Electronic Commerce Research and Applications 32 (2018) 13–22

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Training attractive attribute classifiers based on opinion features extracted from review data

Wei Ou[a,*], Van-Nam Huynh[a], Songsak Sriboonchitta[b]

[a] Japan Advanced Institute of Science and Technology, Asahidai 1-1, Nomi City, Ishikawa, Japan
[b] Centre of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand

ARTICLE INFO

Keywords:
Kano model
Attractive attributes
Review data
Neural networks
Machine learning

ABSTRACT

Researchers have proposed statistical regression models that analyse on-line review data to identify attractive attributes of a product or service. This research has the same aim, but with an approach based on machine learning models instead of statistical models. The proposed approach first extracts attribute-level sentiments from the review text by natural language processing techniques, then derives features that reflect the non-linear relations between attribute performance and customer satisfaction based on the sentiments. The non-linear features are fed to the Support Vector Machine (SVM) model to train predictive attractive attribute classifiers. The proposed approach is evaluated on a hotel review dataset crawled from TripAdvisor. The experiment results indicate that the classifiers reach a precision of 79.3% and outperform the existing statistical models by a margin of over 10%.

1. Introduction

The Kano model (Kano, 1984) has been widely used by researchers and marketing practitioners as a good tool to identify the drivers of customer satisfaction. According to the model, all the attributes of a product or service can be divided into 3 main categories: must-be attributes, one-dimensional attributes, and attractive attributes. Must-be attributes are the minimum requirements that customers expect. Extreme dissatisfaction is caused if their performance is insufficient while satisfaction is not necessarily caused if their performance is sufficient. One-dimensional attributes are also expected by customers but they differ from the must-be aspects in affecting customer satisfaction in a linear way: satisfaction is caused if their performance is sufficient and dissatisfaction is caused if their performance is insufficient. Attractive attributes are usually not expected by customers and their effects are opposite to that of must-be attributes: satisfaction is caused if their performance is sufficient but dissatisfaction is not necessarily caused if their performance is insufficient.

Categorising product attributes into those 3 categories helps people understand better the drivers of customer satisfaction (Mikuli and Prebeac, 2011). In particular, pinpointing the attractive attributes is of strong interest to marketing practitioners because attractive attributes can distinguish a product or service from its competitors. In a mature market where players are homogenous, attractive attributes are usually vital to boost customer satisfaction and expand customer bases. In this research we focus only on the attractive attributes and aim to build a model to classify an attribute of a product or service as whether belonging to the attractive category.

The propensity of an attribute belonging to the attractive category can be decided by using regression techniques to estimate the effects of its performance on customer satisfaction. The first phase in the classification task is to collect customer opinions about their responses to various levels of attribute performance. In the early days of marketing research, researchers usually relied on quantitive questionnaire-based surveys to get the opinion data (Chen, 2012; Tontini and Silveira, 2007; Tan and Shen, 2000; Matzler et al., 2004). Since questionnaire-based surveys use a limited set of predefined questions and can investigate only a trivial fraction of a population, it provides a relatively narrow spectrum of information, especially in today's era of big data. As e-commerce grew rapidly in recent years, on-line reviews have been used to extract the opinions. Because of the accessibility and flexibility customers enjoy on on-line review platforms, the data can cover much broader demographic groups and deliver richer information than the questionnaire-based survey data.

However, there is a problem when trying to use review data for the classification task. Review data usually consists of numerical ratings that indicate overall customer satisfaction levels and free-form text that details customers' evaluation on each attribute. Obviously, the qualitative textual description cannot be processed by regression tools. To address the problem, researchers have proposed approaches (Xu and Li,

[*] Corresponding author.

https://doi.org/10.1016/j.elerap.2018.10.003
Received 7 January 2018; Received in revised form 27 August 2018; Accepted 8 October 2018
Available online 09 October 2018

2016; Lu and Stepchenkova, 2012; Tontini et al., 2017; Zhang and Cole, 2016) that first use natural language processing (NLP) techniques to extract customer sentiments associated with each attribute as the quantitive proxy for the qualitative description, then apply statistical regression models to derive the effects of the attribute-level sentiments on the overall ratings as an approximation for the effects of attribute performance on customer satisfaction.

A major weakness of those works lie in their ineffectiveness in encoding the non-linearity of the effects. Also, those works treat the items to be analysed as isolated and their regression processes are performed on the local data of individual items. The global statistical patterns across the attributes of different items are overlooked. Those limitations lead to severe loss of information. Furthermore, the attribute-level sentiment analysis process usually generates heavy noise as the NLP techniques are still far from being mature. Ignoring the global statistics may also cause those methods to be sensitive to the noise.

In this study, the review data is also used to collect the customer opinions. However, this study differs its research goal from the existing ones by using machine learning models instead of statistical models to train classifiers to determine the Kano categories of attributes. To train such classifiers, numerical features that reflect the aforementioned effects have to be derived from the review data to represent attributes. A very natural form of attribute representation is the empirical probabilities of ratings conditioning on various sentiments associated with each attribute. To compute the representation, the effects of the sentiments have to be assumed as independent. However, in a realistic situation, the sentiments should closely interact with each other in determining the overall ratings; ignoring these interactions may lead to misrepresentation of attributes. Modelling the interactions is difficult as it is essentially a regression problem with a very high degree of nonlinearity. In this research a neural network model is proposed to encode the non-linearity. In the neural network, the input is the attribute-level sentiments of a review and the output is the possibility of each possible rating for the review. The weights of a particular hidden layer are used as the attribute representation incorporating the impacts of the interactions. For the convenience of expression, in the rest of this paper, the former type of representation will be called ‘empirical effect’, and the latter type will be called ‘interactive effect’.

SVM classifiers are trained with the derived attribute representations. Since the majority of the existing statistical models focus on hotel review data (Tontini et al., 2017; Zhang and Cole, 2016; Lu and Stepchenkova, 2012; Xu and Li, 2016), this study also aligns the target on hotel review data for a fair comparative evaluation. The experiment results show that the proposed classifiers outperform the statistical models by a considerable margin of over 10%. This is partly because the classifiers make use of the global statistics and can better encode the non-linearity of the training data. The contributions of this paper are as follows. Firstly, we are the first to use machine learning techniques to train predictive classifiers to identify attractive attributes from on-line review data. Secondly, we propose to use neural network techniques to derive the non-linear interactions among different levels of attribute performance in determining overall customer satisfaction.

A flowchart of the proposed approach is shown in Fig. 1. The rest of this paper is organised as follows. Section 2 reviews the related literature; Section 3 introduces the proposed attribute representation models; Section 4 presents details of the attribute-level sentiment analysis; Section 5 analyses the data and summarises the results; Section 6 provides conclusions and future research directions.

