Expert Systems with Applications 36 (2009) 6527–6535

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Sentiment classification of online reviews to travel destinations by supervised machine learning approaches

Qiang Ye [a,b,*], Ziqiong Zhang [a], Rob Law [b]

[a] School of Management, Harbin Institute of Technology, Harbin, China
[b] School of Hotel and Tourism Management, Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Keywords:
Sentiment classification
Online reviews
Travel destinations
Supervised machine learning algorithm

ABSTRACT

The rapid growth in Internet applications in tourism has lead to an enormous amount of personal reviews for travel-related information on the Web. These reviews can appear in different forms like BBS, blogs, Wiki or forum websites. More importantly, the information in these reviews is valuable to both travelers and practitioners for various understanding and planning processes. An intrinsic problem of the overwhelming information on the Internet, however, is information overloading as users are simply unable to read all the available information. Query functions in search engines like Yahoo and Google can help users find some of the reviews that they needed about specific destinations. The returned pages from these search engines are still beyond the visual capacity of humans. In this research, sentiment classification techniques were incorporated into the domain of mining reviews from travel blogs. Specifically, we compared three supervised machine learning algorithms of Naïve Bayes, SVM and the character based N-gram model for sentiment classification of the reviews on travel blogs for seven popular travel destinations in the US and Europe. Empirical findings indicated that the SVM and N-gram approaches outperformed the Naïve Bayes approach, and that when training datasets had a large number of reviews, all three approaches reached accuracies of at least 80%.

1. Introduction

The importance of word-of-mouth has been widely documented in the existing literature (Anderson, 1998; Goldenberg, Libai, & Muller, 2001; Zhu & Zhang, 2006; Stokes & Lomax, 2002). Although its definition could be slightly varied, the term word-of-mouth generally refers to the personal communications between individuals concerning the perception of goods and services. Anderson (1998) stated that unsatisfied customers do engage in great word-of-mouth than satisfied ones. However, Stokes and Lomax (2002) argued that improved word-of-mouth communications can be an effective marketing strategy for small hospitality businesses. Similarly, Goldenberg et al. (2001) claimed that consumers' decision making process is strongly influenced by word-of-mouth. As such, word-of-mouth communications at the micro-level can influence macro-level phenomena. Recently, the unprecedented growth of Internet applications to travel and tourism has generated a plethora of new opportunities and challenges to consumers and practitioners. Among these new opportunities, the Internet provides a virtual environment for consumers to share their experiences with world-wide travelers via the electronic word-of-mouth communication channel (Pan, MacLaurin, & Crotts, 2007). Zhu and Zhang (2006) as well as Cheung, Shek, and Sia (2004) examined people's behavior for contributing in the virtual community, and stated that such a behavior would be beneficial to consumers and managers. Likewise, Dellarocas (2003) claimed that online word-of-mouth can have important implications for managers to consider their brand building, product development, and quality assurance. In brief, the Internet can serve as a useful platform for personal communications on sharing information of ownership, particular goods and services, and the suppliers and the place of supply.

Travel blogs are a form of digital story telling or word-of-mouth, which are self-published to disseminate travel narratives and adventures (Pan et al., 2007; Pudliner, 2007). In recent years, the wide application of the Internet to tourism has resulted in the availability of a huge amount of travel blogs for publicizing travel-related information. For instance, an attempt of using searching for "Travel Blog" on Google can often return 300 million pages. According to DoubleClick Inc. (2005), travelers generally tend to conduct an online search about their preferred destinations prior to making travel decisions (DoubleClick, 2005). When viewing travel blogs, the electronic word-of-mouth effect exerts

* Corresponding author. Address: School of Management, Harbin Institute of Technology, Harbin, China. Tel.: +86 451 86414022.

