IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 8, 2020, accepted October 18, 2020, date of publication October 26, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033820

Mining Review Unit Model for Online Review Analysis

QINGXI PENG[1], LAN YOU[2], QISHENG LU[2], AND XIANGYU LI[3]
[1]Computer School, Wuhan Donghu University, Wuhan 430212, China
[2]Faculty of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China
[3]Hubei Provincial Library, Wuhan 430071, China

Corresponding author: Lan You (yoyo@huhu.edu.cn)

This work was supported in part by the Hubei Province Natural Science Foundation Item under Grant 2019CFB757, in part by the Fund of Hubei Key Laboratory of Inland Shipping Technology under Grant NHHY2017001, in part by the open fund of National Engineering Research Center for Water Transport Safety under Grant A2019011, and in part by the youth foundation of Wuhan Donghu University under Grant 2017dhzk007.

ABSTRACT An increasing number of people are choosing to shop online; hence, online reviews are an increasingly influential factor in consumer purchasing decisions. However, extracting useful information from online reviews is a challenge in the analysis of consumer sentiment. In this paper, we focus on the automatic discovery of the features evaluated in online reviews and the expression of sentiment. We propose a novel fine-grained topic model called the ''review unit topic model'' (RUTM) to extract semantic meanings and polarities. In this model, a review unit rather than a review sentence is treated as the representational model, and prior knowledge of sentiment is further exploited to identify aspect-aware sentiment polarities. We evaluate RUTM extensively using real-world review data. Experimental results demonstrate that the proposed model outperforms well-established baseline models in sentiment analysis tasks.

INDEX TERMS Review unit, topic model, review analysis, aspect.

I. INTRODUCTION

With the emergence of Web 2.0 and the increase in popularity of review sites, customers prefer to express their opinions on various types of entities, such as the products and services that they have bought. The volume of reviews is exploding, making it difficult for users to quickly find the information that they want. Discovering sentiments and opinions through analysis of a large volume of textual data is extremely difficult. Hence, in recent years, the natural language processing community has dedicated much effort to developing novel text mining approaches for review analysis.

An important task in review analysis is to identify aspects of products that users evaluate in reviews and discover how opinions and sentiments on different aspects are expressed. Aspects are attributes or components of products (e.g., 'price', 'location', etc. for a hotel), and sentiment polarity is a measure of user satisfaction in terms of positive, negative and neutral factors. Automatic sentiment analysis techniques can help users to quickly digest opinions across a large number of reviews. Among sentiment analysis techniques, aspect-level sentiment analysis is especially appealing because this type of analysis could help users effectively navigate through detailed aspect information by organizing opinions in a structured way. For example, a user booking a hotel may want to know what a review says about a room, location, price, and service of a hotel and not only whether the review recommends the hotel. Aspect-based sentiment analysis aims to extract major aspects of a product and predict the sentiment of each aspect discussed in the product reviews.

Probabilistic topic models, which are typically built on a basic latent Dirichlet allocation (LDA) model [1], have been used for aspect-based sentiment analysis [2]–[8], where the semantic aspect can be naturally formulated as one type of latent topic. Much work has extended the traditional topic model from the document level to the sentence level and, in so doing, has extracted more detailed information from reviews. Although these methods improve topic extraction to some extent, they still do not match the actual situation found in reviews. A one-to-one correspondence does not exist between sentence and aspect. In their research, Burns *et al.* found that approximately 19.73% of the sentences on the TripAdvisor

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

website discuss two or more aspects [9]. Zhao *et al.* [10] found that ca. 83% of sentences in hotel and restaurant review data contained only one review aspect, but many sentences contained two or more aspects. This research suggests that sampling from sentences in reviews is sufficiently accurate to reflect the true situation of reviews, causing loss of aspect information. Most current work uses the bag-of-words representation to extract aspects and sentiments from reviews. The relationship between words is ignored, which makes the topic model weak in modeling ability. To handle this problem, a number of methods have been proposed in recent years. For example, Ei-Kishky *et al.* proposed a phrase-LDA model to improve LDA [11]. A phrase is a combination of multiple words obtained by preprocessing the original review. Other methods employ the bag-of-phrase model to improve topic modeling [3], [12], and that work mainly focuses on rating prediction. In this paper, we propose a novel review representation, the review unit, to match the characteristics of a review itself. Our approach divides reviews into review units and incorporates the review unit into the LDA model to improve the extraction. We argue that representation of a review by review unit could be potentially useful for improving aspect and sentiment extraction.

We represent each text review as a bag-of-review unit, where each review unit consists of an aspect word and corresponding sentiment word in the review. We extend the traditional LDA model and construct a probabilistic review unit topic model (RUTM), which simultaneously detects aspects and sentiments. An inference method is presented, and prior knowledge is introduced to the proposed model. Furthermore, we provide a simple but efficient review unit mining algorithm. To our knowledge, no other existing approach exhibits the same merits as our model.

The main contributions of this study are summarized as follows:

1. We present a novel review representation-review unit to handle the intrinsic relationship between review and aspect and apply a simple but efficient review unit mining algorithm to split the aspect and corresponding sentiment from the review.

2. We propose a review unit topic model (RUTM) that incorporates the LDA model with a review unit to improve aspect extraction.

3. We present a detailed inference method for RUTM based on collapsed Gibbs sampling.

4. This work evaluates RUTM against four representative baseline methods and experimentally demonstrates the effectiveness of the RUTM model.

