IEICE TRANS. INF. & SYST., VOL.E102–D, NO.11 NOVEMBER 2019

PAPER

Improved LDA Model for Credibility Evaluation of Online Product Reviews

Xuan WANG[†], **Bofeng ZHANG**[†a)], ***Nonmembers***, **Mingqing HUANG**[†], ***Member***, **Furong CHANG**[††], ***and*** **Zhuocheng ZHOU**[†], ***Nonmembers***

SUMMARY When individuals make a purchase from online sources, they may lack first-hand knowledge of the product. In such cases, they will judge the quality of the item by the reviews other consumers have posted. Therefore, it is significant to determine whether comments about a product are credible. Most often, conventional research on comment credibility has employed supervised machine learning methods, which have the disadvantage of needing large quantities of training data. This paper proposes an unsupervised method for judging comment credibility based on the Biterm Sentiment Latent Dirichlet Allocation (BS-LDA) model. Using this approach, first we derived some distributions and calculated each comment's credibility score via them. A comment's credibility was judged based on whether it achieved a threshold score. Our experimental results using comments from Amazon.com demonstrated that the overall performance of our approach can play an important role in determining the credibility of comments in some situation.

key words: *comment credibility, biterm sentiment latent Dirichlet allocation model, unsupervised method, unequal short text*

1. Introduction

The era of Web 2.0 has witnessed the rapid expansion of online e-commerce platforms that allow consumers to purchase products and share feelings about their purchases. Comments provided for each product become a reference for customers who may want to purchase them. Consequently, it is of great importance to study whether these comments are credible. Some e-commerce platforms, such as Amazon, provide user evaluation mechanisms for creating consumer reviews, and even allow consumers to rate others' comments as useful or not. Still other sites, such as Taobao, do not provide evaluation mechanisms at all. Commonly, even on sites that offer review evaluation mechanisms, there may be no or few evaluations on comments regarding newly listed or low volume products. For popular items that have multiple comments, reviewers tend to rate only the comments that appear first, leaving comments that appear later without a sufficient number of evaluations of their credibility. Given the varied circumstances surrounding available reviews, consumers may wonder which comments are credible. Therefore, the study of comment credibility is of practical significance for both consumers and e-commerce merchants.

In this research area, supervised machine learning models based on the Naive Bayes (NB), Support Vector Machine (SVM), or similar approaches have been applied to judge the credibility of comments or other user-generated information. These methods have the disadvantage of requiring a large quantity of comments or other data with known credibility results to use as training data in the early stages. Then, some features can be extracted to train a classifier to judge comment credibility.

This paper proposes an unsupervised method to judge the credibility of posted comments. We believe that when the topics in a comment are more centralized, the comment tends to be more credible. Online comments from e-commerce platforms usually have some common characteristics. First, online comments are uneven in length. Some of the comments may be less than 20 words, while others may be more than 100 words, but in general, none of them are long. Second, comments usually project clear emotional tones that are reflected in the accompanying ratings. Based on these two characteristics, we developed a topic model called Biterm Sentiment Latent Dirichlet Allocation (BS-LDA). Using this model, we were able to get three distributions over latent topics and sentiments. Based on these distributions, we created a method to calculate a credibility score for each comment. Finally, we established a threshold score to judge the credibility of each comment. The main innovations of this paper are as follows:

- We propose a new unsupervised method to judge the credibility of comments. Training data is not necessary for this method.
- We put forward a new topic model called BS-LDA that takes into account the characteristics of comments, including short text length and sentiment tendency.
- Our proposed credibility score calculation is based on a comprehensive quantification of two indicators: Jensen-Shannon (JS) divergence and topic entropy. We improved the JS divergence in order to compare the degree of dispersion between the three-dimensional distributions.

The remainder of the paper is organized as follows. Section 2 presents an overview of related work. In Sect. 3, we elaborate on our unsupervised credibility judgment model. Section 4 describes an experimental evaluation of

Manuscript received July 9, 2018.
Manuscript revised February 25, 2019.
Manuscript publicized August 22, 2019.

[†]The authors are with the School of Computer Engineering and Science, Shanghai University, 200444, Shanghai, China.
[††]The author is with the School of Computer Science and Technology, Kashgar University, 844006, Kashgar, China
a) E-mail: bfzhang@shu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2018EDP7243

our method in comparison with other methods and presents our results and analysis. In Sect. 5, we offer our conclusions and suggestions for future research.

2. Related Work

In this section, we review the literature regarding credibility assessment methods, including prior topic model research and methods for text mining in short texts.

2.1 Existing Credibility Assessment Method

Credibility is the degree of believability, or the capacity to cause others to believe [1]. In the era of Web 2.0, researchers have studied the credibility of user-generated contents on the Internet, including Twitter, Weibo, and e-commerce online product reviews. Existing studies usually focus on using supervised algorithms to classify comments. T.L. Ngo-Ye et al. [2] proposed a hybrid text regression model for predicting online review helpfulness. F. Yang et al. [3] focused on features extracted from microblogs, and then trained a classifier to detect rumors on the microblogs. Using a topic model, Z. Jin et al. [4] discovered conflicting viewpoints in news tweets, and then built a credibility propagation network to evaluate the credibility of the news. C.C. Chen et al. [5] extracted representative review features and trained an SVM classifier to evaluate the quality of information in product reviews. M.J. Metzger et al. [6] proposed a credibility assessment model based on Naive Bayes algorithm. J. Aigner et al. [7] conducted a controlled web-based study to investigate whether the perception of the credibility of refugee-related tweets can be influenced by cues already reported in the literature for general social media contents. M. Alrubaian et al. [8] proposed a model consisting of six integrated components operating in an algorithmic form to assess the credibility of tweets. J. Ito et al. [9] utilized the "tweet topic" and "user topic" features derived from Latent Dirichlet Allocation(LDA) to assess the tweet credibility.

In summary, some unsupervised methods have been applied to the study of Twitter's credibility, but for online comments, most of the methods are supervised.