E-mail addresses: hmye@inet.polyu.edu.hk (Q. Ye), ziqiong@hit.edu.cn (Z. Zhang), hmroblaw@polyu.edu.hk (R. Law).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.07.035

a strong force that can influence the final decision of customers as well as tourism managers. Consumers can read, and subsequently use, the reviews as references to determine whether a place is their preferred destination. Although the information available on travel blogs could be of use, it is basically not possible for anyone to manually read the thousands, if not millions, of travel blogs. As such, sophisticated sentiment classification techniques that can automatically classify, on the basis of the analyzed travel blogs, whether the overall reviews of a specific destination are positive or negative would certainly be useful to users. Sentiment classification is a class of recently developed web mining techniques that can perform analysis on sentiment or opinions (Liu, Hu, & Cheng, 2005; Morinaga, Yamanishi, Tateishi, & Fukushima, 2002; Pang, Lee, & Vaithyanathan, 2002; Turney, 2002).

Generally speaking, sentiment classification aims at mining text of written reviews from customers for certain products or services, and classifying the reviews into positive or negative opinions. The classification method has been applied to the computing fields of information retrieval and natural language processing (Beineke, Hastie, & Vaithyanathan, 2004; Godbole, Srinivasaiah, & Skiena, 2007; Pang et al., 2002; Turney & Littman, 2003).

Special challenges are associated with mining on tourist reviews. In this domain specific area, word semantics in a particular review could contradict with the overall semantic direction (good or bad) of that review. For instance, an ''unpredictable'' camera implies negative meaning to that camera; whereas a tour with an ''unpredictable'' experience is positive to explorers.

Although some recent studies have begin to conduct content analysis of travel blogs (Choi, Lehto, & Morrison, 2007; Pan et al., 2007), sophisticated web mining technique still need to be integrated into travel blog analysis. To fill in this void in the existing tourism literature, this study makes an attempt to perform automatic classifications based on the sentiment attitudes of online reviews with regards to travel destinations. Additionally, this study compares different supervised machine learning algorithms and their effect on the different amount of training corpus to various performance measurements in terms of accuracy, precision, and recall in the sentiment classification of online reviews about tourist destinations. In this study, three supervised machine learning algorithms, including Naïve Bayes, support vector machine, and the character based N-gram model were incorporated into sentiment classification for the reviews about seven popular travel destinations in Europe and North America. These destinations included New York, Los Angeles, Las Vegas, London, Rome, Paris, and Venice.

2. Background

Travelers generally like to search for information of their preferred destinations (Lo, Cheung, & Law, 2002). In a recent survey that was conduced by DoubleClick (2005), more than half of consumers did an online search before making their online purchases. Among the major categories of products and services, travel had the highest percentage of pre-purchase online searching where 73% of travelers conducted online searching before making their travel decision. In addition, prior studies have demonstrated that opinions in online product reviews could have a significant influence on consumers' purchasing decision (Godes et al., 2005). Zhu and Zhang (2006) investigated the influence of online consumer reviews on the demand for experience goods, and made a similar conclusion. Since travel service is typically experience related (Klein, 1998), opinions in online reviews should have a strong influence to travelers.

As stated, there is presently a huge amount of online reviews on travel blogs which is beyond the visual capacity of any human beings. Hence, there is an urgent need for innovative techniques that can automatically analyze the attitudes of customers in their reviews. As such, sentiment classification (sentiment analysis or opinion mining) can perform the tasks of automatically understanding the online reviews (Liu et al., 2005; Pang et al., 2002; Turney, 2002). Mining opinions from reviews on web pages, however, is a complex process, which requires more than just text mining techniques. The complexity is related to a couple of issues. First, data of reviews are to be crawled from websites, in which web spiders or search engines can play an important role. Moreover, it is necessary to separate the data of reviews from non-reviews. The sentiment classification process can then be conducted. Pang et al. (2002) found text mining algorithms on sentiment classification do not perform as well as that on traditional topic-based categorization. Topics can be identified by keywords but sentiment would be expressed in a more subtle manner. As such, sentiment classification requires more understanding than the usual topic-based classification (Pang et al., 2002).

Sentiment classification aims to extract the text of written reviews of customers for certain products or services by classifying the reviews into positive or negative opinions according to the polarity of the review (Dave, Lawrence, & Pennock, 2003; Okanohara & Tsujii, 2005). With the results of sentiment classification, consumers would know the necessary information to determine which products to purchase and sellers would know the response from their customers and the performances of their competitors. With the wide adoption of computing technology, sentiment classification of reviews has become one of the foci of recent research endeavors. The method has been attempted in different domains such as movie reviews, product reviews, customer feedback reviews, and legal blogs (Beineke et al., 2004; Conrad & Schilder, 2007; Liu et al., 2005; Pang et al., 2002). Other potential applications include extracting opinions or reviews from discussion forums such as blogs, and integrating automatic review mining with search engines to automatically provide useful statistical data of search results or to build sentiment analysis systems for specific products or services. Tourist destinations, naturally, would be one of the good application areas.

