Electronic Commerce Research and Applications 53 (2022) 101141

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Mining multi-brand characteristics from online reviews for competitive analysis: A brand joint model using latent Dirichlet allocation

Yuxuan Guo [a], Feifei Wang [a], Chen Xing [b], Xiaoling Lu [a,*]

[a] *Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China*
[b] *Head Office, Bank of China, Beijing 100818, China*

ARTICLE INFO

Keywords:
Brand competitive analysis
Latent Dirichlet allocation
Sentiment analysis
Reviews mining

ABSTRACT

Online reviews reflect customers' opinions of the products or services they have bought. Analysing online reviews provides a reliable way for e-commerce platforms to understand users' needs and attitudes. To understand the strengths and weaknesses of several competitive brands, we propose a brand joint latent Dirichlet allocation model to analyse multi corpora simultaneously, particul s for multi-brre the general aspects of the online opinions of users concerning multi-brands and specific aspects of user opinions within individual brands. The results can assist the analysis of competitive brands and make meaningful suggestions for brand managers and marketers. We present two case studies to prove the efficiency of the proposed model.

1. Introduction

Assessing the strengths and weaknesses of competitors is the key goal of competitive analysis in the marketing and strategic management field (Fleisher and Bensoussan, 2003). Brand competitive analysis is particularly meaningful because brands in the same product category offer similar products or services to the same target market and to the same target dience. The analysis of competing brands offers deep insights that are useful to managers, such as the unique selling propositions of their brands and customer needs. This type of information is valuable for strategic policy (Choi and Fredj, 2013). Moreover, from the customer's perspective, learning more about brand competitiveness enables the consumer to make informed purchase decisions depending on their individual needs and a brand's features.

Online reviews reveal customers' opinions and sentiment towards the products and brands they have bought (Archak et al., 2011; Decker and Trusov, 2010). Online reviews are, therefore, a promising data source for brand competitive analysis because the customer's voice truly reflects the brand's image. Opinion mining techniques are widely studied to analyse consumer's opinions, sentiments, attitudes, and emotions towards products, services, and events (Liu, 2012). Applying opinion mining techniques to get meaningful brand-related information from online reviews has attracted substantial attention from researchers interested in brand perceptions (Culotta and Cutler, 2016) and brand reputation (Morinaga et al., 2002). Comparative opinion mining is a subfield of opinion mining that identifies and extracts information expressed in a comparative form (e.g. product X is better than product Y) (Varathan et al., 2017). However, comparative opinion mining only uses a small part of the statements belonging to the comparative category within one corpus. Comparative summarization (Sipos and Joachims, 2013) analyses two (or more) corpora separately and then organises the results together. However, comparative summarisation cannot manage contradictory reviews within texts.

Wang et al. (2018) propose a latent Dirichlet allocation (LDA)-based model analysing two competitive brands from a topic perspective. The model could analyse two separate corpora and make full use of review data. There are several advantages to using LDA. First, LDA is an unsupervised method with generalisation. It does not rely on labeled data, and it is easy to transfer the data to different languages and different domains. Second, the LDA model generates semantically related word clusters to represent each topic automatically, which is intuitively explained in the marketing field. Finally, unlike most existing comparative opinion mining techniques, the LDA model makes full use of text data and provides more complete information. The use of LDA is reliable and explicable, so we follow the stream of the LDA model and develop a new generalised and functional model to analyse customers' needs and preferences for multi-brands. The model is called the brand joint latent Dirichlet allocation model (BJ-LDA). It's useful on practice for providing meaningful advice to product manufacturers and managers. Principal component analysis (PCA) and two-tailed t-test are applied on our model

* Corresponding author.
E-mail address: xiaolinglu@ruc.edu.cn (X. Lu).

https://doi.org/10.1016/j.elerap.2022.101141
Received 6 February 2021; Received in revised form 8 March 2022; Accepted 16 March 2022
Available online 21 March 2022

results to provide deeper understandings of the competition between brands and find significant factors which influence the ratings mostly for each brand. This model makes novel contribution to e-commerce research studies, especially for using topic model to comprehend the competitive standings of brands.

BJ-LDA makes two main improvements to Wang et al. (2018). First, their model analyses each brand separately and does not consider different types of topics. BJ-LDA can model multi-brand corpora simultaneously to extract background information, general topics shared by all brands, and specific topics for each brand. Background information is composed of background words that have no meaning for the brands, such as *go* and *buy*. General topics reveal the customers' concerns in common for all brands. Specific topics discover customers' different viewpoints with regard to each unique brand. For example, for facial cleanser products, customers care about cleanliness, smoothness, and scent, which are all general topics. Customers who buy expensive facial cleanser products are swayed more by quality than promotions for cheaper products. These differences form the specific information. Mining general subjects provides managers with a deeper understanding of their product category while mining specific topics for brands enables managers to discover their characteristic specialties, strengths, and weaknesses. The BJ-LDA model can help managers make sense of their relations with other brands and develop purposeful strategies.

Second, by applying a maximum entropy model (MaxEnt) model, our BJ-LDA model could separate aspect words and their corresponding opinion words to describe each brand from a detailed perspective. The semantically related word clusters generated by LDA are valuable because they reflect the subjects or characteristics customers care about most and their opinions towards these aspects. For traditional LDA as well as Wang et al. (2018), the aspects and opinions concerning a topic are mixed, which renders the analysis inconvenient. The task of extracting review aspects and corresponding opinions is a component of aspect-level sentiment analysis (Rana and Cheah, 2016). There is always a strong association between aspects and opinions. Take skincare area as an example, *foam* and *smooth*, *price* and *cheap*. Both *smooth* and *cheap* are opinions in the skincare domain, but each is associated with specific aspects, respectively *foam* and *price*. This type of fine-grained analysis provides useful insights.

The remainder of this article is organised as follows. First, related previous works are reviewed in Section 2. Then the details of the proposed approach are explained in Section 3. In Section 4, we apply BJ-LDA to two case studies. The datasets and results are presented in detail, and this section provides valuable information for product designers and managers. Finally, the conclusions and future work are summarised and discussed in Section 5.

2. Literature review

2.1. Multi-brand opinion competitive analysis

Comparative opinion mining, a subfield of opinion mining, has received substantial attention in recent years. Comparative opinion mining focuses on extracting comparative form information. There are two core tasks in comparative opinion mining. One task is to find entities or features that are compared in each sentence. The other task is to mine the comparative relations. Techniques such as machine learning, rule mining, and natural learning processing are used to tackle these tasks. From the perspective of compared elements, the studies in comparative opinion mining can be roughly categorised into sentence detection, entity detection, and feature detection. These different elements obtain different goals. At the sentence level, Jindal and Liu (2006) first explored extracting comparative sentences using computational methods. The later techniques used in comparative sentence detection included supervised machine learning (Wang et al., 2015), associative rule-based approaches (Liu et al., 2013b), and lexicon-based approaches (Wei et al., 2014). The entity normally represents product names, service names, and the names of individuals. For entity detection, Liu et al. (2013a) try to retrieve two entities in a single comparative statement. Xu et al. (2011) develope a model based on mobile phone data sets. The authors use a lexicon dictionary that contains mobile phone names and attributes to detect entities in a sentence. Feature detection in comparative mining obtains detailed information on products or services. The feature here has the same meaning as the aspect of our paper. In the feature detection area of comparative opinion mining, Sun et al. (2009) build a sentiment-based product feature database to handle features in comparative opinion text. Tkachenko and Lauw (2014) develope a model that can be derived from a sequence of the features by analysing a digital camera dataset.

Comparative opinion mining only uses comparative form statements, and it is not suitable for two or more separate corpora. The comparative summarisation method can compare two or more separate corpora. While facing a set of documents sharing similar topics, we are interested in identifying their differences. Researchers use a comparative summarisation method to solve this problem. Several approaches have been used to tackle comparisons of document text mining, such as discriminative sentence selection (Wang et al., 2013) and linear programming (Huang et al., 2011). For comparative summarisation in review mining, Sipos and Joachims (2013) propose an appropriate objective function to align online review snippets into pairs and ultimately obtain sentences on the same aspect.

Wang et al. (2018) develope a model for mining information from two competitive products from two corpora. The LDA-based model makes full use of linguistic data compared to comparative opinion mining. However, their model analyse multiple corpora separately and does not distinguish aspects and opinions for each topic.

2.2. Aspect-level opinion mining technique based on a topic model

Pioneered by the work of Hu and Liu (2004), aspect-based review mining is a significant research topic. Since aspect identification is a core task of aspect-based analysis, various approaches have been proposed. These approaches can be roughly categorised as supervised approaches and unsupervised approaches (Sun et al., 2017). Although supervised works (Li et al., 2010; Jin et al., 2009; Kobayashi et al., 2007; Cruz et al., 2013) yield good results, the dependence on labelled data renders the supervised method impractical. In contrast, unsupervised aspect detection methods are free of labelled data, which has attracted substantial attention. Typical unsupervised techniques include frequency-based methods (Hu and Liu, 2004; Kushal and Durga, 2013), unsupervised bootstrapping methods (Zhu et al., 2011; Bagheri et al., 2013), and rule-based methods (Liu et al., 2012; Bancken et al., 2014). These approaches have proven effective to extract aspects from reviews. However, these unsupervised methods have the limitation that they do not group semantically related aspects (Bagheri et al., 2014). The use of LDA, an unsupervised probabilistic model, can tackle this problem.