The rest of this paper is organized as follows. We discuss the related work to sentiment analysis and aspect extraction in Section 2. In Section 3, the review unit topic model is proposed, the model is inferred with collapsed Gibbs sampling, and a review unit mining algorithm is proposed. Section 4 presents the empirical experiments to evaluate the proposed model. Finally, the conclusions of our study are given in Section 5.

II. RELATED WORK

Sentiment analysis is a well-studied problem [13]. The most common sentiment analysis problem is classification of a text into either positive or negative polarity. Recently, interest has grown in sentiment analysis using topic models such as LDA [1] and probabilistic latent semantic analysis (pLSA) [14]. The LDA and pLSA models consider documents as a "bag of words" in which a document is represented as a multinomial distribution over a topic, and a topic is represented as a multinomial distribution over words. A document is generated using these distributions. The traditional LDA model can only handle coarse-grained document-level text analysis.

Review analysis research has extended LDA to overcome the shortcomings of standard LDA in review analysis. Titov *et al.* presented a multigrain LDA model (MG-LDA) and multiaspect sentiment model (MAS) [15]. Those models not only extract aspects but also cluster aspects into coherent topics. This method differentiates these approaches from much of the previous work, which only extracts aspect through term frequency analysis, with minimal clustering. However, sentiment analysis was not involved in this research.

Zhao *et al.* proposed a MaxEnt-LDA hybrid model that added a maximum entropy component to the LDA model for joint discovery of both aspects and aspect-specific sentiment words [10]. The MaxEnt component allowed the model to leverage arbitrary features such as POS tags to help separate aspect and sentiment words. In that research, words had the same topic as their sentences. In addition, sentiment polarity was not involved in their model.

Lin *et al.* proposed a fully unsupervised joint sentiment/topic (JST) model [16]. The JST model detects sentiments and topics simultaneously. Various approaches were explored to obtain prior information to improve the sentiment detection accuracy. However, the JST model cannot deliver sentiment polarity detection.

Jo and coworkers proposed two models: the sentence-LDA (SLDA) and aspect sentiment unification models (ASUM) [2]. SLDA and ASUM are constrained by the assumption that all words in a single sentence were generated from one topic. ASUM is an extension of SLDA into which sentiment was incorporated. Compared with previous studies, SLDA and ASUM are more fine-grained sentiment analysis tools. However, the SLDA model assumes that one sentence contains one aspect, which does not reflect the true situation in reviews.

Ma *et al.* proposed a topic and sentiment unification maximum entropy model (TSU) [17] in which a maximum entropy component is added to the TSU model. A sentiment layer was inserted between the topic layer and word layer to extend the proposed model from the traditional three layers to four layers. However, the TSU model also operates under the assumption that each sentence belongs to only one topic and one sentiment.

These efforts have shown that sentence-level analysis improves performance over document-level analysis.

Many researchers have proposed more fine-grained models. Burns *et al.* proposed a two-fold LDA model to identify both aspects and positive or negative sentiments in review sentences [8]. One LDA runs for aspect extraction while another LDA runs for sentiment identification. The twofold-LDA modeled topics and sentiments separately. These authors also proposed an enhanced version that incorporated part-of-speech tagging (POS) into the twofold-LDA modeling process [18].

Ei-Kishky and colleagues proposed a phrase-LDA, which combines a novel phrase mining framework to segment a document into single and multiword phrases and a new topic model that operates on the induced document partition [11]. However, phrase-LDA only focuses on aspects and neglects aspect-specific sentiments.

Hai *et al.* proposed a supervised joint aspect and sentiment model (SJASM) to address the problem in one go under a unified framework [19]. Their model requires labeled rating data for reviews in supervised learning, which is different from our application scenario.

Lu *et al.* argued that the use of preprocessed reviews could improve the ability of models to identify aspects [20]. They assumed that each review can be parsed into an opinion phrase and proposed a probabilistic model based on PLSI to identify major aspects of a product by clustering the head terms.

Moghaddam and Ester proposed interdependent latent Dirichlet allocation (ILDA) [3], which learns a set of product aspects and corresponding ratings from a collection of opinion phrases that have been preprocessed into a collection of opinion phrases. Because our proposed model uses preprocessed reviews as input, this model is the comparative partner for our work.

Alam *et al.* proposed a joint multigrain topic sentiment (JMTS) model [21], which extends MG-LDA [15] by constructing an additional sentiment layer on the presumption that aspects are generated from window-based distributions of topics and sentiment. The JMTS model breaks from the sentence-level modeling assumption, but it combines words from adjacent sentences, resulting in reduced modeling ability.

Sindhu and colleagues applied aspect-based sentiment analysis techniques to the field of education [22] and proposed a two-layered LSTM model for student feedback on faculty teaching performance. The first layer predicted the aspects described within the feedback and later specified the orientation of those predicted aspects. Those researchers also pointed out that the presence of multiple aspects within the review sentence may lead to sentiment misclassification.

Kastrati *et al.* also studied aspect-based sentiment analysis on student reviews of MOOCs [23]. Their framework took advantage of weakly supervised annotation of MOOC-related aspects and automatically identified aspects and sentiment expressed towards a given aspect. This framework was tested and validated on two real-life datasets.

Ma and coworkers found that one opinionated sentence contains multiple aspects [24], and they developed a two-stage paradigm to model the explicit position context between the aspect and its context words and simultaneously process multiple aspects within one opinionated sentence. The assumption underpinning this study is similar to ours. Their solution relies on Gaussian kernel-based position-aware influence propagation. Our review unit mining algorithm is more effective because it is incorporated with a sentiment lexicon.