2. Related work

In the proposed approach, the premise to identify the attractive attributes is devising natural language processing techniques to obtain structured customer opinions from the unstructured review text. In this research, the opinion extraction process is treated as an attribute-level sentiment analysis problem (Pontiki et al., 2016), that is to detect the attributes mentioned in a review and the sentiment orientations on them. This is the first major subproblem that needs to be discussed. Based on the structured customer opinions, associations between attribute performance and customer satisfaction have to be derived to determine the Kano category of each attribute. This is the second subproblem that needs to be investigated. In this section some representative works related to the two problems are presented.

2.1. Attribute-level sentiment analysis

As the name suggests, attribute-level sentiment analysis consists of two sub-classification problems: sentiment classification and attribute classification. The two tasks can be performed either separately or jointly. Various off-the-shelf supervised classification models can be used if they are treated as two separate tasks. A key problem in the classification tasks is deriving features to represent the text. The traditional bag-of-words based feature models, such as one-hot-encoding, tf-idf, etc., have been widely used to represent the review text (Brychcn et al., 2014; Kiritchenko et al., 2014; Pang et al., 2002; Martineau and Finin, 2009; Bespalov et al., 2011). However, the bag-of-words based features are usually very noisy because reviews are very short and the frequencies of informative words in a review are too low to reflect their semantic meaning. To reduce the noise, lexicons of attribute or sentiment keywords can be used and the representation of a review is defined only by the keywords present in the review (Yu et al., 2011; Wang et al., 2010; Hu and Liu, 2004; Long et al., 2010; Toprak et al., 2010; Zhuang et al., 2006; Guo et al., 2009). In recent years, text embedding models that project textual data into a semantic space to derive their semantic representations have been proven to be effective in the two classification tasks (De Boom et al., 2016; Mikolov et al., 2013; Pennington et al., 2014; Kenter et al., 2016; Lilleberg et al., 2015). The elements of an embedding feature vector represent the membership of a word or a text fragment in a limited set of abstract senses. Compared with the bag-of-words based features, the embedding representations are much semantically richer, lower-dimensional, and therefore, easier to be processed by machine learning models.

In the joint approach, the aspect labels and the sentiment labels are included in the same cost functions and the syntactic or semantical dependencies between the two labels serve as the constraints for the cost functions. Optimising such cost functions will generate the two types of labels simultaneously. Methods based on conditional random field (Li et al., 2010a; Marcheggiani et al., 2014; Zhuang et al., 2006 and topic modeling Jo and Oh, 2011; Lin et al., 2009; Mukherjee and Liu, 2012; Li et al., 2010b; Mei et al., 2007) have been proposed for this approach. Compared with the separate approach, the joint approach is naturally more computationally efficient. Furthermore, some of those methods (Jo and Oh, 2011; Mukherjee and Liu, 2012; Li et al., 2010b) are unsupervised and do not require manual labor to annotate the training data. However, the performance of the methods in the joint approach is usually worse than the separate approach because they rely on strong assumptions about the correlations between the attribute and sentiment labels that are not necessarily realistic in the real world.

2.2. Statistical Kano classification models

With the extracted customer opinions, methods based on penalty-reward contrast analysis(PRCA) (Brandt, 1987) can be used to classify each attribute into 1 of the 3 categories. PRCA based methods are focused on estimating the impacts of each attribute on customer satisfaction. The impacts of an attribute consist of the reward impact and the penalty impact. The former represents the impact when the performance of the attribute is sufficient, and the latter indicates the impacts when performance is insufficient. If the reward impact of an attribute is stronger than the penalty impact, then the attribute is very likely to be an attractive attribute. Tontini et al. (2017) first perform attribute-level sentiment analysis on review data, then apply the linear

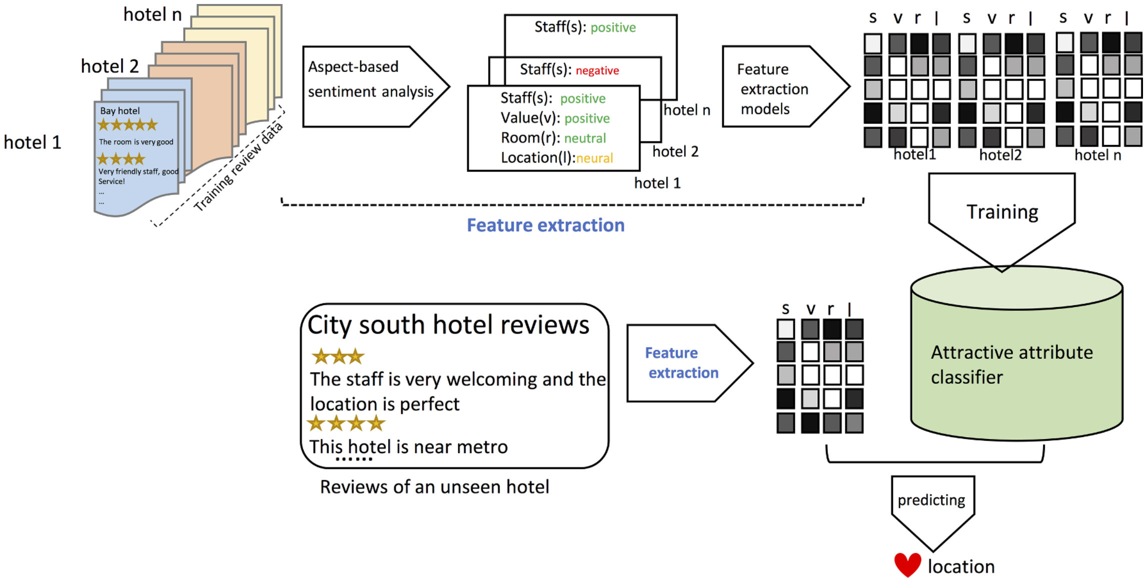

Fig. 1. The workflow of the proposed method.

regression model to derive the relations between the sentiments and ratings as an approximation of the impacts. Zhang and Cole (2016) extract keywords and their sentiment weights for each attribute from the review data to quantise the attribute performance. By conditioning on the derived attribute performance, the authors apply the logistic regression model to derive the reward and penalty impacts.

There are also existing models based on the critical incident technique (CIT) (Gremler, 2004). A critical incident refers to the incidents that cause significant contributions or damages to customer satisfaction. If an attribute frequently causes positive incidents in favourable reviews but rarely causes negative incidents in unfavourable reviews, then the attribute is very likely to be attractive. Lu and Stepchenkova (2012) identify the incidents in each attribute based on the attribute-level sentiments and employ the Krusal-Wallis test to determine the relation between the incidents and the overall ratings. Xu and Li (2016) first divide all reviews into a positive group and a negative group, then represent all the words in each review group by a matrix of TF-IDF weights. The authors employ the latent semantic analysis (LSA) model to decompose the matrices to uncover the keywords and critical incidents for each attribute. The correlation between the incidents and each review group is determined by aggregating the weights of the corresponding keywords.