LDA is a probabilistic topic model based on the concept that a document is a mixture of many latent topics, and each topic is characterised by a probabilistic distribution over words (Blei et al., 2003). LDA-based models can be seen as generative models. These models assume the generative process of corpus and obtain the estimation based on inference methods. LDA can be efficiently used to discover latent topics from a large collection of documents and output clusters of words including aspects. The aspects are naturally gathered for a related topic. Titov and McDonald (2008) first propose an LDA-based model called a multigrain-topic model to extract rateable aspects. Lin and He (2009) develope a joint sentiment/topic model(JST) to detect topics and sentiments from movie reviews. This model extracts aspects and determines the sentiment polarities of reviews. However, the analysis for sentiment polarities in Lin and He (2009) is conducted at the document level. Following these works, many topic model-based methods and their variants have been used to address aspect-level review mining tasks.

An important stream of LDA research in opinion mining concerns extracting aspects and sentiment words simultaneously. Brody and Elhadad (2010) implemente an unsupervised aspect sentiment model for online reviews. The authors first extracte aspects by local LDA and then used mutual information to identify representative sentiment words for each aspect. Their work is the first to consider aspect words and opinion words at the same time and to extract aspect-specific opinions. However, their model is a two-step approach, and they only consider adjective words as opinions. Zhao et al. (2010) improve their model by jointly using model aspect words and aspect-specific opinion words within topic models. This model is called Maxent-LDA. Moreover, Maxent-LDA extends opinion words to non-adjective words. Maxent-LDA is originally used to mine topic information in a single corpus. We follow this promising line of research by extending a jointly-based model for brand difference discovery, which extracts both aspects and opinions from multi brand corpora simultaneously. There are two main differences between our model and Maxent-LDA. First our model is suitable to deal with multi corpora simultaneously while Maxent-LDA can only deal with single corpus. Second, Maxent-LDA is not specifically used for brand comparative. Our BJ-LDA is specially designed for brand competition, so we separate brand opinion for general topics and extract brand specific topics.

3. Model formulation

3.1. Generative process

BJ-LDA aims to solve the problem of extracting general topics and specific topics from multiple corpora on aspect level and their corresponding opinions. As introduced in Section 1, we assume there are one background topics, several general topics and brand specific topics. General topics and brand specific topics are separated into aspects and opinions. Mathematically, we assume there are B multiple brands in a product category. The number of documents in brand $b, b \in \{1, \ldots, B\}$ is denoted as D_b, and V_b represents the number of words in brand b. The whole corpora containing all brands have D review documents with V words. The summation of all brands D_b is equal to $D, \sum_{b=1}^{B} D_b = D$ because the review documents in brands do not overlap. The summation of V_b is larger than V because the vocabularies in brands overlap.

For the whole corpus including all brands, we assume that there is one background topic and K_0 general topics underlying all of the D review documents. Only the general topics are further divided into aspect word clusters and opinion word clusters. The general aspect word clusters are shared with all brands, while the corresponding opinion word clusters are belonged to each brand. It helps get better understandings of brand competition information. For each brand $b, b \in \{1, 2, \ldots, B\}$, we assume there are $K_b, b \in \{1, 2, \ldots, B\}$ brand-specific topics. The number of brand-specific topics can be different for all brands. These K_b brand-specific topics are composed of K_b brand-specific aspect word clusters and K_b brand-specific opinion word clusters. According to the relation $\sum_{b=1}^{B} V_b > V$, the overlapping words are more likely to be background words or general words while the unique words in each brand are more likely to be brand-specific words.

For the brand $b, b \in \{1, 2, \ldots, B\}$, first, we draw several multinomial word distributions from Dirichlet prior with parameters β and β_b. β with dimension V is for background multinomial word distribution Φ^B, K_0 general aspect multinomial word distributions $\Phi_t^{A,g}, t = 1, \ldots, K_0$. These $K_0 + 1$ multinomial distributions are over all V vocabularies. β_b with dimension V_b is for K_0 general opinion multinomial word distributions $\Phi_{t_b}^{O,g}, t_b = 1, \ldots, K_0$, K_b brand-specific aspect multinomial word distributions $\{\Phi_{k_b}^{A,s}\}_{k_b=1}^{K_b}$ and brand-specific opinion multinomial word distributions $\{\Phi_{k_b}^{O,s}\}_{k_b=1}^{K_b}$ for brand b. These $K_0 + 2K_b$ multinomial distributions are over V_b vocabularies. There are $(1 + K_0 + \sum_{b=1}^{B} K_b)$ topics and $(1 + K_0(1 + B) + 2\sum_{b=1}^{B} K_b)$ word clusters in total. Then, for each review document d in brand b, we draw two document topic distributions, a general document topic distribution $\theta_{d_b}^{g} \sim \mathrm{Dir}(\alpha_g)$ and a specific document topic distribution $\theta_{d_b}^{s} \sim \mathrm{Dir}(\alpha_s)$, similar in standard LDA. Because the review data is shorter and the topics are more concentrated, in contrast to drawing a topic for each word in traditional LDA, we draw topics $z_{d_b,s}^{g} \sim \mathrm{Multi}(\theta_{d_b}^{g})$ and $z_{d_b,s}^{s} \sim \mathrm{Multi}(\theta_{d_b}^{s})$ for each sentence s in document d. This concept has proven efficient in Zhao et al. (2010), and we make the same generative assumption in our model. We consider this proper for our model because each sentence will address a brand-specific topic and a general topic.

For each word in sentence s of document d, we have five choices. The nth word $w_{d_b,s,n}$ in sentence s document d may describe a commonly used background word, a general aspect (e.g. *foam*), a general opinion word (e.g. *smooth*), a brand-specific aspect(e.g. for expensive brand customers focus on *price*), or a brand-specific opinion(e.g. *expensive*). Two indicator variables, $y_{d_b,s,n}$ and $u_{d_b,s,n}$, are introduced to distinguish the choice of the word $w_{d_b,s,n}$. $y_{d_b,s,n}$ determines whether $w_{d_b,s,n}$ is a background word, aspect word, or opinion word while $u_{d_b,s,n}$ indicates whether $w_{d_b,s,n}$ is general or brand specific. The indicator $y_{d_b,s,n}$ follows a multicategory distribution over $\{0, 1, 2\}$, parameterised by $\pi_{d_b,s,n}$. $\pi_{d_b,s,n}$ is determined by $x_{d_b,s,n}$ and $\lambda_{d_b,s,n}$, which we discuss in the next subsection. $u_{d_b,s,n}$ is drawn from a Bernoulli distribution over $\{0, 1\}$ parameterised by p. p is drawn from $\mathrm{Beta}(\gamma)$. Then, we draw $w_{d_b,s,n}$ as follows:

$$w_{d_b,s,n} \sim \begin{cases} \mathrm{Multi}(\Phi^B) & \text{if } y_{d_b,s,n} = 0 \\ \mathrm{Multi}(\Phi^{A,s}_{z^s_{d_b,s}}) & \text{if } y_{d_b,s,n} = 1, u_{d_b,s,n} = 0 \\ \mathrm{Multi}(\Phi^{A,g}_{z^g_{d_b,s}}) & \text{if } y_{d_b,s,n} = 1, u_{d_b,s,n} = 1 \\ \mathrm{Multi}(\Phi^{O,s}_{z^s_{d_b,s}}) & \text{if } y_{d_b,s,n} = 2, u_{d_b,s,n} = 0 \\ \mathrm{Multi}(\Phi^{O,g}_{z^g_{d_b,s},b}) & \text{if } y_{d_b,s,n} = 2, u_{d_b,s,n} = 1 \end{cases}$$

The illustration for the generative process is shown in Fig. 1, and meaning of each notation is summarised in Table 1.

3.2. Maximum entropy model

The goal of the maximum entropy (MaxEnt) model is to maximise the entropy of the distribution subject to certain constraints. We use the MaxEnt model as a classifier to distinguish whether a word $w_{d_b,s,n}$ belongs to the background, aspect, or opinion. The efficiency of using MaxEnt on textual data analysis has been proved by Zhao et al. (2010) and Ratnaparkhi (1996). The input features $x_{d_b,s,n}$ in the MaxEnt model

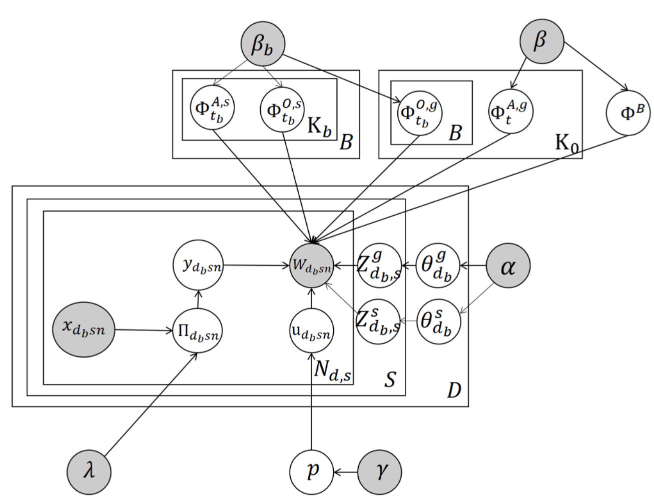

Fig. 1. Illustration for the generative process.