In this work, we extract aspects and sentiments using a review unit topic model (RUTM). Our assumption is that one sentence may correspond to multiple aspects. Hence, we employ the review unit rather than a sentence as the representative model. With prior knowledge, we split the sentences into several review units. RUTM simultaneously identifies aspects and sentiments from reviews.

III. PROPOSED MODEL

We propose a generative model that extends LDA, one of the most widely used probabilistic topic models [1]. Our goal is to discover the aspect and corresponding sentiments in reviews.

A. REVIEW UNIT TOPIC MODEL

In the work of Jo and Oh [2], the sentence-LDA model assumes that a sentence contains only one aspect, which is contrary to the facts. As discussed in the previous section, this property may not always be appropriate. Our assumption is that all the reviews are composed of review units.

Before going further, we make the following definitions.

Aspect Keywords: Words representing aspects. Denoted by w.

Sentiment Keywords: Words representing sentiments. Denoted by q.

Review Unit: The basic unit of a review. Denoted by h. The review unit usually consists of an aspect keyword and a sentiment keyword. $h=<w, q>$.

Fig. 1 shows an example of a review unit.

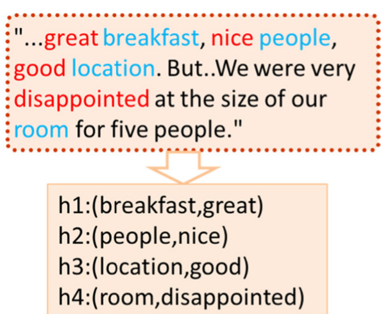

FIGURE 1. Review unit example.

The review text structure in Fig. 1 is relatively fixed with two sentences. According to Jo and Oh [2], the first sentence contains only one aspect. Based on these observations, our approach processes reviews and represents them through a collection of review units and thus splits this text structure into four aspects.

The RUTM model is based on this assumption: A review consists of multiple review units. Each review unit corresponds to one aspect, but one review corresponds to one or more aspects. RUTM acquires prior knowledge through a review unit mining algorithm (RUMA). The representation model of RUTM is not a bag-of-words but a review unit. Therefore, with prior knowledge, RUTM improves modeling ability. Fig. 2 depicts a graph model representation of RUTM. To compare RUTM with ASUM [2], we depict ASUM schematically in Fig. 3. According to the representation of the graphical model, nodes are random variables, edges are dependencies, and plates are replications. In Fig. 2, *a, s, w,* and q represent the aspect, sentiment, aspect keywords and sentiment keywords, respectively; M represents the number of reviews in the review collection; N represents the number of review units in review d; K represents the number of aspects; L represents the number of sentiments; α, β, γ, and λ are the hyperparameters of the Dirichlet distribution; θ is an $M \times K$ matrix sampled from Dirichlet(α), representing the distribution from document to aspect; ψ is a $K \times U$ matrix sampled from Dirichlet(λ), representing the distribution from aspect to aspect keyword; π is a $K \times L$ matrix sampled from Dirichlet(γ), representing the distribution from aspect to sentiment; and ϕ is an $L \times V$ matrix, representing the distribution from sentiment to sentiment keyword.

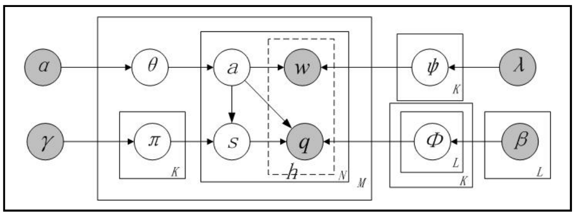

FIGURE 2. Graphical model representation of RUTM.

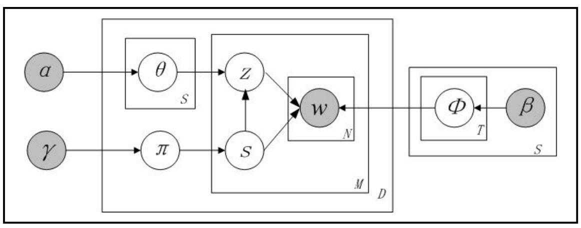

FIGURE 3. Graphical model representation of ASUM.

After preprocessing, all the review units $\{h_1, h_2, \ldots\}$ are generated from the review text. The process for each review unit is listed below.

1. For each aspect k,
   (a) Draw $\psi_k \sim$ Dirichlet(λ)
   (b) For each sentiment polarity l draw $\phi_{kl} \sim$ Dirichlet(β_l)
2. For each review d,
   (a) Draw $\theta_d \sim$ Dirichlet(α)
   (b) For each aspect k in document d, draw $\pi_{dk} \sim$ Dirichlet(γ)
   (c) For each review unit h,
      i Choose an aspect $a \sim$ Multinomial(θ_d)
      ii Choose a sentiment s $\sim$ Multinomial(π_{dk})
      iii Generate aspect keyword $w_{mn} \sim$ Multinomial(ψ_a)
      iv Generate sentiment keyword $q_{mn} \sim$ Multinomial(ϕ_s)

We compared the RUTM model and ASUM model as follows. To extract aspects, ASUM reduces the word co-occurrence information from the document level to the sentence level. The basic assumption is that "a sentence corresponds to an aspect". It can be observed that the central ASUM assumption cannot accurately reflect the actual situation of a review. To solve the problem of "one sentence corresponds to multiple aspects", the RUTM model exploits the review unit model, bringing it in line with the distribution of aspects in the review. The RUTM model also introduces prior knowledge and uses the review unit as a representation model to improve the modeling ability. For each review, RUTM samples an aspect, chooses the aspect keywords and subsequently chooses a sentiment keyword according to aspect and sentiment polarity. Unlike the ASUM model, the modeling process in the RUTM model is more in line with the intuitive understanding of human users.