2.2 Topic Distillation Model

Topic modeling techniques have been used widely in natural language processing to discover latent semantic structures. The earliest topic model was Latent Semantic Analysis (LSA) proposed by S. Deerwester et al. [10]. This model analyzed document collections and built a vocabulary-text matrix. Using Singular Value Decomposition (SVD), researchers can build the latent semantic space. Later, T. Hofmann et al. [11] proposed the Probabilistic Latent Semantic Analysis (PLSA), which improved upon the Latent Semantic Analysis (LSA) model. PLSA considers that documents include many latent topics, and the topics are related to words. Prior to PLSA, Dirichlet distribution was introduced by D.M. Blei et al. [12] and the Latent Dirichlet Allocation (LDA) approach was proposed. Due to the characteristics of LDA generation, this topic model has been improved and used in many different areas.

The drawbacks for using LDA are that topic distribution tends to be less targeted and lacks definite meaning. Researchers have made improvements to the LDA topic model accordingly and applied these models in different areas. For example, D. Ramage et al. [13] improved the unsupervised LDA model by creating a supervised topic model called Labeled-LDA, in which the researchers could attach the topic meaning. Separately, many researchers chose to add a level to the three levels of document-topic-word. I. Titov et al. [14] proposed a multi-grain model that divided the topics into two parts: local topics and global topics. This model was used to extract the ratable aspects of objects from online user reviews. Besides, a range of other approaches have been used as well. H. Chen et al. [15] modeled user's social connections and proposed a People Opinion Topic (POT) model that can detect social communities and analyze sentiment. T. Iwata et al [16] took time into consideration and proposed a topic model for tracking time-varying consumer purchasing behavior. To recommend locations to be visited, T. Kurashima et al. [17] proposed a Geo topic model to analyze the location log data of multiple users. C. Chemudugunta et al. [18] suggested that a model can be used for information retrieval by matching documents both at a general topic level and at a specific level, and C. Lin et al. [19] proposed the Joint Sentiment Topic (JST) model, which can be used to analyze the sentiment tendency of documents. S. Wang et al. [20] proposed the Life Aspect-based Sentiment Topic (LAST) model to mine from other products the prior knowledge of aspect, opinion, and their correspondence.

The differences between these topic models lie mainly in the field of application. Models related to time in the references 16, 18 and 20 can get distributions over time. Topics can also be divided into several parts like the models in the reference 14. Certainly, the user (reference 15) and the location (reference 17) can be considered as the new level to obtain distributions over them. In reference 19 (JST), sentiment can be considered as the new level so we can get distributions over sentiment. As we can see, based on different application areas, different factors are to be considered in different optimization model, such as time, sentiment, location and so on.

Focused on the problem that the LDA algorithm must be given a set number of topics. Y.W. Teh et.al. [21] proposed a nonparametric Bayesian model for clustering problems involving multiple groups of data. T.L. Griffiths et al. [22] proposed Hierarchical Latent Dirichlet Allocation (HLDA), which can determine the number of topics in a corpus automatically. However, because of its computational complexity, this model has not been applied as extensively as LDA.

2.3 Short Text Topic Mining Method

When it comes to short text, the LDA topic model suffers from the data sacristy problem. Many studies of this problem have been conducted.

X.H. Phan et al. [23] used external knowledge to secure enough suitable data as well as to expand the coverage of the classifier to handle future data better. P. Wang et al. [24] enriched the representation of short text, then used the LDA model to extract latent topic information. Next, they combined two classifiers, SVM and Maximum Entropy Model(MaxEnt) to achieve high reliability. Y. Zhu et al. [25] focused on two tasks: making use of different external corpus to identify topics from texts and adding the weight of a few features in texts. L. He et al. [26] presented a novel model for short texts, referred to as the Topic Trend Detection (TTD) model. This model derived more typical terms to represent the topics found in short texts and improved the coherence of topic representations. Y. Zuo et al. [27] proposed the Word Network Topic Model (WNTM), which modeled the distribution over topics for each word instead of learning topics for each document. X. Cheng et al. [28] proposed the Biterm Topic Model (BTM), which enlarged the text content by defining word-pairs in one text as biterms.

In general, some of these methods need prior knowledge to establish relationships between words and reduce sparsity. But the BTM model reduces sparsity by extending the text simply.

3. Credibility Score Calculation Model

The topic model has been used widely in natural language processing. By analyzing latent topics in documents, researchers can mine semantic connotations. However, the LDA model does not have adequate division of potential topics. In reality, some texts, such as comments about products, have distinct sentiment identifiers. Comments that express a positive opinion tend to use positive emotive words, while comments that convey negative opinions prefer to use clearly negative words. When documents contain apparent sentiment, the disadvantages of the LDA model becomes more obvious.

Product reviews are usually short text. Under the circumstances, the number of topics K may exceed the length of the text, which could lead directly to sparser topic distribution and be difficult to mine latent topics.

According to the first two paragraphs, we can conclude that the online comments about products usually have the two characteristics: (1) the comments usually project clear emotional tones and (2) the comments are usually short texts. For the characteristic (1), we take JST [19] model into consideration and add a sentiment level to obtain more accurate distributions. For the characteristic (2), we refer to BTM [28] model because the BTM model does not rely on external knowledge and can effectively reduce the sparsity caused by the short text. Based on the characteristics of product comments and referred topic models, this paper proposed a model called BS-LDA. Via this model, we are able to obtain more accurate topic distributions. In other words, BS-LDA model proposed in this paper is more targeted to the comments corpus.

The general process of our unsupervised method is as follows. First, we use the BS-LDA to gather three distributions (π, θ, φ). Then we use the distributions to calculate the credibility score of each comment. Finally, we compare our result with a threshold, which determines the credibility of this comment. The process is shown in Algorithm 1.

Algorithm 1: The overall process

Input: A,η /*A is the comments collected from e-commerce platform, η is the credibility score threshold.*/
Output: R /*R is comment credibility list, the value is true or false */

preprocess comments, then expand the text according to 3.1 and form a corpus C;
execute Algorithm 2 on corpus C, and get three matrixes: document-sentiment distribution π, document-sentiment-topic distribution θ, sentiment-topic-word distribution φ;
for *document d in corpus C* **do**
    calculate topic entropy $E(d)$ by Eq. (5) to Eq. (7);
    calculate JS divergence $J(d)$ by Eq. (8) to Eq. (10);
    calculate credibility score $Credit(d)$ by Eq. (11) to Eq. (13);
    compare $Credit(d)$ with η and add the result in R;
return R

3.1 Definition of Biterm

Extending text is an effective way to mine latent topics from short texts. This paper refers to the BTM model [28], using biterms to expand texts. "Biterm" refers to disordered word pairs occurring in a short text simultaneously. After exhausting all of the word-pairs in a given sample of short text, an expanded text is formed. For instance, let us assume there are three words in one short text $\{w_1, w_2, w_3\}$. The biterms are $\{(w_1, w_2), (w_1, w_3), (w_2, w_3)\}$. Therefore, the number of biterms in one short text is C_n^2, in which n points to the number of words in the text.