3. Attribute representations

As mentioned in Section 1, two types of representations, the empirical effects and the interactive effects, are derived from the review data to represent attributes to train the classifiers. Assuming there is a set of M review documents for an item, the mth review document is denoted as $d^{(m)}$. The evaluation of the item involves a set of K aspects. There are I different scales of sentiments; the larger a sentiment index i the more positive the sentiment. There are J scales of overall ratings; the larger a rating j the higher the overall customer satisfaction. The notation used throughout this section is shown in Table 1.

3.1. Empirical effect

Let r denote the ratings, r_l and r_h are a rating falling in the lower half of the rating range and a rating in the higher half of the rating range, respectively. Let s_k denote the sentiments on attribute k, s_{k_-} and s_{k_+} are a negative and a positive sentiment on the attribute, respectively.

Table 1
Notation used throughout Section 3.

m	Review index
k	Attribute index
i	Sentiment index
j	Rating index
$d^{(m)}$	The mth review document
$s_k^{(m)}$	The sentiment set on attribute k in $d^{(m)}$
$r^{(m)}$	The rating of $d^{(m)}$
u_k	The empirical effect features of k
v_k	The interactive effect features of k

According to the definition, for must-be attributes, the probability of a negative sentiment leading to a low rating must be greater than the probability of a positive sentiment leading to a high rating, namely, $p(r_l|s_{k_-}) > p(r_h|s_{k_+})$; for attractive attributes, $p(r_h|s_{k_+}) > p(r_l|s_{k_-})$; for one-dimensional attributes, $p(r_l|s_{k_-}) \approx p(r_h|s_{k_+})$. In this research, the conditional probabilities associated with an attribute are concatenated to form a feature vector to represent the attribute:

$$u_k = [p(r = j|s_k = i)]_{i\in[1,I],j\in[1,J]} \tag{1}$$

$$p(r = j|s_k = i) = \frac{\sum_{m=1}^{M} 1\{i \in s_k^{(m)},\ r^{(m)} = j\}}{\sum_{m=1}^{M} 1\{i \in s_k^{(m)}\}} \tag{2}$$

where $p(r = j|s_k = i)$ denotes the probability of rating scale j given sentiment index i associated with aspect k across all M reviews for the item, $s_k^{(m)}$ is the set of sentiments on attribute k in review $d^{(m)}$, and $r^{(m)}$ is the overall rating of the review.

According to the previously presented analysis, the empirical effects of attributes in the same Kano category would share some common patterns. However, there is an obvious weakness in the representation: it assumes aspects are independent of each other and does not reflect the interactions among attributes in deciding the overall ratings. The interactions can also provide information related to the nature of an attribute. For instance, when a basic need is poorly fulfilled, it is very likely that the probability $p(r_l|s_{k_-})$ would be pushed high and the conditional probabilities on other attributes would be lowered significantly. This is because customers may simply ignore the performance of other attributes if a must-be attribute fails to meet their expectations. In this paper we model the interactions with the neural network model and show the details in the following subsection.

3.2. Interactive effect

This subsection assumes that the effects of attribute-level sentiments are dependent and connected in determining the overall ratings. Theoretically, one can hard encode the rules of how the effects interact with each other and use probabilistic models to derive the effects based on those rules. However, given the high variety and complexity of the effects, it is infeasible to construct such a model. Instead, this research uses the neural network model as a mapper that aggregates the effects of the attribute-level sentiments appearing in a review and maps them to the rating of the review. Since the output of the neural network, namely the ratings, is observable in the data, the back-propagation algorithm (LeCun et al., 1990) can be used to derive the mapping function of the neural network and the interactive effects.

In the proposed neural network, the input is a vector indicating the presence or absence of each attribute-level sentiment: $O^{(m)} = [1\{i \in s_k^{(m)}\}]_{k\in[1,K],i\in[1,I]}$. Let $V = [v_{ki}]_{k\in[1,K],i\in[1,I]}$ denote the matrix of interactive effects of all attribute-level sentiments. The activations of the first layer are the weighted average of the effects of the attribute-level sentiments present in review $d^{(m)}$:

$$A^{[1]} = W^{[1]\top}(O^{(m)} \otimes V) + b^{[1]} \tag{3}$$

where $W^{[1]}$ is the vector of weights for the attribute-level sentiments, $b^{[1]}$ is the bias term, $O^{(m)} \otimes V$ is the element-wise product of the indicator vector and the interactive representation matrix. By applying the element-wise product operation, only the attribute-level sentiments present in the review play roles in determining the overall ratings.

The activations in the first hidden layer are fed into subsequent layers in which neurons are fully connected to model the interactions among the attribute-level sentiments in deciding the overall ratings. Assuming there are L hidden layers in the neural network, the activations of a hidden layer $l \in (1, L]$ can be computed as follows:

$$A^{[l]} = g(W^{[l]}A^{[l-1]} + b^{[l]}) \tag{4}$$

where g is the tanh activation function of the neural network, $W^{[l]}$ is the matrix of weights of the connections between layer l−1 and layer l, $b^{[l]}$ is the bias term. In the output layer, the softmax function σ is used to compute the probability of each possible rating for the review:

$$A^{[o]} = \sigma(W^{[o]}A^{[L]} + b^{[o]}) \tag{5}$$

where $A^{[o]} = [p(\hat{r}^{(m)} = 1), p(\hat{r}^{(m)} = 2), ...p(\hat{r}^{(m)} = J)]$, $p(\hat{r}^{(m)} = j)$ is the probability to predict a rating of j for the review. The negative log likelihood is used as the loss function of the proposed model:

$$C^{(m)} = -\sum_{j\in[1,J]} 1\{r^{(m)} = j\}^{*}\log p(\hat{r}^{(m)} = j) \tag{6}$$

The backward propagation method (LeCun et al., 1990) is used to derive the unknown interactive effects and the hidden layer weights and biases. The structure of the neural network is shown in Fig. 2. The hyper-parameters of the structure, including the dimensionality of the interactive effect features, the number of hidden layers, and the number of activations in each hidden layer, are also unknown beforehand and decided by the grid search technique (Bergstra et al., 2011) in the training process. Details of the hyper-parameter search process will be shown in Section 5.

After the training process, the learned interactive effects of all the sentiments associated with an attribute k are concatenated to represent the attribute: $v_k = [v_{ki}]_{i\in[1,I]}$. Furthermore, we concatenate the interactive effects to the aforementioned empirical effects to form a more informative representation for the attribute.

4. Extracting attribute-level sentiments

This section introduces the proposed approach for attribute-level sentiment analysis. The attribute classification and sentiment classification are treated as two separate problems. In other words, two separate classifiers are trained for the attribute classification and the sentiment classification. In the proposed approach, a review is decomposed into sentences and the classifications are performed on each individual sentence. Assuming a sentence of a review $d^{(m)}$ is classified as discussing about attribute k with sentiment i, then the sentiment i is added to the sentiment set for attribute k to make $i \in s_k^{(m)}$ true. In this research, the off-the-shelf sentiment classifier in the StanfordNLP package (Manning et al., 2014) is used for the sentiment classification and the following method is used for the attribute classification.