Table 1
Notation for generative process.

notation	meaning	notation	meaning
B	brand number	K_0	general topic number
K_b	specific topic number	β	Dirichlet prior
p	Bernoulli distribution parameter	γ	Beta distribution parameter
D	document number	S	sentence of document
$N_{d,s}$	word of sentence	α	Dirichlet prior
λ	Maxent model weight	β_b	Dirichlet prior
Φ^B	background word distribution	$\Phi^{A,s}_{t_b}$	specific aspect word distribution
$\Phi^{O,g}_{t_b}$	general opinion word distribution	$\Phi^{O,s}_{t_b}$	specific opinion word distribution
$\Phi^{A,g}_{t}$	general aspect word distribution	$x_{d_b,s,n}$	feature for Maxent model
$\theta^g_{d_b}$	general topic distribution	$\theta^s_{d_b}$	specific topic distribution
$Z^g_{d_b,s}$	general topic for sentence	$Z^s_{d_b,s}$	specific topic for sentence
$w_{d_b,s,n}$	observed word	$u_{d_b,s,n}$	general/specific indicator
$y_{d_b,s,n}$	background/aspect/ opinion indicator	$\Pi_{d_b,s,n}$	multicategory distribution parameter

can include any discriminative information associated with $w_{d_b,s,n}$. Here, it contains two types. The first type is the POS (part-of-speech) tag, including the POS tags for previous, current, and the next word denoted as $\{POS_{d_b,s,n-1}, POS_{d_b,s,n}, POS_{d_b,s,n+1}\}$. According to the observation, opinion words tend to be adjectives while aspects words tend to be nouns. They both play different context roles in sentences. POS tags are important because they distinguish between aspect and opinion words. Considering the connection between the words before and after the current word, the second type is the lexical feature, including the previous, current, and next words$\{w_{d_b,s,n-1}, w_{d_b,s,n}, w_{d_b,s,n+1}\}$. The MaxEnt model is as follows:

$$p(y_{d_b,s,n} = l|x_{d_b,s,n}) = \pi^l_{d_b,s,n} = \frac{exp(\lambda_l x_{d_b,s,n})}{\sum_{l'=0}^{2} exp(\lambda_{l'} x_{d_b,s,n})} \tag{1}$$

where $l \in \{0,1,2\}$, $\{\lambda_l\}_{l=0}^2$ denote the MaxEnt model weights that can be learned from a set of labeled training sentences. The aspects are manually labeled in the training sentence. For the English corpus, the selected opinion words correspond to an integrated sentiment lexicon based on MPQA subjectivity lexicon (Wiebe et al., 2005). For the Chinese corpus, the opinion words can be selected according to the Hownet sentiment word dictionary (Dong and Dong, 2003). Apart from aspect and opinion words, the remaining words are labeled as background words.

3.3. Model inference

The MaxEnt algorithm is first performed to obtain the parameter λ for $\pi_{d,s,n}$. We then use Gibbs sampling (Hastings, 1970) for inference of BJ-LDA. Gibbs sampling is an algorithm for sampling conditional distributions of variables whose distribution over states converges to the true distribution. The goal of our model estimation is to obtain true values of $\Theta^g, \Theta^s, \phi^B, \Phi^{O,g}, \Phi^{A,g}, \Phi^{A,s}$, and $\Phi^{O,s}$. Here, $\boldsymbol{w}$ denotes all the words we observed in the collection, $\boldsymbol{x}$ denotes all the feature vectors for these words, and $\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{z}$ are the hidden variables. Given $\boldsymbol{w}$ and all the hyperparameters, we first derive the full posterior distribution according to the generative process, which is not shown because of space limitations. Based on the posterior distribution, we obtain the following four conditional distributions. We iteratively sample from these four conditional distributions to obtain convergence. That is the final estimation. The conditional distributions are given below.

For the sentence s in document d brand b, given the assignment of all other hidden variables, the conditional distribution of general topic $z^g_{d_b,s}$ is:

$$P(z^g_{d_b,s} = t'|z^g_{\neg(d_b,s)}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{x}) \propto \frac{\boldsymbol{c}^{d_b,g}_{(t')} + \alpha}{\boldsymbol{c}^{d_b,g}_{(.)} + \boldsymbol{T\alpha}}$$
$$\times \left(\frac{\Gamma(\boldsymbol{c}^{A,g,t'}_{(.)} + \boldsymbol{V}\beta)}{\Gamma(\boldsymbol{c}^{A,g,t'}_{(.)} + \boldsymbol{n}^{A,g,t'}_{(.)} + \boldsymbol{V}\beta)} \prod_{\boldsymbol{v}=\boldsymbol{1}}^{\boldsymbol{V}} \frac{\Gamma(\boldsymbol{c}^{A,g,t'}_{(\boldsymbol{v})} + \boldsymbol{n}^{A,g,t'}_{(\boldsymbol{v})} + \beta)}{\Gamma(\boldsymbol{c}^{A,g,t'}_{(\boldsymbol{v})} + \beta)} \right)$$
$$\times \left(\frac{\Gamma(\boldsymbol{c}^{O,g,t',b}_{(.)} + \boldsymbol{V_b}\beta_b)}{\Gamma(\boldsymbol{c}^{O,g,t',b}_{(.)} + \boldsymbol{n}^{O,g,t',b}_{(.)} + \boldsymbol{V_b}\beta_b)} \prod_{\boldsymbol{v}=\boldsymbol{1}}^{\boldsymbol{V_b}} \frac{\Gamma(\boldsymbol{c}^{O,g,t',b}_{(\boldsymbol{v})} + \boldsymbol{n}^{O,g,t',b}_{(\boldsymbol{v})} + \beta_b)}{\Gamma(\boldsymbol{c}^{O,g,t',b}_{(\boldsymbol{v})} + \beta_b)} \right) \tag{2}$$

where $c^{d_b,g}_{(t')}$ is the number of sentences assigned to topic t' in document d_b, $c^{d_b,g}_{(.)}$ is the sentence number in document d_b, $c^{A,g,t'}_{(v)}$ is the number of times word v is assigned as an general aspect word to topic t', and $c^{A,g,t'}_{(.)}$ is the total number of times any word is assigned as an aspect word to topic t'. All these counts, represented by c, exclude sentence s of document d. $n^{A,g,t'}_{(v)}$ is the number of times word v is assigned to an aspect word to topic t' in sentence s of document d, and $n^{A,g,t'}_{(.)}$ is the number of words assigned as an aspect word to topic t' in sentence s document d. The meaning of $c^{O,g,t',b}_{(v)}$ is similar to $c^{A,g,t'}_{(v)}$, and the meaning of $c^{O,g,t',b}_{(.)}$ is similar to $c^{A,g,t'}_{(.)}$ for opinion words, the only difference is that they count in the corpora of brand b.

The conditional distribution of specific topic $z^s_{d_b,s}$ is:

$$P(z^s_{d_b,s} = t|z^s_{\neg(d_b,s)}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{x}) \propto \frac{\boldsymbol{c}^{d_b,s}_{(t)} + \alpha}{\boldsymbol{c}^{d_b,s}_{(.)} + \boldsymbol{T_b}\alpha}$$
$$\times \left(\frac{\Gamma(\boldsymbol{c}^{A,s,t,b}_{(.)} + \boldsymbol{V_b}\beta_b)}{\Gamma(\boldsymbol{c}^{A,s,t,b}_{(.)} + \boldsymbol{n}^{A,s,t,b}_{(.)} + \boldsymbol{V}\beta)} \prod_{\boldsymbol{v}=\boldsymbol{1}}^{\boldsymbol{V_b}} \frac{\Gamma(\boldsymbol{c}^{A,s,t,b}_{(\boldsymbol{v})} + \boldsymbol{n}^{A,s,t,b}_{(\boldsymbol{v})} + \beta_b)}{\Gamma(\boldsymbol{c}^{A,s,t,b}_{(\boldsymbol{v})} + \beta_b)} \right)$$
$$\times \left(\frac{\Gamma(\boldsymbol{c}^{O,s,t,b}_{(.)} + \boldsymbol{V_b}\beta_b)}{\Gamma(\boldsymbol{c}^{O,s,t,b}_{(.)} + \boldsymbol{n}^{O,s,t,b}_{(.)} + \boldsymbol{V_b}\beta_b)} \prod_{\boldsymbol{v}=\boldsymbol{1}}^{\boldsymbol{V_b}} \frac{\Gamma(\boldsymbol{c}^{O,s,t,b}_{(\boldsymbol{v})} + \boldsymbol{n}^{O,s,t,b}_{(\boldsymbol{v})} + \beta_b)}{\Gamma(\boldsymbol{c}^{O,s,t,b}_{(\boldsymbol{v})} + \beta_b)} \right) \tag{3}$$

where the meaning of notation is similar to those explained above, the only difference is that the s in superscript represents the specific word.