For documents which are less than 20 words, the extended text length does not exceed C_{20}^2, which equals to 190. But for longer texts including 100 words, the extended text length can be C_{100}^2, which equals to 4950. Certainly, this procedure consumes too much time to process when thousands of documents are need to be analyzed. Referring to the above theory, we decided to use partial text expansion. For example, for a relatively long text, we use the current word and the following s words to constitute the biterms. When we encounter r (r<s) words at the end of the text, we use the current word and the remaining words to form a biterm, even if the number of biterms containing the current word is less than s.

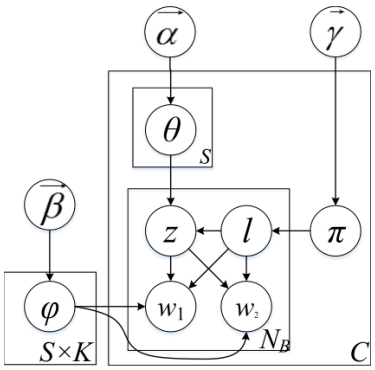

Fig. 1 Generation process of BS-LDA

3.2 BS-LDA Model Description

Given a corpus with M documents denoted by C=$\{d_1, d_2, \ldots, 936d_M\}$, containing V terms denoted by D=$\{w_1, w_2, \ldots, w_V\}$. These terms in the corpus constitute N_B biterms, expressed as B=$\{b_1, b_2, \ldots, b_{N_B}\}$ with $b_i = (w_p, w_q), (p \neq q)$. Each document d can be enlarged as description in Sect. 3.1, creating a new corpus $C' = \{d'_1, d'_2, \ldots, d'_M\}$. The BS-LDA model contains four layers including document layer, latent sentiment layer, latent topic layer and word layer. The entire corpus contains K latent topics distributed over all of the words, and S latent sentiment labels distributed over the latent topics and words. The generation procedure is described as follows. A graphical representation is shown in Fig. 1.

- For each document d', draw a document-sentiment distribution $\pi_d \sim Dir(\vec{\gamma})$
- For each sentiment label l under document d', draw a document-sentiment-topic distribution $\theta_{d,l} \sim Dir(\vec{\alpha})$
- For each topic z under sentiment label l, draw a sentiment-topic distribution $\varphi_{l,k} \sim Dir(\vec{\beta})$
- For each biterm b_i in document d'
  - Choose a sentiment label $l_i \sim Mult(\pi_d)$
  - Choose a topic $z_i \sim Mult(\theta_{d,l})$
  - Choose two words $w_{i,1}, w_{i,2} \sim Mult(\varphi_{l_i,z_i})$, and $w_{i,1}, w_{i,2}$ constitute biterm b_i

3.3 Model Inference

A challenge for text mining is that documents and words are visible while the distributions are invisible. Therefore, the parameter distributions, including π, θ, φ, need to be estimated. Similar to LDA, this paper uses the Gibbs Sampling algorithm to estimate these parameter distributions. For one biterm, the two words share the same latent sentiment label and same latent topic. If other biterms' latent sentiment labels and latent topics are known, we can use Eq. (1) to estimate this biterm's existence probability in each of the sentiments and topics.

Table 1 Descriptions of elements in Eq. (1)

Elements	Meaning
$N_{d,s,\neg i}$	the number of biterms in document d', for which the sentiment label is s, excluding biterm i
$N_{d,\neg i}$	the number of biterms in document d', excluding biterm i
$N_{d,s,k,\neg i}$	the number of biterms in document d', for which the sentiment label is s and the topic is k, excluding biterm i
$N_{s,k,w_{i,1},\neg i}$	the number of words $w_{i,1}$ in corpus, for which the sentiment label is s and the topic is k, excluding word $w_{i,1}$
$N_{s,k,w_{i,2},\neg i}$	the number of $w_{i,2}$ in corpus, for which the sentiment label is s and the topic is k, excluding word $w_{i,2}$
$N_{s,k,\neg i}$	the number of words in corpus, for which the sentiment is s and the topic is k, excluding this word

$$p(z_i = k, l_i = s|\mathbf{B}, \overrightarrow{z_{\neg i}}, \overrightarrow{l_{\neg i}}, \overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\gamma}) = \frac{N_{d,s,\neg i} + \gamma_i}{N_{d,\neg i} + \sum_{t=1}^{S} \gamma_t} \times \frac{N_{d,s,k,\neg i} + \alpha_i}{N_{d,s,\neg i} + \sum_{t=1}^{K} \alpha_t} \times \frac{N_{s,k,w_{i,1},\neg i} + \beta_i}{N_{s,k,\neg i} + \sum_{t=1}^{V} \beta_t + 1} \times \frac{N_{s,k,w_{i,2},\neg i} + \beta_i}{N_{s,k,\neg i} + \sum_{t=1}^{V} \beta_t} \quad (1)$$

We can use the Gibbs sampling procedure to update each biterm's latent sentiment label and topic. First, sentiment label and topic are assigned randomly to each biterm in the corpus. In every iteration, elements in Table 1 are counted. Then, Eq. (1) is used to update each biterm's sentiment label and topic. When the process reaches the specified number of iterations, it stops. The Gibbs sampling procedure is shown in Algorithm 2.