Word-to-vector (W2V) (Mikolov et al., 2013), a representative text embedding model, is used to represent the review text because it reflects the semantic information and allow for computing the similarity between any pair of words in the vocabulary. Assuming there exists a set of keywords ϕ_k for each aspect k, the similarity between a review sentence and an given aspect k is computed as follows:

$$sim(d^{(m,n)}, k) = \max_{e_1\in d^{(m,n)}, e_2\in\phi_k} Cos(f_{e_1}, f_{e_2}) \tag{7}$$

where $d^{(m,n)}$denotes the nth sentence in review $d^{(m)}$, $e_1 \in d^{(m,n)}$ denotes each word in the review sentence, $e_2 \in \phi_k$ denotes each keyword for aspect k, f denotes the word-to-vector features of a word, Cos denotes the cosine similarity between a pair of feature vectors. The equation indicates that we compute the similarity between each word of a sentence and each keyword of an aspect and use the maximum similarity as the similarity between the sentence and the aspect.

If the similarity between a sentence and an aspect exceeds a predefined threshold γ, then the sentence is deemed to be about the aspect:

$$z^{(m,n)} = \arg_k sim(d^{(m,n)}, k) > \gamma \tag{8}$$

where $z^{(m,n)}$ is the set of aspects the sentence may discuss about.

5. Evaluation

A dataset containing around 60,000 reviews of 400 hotels in New York city is collected from TripAdvisor (Wang et al., 2010). In this dataset, each hotel has at least 100 reviews. Each review has an integer rating in the range [1, 5] and a piece of review text. The distributions of customer opinions may vary significantly by room rate, therefore, all the hotels are divided into 4 categories according to their room rates: cheap, budget, medium, and luxury. Statistics of each category of hotels are shown in Table 2.

To get the ground truth labels for the training process, all the reviews of each hotel are evenly divided into 15 portions. 15 volunteers with background in hotel management are employed to read the reviews, each volunteer is assigned one portion of the reviews. Based on their understanding from the reading, each volunteer is asked to fill in the Kano matrix shown in Table 3 for each attribute discussed in the reviews. According to the Kano theory (Kano, 1984), if the answer of a volunteer for an attribute falls in the cells marked with ‘A’ then the attribute is attractive. If more than 7 of the 15 volunteers give the attractive response then the attribute is labelled as attractive, otherwise as non-attractive.

In the attribute-level sentiment analysis, as previously mentioned in Section 4, the sentiment classifier provided by StanfordNLP (Manning et al., 2014) is used to determine the sentiment of each review sentence. In this classifier each sentiment is indicated by an integer in the range [1, 4], such that close to 4 indicates a strong positive sentiment while close to 1 indicates a strong negative sentiment. The proposed method described in Section 4 is used to detect the attributes in each review sentence. To use the method, the word-to-vectors features have to be learned from the review text. Since the aspects of a sentence are mostly

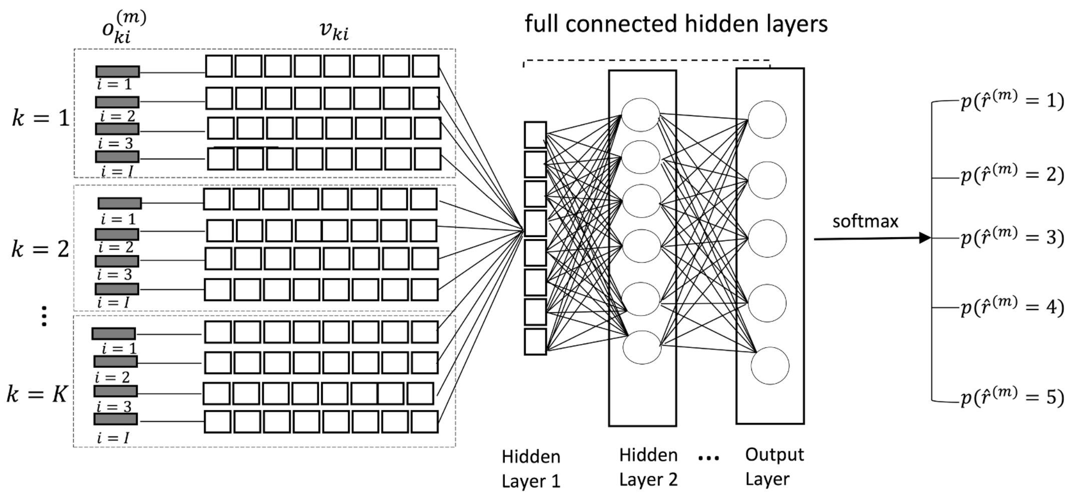

Fig. 2. The structure of the proposed neural network.

Table 2
Statistics of each hotel category.

Hotel type	Price range(USD)	Number of hotels	Number of reviews	Average rating	Standard deviation of ratings
Cheap	0–100	86	18567	3.23	1.75
Budget	100–200	114	20463	3.05	1.64
Medium	200–400	120	26742	3.39	1.67
Luxury	Above 400	85	15064	3.71	1.26

Table 3
The Kano matrix.

		Dysfunctional question: how do you feel if the performance of the attribute is insufficient?		
		I am neutral	I can live with it that way	I dislike it that way
Functional question: how do you feel if the performance of the attribute is sufficient?	I like it that way	A	A	
	It must be that way			
	I am neutral			

Table 4
The rating attributes of the reviews.

General hotel attributes	Lobby, accessibility, parking, shuttle services, check-in/out, staff, price, reservation, nearby leisure facility, taxi-calling
General room attributes	Cleanliness, quietness, room size, furniture, bedding, AC, view, bathroom, decoration, internet, pet-friendliness
Room appliances	Fridge, microwave, coffee-maker, computer
Food related attributes	Breakfasts, affiliated restaurants, complimentary food and drinks
Additional facilities	Conference rooms, gym, pool

characterised by the nouns, verbs, adjective, and adverbs, the part-of-speech tagger (Manning et al., 2014) is used to select those words from each sentence as the input of the word-to-vector model. The number of dimensions of the representation is set to be 100.

A set of attributes shown in Table 4 that are frequently discussed in the existing hospitality research papers (Tontini et al., 2017; Zhang and Cole, 2016; Lu and Stepchenkova, 2012; Xu and Li, 2016) are used as the possible labels the proposed method would assign to a review sentence. Also, 10 keywords for each attribute are obtained by consulting a volunteer expert in the hospitality industry. We set the threshold γ in Eq. (8) to 0.6, on which the best performance is observed.