B. MODEL INFERENCE

We adopted the collapsed Gibbs sampling algorithm [25] to derive the model and estimate the four parameters in the model. We use $h_i =< w_i, q_i>$ to represent the review unit. To solve the conditional probabilities,

$$p\,(a, s \,|\, w, q, \alpha, \beta, \gamma, \lambda) = \frac{p\,(a, s, q, w \,|\, \alpha, \beta, \gamma, \lambda)}{p\,(q, w \,|\, \alpha, \beta, \gamma, \lambda)} \quad (1)$$

we use the following equation:

$$p\left(a_i = t, s_i = l \,|\, a^{\neg i}, s^{\neg i}, w, q, \alpha, \beta, \gamma, \lambda\right) \quad (2)$$

$a^{\neg i}$ means that all aspects are allocated except aspect i-th.
$s^{\neg i}$ means that all sentiments are allocated except i-th.
We expand Equation (2) as follows:

$$\begin{aligned} p&\left(a_i = t, s_i = l \,|\, a^{\neg i}, s^{\neg i}, w, q, \alpha, \beta, \gamma, \lambda\right) \\ &= \frac{p\,(a, s, w, q \,|\, \alpha, \beta, \gamma, \lambda)}{p\left(a^{\neg i}, s^{\neg i}, w, q \,|\, \alpha, \beta, \gamma, \lambda\right)} \\ &\propto p\,(a, s, w, q \,|\, \alpha, \beta, \gamma, \lambda) \end{aligned} \quad (3)$$

According to the dependency of the RUTM model,

$$\begin{aligned} &p\,(a, s, w, q \,|\, \alpha, \beta, \gamma, \lambda) \\ &\quad = p\,(a \,|\, \alpha)\, p\,(s \,|\, \alpha, \gamma)\, p\,(w \,|\, \alpha, \lambda)\, p\,(q \,|\, a, s, \beta) \end{aligned} \quad (4)$$

Finally, we obtain Equation (5):

$$\begin{aligned} p&\left(a_i = t, s_i = l \,|\, a^{\neg i}, s^{\neg i}, w, q, \alpha, \beta, \gamma, \lambda\right) \\ &\propto \frac{N_{m,k}^{\neg i} + \alpha}{N^{\neg i} + \sum_k' \alpha_{k'}} \cdot \frac{N_{m,k,l}^{\neg i} + \gamma}{N_{m,k}^{\neg i} + L\gamma} \\ &\quad \cdot \frac{N_{k,u}^{\neg i} + \lambda}{N_k^{\neg i} + |U|\,\lambda} \cdot \frac{N_{k,l,v}^{\neg i} + \beta_{l,v}}{N_{k,l}^{\neg i} + \sum_v' \beta_{l,v'}} \end{aligned} \quad (5)$$

TABLE 1. Sentiment lexicon description.

Lexicon Name	# of Positive	# of Negative	Description
HowNet	4332	4574	English Sentiment Lexicon of Chinese words
SentiWordNet	2290	4800	Words with positive or negative score above 0.6
MPQA	2304	4152	MPQA subjectivity lexicon
UIC	2006	4783	Sentiment Lexicon from Bing Liu[26]

In Equation (5), N represents the total number of words in reviews, $N_{m,k}$ represents the number of words that belong to aspect k in review d_m, $N_{m,k,l}$ represents the number of words that belong to aspect k and sentiment l in review d_m, and $N_{k,u}$ represents the number of aspect keywords u that belong to aspect k in dictionary U. After Gibbs sampling, we obtain the parameter values,

$$\theta_{m,k} = \frac{N_{m,k} + \alpha_k}{N + \sum_{k'=1}^{K} \alpha_{k'}} \tag{6}$$

$$\pi_{m,k,l} = \frac{N_{m,k,l} + \gamma}{N_{m,k} + L\gamma} \tag{7}$$

$$\psi_{k,u} = \frac{N_{k,u} + \lambda}{N_k + |U|\,\lambda} \tag{8}$$

$$\phi_{k,l,v} = \frac{N_{k,l,v} + \beta_{l,v}}{N_{k,l} + \sum_{v'=1}^{|V|} \beta_{l,v'}} \tag{9}$$

C. PRIOR KNOWLEDGE

The existing sentiment lexicons are the prior knowledge in our model. We collected four sentiment lexicons and used their union set as a sentiment dictionary in the RUTM model.

D. REVIEW UNIT MINING ALGORITHM

Wang *et al.* proposed a review aspect splitting algorithm based on boot strapping [27]. This algorithm has been widely used because it is simple and efficient. We modify this algorithm with sentiment lexicons for review unit mining. We manually input some seed aspect keywords w_i and continuously extend the aspect keyword in the process of identifying the review unit. The review unit mining algorithm is shown below.

For step 1, we first split the reviews into sentences and then split each sentence into several subsentences using punctuation (such as commas) and sentential connectives (such as "but, and"), making sure that the length of subsentence is no less than a threshold value δ. Although syntactic analysis can make subsentence splitting more accurate, we did not employ it because this step is a filter. The boot-strapping algorithm can identify the review unit in the following process.