The equations to estimate the parameters π, θ, φ are shown as Eqs. (2), (3), and (4), in which π is a M×S matrix and represents sentiment label distribution over each document; θ is a M×S×K matrix and represents the topic distribution over each sentiment label in each document; and φ is a S×K×V matrix and represents word distribution over each topic in each sentiment label.

$$\pi_{d,s} = \frac{N_{d,s} + \gamma_s}{N_d + \sum_{t=1}^{S} \gamma_t} \quad (2)$$

$$\theta_{d,s,k} = \frac{N_{d,s,k} + \alpha_k}{N_{d,s} + \sum_{t=1}^{K} \alpha_t} \quad (3)$$

$$\varphi_{s,k,i} = \frac{N_{s,k,i} + \beta_i}{N_{s,k} + \sum_{t=1}^{V} \beta_t} \quad (4)$$

3.4 Credibility Calculation Method

In this paper, assessing the credibility of a document depends on the degree of dispersion of topics in the document. We use a mixture of topic entropy and JS divergence to calculate the credibility score for each document.

Algorithm 2: Gibbs Sampling procedure

Input: corpus C', biterms set B, sentiment label number S, topic number K, hyper-parameters $\vec{\alpha}, \vec{\beta}, \vec{\gamma}$
Output: document-sentiment distribution π, document-sentiment-topic distribution θ, sentiment-topic-word distribution φ

initialize each biterm's the sentiment label and topic randomly;
for *iter = 1 to iteration number* **do**
  for *each document d′ in corpus C′* **do**
    for *each biterm b in document d′* **do**
      calculate the probability of each sentiment label and topic of *b*' by Eq. (1);
      sample *b*'s sentiment label and topic based on the result of step 5;
calculate the parameter matrixs π, θ, φ by Eqs. (2), (3), (4);
return π, θ, φ;

The concept of topic entropy proposed by E. Momeni [29] relies on the document-topic distribution to determine the degree of dispersion of topics in the document. If a topic is related to many documents, it is treated as a noisy topic that contains less valuable information. If a topic exists in only a few documents, the topic is considered to contain more information that is valuable. Based on this theory, the entropy value of each topic in the corpus is required to be calculated. Thus, we define Eq. (5) to calculate each topics entropy value. Since entropy value for each topic has been obtained, we define Eq. (6) to calculate the topic entropy value of each document and T means the set of topics.

$$H(t) = -\sum_{d \in C} p(d|t) \log p(d|t) \tag{5}$$

$$E(d) = \sum_{t \in T} H(t) p(t|d) \tag{6}$$

$$p(d|t) = \frac{p(d)p(t|d)}{p(t)} \tag{7}$$

In Eq. (5), $p(d|t)$ represents the probability of the distribution of document d in the case of known topic t, which can be derived from the Bayesian formula shown as Eq. (7). In order to get document-topic distribution, we multiple π and θ to get a matrix denoted as σ wihich size is $M \times K$. Based on the definition in 3.3, the distribution π can be considered as M matrices with dimension $1 \times S$, while θ can be regarded as M matrices with dimension $S \times K$. Multiplying the matrix with dimension $1 \times S$ by the matrix with dimension $S \times K$, a matrix with dimension $1 \times K$ can be obtained. After similar operation for M times, M matrices with dimension $1 \times K$ can be obtained. Finally, we can obtain a matrix with dimension $M \times K$ by merging these matrices. The whole progress can be considered as the generation theory of σ, which can be regarded as document-topic distribution of the corpus. The value of $p(t|d)$ is the value in row d and column t of matrix σ.

JS divergence is often used to measure the degree of discrepancies between different distributions. In our method, we choose high frequency topics from each document and compare the dispersion between them via JS divergence. In the first place, we need to reduce the dimension of φ to simplify the calculation. We develop a two-dimension matrix called φ' of size $K \times V$. The elements in φ' are defined as shown in Eq. (8), where $\varphi'_{k,v}$ represents the sentiment label number corresponding to the maximum probability for the topic k and word v ($k \in \{1, K\}, v \in \{1, V\}$).

$$\varphi'_{k,v}=p, \text{ when } \varphi_{p,k,v} = \max \varphi_{s,k,v} (1 \le s \le S) \tag{8}$$

In Eq. (9) and Eq. (10), we define an improved JS divergence calculation method for each document. In our improved approach, if the JS divergence between the high-frequency topics in a document is high, we consider that the content of the high-frequency topic is similar across the instances, and the content is more concentrated. If the JS divergence is low, then the content is sparser.

$$J(d) = \frac{1}{|top\_k|(|top\_k| - 1)} \times \sum_{i,j \in top\_k} \sum_{v \in V} [\tau_{i,j,v}(p_{i,v} log \frac{2p_{i,v}}{p_{i,v} + p_{j,v}} + p_{j,v} log \frac{2p_{j,v}}{p_{i,v} + p_{j,v}})] \tag{9}$$

$$\tau_{i,j,v} \begin{cases} 1 & (\varphi'_{i,v} = \varphi'_{j,v}) \\ -1 & (\varphi'_{i,v} \neq \varphi'_{j,v}) \end{cases} \tag{10}$$

where top_k represents a specified number of high frequency topics set, and $| \ top\_k \ |$ represents the size of the set. Generally, the latent topics that appear more frequently in a document are usually limited to a certain range, so we choose $\lceil K/10 \rceil$ as the value of $| \ top\_k \ |$. $p_{i,v}$ and $p_{j,v}$ represent $\varphi_{\varphi'_{i,v},i,v}$ and $\varphi_{\varphi'_{j,v},j,v}$ respectively.

We need to normalize $E(d)$ and $J(d)$ via Eq. (11) and Eq. (12). Based on topic entropy and improved JS divergence, we can compute the credibility score using the calculation method shown in Eq. (13).

$$E'(d) = E(d)/\max\{E(d)\} \quad (d \in C) \tag{11}$$

$$J'(d) = J(d)/\max\{J(d)\} \quad (d \in C) \tag{12}$$

$$Credit(d) = \lambda E'(d) + (1 - \lambda) J'(d) \tag{13}$$

We define threshold η_e for topic entropy aspect ($E'(d)$), threshold η_j for JS divergence aspect ($J'(d)$) and η for credibility score ($Credit(d)$), which is the final criterion for the credibility of different comments. For document d, if the value of $E'(d)$ is greater than η_e, the document is credible in entropy topic aspect. Conversely, it is not credible in terms of topic entropy. The situations of JS divergence and credibility score are similar. For $E'(d)$ and $J'(d)$ in document d, we believe that the one with a larger difference from the

corresponding threshold is more extreme and that means the one has greater reference value. The weighted sum of $E'(d)$ and $J'(d)$ is obtained by using the parameter λ to obtain the final credibility score, shown in Eq. (13). Therefore, the weight of the one with larger threshold gap must be larger. This is the basic idea of parameter λ value setting.