5.1. Aspect identification

The performance of the proposed attribute detection method is compared with the following methods: the supervised Naive Bayesian (Zhai et al., 2010) and SVM model (Kiritchenko et al., 2014) trained with the bag-of-words based tf-idf features, the lexicon based Bootstrapping method (Wang et al., 2010), and the SVM model trained with the word-to-vector features (Lilleberg et al., 2015). The comparison results are shown in Table 5. The results indicate that the proposed method outperforms the best among the first 3 methods by a margin of 5% in terms of the F1 score. The SVM classifier trained with word-to-vector features outperforms the proposed method, but the advantage comes at the price of tedious manual labor for labelling the training data. Also, since the threshold γ in Eq. (8) allows the proposed method to assign aspect labels to an instance only when it is ‘extremely’ confident about the decision, the proposed method usually assigns no label, while other models usually assign wrong labels in the false negative cases. This may result in less noise in analysing the correlations between attribute-level sentiments and the overall ratings, especially when the training data comes in very large volume.

Table 5
Performance of the attribute classification models.

	SVM + tf-idf	Naive Bayesian + tf-idf	Lexicon-based boots-strapping	SVM + W2V	Our model
Average precision	0.80	0.72	0.82	0.88	0.85
Average recall	0.66	0.62	0.60	0.79	0.70
Average F1 score	0.72	0.67	0.69	0.83	0.77

We also show the number of review sentences related to each aspect across the whole dataset in Fig. 3. The results indicate that staff is the most frequently discussed attribute in the dataset. This is consistent with the fact that professional ethics and competence of staff members are crucial elements to increase customer satisfaction in the hospitality industry.

5.2. *Empirical effect*

As previously mentioned, the dimensionality of the empirical effects of an aspect is $I^* J = 4 * 5 = 20$. Each dimension represents the conditional probability of an overall rating given a sentiment index detected on the aspect. We compute the features for each aspect of each hotel and show a snippet of them in Fig. 4. The features in the first two rows are for two attractive aspects of a hotel: gym and shuttle service. The waves of the two features have relatively low values when the sentiments are not positive and have sharp peaks at $p(r = 4|s = 4)$ and $p(r = 5|s = 4)$. This indicates that customers are prone to give high ratings when the performance of those aspects is sufficient, but are insensitive to them when their performance is insufficient. The two features in the middle row are for two must-be aspects of the same hotel: check-in/out and room quietness. The waves are opposite to that of the attractive aspects. This indicates that customers can be easily dissatisfied when the performance of the aspects is insufficient but less possible to be satisfied when the performance is sufficient. The features in the third row are for two one-dimensional attributes: staff and cleanliness. Those waves have peaks at both ends, reflecting the linear relation between the performance of those aspects and the overall customer satisfaction.

5.3. *Interactive effect*

In this research, the Tensorflow package (Abadi et al., 2015) is used to implement the proposed neural network to derive the interactive effects. As previously mentioned in Section 3, there are several hyper parameters of the neural network that need to be determined by grid search: the dimensionality of the interactive effect features, the number of hidden layers, and the number of activations in each hidden layer. In the grid search, the dimensionality of the interactive effect features is empirically set in the interval [5, 20] with a step size of 2; the number of layers in the interval [2, 10] with a step size of 1; the activation size in the interval [10, 30] with a step size of 2. The grid search tries all possible parameter combinations and finds the neural network that has the best performance when the dimensionality of the features is 9, the number of hidden layers is 3, the first hidden layer size is 9 (the same as the dimensionality of the features), the second hidden layer size is 20, and the third hidden layer size is 12.

The interactive effect features of sentiments associated with each attribute are visualised by the Hinton diagram (Bremner et al., 1994). There are 4 rows in the diagram, each row for a sentiment indicator. The white blocks in the diagram represent positive values, and the black blocks negative values. The areas of those blocks indicate the magnitude of the values. We find that the features of attractive attributes have appearance distinctively different from that of the must-be and the one-dimensional attributes. A snippet of the results is shown in Figs. 5 and 6. In these figures, the features of attractive attributes are usually preoccupied with tiny blocks after the second column, while in the features of non-attractive attributes relatively large blocks appear across the whole matrix.

To further demonstrate the discriminatory power of the features, K-

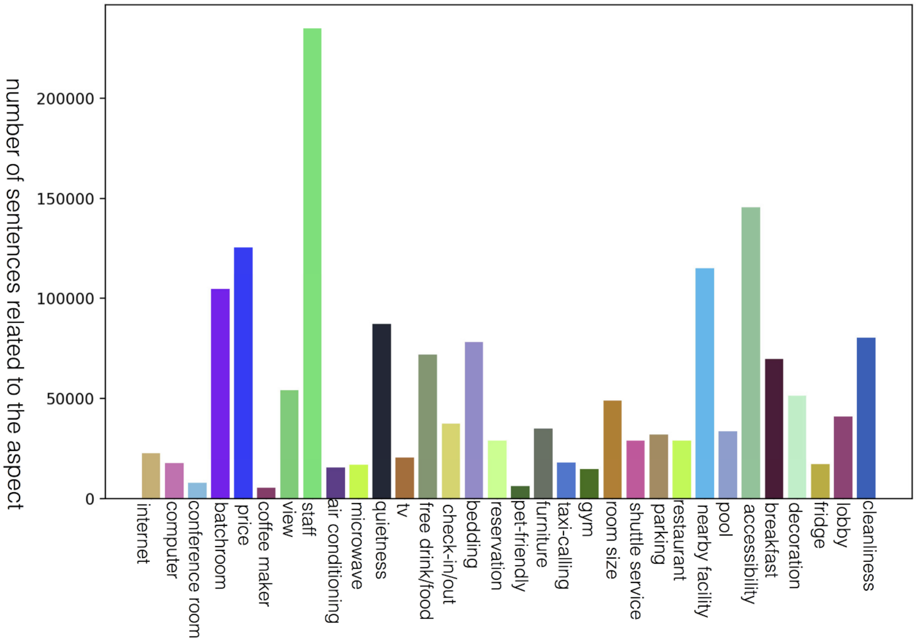

Fig. 3. Number of sentences related to each aspect.

