

[8] Y. He, C. Lin, and H. Alani. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of HLT*, pages 123–131. ACL, 2011.

[9] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of SOMA*, pages 80–88. ACM, 2010.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of KDD*, pages 168–177. ACM, 2004.

[11] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of WSDM*, pages 815–824. ACM, 2011.

[12] C. Li, J. Zhang, J.-T. Sun, and Z. Chen. Sentiment topic model with decomposed prior. In *Proceedings of SDM*, pages 767–775. SAIM, 2013.

[13] F. Li, S. Wang, S. Liu, and M. Zhang. Suit: A supervised user-item based topic model for sentiment analysis. In *Proceedings of AAAI*, pages 1636–1642. AAAI, 2014.

[14] K. W. Lim and W. Buntine. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of CIKM*, pages 1319–1328. ACM, 2014.

[15] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM*, pages 375–384, Hong Kong, China, 2009. ACM.

[16] T. Lin, W. Tian, Q. Mei, and H. Cheng. The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings of WWW*, pages 539–550. ACM, 2014.

[17] B. Liu. *Sentiment Analysis And Opinion Mining*. Morgan Claypool Publishers, 2012.

[18] Y. Lu, H. Wang, C. X. Zhai, and R. Dan. Unsupervised discovery of opposing opinion networks from forum discussions. In *Proceedings of CIKM*, pages 1642–1646. ACM, 2012.

[19] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of WWW*, pages 131–140. ACM, 2009.

[20] S. Matsumoto, H. Takamura, and M. Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD*, pages 301–311. Springer, 2005.

[21] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of SIGIR*, pages 889–892. ACM, 2013.

[22] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW*, pages 171–180. ACM, 2007.

[23] S. Moghaddam and M. Ester. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of SIGIR*, pages 665–674. ACM, 2011.

[24] S. Moghaddam and M. Ester. On the design of lda models for aspect-based opinion mining. In *Proceedings of CIKM*, pages 803–812. ACM, 2012.

[25] S. Moghaddam and M. Ester. The flda model for aspect-based opinion mining: addressing the cold start problem. In *Proceedings of WWW*, pages 909–918. ACM, 2013.

[26] S. Mukherjee, G. Basu, and S. Joshi. Joint author sentiment topic model. In *Proceedings of SDM*, pages 370–378. SIAM, 2014.

[27] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124. ACL, 2005.

[28] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[29] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of WWW*, pages 91–100. ACM, 2008.

[30] X. Quan, C. Kit, Y. Ge, and S. J. Pan. Short and sparse text topic modeling via self-aggregation. In *Proceedings of IJCAI*, pages 2270–2276. AAAI, 2015.

[31] E. Riloff, S. Patwardhan, and J. Wiebe. Feature subsumption for opinion analysis. In *Proceedings of EMNLP*, pages 440–448. ACL, 2006.

[32] J. Tang, M. Zhang, and Q. Mei. One theme in all views: Modeling consensus topics in multiple contexts. In *Proceedings of KDD*, pages 5–13. ACM, 2013.

[33] I. Titov and R. T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL*, pages 308–316. ACL, 2008.

[34] I. Titov and M. Ryan. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW*, pages 111–120. ACM, 2008.

[35] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pages 417–424. ACL, 2002.

[36] H. Wang and M. Ester. A sentiment-aligned topic model for product aspect rating prediction. In *Proceedings of EMNLP*, pages 1192–1202. ACL, 2014.

[37] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of KDD*, pages 618–626. ACM, 2011.

[38] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of KDD*, pages 424–433. ACM, 2006.

[39] Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng. Hashtag graph based topic model for tweet mining. In *Proceedings of ICDM*, pages 1025–1030. ACM, 2014.

[40] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of WSDM*, pages 261–270. ACM, 2010.

[41] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37(""):186 – 195, 2013.

[42] P. Xie and E. P. Xing. Integrating document clustering and topic modeling. *Proceedings of UAI*, pages 694–703, 2013.

[43] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *Proceedings of WWW*, pages 1445–1456. ACM, 2013.

[44] P. Yang, W. Gao, Q. Tan, and K.-F. Wong. A link-bridged topic model for cross-domain document classification. *Information Processing & Management*, 49(6):1181 – 1193, 2013.

[45] T.-J. Zhan and C.-H. Li. Semantic dependent word pairs generative model for fine-grained product feature mining. In *Proceedings of PAKDD*, pages 460–475. Springer, 2011.

[46] Y. Zhang, D.-H. Ji, Y. Su, and C. Sun. Sentiment analysis for online reviews using an author-review-object model. In *Proceedings of AIRS*, pages 362–371. Springer, 2011.

[47] Y. Zhang, D.-H. Ji, Y. Su, and H. Wu. Joint na ve bayes and lda for unsupervised sentiment analysis. In *Proceedings of PAKDD*, pages 402–413. Springer, 2013.

[48] Q. G. F. W. Zhang L. Sentiment analysis based on light reviews. *Ruan Jian Xue Bao/ Journal of Software*, 12(25):2790–2807, 2014.

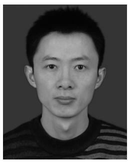
[49] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of ECIR*, pages 338–349. Springer Berlin Heidelberg, 2011.

[50] W. X. Zhao, J. Jiang, H. Yan, and X. Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of EMNLP*, pages 56–65. ACL, 2010.

[51] X. Zheng, Z. Lin, X. Wang, K.-J. Lin, and M. Song. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowledge-Based Systems*, 61(""):29 – 47, 2014.

[52] G. Zhou, J. Zhao, T. He, and W. Wu. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. *Knowledge-Based Systems*, 66(""):136 – 145, 2014.

Biography



Shufeng Xiong received his Ph.D. degree in Computer Science from Wuhan University, China. He is an associate professor in Computer School, Pingdingshan University. His research interests includes sentiment analysis and topic model.

Kuiyi Wang is a teacher in Computer School, Pingdingshan University. He has worked in the areas of data mining and machine learning.

Donghong Ji is currently a Professor in School of Computer of Wuhan University. He received his Ph.D., M.Sc. and B.Sc. degrees from Wuhan University in Computer Science in 1995, 1992 and 1989 respectively. He also received his M.Sc. degrees from Oxford University in Linguistics in 2005. His main research interests include natural language processing and semantic information retrieval.

He received his Ph.D degree from Tsinghua University, China. He is an associate professor in Computer School, Pingdingshan University. His research interests includes sentiment analysis and machine learning.