Algorithm Review Unit Mining Algorithm

Input:	$R = \{r\}$//review collection
	$W = \{w_i\}$//aspect keyword list, w_i are seeds
	D//sentiment Lexicon
	p//selection threshold
	$iter^{max}$//iteration limit
Output:	H//review unit collection
1	Split all reviews into subsentences X=$\{x_1, x_2, \ldots x_M\}$, put all words into vocabulary V after removing duplicates
2	Match the aspect keywords w_i in each subsentences of X and record the matching hits for each aspect i in count(i);
3	Assign the subsentence an aspect label by $x_i =$ *argmax count(i)*. If there is a tie, assign the subsentence with multiple aspects.
4	Find sentiment keyword q_i in subsentence and D, get review unit h_i=<w_i, q_i>
5	$H = H \cup h_i$//add to review unit collection
6	Compute χ^2 measures of each word w (in V)
7	Sort each word w according to χ^2 value to each review unit h_r
8	If aspect keyword list W is unchanged or iteration exceeds p, go to **step 9**, else go to **step 2**
9	Output H

The χ^2 statistic computes the dependencies between a term w and review unit h_r *and* is defined as follows:

$$\chi^2(w, h_r) = \frac{C \times (C_1C_4 - C_2C_3)^2}{(C_1{+}C_3) \times (C_2{+}C_4) \times (C_1{+}C_2) \times (C_3{+}C_4)} \tag{10}$$

where C_1 is the number of times w occurs in subsentences belonging to review unit h_r, C_2 is the number of times w occurs in subsentences not belonging to review unit h_r, C_3 is the number of subsentences of review unit h_r that do not contain w, C_4 is the number of subsentences that neither belong to review unit h_r nor contain word w, and C is the total number of word occurrences.

IV. EXPERIMENTS AND ANALYSIS

To assess the effectiveness of our RUTM model, we used several product review datasets to perform experiments.

A. DATASETS AND PREPROCESSING

Our machine configuration consisted of an Intel(R) Core(TM) i5-2450 M CPU, 2.50 GB memory, Windows Server 2012, Python 3.5.2, Numpy 1.11.2, SciPy 0.17.0, and Scikit-learn 0.19.1.

Our datasets come from product reviews, including a book dataset from Amazon [28], a hotel dataset from TripAdvisor [27], and a restaurant dataset from Yelp. Table 2 shows the dataset statistics.

TABLE 2. Dataset statistics.

Dataset	Book	Hotel	Restaurant
#of reviews	1254	2693	3037
#of review unit	3782	10074	8654
Review unit/sentence	1.43	2.87	2.44

We performed preprocessing in the following steps.

1. Remove the nontext characters
2. Check misspelled words
3. Lemmatization
4. Convert all words to lowercase
5. Calculate the number of stopwords

B. EXPERIMENTAL SETTINGS

This paper chooses four baselines:

ASUM: ASUM extracts the aspect and sentiment simultaneously [2]. The ASUM model assumes that each sentence corresponds to one aspect. The model first extracts sentiment and then extracts the corresponding aspect.

ILDA: The ILDA model also realizes the simultaneous extraction of aspect and sentiment [3]. The representative model of ILDA is phrase, employing the frequency of nouns in reviews to process the review to obtain a phrase collection. The ILDA model assumes that a sentiment corresponding to an aspect is determined by the latent rating, extracts the aspects and the ratings of the aspects, and obtains the sentiment keywords from the ratings. Compared with the ILDA model, the ASUM model is closer to the real situation.

JMTS: JMTS incorporates window-based sentiment distributions and window-based topic distributions [21]. The JMTS model is an extension of JST [16] and the MG-LDA model [15].

TSU: The TSU model incorporates a maximum entropy component [17]. A sentiment layer was inserted between the topic layer and word layer to extend the traditional three layers to four layers. The model also assumes that each sentence is associated with only one topic and sentiment and that each word has the same topic and sentiment as its sentence.

We set the configuration parameters of the RUTM model. The number of aspects $k = 5$; the number of sentiments $L = 2$, representing positive and negative; super parameter $\alpha = 1$; super parameter β are 0.35 to positive and 0.65 to negative; $\gamma = 0.5$; and $\lambda = 0.01$. For baseline models, we use their default settings.

C. EXPERIMENTAL RESULT

1) ASPECT IDENTIFICATION

In this section, we evaluate the accuracy when identifying the k major aspects in the given test sets. The accuracy of aspect identification can be verified by the Rand index [29]. The Rand index is a metric used to evaluate the similarity of two clusters.

For n elements $s = \{o_1, o_2, \ldots o_n\}$, two clustering models X and Y run on the same dataset s to obtain $X = \{X_1, X_2, \ldots X_r\}$ and $Y = \{Y_1, Y_2, \ldots Y_t\}$.

n_1 is the number of elements that are assigned to the same cluster in X and the same cluster in Y.

n_2 is the number of elements that are assigned to different clusters in X but the same cluster in Y.

n_3 is the number of elements that are assigned to the same cluster in X but different clusters in Y.

n_4 is the number of elements that are assigned to different clusters in X and different clusters in Y.

Thus,

$$RandIndex\,(X, Y) = \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4} \tag{11}$$

In Equation (11), X and Y represent the clustering as generated by the topic model and manually, respectively. We randomly selected 500 reviews from three datasets and annotated them manually as model Y. Higher Rand index values indicate higher clustering similarity of the topic model X and Y, which implies the efficiency of the model that identifies the aspects. We compare the models in Fig. 4.