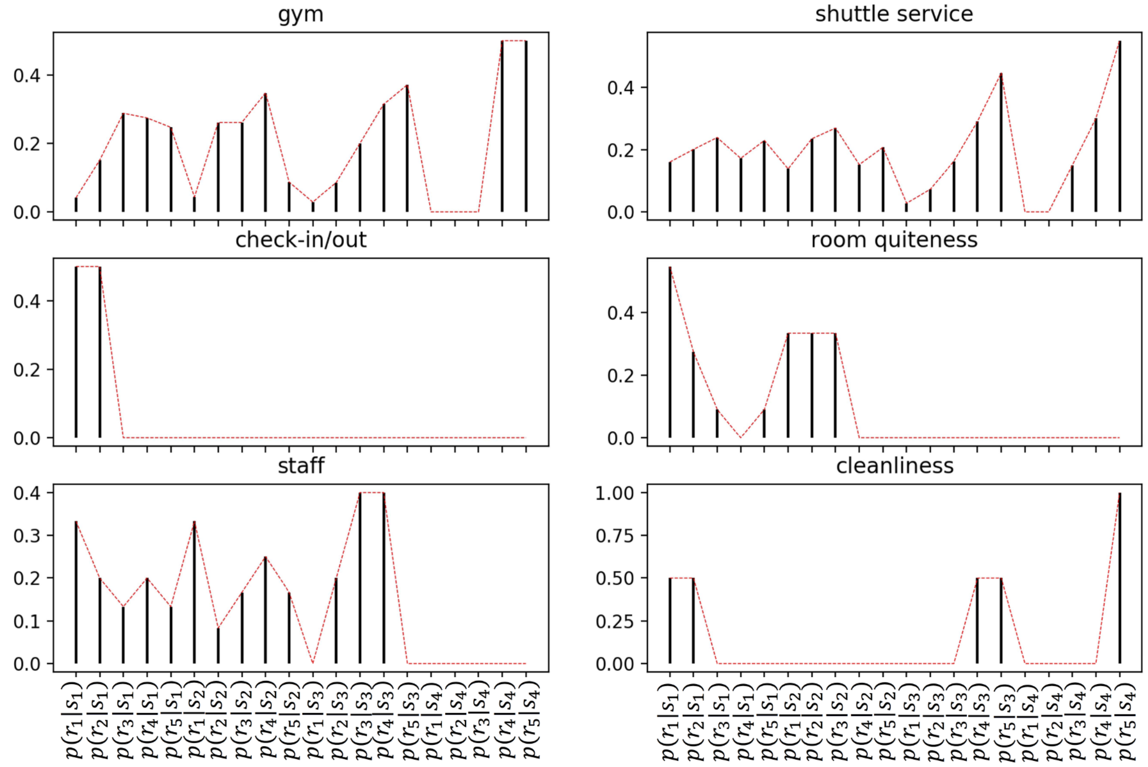

Fig. 4. A snippet of the empirical effect features. The tick $p(r_j|s_i)$ represents the conditional probability of rating j given sentiment i associated with an aspect.

means clustering is performed on the features. The number of clusters is set to 2, one cluster represents attractive attributes, and the other non-attractive attributes. In the results, the clustering purities for the cheap, budget, medium, and luxury categories are 0.63, 0.62, 0.75, and 0.68, respectively. We visualise the clustering results on 20 hotels randomly chosen from each hotel category in Fig. 7. As the 4 plots indicate, the majority of the attractive attributes fall in the cluster with yellow background, and the majority of the non-attractive aspects in the

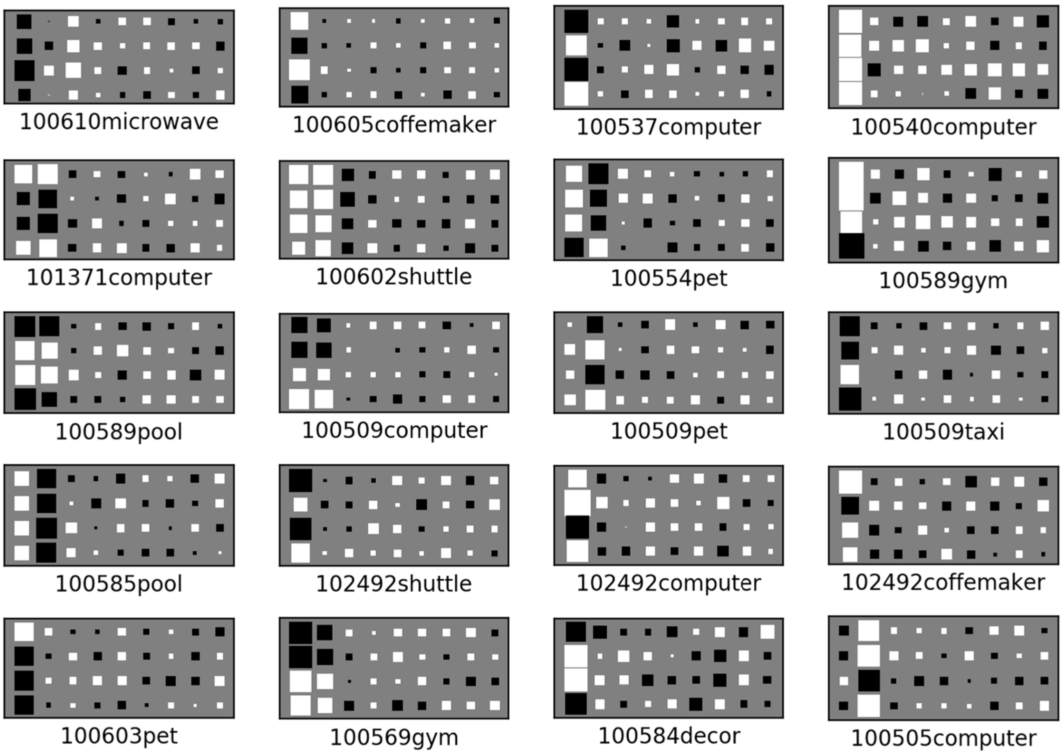

Fig. 5. A snippet of the interactive effects of attractive attributes (the number below each diagram is the hotel ID used during the experiment).

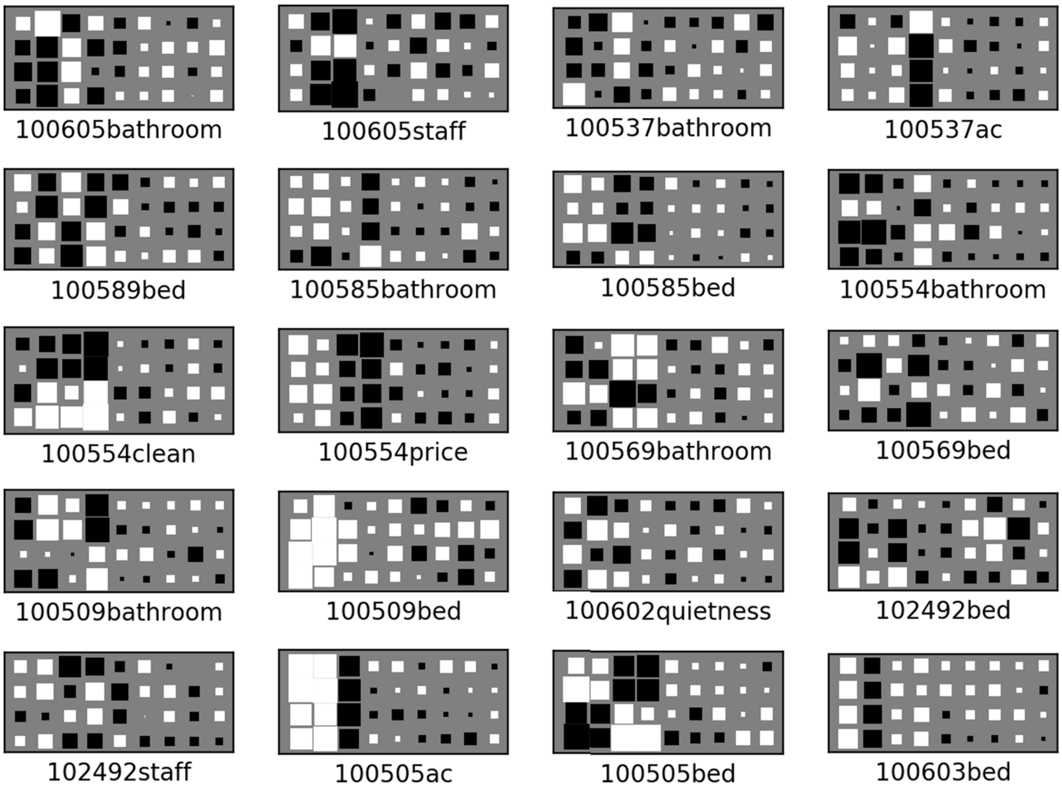

Fig. 6. A snippet of the interactive effects of non-attractive attributes.