4. Experimental Evaluation

We conducted some experiments to evaluate the BS-LDA method proposed in Sect. 3. The experiment consisted of three different parts: (1) an examination of the effectiveness of our model for documents of different lengths; (2) a comparison of our model with some unsupervised methods; and (3) a comparison of our approach with some classic supervised machine learning algorithms.

4.1 Data Collection

The data set used in this experiment was taken from the Amazon comment data collected by J.J. McAuley et al.[30], which spanned 2012 to 2016, targeted 75,000 kinds of food, and included approximately 500,000 comments. The details about the documents are shown in Table 2.

On the Amazon website, users who purchase a product can evaluate the product, and other customers can evaluate these comments by clicking Agree or Disagree. Most of the contents in Table 2 were easy to understand, we only needed to explain the meaning of "helpfulness". Each element was divided into two parts: the number before the slash denoted the number of Agree votes, and the number after the slash denoted the total Vote number (which contained the number of Agree and Disagree votes). This element was the basic criterion for the credibility of comment.

4.2 Data Preprocessing

Although the data collected contained almost 500,000 comments, most of the comments were not usable for various reasons. Some of them were duplicated, and many comments had few or no votes, which influenced us to judge the comments credibility, so we had to discard them. To obtain more accurate evaluation criteria, we chose comments that had more than 10 votes. Based on the above standards, almost 90% of the comments were filtered out. Finally, after further data cleansing, we had 18,982 useful comments.

Regarding the standard for different lengths of text, the comments were divided into four groups based on document length: less than 20 words, between 21 and 50 words, between 51 and 100 words, and more than 100 words. The details for each group are shown in Table 3. "Length" meant document length group division. "Enlarge" meant the following word numbers needed to constitute a biterm for each word in current document, which had been mentioned in 3.1. N_{Doc} meant document numbers in this length group, N_{Biterm} meant biterm numbers in this length group, N_{Term} meant term numbers in this length group.

We took the following steps to deal with noise in the data. First, we performed word segmentation and removed the punctuation, numbers, and other non-alphabet characters. Then, we removed any meaningless words or symbols. Next, to reduce the vocabulary size, we extracted stem words using a stemming algorithm. Finally, based on the definition of biterm, we enlarged the remaining content.

Table 2 The data format

Data item	Content
productId	B001E4KFG0
userId	A3SGXH7UHU8GW
profileName	tdemartian
helpfulness	1/1
score	5.0
time	1303862400
summary	Good Quality Dog Food
text	I have bought several of the Vitality canned dog food products and have found them all to be of good quality.

4.3 Parameters Setting

Based on previous research regarding the LDA topic model, in the BS-LDA model, we set α as $50/K$ (K is the topic number) and β as 0.01. The parameter γ influences the sentiment label distribution, so we chose three different values for γ: 20, 1, 0.01, representing positive sentiment, neutral sentiment, and negative sentiment, respectively. We ran the Gibbs sampling 1,500 times.

For calculating the credibility, we had the previously normalized values $E'(d)$ and $J'(d)$, which were between 0 and 1. For $E'(d)$, we set the threshold η_e to be 0.5, and for $J'(d)$ we set η_j to be 0.5. Considering that $Credit(d)$ is a comprehensive indicator of $E'(d)$ and $J'(d)$ and the values of $E'(d)$ and $J'(d)$ are both between 0 and 1, the value of $Credit(d)$ is also between 0 and 1. Thence, for $Credit(d)$ we set the threshold η to be 0.5. For each document d, the parameter λ was set as shown in Table 4.

Based on the relationship between $E'(d)$, $J'(d)$, $Credit(d)$ and their corresponding thresholds, the credibility score of documents can be calculated and judged. For one document d, if the value of $E'(d)$ is greater than the threshold, we believe that this comment is credible

Table 3 Document information of our data

Length	Enlarge	N_{Doc}	N_{Biterm}	N_{Term}
3-20	all	4737	220140	5896
21-50	5	6530	512160	13631
51-100	4	4684	591578	17019
>100	3	3031	632280	23659

Table 4 Parameter setting of λ

Condition	λ
$(E'(d)-\eta_e)(J'(d)-\eta_j)>0$	0.5
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta J(d)/\Delta E(d)>5$	0.1
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta J(d)/\Delta E(d)>4$	0.2
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta J(d)/\Delta E(d)>3$	0.3
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta J(d)/\Delta E(d)>2$	0.4
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta J(d)/\Delta E(d)>1$	0.5
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta E(d)/\Delta J(d)>1$	0.5
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta E(d)/\Delta J(d)>2$	0.6
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta E(d)/\Delta J(d)>3$	0.7
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta E(d)/\Delta J(d)>4$	0.8
$(E'(d)-\eta_e)(J'(d)-\eta_j)<0, \Delta E(d)/\Delta J(d)>5$	0.9

in terms of topic entropy. Conversely, it is not credible in terms of topic entropy. The situation of JS divergence is similar. On the basis of this theory, for document d, if the value of $E'(d)$ and $J'(d)$ are both greater or less than corresponding threshold, the credibility of this document has been decided so the weights of the two values are same and are set to be 0.5. To describe the two situations more concisely, we combine these two cases into $(E'(d)-\eta_e)(J'(d)-\eta_j)>0$. On the contrary, if $E'(d)$ and $J'(d)$ have different credibility situation compared to corresponding threshold, these situations are combined into $(E'(d)-\eta_e)(J'(d)-\eta_j)<0$. In the circumstances, the credibility of this document cannot be judged by only one aspect, so we define $\Delta E(d)$ to indicate the difference between $E'(d)$ and η_e, $\Delta J(d)$ to indicate the difference between $J'(d)$ and η_j. The definition of $\Delta E(d)$ and $\Delta J(d)$ are shown in Eqs. (14)–(15). Under the circumstances, we are required to compare the relative values of $\Delta E(d)$ and $\Delta J(d)$. Between $\Delta E(d)$ and $\Delta J(d)$, the one with a relatively larger value has a higher weight. For a document d, the larger the value of λ is, the larger the weight of $\Delta E(d)$ is, and the smaller the weight of $\Delta J(d)$ is.

$$\Delta E(d) = |E'(d) - \eta_e| \tag{14}$$

$$\Delta J(d) = |J'(d) - \eta_j| \tag{15}$$

4.4 Experiment Evaluation Standard

4.4.1 Four Base Standard

Essentially, the credibility determination problem is a classification problem, so we chose four evaluation indexes: *Accuracy*, *Precision*, *Recall*, and *F-measure*. Table 5 shows the base elements of the four indexes. The equations of the four indexes are shown as Eqs. (16)–(19).