In relation to opinion mining applications, the extant literature indicates two types of techniques have been utilized, including machine learning and semantic orientation (Turney, 2002). The machine learning approach that is applicable to this problem mostly belongs to supervised classification in general, and text classification techniques in particular, for opinion mining. Thus, it is called ''supervised learning''. In contrast, using a semantic orientation approach to opinion mining is ''unsupervised learning'' because it does not require prior training in order to mine the data. Instead, it measures how far a word is inclined towards positive and negative. Chaovalit and Zhou (2005) compared the SO approach with the machine learning approach by applying the SO approaches to movie reviews and found the machine learning approach was more reliable than the unsupervised semantic orientation approach.

The most well-known[cite] machine learning methods in the natural language processing area is the support vector machine (SVM), Naïve Bayes, and the N-gram model. The support vector machine is a statistical classification method proposed by Vapnik (1995). This model can be used for both binary and multiple text category classifications. SVM can meet significant success in numerous real-world learning tasks. Joachims (1998) published the results on a set of binary text classification experiments using SVM, and showed that SVM generated lower error levels than other classification techniques. Besides, Yang and Liu (1999) compared SVM with Linear Least-squares Fit (LLSF), Neural Network (NNet), Naive Bayes (NB) and *k*-nearest neighbors (kNN), and found that SVM achieved an equal performance like other classifiers in their experiments. When movie reviews were analyzed, Pang et al.

(2002) applied SVM to conduct sentiment classification on customer reviews, and Chaovalit and Zhou (2005) applied the N-gram approach on the sentiment classification.

As tourists' review mining is very much domain specific, it would be necessary to test the performance of sentiment classification techniques on reviews about travel destinations. According to publications in data mining and classification, Naïve Bayes, SVM and character based N-gram are the three most important approaches in text mining and sentiment classification (Carpenter, 2005; Joachims, 1998; Lewis, 1998; McCallum & Nigam, 1998; Pang et al., 2002; Yang & Liu, 1999). This study therefore makes an initial attempt to apply supervised machine learning algorithms of Naïve Bayes, SVM and the character based N-gram model to the reviews of some of the world's most popular travel destinations.

3. Methodology

The basic mechanism of sentiment classification by supervised machine learning algorithms is depicted in Fig. 1.

In this research, we applied three supervised machine learning models for sentiment classification of reviews for the selected travel destinations. These models are namely Naïve Bayes classifier, SVM classifier, and dynamic language model classifier.

3.1. Naïve Bayes model for sentiment classification

Despite its simplicity, the Naive Bayes classifier is a popular machine learning technique for text classification, and it performs well in many domains (Domingos & Pazzani, 1997). During its operation, Naive Bayes assumes a stochastic model of document generation. Using Bayes' rule, the model is inverted in order to predict the most likely class for a new document.

In this research, we assume that documents are generated according to a multinomial event model (McCallum & Nigam, 1998). Hence, a document is represented as a vector $d_i = (x_{i1}, \cdots, x_{i|V|})$ of word counts where V is the size of the vocabulary (vol.) for all documents under an experiment, here vol $= \{w_1, \cdots, w_{|V|}\}$. Each $x_{it} \in \{0, 1.2, \cdots\}$ indicates how often w_t presents in a certain document D_i. Given model parameters $p(w_t|c_j)$ and class prior probabilities $p(c_j)$, and assuming an independence of the words, the most likely class for a document d_i is computed as

$$\begin{aligned} c^*(d_i) &= \operatorname*{argmax}_j p(c_j)p(d|c_j) \\ &= \arg\max_j p(c_j) \prod_{t=1}^{|V|} p(w_t|c_j)^{n(w_t,d_i)} \end{aligned} \tag{1}$$