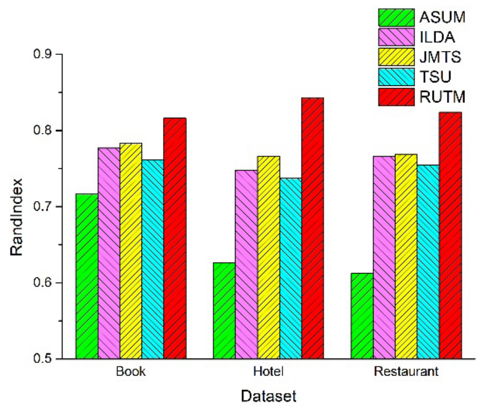

FIGURE 4. Rand index of different models vs. datasets.

Fig. 4 illustrates that RUTM improved the ability to identify aspects. RUTM achieved the highest Rand index value of 84.3% on the hotel dataset, which is 10.5%, 7.7%, 9.5% and 21.7% higher than those of TSU, JMTS, ILDA and ASUM, respectively. On the restaurant dataset, the RUTM model also achieved the highest Rand index value of 82.4%, which was 6.9%, 5.5%, 5.8% and 21.2% higher than those of TSU, JMTS, ILDA and ASUM, respectively. Similarly, on the book dataset, RUTM also obtained a high modeling ability score, with a Rand index value of 81.6%, which is 5.5%, 3.3%, 3.9% and 10.5% higher than those of TSU, JMTS, ILDA and ASUM, respectively. Compared with the baseline models, RUTM shows a strong modeling ability on different datasets.

Among these baseline models, ASUM and TSU follow the assumption that one sentence corresponds to one aspect. JMTS did not follow this assumption, instead following window-based sentiment distributions and window-based topic distributions, which improves its modeling ability.

The ASUM model results differ greatly for the three datasets, as the modeling ability on the hotel and restaurant

TABLE 3. Sampling results of RUTM.

value				room				location			
positive		negative		positive		negative		positive		negative	
aspect	opinion	aspect	opinion	aspect	opinion	aspect	opinion	aspect	opinion	aspect	opinion
value	good	hotel	expensive	decor	great	balcony	little	location	great	parking	problem
price	affordable	store	overpriced	furniture	beautiful	table	tiny	station	near	scene	annoying
rate	cheap	price	outrageous	room	comfortable	shape	oddly	view	lovely	gym	tiny
money	worth	money	waste	bathroom	comfortable	room	small	place	fine	hotel	noisy
pay	cheap	restaurant	expensive	internet	free	lighting	poor	restaurant	decent	neighborhood	bad
eat	cheap	parking	nightmare	bed	fabulous	carpet	worn	landscape	enjoyable	location	disgusting
charge	normal	breakfast	expensive	window	large	pillow	tiny	setting	quiet	station	far

datasets was low because users may mention more than one aspect in one sentence. Generally, the language adopted by users on the book dataset was more standardized, which makes the ASUM model perform more effectively.

With the RUTM model, the aspect and opinion were identified. The sampling results of aspect words and opinion words in RUTM are shown in Table 3.

2) SENTIMENT CLASSIFICATION

We evaluated sentiment classification results produced by the four different models. Among these models, only the RUTM model performs aspect identification and sentiment classification simultaneously. The JMTS model could not identify sentiment, and ASUM, ILDA, and TSU could only identify sentiment indirectly. ILDA can predict the specific aspect rating, whereas ASUM and TSU can predict the review sentiment polarity, as the number of aspects can influence the accuracy of sentiment classification.

To address this problem, we evaluated the overall sentiment of reviews. This method has been widely adopted [17], [30]. We labeled the sentiment polarity in the reviews, and all the reviews are labeled by five stars in the book, hotel and restaurant datasets. We selected reviews with four and five stars as positive reviews and reviews with one and two stars as negative reviews. The reviews with three stars were considered neutral and were not included in this experiment.

ASUM: The parameter π of the ASUM model determines the sentiment, where $\pi > 0$ means positive reviews, and $\pi < 0$ means negative reviews.

ILDA: ILDA does not directly determine the aspect-specific sentiment but generates the sentiment word by predicting the aspect-specific rating. We employ the aspect-specific rating r to estimate the sentiment polarity. We set this rule: $r > 3$ means positive reviews, and $r < 3$ means negative reviews.

JMTS: JMTS could not identify sentiment. Therefore, it is not included in this experiment.

TSU: The TSU model incorporates a maximum entropy component. A sentiment layer is inserted between the topic layer and the word layer to extend the traditional three layers to four layers. This model also assumes that each sentence is associated with one topic and sentiment and that each word has the same topic and sentiment as its sentence.

RUTM: The parameter π in RUTM determines the aspect-specific sentiment. If the aspect distribution θ_{mk} of review d and the aspect-specific sentiment distribution π_{ml} are known, then we can obtain the review sentiment by the weighted average value:

$$P(l = l_i \mid d) = \sum_{m=1}^{k} \theta_{dm} \cdot \pi_{ml} \tag{12}$$

We used accuracy as the evaluation standard. The accuracy is defined as follows:

$$Accuracy = \frac{M}{N} \tag{13}$$

where N is the total number of reviews in our collection, and M is the number of correctly predicted reviews. A higher accuracy value indicates more effective performance.

To study how the number of aspects influences the sentiment classification, we set the number of aspects k=5, 10, 15, 20, 25, and 30 when conducting the experiment.