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{16}$$

Table 5 The definition of tp fp fn tn

	actual: credible	actual: not-credible
predicted: credible	tp	fp
predicted: not-credible	fn	tn

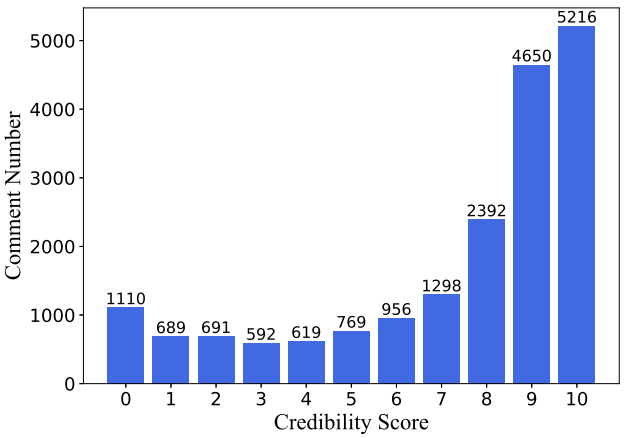

Fig. 2 Distribution of credibility score

$$Precision = \frac{tp}{tp + fp} \tag{17}$$

$$Recall = \frac{tp}{tp + fn} \tag{18}$$

$$F - measure = \frac{2 \times Precsion \times Recall}{Precision + Recall} \tag{19}$$

4.4.2 Credibility Criteria

In this paper, we used the "helpfulness" value for each comment to build the credibility criteria. For each comment, we needed to calculate the rate of Agree number and total Vote number. For this purpose, we divided the credibility score into 11 levels (0-10). The calculation equation is shown as Eq. (20). The distribution of credibility scores in the data set is displayed graphically in Fig. 2.

$$credibility\_score = 10 \times n_{Agree}/n_{Vote} \tag{20}$$

Based on Fig. 2, we defined scores greater than or equal to 5 as credible; scores less than 5 were not credible. The results showed 15,281 credible comments and 3,701 not-credible comments. Clearly, most of the comments tended to be credible.

4.4.3 Feature Extraction

For the supervised classification algorithm, it was necessary to extract features from the data. In our experiment, there

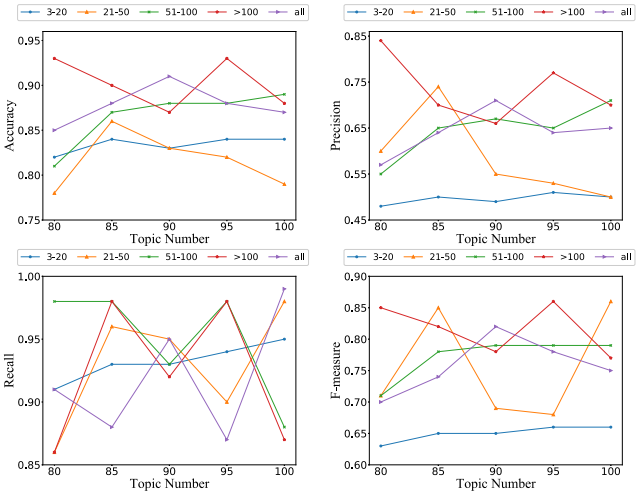

Fig. 3 *Accuracy*, *Precision*, *Recall* and *F-measure* comparison for different document length

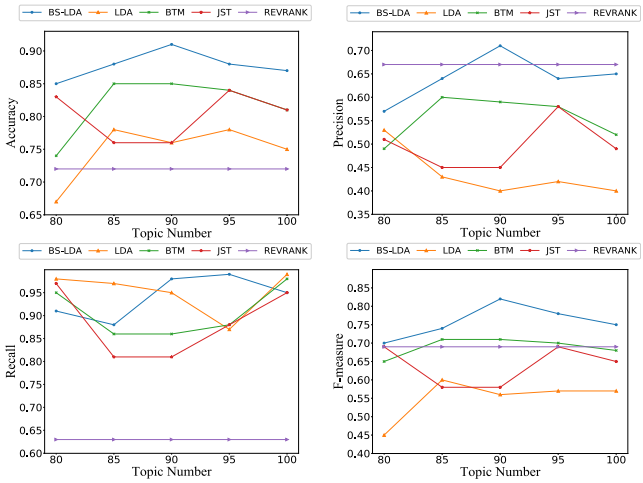

Fig. 4 *Accuracy*, *Precision*, *Recall* and *F-measure* comparison for different unsupervised method

were three features as follows.

1) The linguistic feature

The linguistic feature included apparent document content and grammar information, including the word count, sentence count, word count per sentence, punctuation count, error count, and entity count.

2) The user feature

A user's historical behavior usually has an influence on the credibility of recently released comments. The user feature contained the number of comments offered by the user, the total number of votes received by a user, the number of Agree votes achieved by a user, and the Agree rate.

3) The sentiment feature

The sentiment feature reflected some of the emotional content of the comment, including the positive word count, negative word count, neutral word count, and sentiment score.

4.5 Result and Analysis

4.5.1 Comparison of the Effectiveness of BS-LDA for Documents of Different Lengths

In this portion of the research, our goal was to study the effectiveness of our method for documents of different lengths. In accordance with our work in Sect. 4.2, we divided the documents into four groups by document length. Figure 3 shows the results for our method with different numbers of topic in texts with different document lengths. The topic numbers were set from 80 to 100, stepped by 5. In Fig. 3, different lines represent different text lengths, and "all" means all of the texts in the corpus.