where $n(w_t, d_i)$ is the number of occurrences of w_t in d_i. $p(w_t|c_j)$ and $p(c_j)$ are estimated from training documents with known category, using maximum likelihood estimation with a Laplacean prior:

$$p(w_t|c_j) = \frac{1 + \sum_{d_i \in c_j} n(w_t, d_i)}{|V| + \sum_{t=1}^{|V|} \sum_{d_i \in c_j} n(w_t, d_i)}$$

$$p(c_j) = \frac{|c_j|}{\sum_{r=1}^{|C|} |c_r|}$$

We developed a tool of classifier by *VC 6.0* according to the above functions to perform sentiment classification of online reviews about travel destinations. The information gain (IG) method was then applied as the feature selection technique in this study.

3.2. Support vector machines for sentiment classification

SVMs have been shown to be highly effective at traditional text categorization, which generally outperform Naive Bayes (Joachims, 1998). SVMs seek a hyperplane represented by vector $\vec{w}$ that separates the positive and negative training vectors of documents with maximum margin (Fig. 2).

Findings in this hyperplane can then be translated into a constrained optimization problem. Let y_i equal +1(−1), if document d_i is in class +(−). The solution can be written as

$$\vec{w} = \sum_{i=1}^{n} \alpha_i^* y_i \vec{d_i} \quad \alpha_i \geqslant 0 \tag{2}$$

where α_i^* are obtained by solving a dual optimization problem. Eq. (2) shows that the resulting weight vector of the hyperplane is constructed as a linear combination of $\vec{d_i}$. Only those examples that contribute to which the coefficient αi is greater than zero. Those vectors are called *support vectors*, since they are the only document vectors contributing to $\vec{w}$.

We applied SVM classifier with information gain (IG) as a feature selection method. In the experiments, we chose the word frequency to present a document rather than word presence for probability estimation.

3.3. N-gram based character language model for sentiment classification

The N-gram based character language model is a new model in natural language processing (Carpenter, 2005). It is derived from the N-gram language models. Instead of taking words as the basic unit, this model takes characters (letters, space, or symbols) as the basic unit in the algorithm.

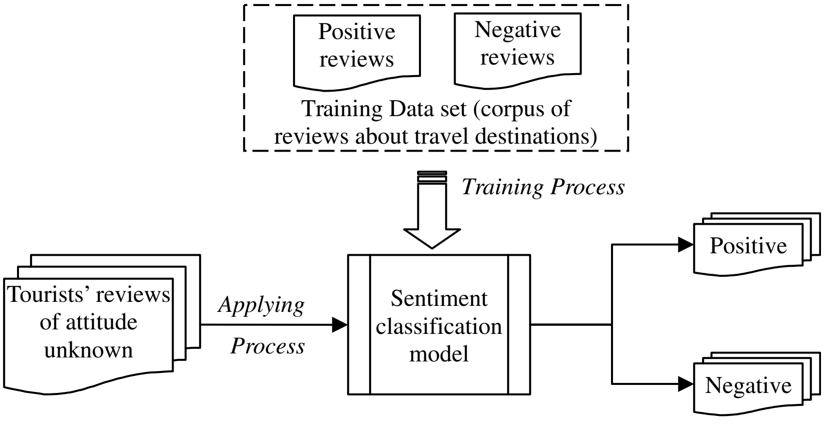
Positive reviews
Negative reviews
Training Data set (corpus of reviews about travel destinations)
Training Process
Tourists' reviews of attitude unknown
Applying Process
Sentiment classification model
Positive
Negative

Fig. 1. The mechanism of sentiment classification by supervised machine learning algorithms.

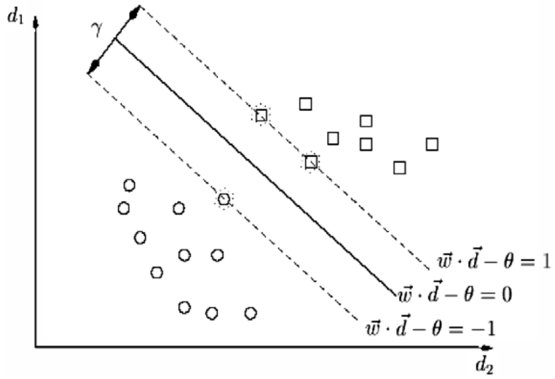
d_1
γ
$\vec{w} \cdot \vec{d} - \theta = 1$
$\vec{w} \cdot \vec{d} - \theta = 0$
$\vec{w} \cdot \vec{d} - \theta = -1$
d_2

Fig. 2. A maximum margin classifier of SVM.