The conditional probability for $y_{d_b,s,n}$ is:

$$P(y_{d_b,s,n} = 0|z, \boldsymbol{y}_{\neg(\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n})}, \boldsymbol{u}_{\neg(\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n})}, \boldsymbol{w}, \boldsymbol{x})$$
$$\propto \frac{\exp(\lambda_0 \boldsymbol{x}_{\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n}})}{\sum_{l'} \exp(\lambda_{l'} \boldsymbol{x}_{\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n}})} \frac{\boldsymbol{c}^{\boldsymbol{B}}_{(\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n})} + \beta}{\boldsymbol{c}^{\boldsymbol{B}}_{(.)} + \boldsymbol{V}\beta}, \tag{4}$$

and the conditional probability for $y_{d_b,s,n}$ and $u_{d_b,s,n}$ is:

$$P(y_{d_b,s,n} = l, u_{d_b,s,n} = b|z, \boldsymbol{y}_{\neg(\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n})}, \boldsymbol{u}_{\neg(\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n})}, \boldsymbol{w}, \boldsymbol{x})$$
$$\propto \frac{\exp(\lambda_l \boldsymbol{x}_{\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n}})}{\sum_{l'} \exp(\lambda_{l'} \boldsymbol{x}_{\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n}})} \boldsymbol{g}(\boldsymbol{w}_{\boldsymbol{d_b},\boldsymbol{s},\boldsymbol{n}}, z^s_{\boldsymbol{d_b},\boldsymbol{s}}, z^g_{\boldsymbol{d_b},\boldsymbol{s}}, \boldsymbol{l}, \boldsymbol{b}), \tag{5}$$

where the function $g(v, t, t', l, b)$ is defined as:

$$g(v,t,t',l,b)=\begin{cases}\dfrac{c_{(v)}^{A,s,t,b}+\beta_b}{c_{(.)}^{A,s,t,b}+V_b\beta_b}\dfrac{c_{(0)}+\gamma}{c_{(.)}+2\gamma} & \text{if} l=1,b=0\\ \dfrac{c_{(v)}^{O,s,t,b}+\beta_b}{c_{(.)}^{O,s,t,b}+V_b\beta_b}\dfrac{c_{(0)}+\gamma}{c_{(.)}+2\gamma} & \text{if} l=2,b=0\\ \dfrac{c_{(v)}^{A,g,t'}+\beta}{c_{(.)}^{A,g,t'}+V\beta}\dfrac{c_{(1)}+\gamma}{c_{(.)}+2\gamma} & \text{if} l=1,b=1\\ \dfrac{c_{(v)}^{O,g,t',b}+\beta_b}{c_{(.)}^{O,g,t',b}+V_b\beta_b}\dfrac{c_{(1)}+\gamma}{c_{(.)}+2\gamma} & \text{if} l=2,b=1\end{cases} \tag{6}$$

where $c_{(0)}$ is the number of words belonging to brand-specific topics, $c_{(1)}$ is the number of words belonging to the general topic, and $c_{(.)}$ is the total number of words. Here, the variable c denotes various counts excluding the nth word in sentence s of document d.

3.4. Selection of topic numbers

In our BJ-LDA, we need to choose the K_0 and $K_b, b=1,\dots,B$ to get the best performance. Here, we propose a forward selection algorithm, the idea of which is widely used in feature selection problem for multivariate analysis. Two evaluation metrics are model perplexity and topic coherence.

Model perplexity is first proposed by Blei et al. (2003). It is a commonly used measure to predict likelihood and can be treated as an indicator of the model's predictive ability. The lower the perplexity, the better the model. The model perplexity for BJ-LDA is defined as follow, where $p_w^b, p_w^{A,g}, p_w^{O,g}, p_w^{A,s}, p_w^{O,s}$ represent the probability of word w belongs to the background words, general aspects, general opinion, specific aspects and specific opinion.

$$p(w)=p_w^b\phi_w^b+\sum_{k=1}^{K}\theta_{dk}^g(p_w^{A,g}+p_w^{O,g}\phi_{wb}^{O,g})+\sum_{k=1}^{K_b}\theta_{dk}^s(p_w^{A,s}\phi_{wb}^{A,s}+p_w^{O,s}\phi_{wb}^{O,s}) \tag{7}$$

$$\text{perplexity}=\exp\left(-\frac{\sum_{b=1}^{B}\sum_{d=1}^{D_b}\sum_{w=1}^{W_d}\log(p(w))}{\sum_{b=1}^{B}\sum_{d=1}^{D_b}N_d}\right) \tag{8}$$

The topic coherence is a commonly used index to measure topic quality extracted by LDA based model (Mimno et al., 2011). If top word pairs of topics appear in a document frequently, the topics are thought to have high quality and the coherence is high. For topic t, the topic coherence is defined as:

$$\text{coherence}(t)=\sum_{m=1}^{M}\sum_{l=1}^{m-1}\log\frac{D(v_m^{(t)},v_l^{(t)})+1}{D(v_m^{(t)})}$$

where M is the number of top words we concentrate on, $D(v_m^{(t)},v_l^{(t)})+1$ is the document number that words $v_m^{(t)}$ and $v_l^{(t)}$ co-occured.

There are three steps in our forward selection algorithm. First, we start from simplest model whose $K_0=1$ and $K_b=1, b=1,\dots,B$, record its model perplexity. Second, separately add 1 to K_0 and $K_b, b=1,\dots,B$, record $B+1$ models' perplexity. Third, choose the model with the most decrease of perplexity as the current optimal model. Repeat these three steps until model perplexity doesn't decrease and finally get the optimal K_0 and $K_b, b=1,\dots,B$. In the whole procedure, topic coherence is used as an auxiliary index. The higher the value of average topic coherence, the better the model.

4. Case studies

To apply our proposed model, we use two sets of data in two different domains, skincare for products and Japanese restaurants in Beijing for service. Each store or shop can be treated as a brand in the service domain. We chose these two different review datasets to reflect the applicability of our model for different domains.

4.1. Skincare products

4.1.1. Data collection and preprocessing

We write web crawler code to grab review data from China's Jingdong Mall (http://www.jd.com) platform. Jingdong Mall is one of the largest e-commerce platforms in China. Many individuals pay attention to their skincare and invest significant time commenting online with regard to skincare products. Consumers are willing to share their opinions about the skincare products they buy online, and they are interested in the comments of others. Therefore, the opinions expressed in online skincare reviews are rich. We select six well-known brands for our analysis, according to price. They are Ponds, Dove, Curel, Cetaphil, Clarins, and Sulwasoo. Based on the information on their official online websites, these six brands represent two discount brands, two medium-priced brands, and two luxury brands. The brands are manufactured in China, Japan, etc. Each brand is designed for a specific purpose, for example, Curel is designed for sensitive skin, and Sulwasoo is a herbal, luxury product. We summarise the product information in Table 2. Brands at the same price level have similar product characteristics and target similar customers. This explains the competitive relationship. Skincare products are daily consumables, and customers may occasionally change brands. Competition exists among the different product price levels.

The study period is from December 2018 to February 2019. The raw number of reviews is 3,992. Preprocessing is a critical procedure for text mining. This procedure helps clean the data before analysis. Unstructured text information, such as reviews, often contains substantial redundant noise. Our pre-processing procedures contain the following. The first step is filtering. We filter duplicate reviews with the same buyer ID or with the same content. Additionally, we eliminate some reviews with pure numbers or punctuation. The second step is tokenisation. We apply the 'Jieba' package in Python to accomplish this task. Moreover, we add a manually summarised domain dictionary to ensure the quality of tokenisation. Thus, 'sensitive skin' will not be cut as 'sensitive' and 'skin'. The third step is stop-words removal. In this work, we employ an augmented version of the stop-word list to include common words that occur frequently such as 'of', 'uh', and 'ha', and some useless words referring to the irrelevant text abbreviations.

Before modeling, we carry out high-frequency word analysis as data descriptive analysis. There are general high-frequency words that appeared for every brand, such as 'feeling' and 'very good'. These are likely to be general information. The high-frequency words for each brand are brand-specific words, such as 'amino acids' and 'cheap' for Ponds, 'sensitive skin' and 'mild' for Curel, and 'pregnant' for Clarins. The general and specific words indicate the need to mine general and specific information with our proposed model.

Table 2
Brand information and descriptive statistics.

Brand	Price	Origin place	Orientation	Reviews number	Average length
Ponds	30	China	Mild, Skin whitening	499	37.83
Dove	45	Japan	Amino acid cleansing	643	31.28
Curel	108	Japan	Sensitive skin	520	25.78
Cetaphil	109	Canada	Sensitive skin	1170	44.60
Clarins	250	France	Soften skin, High-end	545	30.07
Sulwasoo	320	Korea	Herbal, High-end	561	27.44

4.1.2. Model results

First, we train the MaxEnt model with randomly chosen 10% of reviews, as explained in Section 3.2, to obtain an estimation for $\pi_{d_b,s,n}$.

Second, to choose a proper K_0 and $K_b, b = 1, \dots, 6$ for the skincare datasets, we use the forward selection algorithm discussed in Section 3.4. We begin with $K_0 = 1$ and $K_b = 1, b = 1, \dots, 6$. After 13 steps, the perplexity decreases to a stable level, and the topic coherence reaches a relatively high level. Then the most proper topic numbers for skincare datasets are decided corresponding to model 12. $K_0 = 2$ and the specific topic number for Ponds, Dove, Curel, Cetaphil, Clarins and Sulwasoo are $3, 2, 3, 4, 3, 3$. The variation tendency of perplexity and model coherence are shown in Fig. 2.

There are two general topics. The first general topic is about store information, including the aspect words ‘coupon’, ‘review’, and ‘quality’. The second general topic talks about product usage, including aspect words ‘price’, ‘usage’ and ‘ingredient’. These aspects are what all brands need to improve because they are considered as core competitive elements. The opinions for general topics are summarised in Table 3. The customers’ opinions on these two aspects are different for different brands. For example, customers are satisfied with the express service for Dove, and consider the score card of Cetaphil as a surprise. Dove is popular for its oil-control function and Sulwasoo smells good.

The brand-specific characteristics with their strengths and weaknesses are revealed in our results. We summarise the information for each brand.

- **Ponds**

  –**Strengths: high-cost performance, amino acid ingredient** The two strengths are cost performance and express. Customers are satisfied with the Ponds product w.r.t. the price. The aspect ‘amino acid’ is mentioned in specific topic.