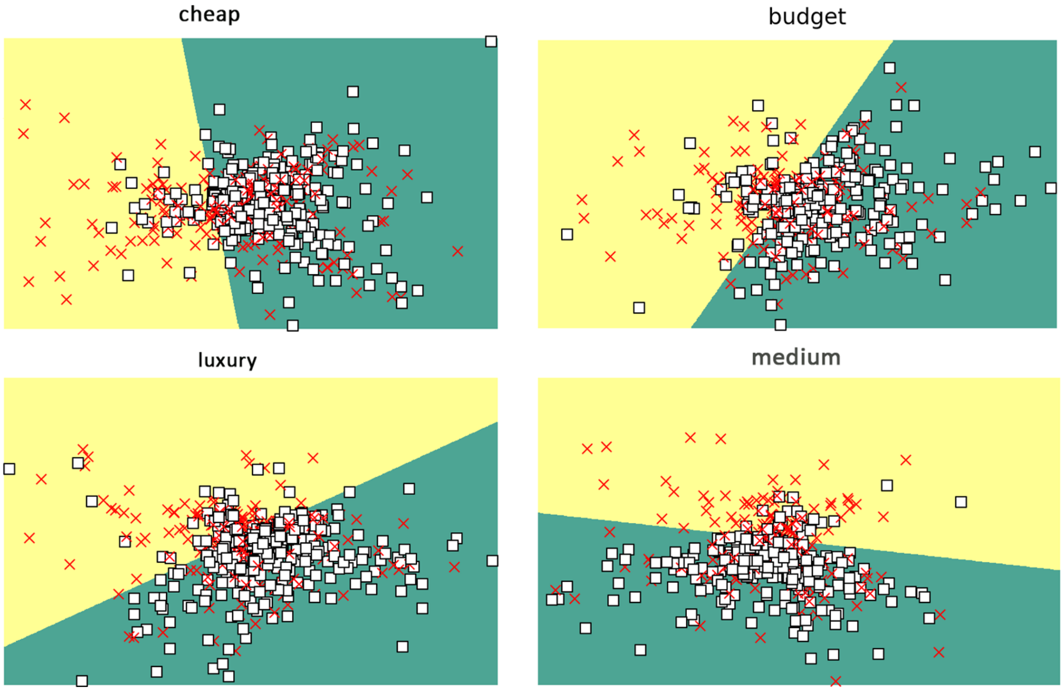

Fig. 7. The results of K-means clustering with the interactive effect features. The red cross marks represent attractive attributes, and the white squares non-attractive attributes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cluster with green background. It is noteworthy that the clustering quality on the medium category data is the best. This is partly because the number of favourable and unfavourable reviews in this category is relatively balanced, providing evenly distributed information about the attractive attributes for the neural network.

5.4. The attractive attribute classifiers

Each detected aspect of each hotel is treated as a data instance, and the following 3 features are used to train 3 separate SVM classifiers: the empirical effect features, the interactive effect features, and the

Table 6
Performance of the proposed classifiers and the PRAC/CIT based methods. Prec stands for 'Precision', Rec stands for Recall'.

	SVM + the empirical effects			SVM + the interactive effects			SVM + the concatenation features			PRCA(Tontini et al., 2017)			CIT (Lu and Stepchenkova, 2012)		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Luxury	0.76	0.69	0.72	0.86	0.80	0.83	0.88	0.85	0.86	0.65	0.55	0.60	0.65	0.61	0.63
Medium	0.71	0.51	0.59	0.76	0.59	0.66	0.80	0.61	0.69	0.46	0.62	0.53	0.70	0.50	0.58
Budget	0.69	0.89	0.78	0.73	0.91	0.81	0.77	0.91	0.83	0.59	0.74	0.66	0.67	0.59	0.63
Cheap	0.82	0.65	0.73	0.85	0.69	0.76	0.87	0.74	0.80	0.68	0.63	0.65	0.79	0.62	0.69
Average	0.75	0.68	0.70	0.80	0.75	0.76	0.83	0.78	0.79	0.60	0.64	0.61	0.66	0.61	0.63

concatenation of the two types of features. We use the 5-fold cross validation to evaluate the classifiers. We also evaluate the PRCA (Tontini et al., 2017) and CIT (Lu and Stepchenkova, 2012) methods on the data and compare their performance with the classifiers. The results are shown in Table 6.

Firstly, the results show that the average precision of the 3 classifiers is 79.3% and the average recall is 73.6%. This can be interpreted as that around 80% of the attractive attributes classified by the classifiers are actually attractive in the ground truth and the classifiers can retrieve more than 70% of the attractive attributes in the ground truth. Among the 3 classifiers, the one trained with the concatenation features is the best, followed by the one trained with the interactive effect features. The classifier trained with the empirical effect features is the worst. It is noteworthy that the concatenation features and the interactive effect features have more clear advantages over the empirical effect features on the luxury category. One possible reason is that the favourable reviews dominantly outnumber the unfavourable reviews in this category, therefore, information concerning negative opinions is severely inadequate. The empirical effect features are more sensitive to such lack of information.

Secondly, the performance of the classifiers is over 10% higher than that of the statistical models in terms of the F1 score. According to the analysis in Section 1, the low performance of the statistical models is partly caused by the noisy nature of the derived attribute-level sentiments and their ineffectiveness in modelling the non-linearity of the data.

In the experiment, the average rating of the hotels on which attractive attributes are detected is higher than that of the hotels on which no attractive attribute is detected across the 4 hotel categories. The averag rating of the hotels with attractive attributes is 0.27 higher than that of the hotels without the attributes in the cheap category, 0.31 higher in the budget category, 0.42 higher in the the medium category and 0.21 higher in the luxury category. The highest difference margin occurs in the medium category. This may indicate that there exists very intense competition in this category and the clients of those hotels are more sensitive to the attractive attributes than the clients of other hotel categories.