According to Fig. 3, we can conclude that with the growth of document length, the performance of our model became better in terms of all of the evaluation standards. For *Accuracy* and *Recall*, there was no significant direct difference between the results for texts of different lengths. For *Precision* and *F-measure*, it was apparent that longer documents performed well. Furthermore, we can also see that the results for all of the corpus were better than for shorter texts, and were worse than for longer texts, which proved that the length of documents influenced the result.

4.5.2 Comparison of BS-LDA with Other Unsupervised Models

We also assessed the performance of our model in comparison with other unsupervised topic model including the BTM and JST that we referred and traditional topic model LDA for the entire corpus. For this testing, the credibility calculation method had to be adjusted. Through LDA and BTM, we can achieve document-topic distribution θ and topic-word distribution φ. For topic entropy, we did not need to make many changes. For JS divergence, Eq. (9) was simplified to Eq. (21) in BTM and LDA model, while the JST model use the Eq. (9). The meaning of each symbol was the same as for Eq. (9) and Eq. (10).

$$J(d) = \frac{1}{|top\_k|(|top\_k|-1)} \times \sum_{i,j \in top\_k} \sum_{v \in V} (\varphi_{i,v} log \frac{2\varphi_{i,v}}{\varphi_{i,v}+\varphi_{j,v}} + \varphi_{j,v} log \frac{2\varphi_{j,v}}{\varphi_{i,v}+\varphi_{j,v}}) \tag{21}$$

Figure 4 shows the comparison between our proposed BS-LDA model, JST model, BTM model and the LDA model in all four aspects. In this experiment, the topic numbers were set from 80 to 100, stepped by 5.

As shown in Fig. 4, generally, BS-LDA performed better than BTM and JST model in all the four aspects. Besides, both BTM and JST performed much better than LDA in *Accuracy*, *Precision* and *F-measure* aspects. Regarding *Recall*, nearly all methods were comparable and were stable at around 90%. Based on these findings, we can conclude that it was effective to improve BTM, JST and LDA. Besides, both BTM and JST improved the LDA model from different aspects. BTM model reduced the sparsity of distributions

from short text and JST model added the division of sentiment. Therefore, the BS-LDA model that combined these two advantages can achieve the best experiment results.

In addition to these unsupervised methods based on topic models, we also conducted another experiment on REVRANK [31]. In REVRANK, comments were converted to representation vectors and given scores according to their distance from a virtual core review vector. This method ranked the comments based on the score. The results of *Accuracy*, *Precision*, *Recall* and *F-measure* were also shown in Fig. 4. Because the REVRANK method was independent of the number of topics, the result of REVRANK was shown as a straight line rather than a polyline. As can be seen from Fig. 4, we can conclude that our method performed better than REVRANK under all the four evaluation criteria in general. Under the standard *Accuracy*, *Recall* and *F-measure*, our model performed much better than the REVRANK. Under the standard *Precision*, our model had similar results with the REVRANK model. In summary, we can say that the overall effect of our model was superior to REVRANK. From my point of view, the main reason for result was that our model took more account of semantics of the text while the REVRANK only built represent vector.

4.5.3 Comparison with Supervised Method

Based on the features extracted (as described in Sect. 4.4.3), we implemented three supervised algorithms: a NB algorithm, an SVM algorithm, and a Decision Tree (J48) algorithm. These algorithms all used 10-fold cross-validation for training and testing. We conducted experiments using the five different document lengths detailed in Sect. 4.2 and 4.5.1. For our model, we set the topic number to be 90, in which case our model performed best generally according to the experiment results of 4.5.1.

Considering the fact that the standard value of credibility of each comment and credibility threshold offered by our model (η) were all set to be median value, 5 and 0.5. Therefore, we thought it was a reasonable choice and it was theoretically feasible to compare our method with supervised method.

Based on the results shown in Fig. 5, we can draw the following conclusions. For *Accuracy*, *Precision*, and *F-measure*, the Decision Tree method performed better. Regarding *Recall*, the SVM performed better. For *Accuracy* and *Recall*, the differences between our method and the supervised methods were very small. However, in terms of *Precision* and *F-measure*, there was a gap between our method and the supervised method.

Considering all three of the experiments, our method did not perform very well in *Precision*, which had a detrimental effect on the *F-measure*. A low *Precision* value meant our method tends to judge some negative examples as positive examples. This outcome might arise because our judgment of credibility was based on the degree of concentration of the topics. Since there were situations in which some untrustworthy comment topics were concentrated, some not-credible comments could be incorporated into the credible comments.

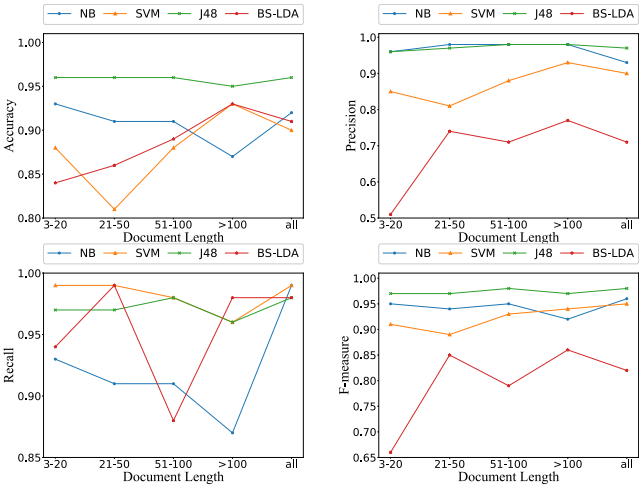

Fig. 5 *Accuracy*, *Precision*, *Recall* and *F-measure* comparison with supervised method

As we can see, our model performs relatively well for *Recall*, which meant our method is less error-prone for credible text. This finding also explained that the topics of credible comments were relatively concentrated. The number of credible comments in the corpus was obviously high, and there were few judgment errors for the authentic samples.

Since our method was unsupervised, both theoretically and practically, it was difficult for our method to perform better than classical supervised methods. In general, the overall gap between our unsupervised method and supervised methods was not particularly large, and the advantage of our method was that it did not require pre-preparation of training data. The experiments proved that in the absence of training data, our method had a certain practical value.