  –**Weaknesses: not apparent**
- **Dove**

  –**Strengths: quick express, ability of oil-control** Dove’s strengths are ‘logistics’ and ‘oil-control’. The corresponding opinions ‘speed’, ‘quick’ and ‘cheap’ express the customers’ preferences.

  –**Weaknesses: dirty packaging** Dove’s weakness is ‘packaging’. The word ‘weep’ implies that the bottle may be broken when customers receive the packages.

Table 3
Opinions for general topics for skincare dataset.

	general topic1: store information	general topic2: product usage
Ponds	cheap and fine	cheap,smell
Dove	feedback,express	moist,oil-control
Curel	suitable,cost-efficient	description,good
Cetaphil	score card,surprise	loyalty,quality
Clarins	expensive,good	health,try
Sulwasoo	comfortable,gift box	comfort,smell

- **Curel**

  –**Strengths: nice discount, free gift** The aspect ‘double 11’, a shopping carnival for Chinese e-commerce is appeared, and the opinion ‘favorable price’ expresses customers’ satisfaction to discount. Curel is the only brand that mentions ‘free gift’ among medium-priced brands.

  –**Weaknesses: not apparent**
- **Cetaphil**

  –**Strengths: sensitive skin-friendly, fast express** Cetaphil is designed for sensitive skin. The opinions ‘mild’ and ‘clean’ imply that this product is suitable for sensitive skin. The main strength of Cetaphil is given by the word ‘express’ along with its opinion ‘fast’.

  –**Weaknesses: pasty texture** Cetaphil’s weakness is texture. Many customers complained that this product is ‘pasty’, which causes dissatisfaction among some customers who are using the brand.
- **Clarins**

  –**Strengths: preferred by pregnant women** The word ‘pregnant’ reveals that Clarins is preferred by pregnant women. The words ‘freshness’ and ‘healthy’ reveal Clarins is suitable for expected mother.

  –**Weaknesses: high price** The weakness of Clarins is that some customers consider the brand too ‘expensive’.
- **Shulwasoo**

  –**Strengths: herbal ingredient, good smell** Herbal ingredients are one of the competitive advantages of Sulwasoo, the herbal flavor is very popular, many customers express preference toward it.

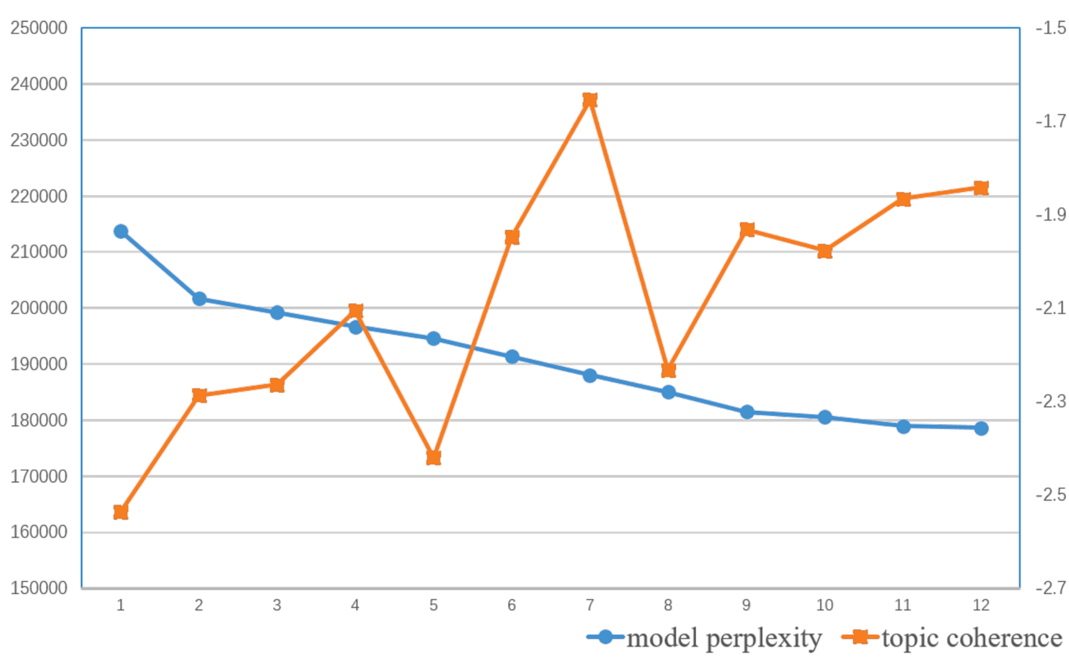

Fig. 2. The variation tendency of perplexity and model coherence.

–**Weaknesses: dust packaging** Shulwasoo is the most expensive of the six brands, customers have high expectations in many aspects. The 'dust' packaging will leave a bad impression on customers.

In addition to determining the strengths and weaknesses of brands, our model provides additional information assisting brand development. First, the selling points from customers' perspective for each specific brand can be mined. Cetaphil is designed for sensitive skin. The top words of this brand contain 'sensitive skin', and the opinion words contain the word 'mild', which reflects that the products are suitable for sensitive skin. The Ponds product contains the essence of rice, which is a unique product feature, and our model reflects this with the results showing the words ' essence of rice'.

Second, our model mined several customer concerns regarding specific brands. Sulwasoo and Clarins are two high-end skincare brands, and their products are more expensive. Customers who buy these expensive products online are concerned that the products they receive may not be authentic. Consumers often compare online products with products bought from physical stores. The managers for these brands should pay more attention to their brands' reputations.

Finally, the results of the model also identify potential loyal customers. For the brand Clarins, the word 'pregnant' is a top word among topics. This implies that pregnant women prefer this brand's products. Although the facial cleaning product is not designed for pregnant women, this brand does provide other products for expectant mothers. The facial cleaning product is preferred by pregnant women because of the brand's other pregnancy-related products. The manager of this brand could launch more pregnancy-related products, for example, facial cleansing products for expectant mothers.

4.1.3. Model comparison

In order to verify the effectiveness of our model, we compare it with the Maxent-LDA model (Zhao et al., 2010). As reviewed in Section 2.2, Maxent-LDA model separates the aspect words and opinion words, and it has the structure of general and specific topics.

To make an appropriate comparison, we first run Maxent-LDA on the whole corpus. We set the number of general topic as 1 and the number of specific topics as $K_0 = 2$, same as BJ-LDA.

The results of Maxent-LDA on the whole skincare corpus are shown in Table 4. The general topic of Maxent-LDA on the whole corpus extracts the commonly used aspect in skincare area, such as 'foam' and 'smell', and the corresponding opinion such as 'good' and 'favorite'. However, the Maxent-LDA can not extract brand information separately. The information for each brand is mixed in general topic. For example, the opinion word 'mild' is usually used in sensitive skin products. It appears as an opinion word of sensitive brand Chetaphil in BJ-LDA, but it is mixed with general opinion words 'not bad' and 'good' in general opinion topic within Maxent-LDA. As for two specific topics, the information for each brand is also mixed. Take an example, 'oil-control' is the specific aspect for brand Dove extracted by BJ-LDA, it is mixed with other aspects in a specific topic in Maxent-LDA. The 'pregnant' is the specific opinion for brand Clarins extracted by BJ-LDA, it is mixed with other opinions in Maxent-LDA.

Then we apply Maxent-LDA on each brand corpus separately by setting the general topic as 1 and the specific topic numbers as $K_b, b = 1, 2, \ldots, 6$, same as BJ-LDA. The topics of Maxent-LDA on each brand corpus can extract brand-specific information without overall domain customers' focus. The results for brand Sulwasoo are shown in Table 5. We have similar results for other brands, which are not shown here for sake of space. We can get that customers for Sulwasoo care much about 'genuine guarantee', because of the high price. The products of Sulwasoo have a herbaceous smell, this is one of the selling points of this brand and can be mined as topic 1's opinion, 'herbal' and 'smelly'. Although this brand's own characteristics can be explored, users' opinions towards other comparative brands are difficult to compare. In summary, the model Maxent-LDA is not suitable for brand competition analysis, especially in multi corpora situations. Our BJ-LDA is well suited to solving this problem, as shown in Table 3.

Finally, we compare the two models quantitatively. As introduced in Section 3.4, topic coherence is used to measure quality of topics and we choose it as the comparison metric. Our BJ-LDA gets a better topic coherence as -1.6719 on the whole corpus, Maxent-LDA has worse value as -2.5223 on the whole corpus. It verifies that BJ-LDA can extract higher quality topics than Maxent-LDA.

4.1.4. Further exploration

To further map competitive landscape and develop competitive strategies, we use the aspects extracted by BJ-LDA and compute importance value of each aspect for six brands. After applying a principal component analysis, competitive landscape of these six brands is visualized in multi-dimension graphs.

We use the Jaccard coefficient to detect unique brand associations, which can be regarded as the aspect similarity between brands (Romesburg, 2004; Hu and Trivedi, 2020). The Jaccard coefficient at occurrence level is defined as $\text{Jaccard} = \frac{a}{F_1+F_2-a}$, where a is the frequency of documents including specified word in certain brand. F_1 is the document frequency of this certain brand, and F_2 is the frequency of documents including the specified aspect in whole data. According to Hu and Trivedi (2020), the Jaccard ranking is used instead of Jaccard coefficient to get performance. We select 13 frequently appeared aspects in specific aspect sets (the second column in Table 4). After deriving 13 aspect importance coefficient for six skincare brands, we perform a principal component analysis (PCA) of the brand associations in a 3-dimension solution with 84.55% cumulative variance.