6. Conclusion

This paper aims to develop predictive classifiers based on machine learning techniques that analyse on-line review data to identify the attractive attributes of a product or service. A critical problem in training such classifiers is deriving discriminative features that can reflect the non-linear relations between attribute-level opinions and overall ratings. Two types of features, the empirical effect feature that is the empirical probabilities of ratings conditioning on various attribute-level opinions, and the interactive effect feature that encodes the interactions among attribute-level opinions in deciding the overall ratings, are used together to train the classifiers. Compared with the existing methods based on statistical models, the proposed classifiers not only encode a much higher degree of non-linearity in customer opinions, but also make use of the global statistics of the training data. The proposed classifiers are evaluated on a hotel review dataset crawled from TripAdvisor. The experiment results indicate that the classifiers reach a precision of 79.3% and outperform the existing statistical models by a margin of over 10%.

One drawback with the classifiers is that the information encoded by the empirical effect feature and the interactive effect feature may overlap to some degree. The information redundancy can result in severe overfitting. In the future, we will improve the attribute representations and investigate the possibilities of applying deep learning techniques to build more reliable classifiers.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org.https://www.tensorflow.org/.

Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, pp. 2546–2554.

Bespalov, D., Bai, B., Qi, Y., Shokoufandeh, A., 2011. Sentiment classification based on supervised latent n-gram analysis. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, pp. 375–382.

Brandt, R.D., 1987. A Procedure for Identifying Value-Enhancing Service Components Using Customer Satisfaction Survey Data, Add Value to Your Service. American Marketing Association, Chicago, pp. 61–65.

Bremner, F.J., Gotts, S.J., Denham, D.L., 1994. Hinton diagrams: Viewing connection strengths in neural networks. Behav. Res. Methods Instr. Comput. 26 (2), 215–218. https://doi.org/10.3758/BF03204624.

Brychcn, T., Konkol, M., Steinberger, J., 2014. UWB: Machine learning approach to aspect-based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp. 817–822. URL:http://www.aclweb.org/anthology/S14-2145.

Chen, L.-F., 2012. A novel approach to regression analysis for the classification of quality attributes in the Kano model: an empirical test in the food and beverage industry. Omega 40 (5), 651–659.

De Boom, C., Van Canneyt, S., Demeester, T., Dhoedt, B., 2016. Representation learning for very short texts using weighted word embedding aggregation. Pattern Recogn. Lett. 80, 150–156.

Gremler, D.D., 2004. The critical incident technique in service research. J. Service Res. 7 (1), 65–89. https://doi.org/10.1177/1094670504266138.

Guo, H., Zhu, H., Guo, Z., Zhang, X., Su, Z., 2009. Product feature categorization with multilevel latent semantic association. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, pp. 1087–1096.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the Ttenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 168–177.

Jo, Y., Oh, A.H., 2011. Aspect and sentiment unification model for online review analysis. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, pp. 815–824.

Kano, N., 1984. Attractive quality and must-be quality. J. Japanese Soc. Quality Control 14, 39–48.

Kenter, T., Borisov, A., de Rijke, M., 2016. Siamese CBOW: optimizing word embeddings for sentence representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, pp. 941–951. URL:http://www.aclweb.org/anthology/P16-1089.

Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S., 2014. NRC-Canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 437–442.

LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., et al., 1990. Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems, pp. 396–404.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., et al., 2010a. Structure-aware review mining and summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 653–661.

Li, F., Huang, M., Zhu, X., 2010b. Sentiment analysis with global topics and local dependency. In: AAI, pp. 1371–1376 vol. 10.

Lilleberg, J., Zhu, Y., Zhang, Y., 2015. Support vector machines and word2vec for text classification with semantic features. In: Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on. IEEE, pp. 136–140.

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

Long, C., Zhang, J., Zhut, X., 2010. A review selection approach for accurate feature rating estimation. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, pp. 766–774.

Lu, W., Stepchenkova, S., 2012. Ecotourism experiences reported online: classification of satisfaction attributes. Tourism Manage. 33 (3), 702–712. https://doi.org/10.1016/j.tourman.2011.08.003.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations), pp. 55–60.

Marcheggiani, D., Tckstrm, O., Esuli, A., Sebastiani, F., 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: ECIR. Springer, pp. 273–285.

Martineau, J., Finin, T., 2009. Delta TFIDF: an improved feature space for sentiment analysis. ICWSM 9, 106.

Matzler, K., Bailom, F., Hinterhuber, H.H., Renzl, B., Pichler, J., 2004. The asymmetric relationship between attribute-level performance and overall customer satisfaction: a reconsideration of the importance performance analysis. Ind. Mark. Manage. 33 (4), 271–277.

Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C., 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th International Conference on World Wide Web. ACM, pp. 171–180.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikuli, J., Prebeac, D., 2011. A critical review of techniques for classifying quality attributes in the Kano model. Manag. Service Qual.: Int. J. 21 (1), 46–66. https://doi.org/10.1108/09604521111100243.

Mukherjee, A., Liu, B., 2012. Aspect extraction through semi-supervised modeling. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, pp. 339–348.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, pp. 79–86.

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., et al., 2016. SemEval-2016 task 5: aspect based sentiment analysis. In: ProWorkshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 19–30.

Tan, K.C., Shen, X.-X., 2000. Integrating Kano's model in the planning matrix of quality function deployment. Total Qual. Manag. 11 (8), 1141–1151.

Tontini, G., Silveira, A., 2007. Identification of satisfaction attributes using competitive analysis of the improvement gap. Int. J. Oper. Prod. Manage. 27 (5), 482–500.

Tontini, G., Bento, G.d.S., Milbratz, T.C., Volles, B.K., Ferrari, D., 2017. Exploring the nonlinear impact of critical incidents on customers general evaluation of hospitality services. Int. J. Hospitality Manage. 66 (Suppl. C), 106–116. https://doi.org/10.1016/j.ijhm.2017.07.011.

Toprak, C., Jakob, N., Gurevych, I., 2010. Sentence and expression level annotation of opinions in user-generated discourse. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 575–584.

Wang, H., Lu, Y., Zhai, C., 2010. Latent aspect rating analysis on review text data: a rating regression approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 783–792.

Xu, X., Li, Y., 2016. The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: a text mining approach. Int. J. Hospitality Manage. 55 (Suppl. C), 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003.

Yu, J., Zha, Z.-J., Wang, M., Chua, T.-S., 2011. Aspect ranking: Identifying important product aspects from online consumer reviews. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 1496–1505.

Zhai, Z., Liu, B., Xu, H., Jia, P., 2010. Grouping product features using semi-supervised learning with soft-constraints. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 1272–1280.

Zhang, Y., Cole, S.T., 2016. Dimensions of lodging guest satisfaction among guests with mobility challenges: a mixed-method analysis of web-based texts. Tourism Manage. 53, 13–27. https://doi.org/10.1016/j.tourman.2015.09.001.

Zhuang, L., Jing, F., Zhu, X.-Y., 2006. Movie review mining and summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. ACM, pp. 43–50.