The N-gram character language model provides a probability distribution $p(s)$ defined for strings $s \in \Sigma^*$ over a fixed alphabet of characters Σ. The chain rule factors $p(sc) = p(s) \cdot p(c|s)$ for a character c and string s. The N-gram Markovian assumption restricts the context to the previous n − 1 characters, taking

$$p(c_n|s_{c1} \cdots c_{n-1}) = p(c_n|c_1 \cdots cn - 1)$$

The maximum likelihood estimator for N-grams is thus

$$\hat{p}_{\text{ML}}(c|s) = \frac{C(sc)}{\sum_c C(sc)} \tag{3}$$

where $C(sc)$ is the number of times the sequence sc was observed in the training data and $\sum_c C(sc)$ is the number of single-character extensions of sc.

In this paper, we used the LingPipe DynamicLMClassifier for our data experiment (Alias-I, 2006). This classifier depends on an N-gram based character language model with a generalized form of Witten-Bell smoothing (Carpenter, 2005). The DynamicLMClassifier accepts training events of categorized character sequences. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators. In this research, we used the default value of $N = 8$ in the N-gram statistic model.

3.4. Corpus of reviews about travel destinations

To conduct the research, we established a data set by retrieving corpus from tourists' reviews in the travel column of Yahoo.com (http://www.travel.yahoo.com). We focused on the seven most popular travel destinations in the US and Europe, including New York, Los Angeles, Las Vegas, London, Rome, Paris, and Venice. The reviews about these travel destinations by tourists were downloaded and analyzed. On Yahoo's travel reviews, each review was labeled with a number of stars on scales of 1–5 by the reviewer. More stars could be interpreted as having stronger positive opinions of the destination. In this study, we concentrated on reviews with four or five stars as positive corpus and reviews with one or two stars were considered as negative corpus. The download was conducted in September 2007. After checking the reviews in multiple rounds, they yielded a corpus of 591 negative and 600 positive reviews to the cities as indicated in Table 1.

Table 1
Number of reviews for the seven destinations

ID	Region	Destinations	Number of reviews	Positive	Negative
1	US	New York	333	168	165
2		Los Angeles	67	35	32
3		Las Vegas	302	148	154
4	Europe	London	109	56	53
5		Rome	102	58	44
6		Paris	190	88	102
7		Venice	88	47	41
All destinations			1191	600	591

3.5. Experiments

We conducted K-fold-cross-validation (Kohavi, 1995) in the experiments. In this research, $K = 3$ was adopted. The 600 positive and 591 negative reviews were applied to make a 3-fold cross validation in the data experiments. The data were partitioned randomly into three folds, each including 200 positive and 197 negative documents. On each round of experiment, two folds were used a training data set, and the remaining fold was used as the testing data set. Since a jackknife approach was used, each review was in a test set once and in a training set twice.

As stated, an objective of this study was to examine how a classifier works with various sizes of training data set. It was, therefore, important to create small subsets from a given large training set. Let A and T, respectively, be the training and testing datasets of each round. We further split training data set A into 10 disjoint sets (A_1,A_2, A_3, ..., A_{10}), not necessarily of equal size, and then 10 new training sets (AA_1,AA_2,AA_3,...,AA_{10}) are constructed, where $AA_1 = A_1$, $AA_i = AA_{i-1} + A_i$ ($i = 2,\ldots,10$). The performance of a classifier could be assessed based on the results of 10 experiments conducted on 10 train-test pairs (AA_i,T).

We conducted each round of experiment by increasing the number of training examples with each experiment; this was repeated 3-fold. Table 2 shows the number of training examples of the categories in each round.

To prepare the documents, we converted all characters to lowercase in both training and testing sets. Then the data were fed to LingPipe 3.0.0. For SVMs and Naïve Bayes experiments, we treated punctuation as separate lexical items, and no stemming or stop-word lists were used.