Table 6 summarizes the factor loading of selected 13 aspects for three dimensions. PC1 relates to usage (e.g. foam and amino acid) and smell (e.g. smell), PC2 presents feel (e.g. sensitive skin and feel) and promotion (e.g. double 11, free gift, genuine guarantee), PC3 shows logistics (such as packaging and express) and price(e.g. price and cost performance). We further display PC1 and PC2 in two-dimension plots to get more intuition. As shown in Fig. 3, we group these attributes into several categories according to their loading values.

From Fig. 3, we can see that there is competition between the two high-end products, Clarins and Sulwasoo, because they are relatively close to each other. These two brands compete mainly on feel and foam. The competition between Dove and Pond's is also obvious, the closest brand to Pond's is Dove. That's because both are affordable brands. They're competing at effect and packaging. Curel and Cetaphil are competing on free gift and cost performance.

Table 4
Results of Maxent-LDA for whole skincare corpus.

general topic	aspect	**foam**,feel,nothing,logistics,packaing,**smell**
	opinion	not bad,**good**,**mild**,clean,**favorite**,useful
topic 1	aspect	cleaness,**oil-control**,plentiful,discount,buy,quick
	opinion	very,deserve,sensitive,moist,watery,**pregnant**
topic 2	aspect	service,express,discount,friend,brand,shopping
	opinion	quick,speed,disappointed,too much,expect,bad

Table 5
Results of Maxent-LDA for Sulwasoo corpus.

general topic	aspect	effect,feel,price,**genuine guarantee**,smell,logistics
	opinion	not bad,good,like,quick,satisfied,comfortable
topic 1	aspect	shop,genuine guarantee,quality,select,oil control,price
	opinion	opinion,thank,careful,**herbal**,share,**smelly**
topic 2	aspect	genuine guarantee,Sulwasoo,receive,smell,press,total
	opinion	oily,special,good,real,dry,quick
topic 3	aspect	promotion,guarantee,package,skincare,JD,product
	opinion	reply,clean,open,trust,doubt,liquid

Table 6
Structure Coefficients of brand attributes in 3-dimensions.

Index	Aspects	PC1	PC2	PC3
1	foam	0.3652	0.2336	−0.1921
2	packaging	−0.0124	0.0884	−0.4720
3	logistics	0.0394	0.0923	−0.2636
4	feel	0.0915	0.2378	0.2024
5	effect	−0.3104	0.1982	0.2988
6	price	0.1294	−0.1212	0.0329
7	smell	−0.2973	−0.3290	−0.5080
8	cost performance	0.2612	−0.1658	0.2790
9	sensitive skin	0.0198	0.6447	0.0759
10	genuine guarantee	−0.5073	−0.1874	0.3158
11	free gift	−0.1556	−0.1824	0.2153
12	amino acied	0.5303	−0.4406	0.1686
13	double 11	−0.1538	−0.0686	−0.1531

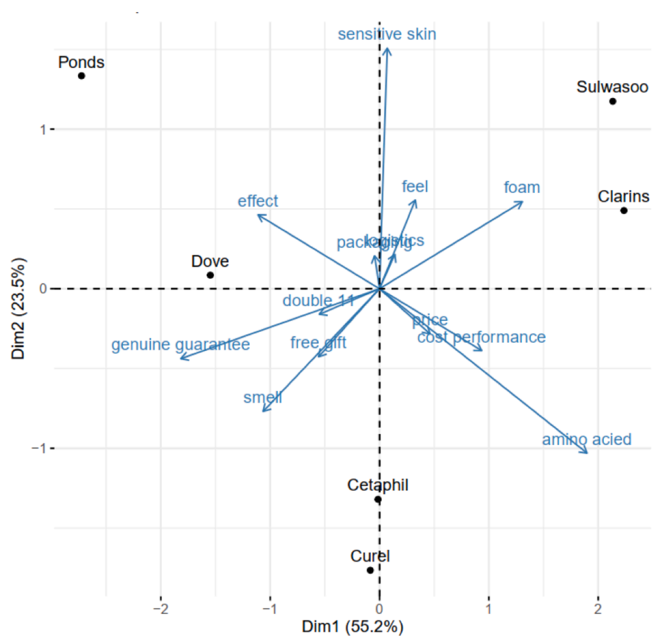

Fig. 3. Competitive landscape of 6 skincare brands on 13 aspects.

4.2. Japanese restaurants

4.2.1. Data collection and description

The Japanese restaurant data are crawled from the Chinese Dianping platform (https://www.dianping.com/), one of the leading food service industry e-commerce platforms. Dianping is China's leading local lifestyle information and trading platform, and the first independent third-party consumer review website in the world. Dianping hosts information on many restaurants including substantial review data. Japanese foods have gained popularity in China in recent years, and Chinese people are willing to try Japanese food. Many Japanese restaurants are chain stores. To explore competitive relationships within a brand, we choose a chain restaurant (Shota Muni Sushi and Grill) and analyse the reviews from different locations using our BJ-LDA model. The selected locations are Chaoyang Joy City Branch (CJCB), Wanda Plaza Branch (WZB), Hengtai Plaza Branch (HPB), Euramerican Shopping Center Branch (ECB), Swiss Apartment Restaurant Branch (SARB), and Xidan Hanguang Branch (XHB). The data collection and preprocession procedures are the same as in Section 4.1.1. The descriptive information is shown in Table 7. The study period is from April 2006 to June 2020. The number of reviews for each restaurant ranges from 1,413 to 11,280. We see that consumption per person and the customer ratings are similar. We look for some detailed differences according to the reviews with the BJ-LDA model.

In this dataset, we collect customer ratings for each review. The rating system uses a scale of 1 to 5 points for four aspects including overall, taste, atmosphere, and service, with 1 point as the lowest rating and 5 as the highest rating. The descriptive statistics are shown in Table 7. To discover the strengths and weaknesses from different customers' perspective, we divide the reviews into two parts, positive reviews and negative reviews. The negative reviews include reviews with points 1 to 3, which reflect dissatisfaction. Positive reviews are rated at point 4 or above.

The number of positive and negative reviews is shown in Fig. 4. For all six stores, the negative reviews are fewer in number than the positive reviews. The store with the highest number of positive comments is ECB, for which there are 12,837 positive reviews. The ratio of positive reviews to negative reviews is almost 6. The brand with the highest number of negative comments is SARB, for which there were 3283 negative reviews. The ratio of positive reviews to negative reviews is only 2.

We apply the same pre-procession procedures as for the skincare product dataset, and the high-frequency word analysis.

4.2.2. Model results

The proper K_0 and $K_b, b = 1, \ldots, 12$ (because the separation of positive reviews and negative reviews) for Japanese restaurants are also selected by forward selection algorithm. The variation tendency of perplexity and model coherence are similar to that of skincare dataset, as shown in Fig. 5. The final model is model 12, and the corresponding topic number is shown in Fig. 6.

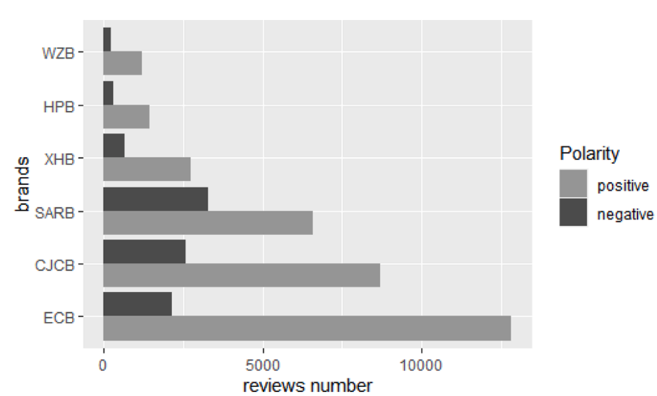

Fig. 4. Number of positive and negative reviews for chain restaurants.

Table 7
Descriptive statistics for six chain restaurants.

	Consumption per person	Overall rating	Taste rating	Environment rating	Service rating	Number of reviews
CJCB	155	4.65	4.57	4.56	4.39	11280
WZB	147	4.70	4.71	4.78	4.71	1413
HPB	138	4.69	4.75	4.77	4.69	1741
ECB	144	4.78	4.76	4.77	4.67	14992
SARB	126	4.67	4.48	4.05	4.05	9860
XHB	145	4.66	4.71	4.64	4.59	3389

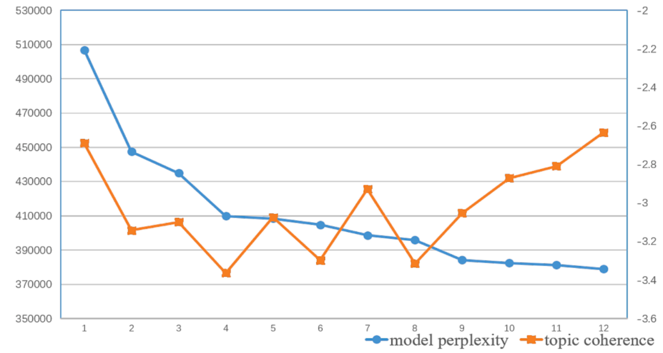

Fig. 5. Perplexity and model coherence variation tendency of Japanese restaurant datasets.

The eight general topics talk about 'lamian', 'branch store', 'hot pot', 'atmosphere', 'snacks', 'promotion', 'dishes' and 'sea food'. The opinions for six stores are summarised in Table 8.