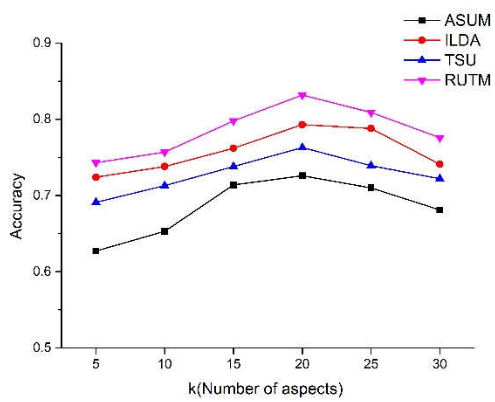

FIGURE 5. Sentiment classification vs. number of aspects on book dataset.

Figures 5-7 compare the number of aspects and accuracy of sentiment classification, including similar accuracy trends for all three datasets. An increase in the number of aspects is associated with an increase in sentiment classification accuracy. Accuracy decreases with an increasing number of aspects when a certain value is reached.

In Fig. 5, the models are ranked according to the accuracy of sentiment classification for the book dataset from high to low as RUTM, ILDA, TSU and ASUM. Among all models, RUTM outperformed the baseline models. The accuracy of RUTM increased as the number of aspects increased from

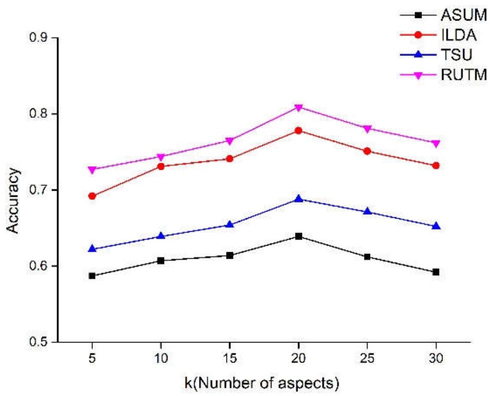

FIGURE 6. Sentiment classification vs. number of aspects on the restaurant dataset.

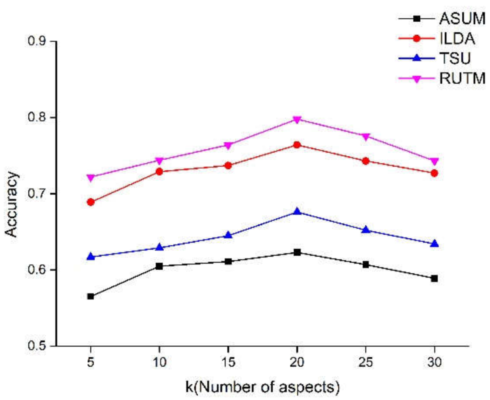

FIGURE 7. Sentiment classification vs. number of aspects on hotel dataset.

5 to 20. For the book dataset, the RUTM model acquires the highest accuracy of 83.2% when the number of aspects k=20.

For the restaurant review dataset, the RUTM model also yielded the highest accuracy of 80.9% when the number of aspects k=20.

As shown in Fig. 6 and Fig. 7, although the trends in growth and decrease in accuracy are similar to those in Fig. 5, the gaps are different in accuracy for several of the models. For the ASUM and TSU models, the accuracy decreased significantly in the restaurant and hotel review datasets. As shown in Table 2, the average numbers of aspects in each sentence in the restaurant and hotel datasets are 2.87 and 2.44, respectively, which are greater than the 1.43 in the book dataset. Because the ASUM and TSU models are all based on the assumption that one sentence only contains one aspect, their modeling ability becomes weaker in the restaurant and hotel datasets. This analysis also verifies the robustness of RUTM. The ILDA employs phrase-level rather than sentence-level modeling, making it more accurate than ASUM and TSU, but RUTM uses a more fine-grained modeling technique. The use of prior knowledge yields a more powerful modeling ability than the other tested baseline models.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel fine-grained topic model, RUTM. RUTM extracts aspects and corresponding sentiments simultaneously and employs a review unit rather than review sentence as the representation model. By incorporating prior knowledge, RUTM delivers powerful modeling ability. To assess our proposed methods, we conducted two experiments on three datasets and compared four baseline models. The experimental results demonstrate that the proposed model representing reviews with aspect-based sentiment is effective and outperforms the existing approaches.

For future work, it is vital to extend our model to fit more cross-domain applications. Additionally, we will study whether the application of word embedding could have a positive impact on fine-grained sentiment analysis.

REFERENCES

[1] M. D. Blei, Y. A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003, doi: 10.1162/jmlr.2003.3.4-5.993.

[2] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. 4th ACM Int. Conf. Web Search Data Mining - WSDM*, 2011, pp. 815–824.

[3] S. Moghaddam and M. Ester, "ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. SIGIR*, 2011, pp. 665–674.

[4] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012, doi: 10.1109/TKDE.2011.48.

[5] S. Kim, J. Zhang, Z. Chen, A. Oh, and S. Liu, "A hierarchical aspect-sentiment model for online reviews," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 526–533.

[6] M. Dermouche, J. Velcin, L. Khouas, and S. Loudcher, "A joint model for topic-sentiment evolution over time," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 773–778, doi: 10.1109/ICDM.2014.82.

[7] L. You, Q. Peng, Z. Xiong, D. He, M. Qiu, and X. Zhang, "Integrating aspect analysis and local outlier factor for intelligent review spam detection," *Future Gener. Comput. Syst.*, vol. 102, pp. 163–172, Jan. 2020, doi: 10.1016/j.future.2019.07.044.