3.6. Performance evaluations

To evaluate the performance of sentiment classification, we adopted three indexes that are generally used in text categorization: Recall, Precision, and Accuracy.

The indexes can be calculated according to the figures in Table 3 and the following formulas, respectively,

$$\text{Accuracy} = \frac{a+d}{a+b+c+d},$$
$$\text{Recall(pos)} = \frac{a}{a+c}, \quad \text{Precision(pos)} = \frac{a}{a+b},$$
$$\text{Recall(neg)} = \frac{d}{b+d}, \quad \text{Precision(neg)} = \frac{d}{c+d}$$

Table 2
Numbers of reviews in training data sets

Round of experiments	Corpus Numbers of reviews in each round of training		
	Positive	Negative	All
1	20	20	40
2	50	50	100
3	80	80	160
4	120	120	240
5	160	160	320
6	200	200	400
7	250	250	500
8	300	300	600
9	350	350	700
10	400	394	794

Table 3
Contingency table for performance evaluations

	Actual positive reviews	Actual negative reviews
Predict positive	*A*	*B*
Predict negative	*C*	*D*

Here, Recall(pos) and Precision(pos) are the recall ratio and precision ratio for actual positive reviews. Recall(neg) and Precision(neg) are the recall ratio and precision ratio for actual negative reviews. Accuracy is the overall Accuracy of certain sentiment classification models.

4. Findings

The overall accuracies of the three algorithms in 10 rounds of experiments are indicated in Table 4 and Fig. 3.

The result indicated that the SVM approach and N-gram approach had better accuracies than the Naïve Bayes approach. The difference in accuracies among the three approaches was very significant ($p < 0.01$) when the training data set had 100 or less reviews. For all three approaches, accuracies increased with more reviews in the training data sets. When training data sets had 500 or more reviews, all three approaches could reach accuracies of more than 80%.

The precisions for positive corpus in the testing data set were showed in Table 5 and Fig. 4.

Table 6 and Fig. 5 indicated the precisions for negative corpus in the testing data set.

Table 7 and Fig. 6 showed the result of recalls for positive corpus in the testing data set in the three approaches.

Table 8 and Fig. 7 indicated the result of recalls for negative corpus in the testing data set in the three approaches.

5. Implications and conclusions

This study has applied three supervised machine learning algorithms of Naïve Bayes, SVM and the character based N-gram model to the online reviews about seven popular travel destinations in the world. Different to the few previous studies, we found that well trained machine learning algorithms can perform very good classifications to the sentiment polarities of reviews about travel destinations. In terms of accuracy, all three algorithms can reach more than 80% of classification correctly.

Generally speaking, the SVM model and character based N-gram model had achieved better performance than the Naïve Bayes method. When the training data set was as small as 40 or 100 reviews, the difference among the algorithms was extremely significant. However, with the increasing of the size in the training data set, the difference becomes not so significant. A conclusion thus drawn is that a larger training data set with 300 to 800 reviews will

Table 4
Accuracies in testing data set

Round of experiments	Numbers of reviews in training dataset	Numbers of Reviews in Testing dataset (positive/negative)	Accuracy				
			Na Bayes model (%)	SVM classifier (%)	Character based N-gram model (%)	χ^2	*P*
1	40	200/197	51.39	64.57	66.16	22.1517	0.0000**
2	100	200/197	55.33	73.97	72.71	39.1760	0.0000**
3	160	200/197	72.54	75.99	76.15	1.5146	0.4689
4	240	200/197	74.98	80.94	78.34	4.1445	0.1259
1	320	200/197	78.67	82.54	81.78	2.1613	0.3394
6	400	200/197	79.93	82.20	83.38	1.6356	0.4414
7	500	200/197	80.18	84.63	83.46	2.9593	0.2277
8	600	200/197	80.44	85.31	83.71	3.4933	0.1744
9	700	200/197	80.27	86.06	83.29	4.7712	0.0920
10	794	200/197	80.71	85.14	84.05	1.5312	0.4651