5. Conclusion and Future Work

This research proposes an unsupervised method, BS-LDA, to judge the credibility of consumer comments regarding products they want to purchase online. Based on our experiments, we can present three conclusions. First, the usefulness of this method increases with the length of the text. Second, the overall performance of this method is better than other unsupervised models. Third, although the final results for our proposed BS-LDA method were not as strong as for the traditional supervised methods, the overall performance was not much different. In the absence of training data, the proposed approach can play an important role in determining the credibility of consumer comments about products.

In future studies, we will explore the possibility of updating the hyper-parameters dynamically during the BS-LDA process to get a more precise topic distribution. Moreover, we can also use the HLDA model to avoid the deviation caused by the number of topics given artificially.

Acknowledgments

This study was partially sponsored by the National Key Research and Development Program of China (No. 2017YFC0907505), and the Xinjiang Social Science Foundation (No. 2015BGL100).

References

[1] M.J. Metzger, A.J. Flanagin, K. Eyal, D.R. Lemus, and R.M. Mccann, "Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment," Annals of the International Communication Association, vol.27, no.1, pp.293–335, 2003.

[2] T.L. Ngo-Ye and A.P. Sinha, "The influence of reviewer engagement characteristics on online review helpfulness: A text regression model," Decision Support Systems, vol.61, no.4, pp.47–58, May 2014.

[3] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," pp.1–7, 2012.

[4] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," AAAI, pp.2972–2978, 2016.

[5] C.C. Chen and Y.D. Tseng, "Quality evaluation of product reviews using an information quality framework," Decision Support Systems, vol.50, no.4, pp.755–768, March 2011.

[6] M.J. Metzger, A.J. Flanagin, and R.B. Medders, "Social and heuristic approaches to credibility evaluation online," J. communication, vol.60, no.3, pp.413–439, 2010.

[7] J. Aigner, A. Durchardt, T. Kersting, M. Kattenbeck, and D. Elsweiler, "Manipulating the perception of credibility in refugee related social media posts," Proc. 2017 Conference on Conference Human Information Interaction and Retrieval, pp.297–300, ACM, 2017.

[8] M. Alrubaian, M. Al-Qurishi, M. Al-Rakhami, and A. Alamri, "A credibility assessment model for online social network content," in From Social Data Mining and Analysis to Prediction and Community Detection, pp.61–77, Springer, 2017.

[9] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama, "Assessment of tweet credibility with lda features," pp.953–958, 2015.

[10] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J. American society for information science, vol.41, no.6, p.391, 1990.

[11] T. Hofmann, "Probabilistic latent semantic analysis," Proc. 15th Conference on Uncertainty in Artificial Intelligence, pp.289–296, 1999.

[12] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Machine Learning Research, vol.3, no.Jan, pp.993–1022, 2003.

[13] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," Proc. 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pp.248–256, Association for Computational Linguistics, 2009.

[14] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," Proc. 17th Int. Conf. World Wide Web, pp.111–120, ACM, 2008.

[15] H. Chen, H. Yin, X. Li, M. Wang, W. Chen, and T. Chen, "People opinion topic model: opinion based user clustering in social networks," Proc. 26th Int. Conf. World Wide Web Companion, pp.1353–1359, International World Wide Web Conferences Steering Committee, 2017.

[16] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," IJCAI, pp.1427–1432, 2009.

[17] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura, "Geo topic model: joint modeling of user's activity area and interests for location recommendation," Proc. sixth ACM International Conference on Web search and data mining, pp.375–384, ACM, 2013.

[18] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," Advances in neural information processing systems, pp.241–248, 2007.

[19] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," Proc. 18th ACM conference on Information and knowledge management, pp.375–384, ACM, 2009.

[20] S. Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," Proc. 25th Int. Conf. world wide web, pp.167–176, International World Wide Web Conferences Steering Committee, 2016.

[21] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," Advances in Neural Information Processing Systems, pp.1385–1392, 2005.

[22] T.L. Griffiths, M.I. Jordan, J.B. Tenenbaum, and D.M. Blei, "Hierarchical topic models and the nested Chinese restaurant process," Advances in Neural Information Processing Systems, pp.17–24, 2004.

[23] 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," Proc. 17th Int. Conf. World Wide Web, pp.91–100, ACM, 2008.

[24] P. Wang, H. Zhang, Y.F. Wu, B. Xu, and H.W. Hao, "A robust framework for short text categorization based on topic model and integrated classifier," 2014 International Joint Conference on Neural Networks (IJCNN), pp.3534–3539, IEEE, 2014.

[25] Y. Zhu, L. Li, and L. Luo, "Learning to classify short text with topic model and external knowledge," Int. Conf. Knowledge Science, Engineering and Management, pp.493–503, Springer, 2013.

[26] L. He, Y. Du, and Y. Ye, "Tracking topic trends for short texts," China Conference on Knowledge Graph and Semantic Computing, pp.117–128, Springer, 2017.

[27] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," Knowledge and Information Systems, vol.48, no.2, pp.379–398, Aug. 2016.

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

[29] E. Momeni, K. Tao, B. Haslhofer, and G.J. Houben, "Identification of useful user comments in social media: a case study on flickr commons," Proc. 13th ACM/IEEE-CS joint conference on Digital libraries, pp.1–10, ACM, 2013.

[30] J.J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," Proc. 22nd Int. Conf. World Wide Web, pp.897–908, ACM, 2013.

[31] O. Tsur and A. Rappoport, "Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews," ICWSM, 2009.

Xuan Wang is an undergraduate student at School of Computer Engineering and Science, Shanghai University. Her research interest covers natural language processing and machine learning.

Bofeng Zhang is a professor at School of Computer Engineering and Science, Shanghai University. His main research interest covers machine learning, social network and data mining. Corresponding author of this paper

Mingqing Huang is a Ph.D. candidate with the School of Computer Engineering and Science at Shanghai University, China. His major research interests include user modeling, complex network and social computing. He is a student member of the IEICE.

Furong Chang received the M.S. degree in 2012 from Information Technology Academe, North-West Minorities University, China. Chang is currently a lecturer in Kashgar University and is also a doctoral student in Shanghai University, China. Her research interest mainly focuses on complex network

Zhuocheng Zhou is an undergraduate student at School of Computer Engineering and Science, Shanghai University. His research interest covers machine learning algorithm and Generative Adversarial Nets