[8] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "A practical guide to sentiment analysis," in *A Practical Guide to Sentiment Analysis*, vol. 5. Cham, Switzerland: Springer, 2017, pp. 1–10, doi: 10.1007/978-3-319-55394-8_1.

[9] N. Burns, Y. Bi, H. Wang, and T. Anderson, "Extended twofold-LDA model for two aspects in one sentence," in *Advances in Computational Intelligence*. Berlin, Germany: Springer, 2012, pp. 265–275, doi: 10.1007/978-3-642-31715-6_29.

[10] W. X. Zhao, J. Jing, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Cambridge, MA, USA, 2010, pp. 56–65.

[11] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable topical phrase mining from text corpora," *Proc. VLDB Endowment*, vol. 8, no. 3, pp. 305–316, Nov. 2014.

[12] P. Lv, X. Wang, Y. Luo, and C. Ji, "(Aspect,Rating) summarization based on topic model," *Acta Electronica Sinica*, vol. 44, no. 12, pp. 3036–3043, 2016, doi: 10.3969/j.issn.0372-2112.2016.12.032.

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2002.

[14] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, 2001, doi: 10.1023/a:1007617005950.

[15] I. Titov and R. McDonald, "Modeling online reviews with multigrain topic models," in *Proc. Int. Conf. Worldwide Web*, Beijing, China, 2008, pp. 111–120, doi: 10.1145/1367497.1367513.

[16] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. - CIKM*, 2009, pp. 375–384.

[17] Changlin Ma, Meng, Wang, and Xuewen Chen, "Topic and sentiment unification maximum entropy model for online review analysis," in *Proc. 24th Int. Worldwide Web Conf.*, Florence, Italy, 2015, pp. 649–654, doi: 0.1145/2740908.2741704.

[18] N. Burns, Y. Bi, H. Wang, and T. Anderson, "Enhanced twofold-LDA model for aspect discovery and sentiment classification," *Int. J. Knowl.-Based Org.*, vol. 9, no. 4, pp. 1–20, Oct. 2019, doi: 10.4018/IJKBO.2019100101.

[19] Z. Hai, G. Cong, K. Chang, P. Cheng, and C. Miao, "Analyzing sentiments in one go: A supervised joint topic modeling approach," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 6, pp. 1172–1185, Jun. 2017, doi: 10.1109/TKDE.2017.2669027.

[20] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarization of short comments," in *Proc. 18th Int. Conf. World Wide Web WWW*, 2009, pp. 131–140.

[21] M. H. Alam, W.-J. Ryu, and S. Lee, "Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews," *Inf. Sci.*, vol. 339, pp. 206–223, Apr. 2016, doi: 10.1016/j.ins.2016.01.013.

[22] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, "Aspect-based opinion mining on Student's feedback for faculty teaching performance evaluation," *IEEE Access*, vol. 7, pp. 108729–108741, 2019, doi: 10.1109/ACCESS.2019.2928872.

[23] Z. Kastrati, A. S. Imran, and A. Kurti, "Weakly supervised framework for aspect-based sentiment analysis on Students' reviews of MOOCs," *IEEE Access*, vol. 8, pp. 106799–106810, 2020, doi: 10.1109/ACCESS.2020.3000739.

[24] X. Ma, J. Zeng, L. Peng, G. Fortino, and Y. Zhang, "Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis," *Future Gener. Comput. Syst.*, vol. 93, pp. 304–311, Apr. 2019, doi: 10.1016/j.future.2018.10.041.

[25] M. W. Darling, "A theoretical and practical implementation tutorial on topic modeling and Gibbs sampling," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 642–647.

[26] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the Web," in *Proc. 14th Int. Conf. World Wide Web WWW*, 2005, pp. 342–351.

[27] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data: A rating regression approach," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2010, pp. 783–792.

[28] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. 7th ACM Conf. Recommender Syst. - RecSys*, 2013, pp. 165–172, doi: 10.1145/2507157.2507163.

[29] W. M. Rand,"Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

[30] F. Li, M. Huang, and X. Zhu, "Sentiment analysis with global topics and local dependency," in *Proc. 24th AAAI Conf. Artif. Intell.*, Atlanta, GA, USA, 2010, pp. 1371–1376.

QINGXI PENG received the Ph.D. degree in computer software and theory from Wuhan University.

He is currently a Professor with the Computer School, Wuhan Donghu University. His research interests focus on machine learning, natural language processing, and information retrieval. He also specializes in the design and implementation of large-scale distributed web extraction. He is the author of one book, more than 50 articles, and seven inventions.

Dr. Peng is member of China Computer Federation, and member of the International Association of Engineers (IAENG).

LAN YOU was born in Wuhan, China, in 1978. She received the Ph.D. degree from Wuhan University, LIESMARS, in 2015.

She is currently an Associate Professor with the Computer Science and Information Technology Faculty, Hubei University. She completed her internship at Microsoft Azure, Shanghai, China, in 2012. She worked as a Research Associate with the Chinese University of Hong Kong, from 2015 to 2017. Her major studies are spatiotemporal data mining, knowledge graph, and virtual geographical environment. She has published widely in the field with more than 20 publications.

QISHENG LU received the B.S. degree in computer school from the Huazhong University of Science and Technology, in 2017. He works as a Research Assistant with Hubei University, specializing in machine learning and distributed systems development.

XIANGYU LI received the Ph.D. degree in computer architecture from Wuhan University, China. He is currently with Hubei Provincial Library as an Associate Professor. He is especially interested in file and storage systems, high performance computing, distributed systems, and computer networks.