One interesting finding we can see that, in positive (negative) reviews, the opinion of aspect can be opposite. It's because the customers are satisfied (unsatisfied) with the dining experience overall. But they show complaint (praise) on particular aspects.

Let's look at the 'atmosphere' and 'lamian' topics, other topics can be explained similarly. The 'atmosphere' is the evaluation of the restaurant environment, and the aspect words include 'environment', 'decoration', 'location', 'style' and other key words. The following information can be extracted from the opinions of this topic:

- Customers, regardless of positive reviews or negative reviews, of SARB and XHB generally believe that the restaurant environment is poor, cramped, crowded and noisy.
- The opinions of customers in CJCB, WZB, HPB and ECB are mixed. The majority of customers (with good reviews) think the restaurant is bright, elegant and comfortable; some customers (with negative reviews) thought the restaurant was noisy, shabby, crowded.

The 'lamian' topic is about the evaluation of the of lamian dishes. The aspect words include 'lamian', 'noodle', 'bone soup', highly related with lamian. The following information can be extracted from the opinion words of this topic:

- Customers of CJCB and WZB generally think this dish is unpalatable and do not recommend it.
- The customers of SARB generally think that this dish is delicious, spicy but fragrant, and recommend it.
- The customers of HPB, ECB and XHB have mixed opinions on the dish. Part of the customers (with positive reviews) think the food is very good; some of the customers (with negative reviews) thought the food was very spicy and unpalatable

4.2.3. Model comparison

Similar to the skincare dataset, we compare BJ-LDA with Maxent-LDA on this dataset. The comparison results are also very similar to the skincare dataset, which are briefly introduced.

To apply Maxent-LDA on the whole corpora, we set the number of general topic as 1, and the specific topic number is set as $K_0 = 8$, same as BJ-LDA. The results are shown in Table 9. The characteristics of the

Table 8
Detail of general opinion for Japanese restaurant dataset.

		lamian	branch store	hot pot	atmosphere
CJCB	pos	not recommend	expensive	beef	elegant
	neg	terrible	famous	warm	crowded
WZB	pos	cheap	disappointed	spicy	comfortable
	neg	taste bad	decrease	strange	roomy
HPB	pos	spicy	decrease	a little	private
	neg	remember	elegant	taste bad	noisy
ECB	pos	lamian	Canadian	beef	comfortable
	neg	full-bodied	disappointed	fail	noisy
SARB	pos	taste	update	bad	crowded
	neg	smell good	creative	warm	crowded
XHB	pos	good	expensive	full-bodied	noisy
	neg	cheap	famous	pity	uncomfortable
		snacks	promotion	dishes	sea food
CJBJ	pos	sweet	stable	surprise	fresh
	neg	good	bad	satisfied	fishy smell
WZB	pos	full-bodied	delet	light	cooked
	neg	abundant	speechless	good	deserve
HPB	pos	crispy	convenient	sweet	fresh and tender
	neg	abundant	free	encourage	thick
ECB	pos	egg	enjoyable	super good	fresh
	neg	oily	automatic	surprise	excellent
SARB	pos	fresh	improvement	hard	fresh and sweet
	neg	enjoyable	ordinary	good	fat
XHB	pos	creative	decrease	tender	soft
	neg	bad	nice	fresh	best

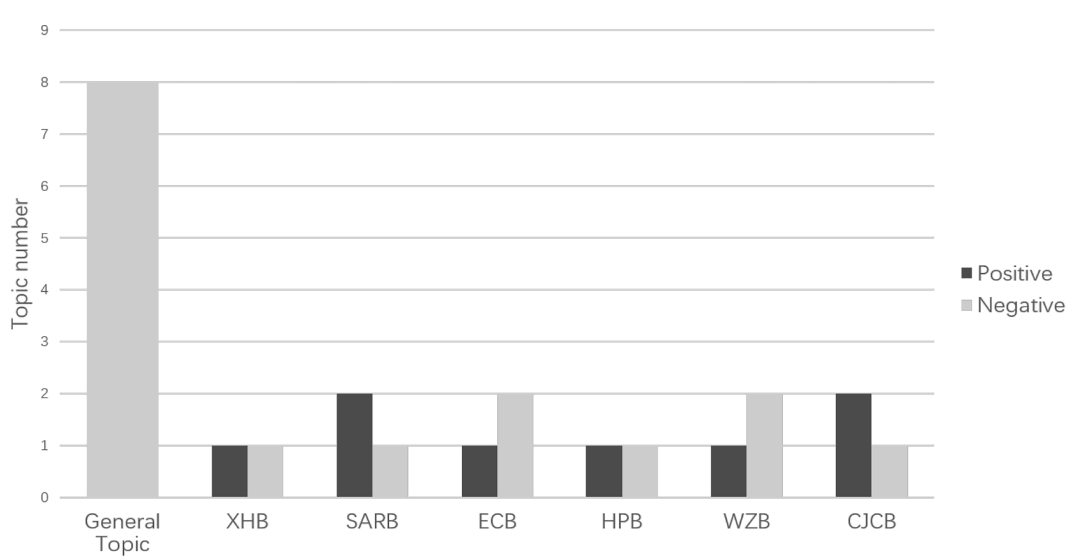

Fig. 6. Number of optimal topic number for Japanese restaurants.

stores are mixed. For example, ‘Chaoyang’, the specific location information for CJCB is mixed in general aspect topic, while BJ-LDA model can discover it precisely. The word ‘private’ is the specific aspect for HPB extracted by BJ-LDA, it is mixed with other aspects in a specific topic in Maxent-LDA. The word ‘noisy’ is the specific opinion for XHB extracted by BJ-LDA, it is mixed with other opinions in Maxent-LDA.

Then we apply Maxent-LDA on each store corpus separately. The results are similar. The characteristics of each brand can be revealed, but it is difficult to compare the opinions of different brands for common aspects like BJ-LDA shown in Table 8. From quantitative perspective, our BJ-LDA achieves the topic coherence -2.4226, that of Maxent-LDA is -2.4228 on the whole corpora, our model is slightly better on topic quality.

4.2.4. *Further exploration*

Similar to skincare dataset, we get six chain brand competition relationships according to Jaccard score and PCA. The result is shown in Fig. 7. The competition between HPB and WZB is close on PC1. It is because they have similar environment and dishes. For example, in the ‘atmosphere’ topic, both customers mention ‘elegant’ and ‘private’. Their main points of competition are price, cost performance and hot pot.

In Japanese data set, store rating can be used as an external variable to assist the analysis. In this subsection, we want to take advantage of the results for BJ-LDA, especially the eight general topics, to analyze the factors that influence the store ratings.

The document-topic distribution θ reflects the focus of the review. The higher value of θ on a topic, the more important the customer focus on it. We performed two-tailed t-test on the two sets of estimation of θ with positive reviews and negative reviews. The results are shown in Table 10. P (N) means the value of θ is significantly higher for positive (negative) reviews than that of negative (positive) one. There is no difference between two groups for the rest.

For example, store WZB has a higher value of θ for their positive reviews at the aspect of dishes, it means the customers with positive reviews focus more on dishes. SARB has higher value of θ for their negative reviews at the aspect of branch store. The customers with negative reviews talk more about this topic. Each store manager should pay attention to all these particular points that influence their store ratings and make corresponding improvements.

5. Conclusion

In this paper, we propose a BJ-LDA approach to analyse the advantages and weaknesses of multiple competing brands using online product reviews. Compared with the extant methods, our method models the corpus of all brands simultaneously, which simplifies the analysis of a large number of brands. Additionally, our model makes full use of the whole corpus. The BJ-LDA model mines general aspects and opinions shared across the brands and discovers brand-specific information. Each word cluster generated by BJ-LDA can be separated into aspects words and opinion words, which helps managers understand the concerns and opinions of consumers. We used the perplexity and topic coherence to select the number of topics. The topics generated by our model are meaningful and provide a variety of information. The BJ-LDA is applied to two datasets, skincare product reviews and Japanese restaurant reviews. The results show that the BJ-LDA model is effective in discovering brand strengths and weaknesses. It performs better than Maxent-LDA model. Principal component analysis (PCA) and two-tailed t-test are applied to provide deeper understanding of the competition between brands and find the significant factors which influence the ratings mostly for each brand. Besides brands, practice use of our model can be extended to any corpora with subgroups such as different regions or prices.

One of the limitations is that we do not use the sparsity technique, and the topics’ distributions span all vocabularies. In future work, we should use the sparsity LDA framework in our model to obtain more interpretable word clusters. We can also try embedding LDA which

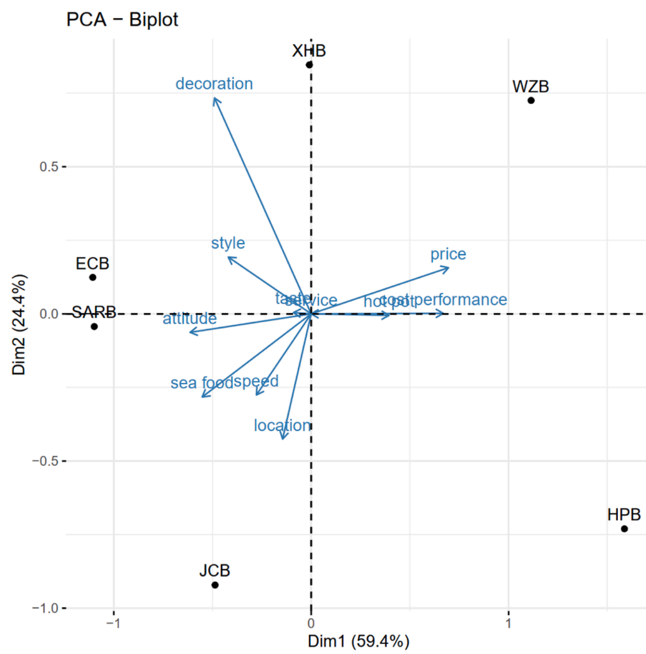

Fig. 7. Competitive landscape of 6 chain brands on 11 aspects.

Table 9
Results of Maxent-LDA for whole chain restaurant corpus.

general topic	aspect	sushi,taste,environment,service,**Chaoyang store**,beef
	opinion	not bad,tasty,nice,like,special,store
topic 1	aspect	oil,bill,Canada,queue,food,wait
	opinion	small,roast food,dirty,soap,fresh
topic 2	aspect	**private**,food,eel,price,super,go
	opinion	totally,deep,dry,not bad,early,reason
topic 3	aspect	weekend,worry,mango,expensive,food,raspberry
	opinion	good,opinion,hot,advantage,not bad,fragrance
topic 4	aspect	taste,chicken,location,rice,promotion,roast food
	opinion	hard,many,thankful,quiet,good,delay
topic 5	aspect	noise,lunch,watermelon,elevator,oden,telephone
	opinion	terrible,expensive,pretty,old,smelly,sorry
topic 6	aspect	lamian,soup,sauce,table,service,preference
	opinion	like,satisfied,tight,active,cool,excellent
topic 7	aspect	fish,arctic shellfish,ice,egg,roast food,people
	opinion	good,many,depress,fresh,upgrade,**noisy**
topic 8	aspect	brand,price,tea,toast food,octopus,tomato
	opinion	mild,tasty,like,arrange,calssical,remember

Table 10
t-test results for each brand on positive reviews and negative reviews.

	CJCB	WZB	HPB	ECB	SARB	XHB
lamian	–	–	–	–	–	–
snacks	–	–	–	–	–	–
branch store	–	–	–	–	N^{**}	N^{**}
hot pot	–	–	P^{*}	P^{*}	–	–
sea food	–	–	–	–	–	–
promotion	–	–	N^{**}	N^{*}	–	–
dishes	–	P^{*}	–	P^{**}	–	P^{*}
atmosphere	–	–	–	–	P^{**}	–

P (N) means the value of the θ is significantly higher for positive (negative) reviews.
*: significant at 0.1.
**: significant at 0.01.

combines word2vec technique. The embedding result will allow us to further explore the brand relationship.

CRediT authorship contribution statement

Yuxuan Guo: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. **Feifei Wang:** Conceptualization, Methodology, Supervision. **Chen Xing:** Formal analysis, Investigation, Methodology. **Xiaoling Lu:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Xiaoling Lu's research is supported by National Natural Science Foundation of China No.72171229 and fund for building world-class universities (disciplines) of Renmin University of China. Feifei Wang's research is supported by National Natural Science Foundation of China (No.72001205, 11971504)

References

Archak, N., Ghose, A., Ipeirotis, P.G., 2011. Deriving the pricing power of product features by mining consumer reviews. Manage. Sci. 57 (8), 1485–1509.

Bagheri, A., Saraee, M., de Jong, F., 2013. An unsupervised aspect detection model for sentiment analysis of reviews. In: International conference on application of natural language to information systems. Springer, pp. 140–151.

Bagheri, A., Saraee, M., de Jong, F., 2014. Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. J. Inf. Sci. 40 (5), 621–636. https://doi.org/10.1177/0165551514538744.

Bancken W., Alfarone D., Davis J., 2014. Automatically detecting and rating product aspects from textual customer reviews. In: Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD. vol. 1202. pp. 1–16.

Blei, D.M., Ng, A.Y., Jordan, M., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Brody, S., Elhadad, N., 2010. An unsupervised aspect-sentiment model for online reviews. In: Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, pp. 804–812.

Choi, S., Fredj, K., 2013. Price competition and store competition: Store brands vs. national brand. Eur. J. Oper. Res. 225 (1), 166–178.

Cruz, F.L., Troyano, J.A., Enriquez, F., Ortega, F.J., Vallejo, C.G., 2013. 'long autonomy or long delay?' The importance of domain in opinion mining. Expert Syst. Appl. 40 (8), 3174–3184.

Culotta, A., Cutler, J., 2016. Mining brand perceptions from twitter social networks. Market. Sci. 35 (3), 343–362. https://doi.org/10.1287/mksc.2015.0968.

Decker, R., Trusov, M., 2010. Estimating aggregate consumer preferences from online product reviews. Int. J. Res. Mark. 27 (4), 293–307.

Dong Z., Dong Q., 2003. Hownet – a hybrid language and knowledge resource. In: Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on.

Fleisher, C.S., Bensoussan, B.E., 2003. Strategic and competitive analysis: methods and techniques for analyzing business competition. Prentice Hall Upper Saddle River, NJ.

Hastings W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1), 97–109. doi: 10.1093/biomet/57.1.97.

Hu, F., Trivedi, R.H., 2020. Mapping hotel brand positioning and competitive landscapes by text-mining user-generated content. Int. J. Hospitality Manage. 84, 102317.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168–177.

Huang, X., Wan, X., Xiao, J., 2011. Comparative news summarization using linear programming. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 648–653.

Jin, W., Ho, H.H., Srihari, R.K., 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In: Proceedings of the 26th annual international conference on machine learning. Citeseer; vol. 10.

Jindal, N., Liu, B., 2006. Mining comparative sentences and relations. In: Aaai. vol. 22. p. 9.

Kobayashi, N., Inui, K., Matsumoto, Y., 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 1065–1074.

Kushal, B., Durga, T., 2013. Feature based summarization of customers' reviews of online products. Proc. Comput. Sci. 22, 142–151.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y., Zhang, S., Yu, H., 2010. Structure-aware review mining and summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 653–661.

Lin, C., He, Y., 2009. Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 375–384.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. vol. 5.

Liu, C., Xu, R., Liu, J., Qu, P., Wang, H., Zou, C., 2013. Comparative opinion sentences identification and elements extraction. In: 2013 International Conference on Machine Learning and Cybernetics. IEEE, vol. 4. pp. 1886–1891.

Liu, K., Xu, L., Zhao, J., 2012. Opinion target extraction using word-based translation model. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp. 1346–1356.

Liu, Q., Huang, H., Zhang, C., Chen, Z., Chen, J., 2013b. Chinese comparative sentence identification based on the combination of rules and statistics. In: International Conference on Advanced Data Mining and Applications. Springer, pp. 300–310.

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272.

Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T., 2002. Mining product reputations on the web. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 341–349.

Rana, T.A., Cheah, Y.N., 2016. Aspect extraction in sentiment analysis: Comparative analysis and survey. Artif. Intell. Rev. 46 (4), 459–483, 10.1007/s10462-016-9472-z.

Ratnaparkhi, A., 1996. A maximum entropy model for part-of-speech tagging. In: Conference on empirical methods in natural language processing.

Romesburg, C., 2004. Cluster analysis for researchers. Lulu. com.

Sipos, R., Joachims, T., 2013. Generating comparative summaries from reviews. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 1853–1856.

Sun, J., Long, C., Zhu, X., Huang, M., 2009. Mining reviews for product comparison and recommendation. Polibits 39, 33–40.

Sun, S., Luo, C., Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. Inf. Fusion 36, 10–25.

Titov, I., McDonald, R., 2008. Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th international conference on World Wide Web, pp. 111–120.

Tkachenko, M., Lauw, H.W., 2014. Generative modeling of entity comparisons in text. In: Proceedings of the 23rd acm international conference on conference on information and knowledge management, pp. 859–868.

Varathan, K.D., Giachanou, A., Crestani, F., 2017. Comparative opinion mining: a review. J. Assoc. Inf. Sci. Technol. 68 (4), 811–829.

Wang, D., Zhu, S., Li, T., Gong, Y., 2013. Comparative document summarization via discriminative sentence selection. ACM Trans. Knowl. Discovery Data 7 (1), 1–18.

Wang, W., Feng, Y., Dai, W., 2018. Topic analysis of online reviews for two competitive products using latent dirichlet allocation. Electron. Commer. Res. Appl. 29, 142–156, 10.1016/j.elerap.2018.04.003.

Wang, W., Zhao, T., Xin, G., Xu, Y., 2015. Exploiting machine learning for comparative sentences extraction. Int. J. Hybrid Inf. Technol. 8 (3), 347–354.

Wei, X., Zhang, Q., Yuan, Y., Chi, Z., 2014. Conceptual collocations of words in comparative sentences. In: Proceedings of the 15th chinese lexical semantics workshop.

Wiebe, J., Wilson, T., Cardie, C., 2005. Annotating expressions of opinions and emotions in language. Language Resour. Eval. 39 (2–3), 165–210.

Xu, K., Liao, S.S., Li, J., Song, Y., 2011. Mining comparative opinions from customer reviews for competitive intelligence. Decis. Support Syst. 50 (4), 743–754.

Zhao, X., Jiang, J., Yan, H., Li, X., 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. ACL.

Zhu, J., Wang, H., Zhu, M., Tsou, B.K., Ma, M.Y., 2011. Aspect-based opinion polling from customer reviews. IEEE Trans. Affect. Comput. 2 (1), 37–49.