** Significant at a 0.01 level.

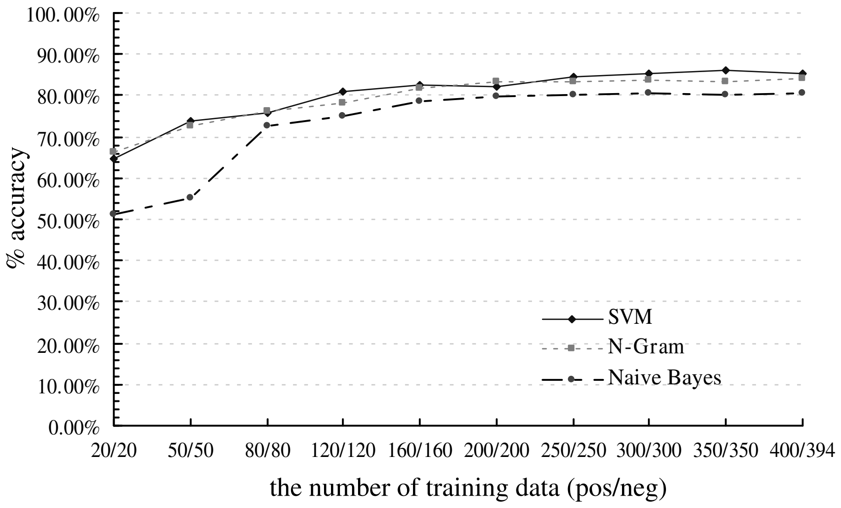

Fig. 3. Diagrammatic presentation of accuracies in the experiments.

Table 5
Precisions for positive corpus in testing data set

Round of experiments	Numbers of reviews in training dataset	Numbers of reviews in testing dataset (positive/negative)	Precision on positive corpus				
			Na Bayes model (%)	SVM classifier (%)	character based N-gram model (%)	χ^2	*P*
1	40	200/197	50.91	63.86	66.12	16.1132	0.0003**
2	100	200/197	53.57	75.35	69.61	28.8996	0.0000**
3	160	200/197	69.87	75.74	73.94	2.0124	0.3656
4	240	200/197	72.74	81.24	76.47	4.2374	0.1202
1	320	200/197	78.74	81.51	80.16	0.4931	0.7815
6	400	200/197	80.54	81.60	82.21	0.1904	0.9092
7	500	200/197	81.49	83.79	82.87	0.3736	0.8296
8	600	200/197	81.26	84.67	82.64	0.8292	0.6606
9	700	200/197	80.57	85.23	82.18	1.5852	0.4527
10	794	200/197	83.71	85.07	82.23	0.6101	0.7371

** Significant at a 0.01 level.

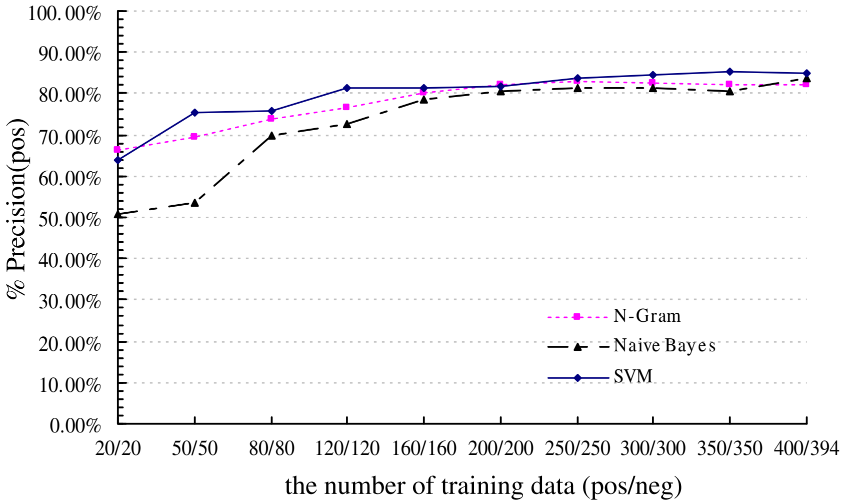

Fig. 4. Diagrammatic presentation of precisions for positive corpus.

Table 6
Precisions for negative corpus in testing data set

Round of experiments	Numbers of reviews in training dataset	Numbers of reviews in testing dataset (positive/negative)	Precision on negative corpus				
			Na Bayes model (%)	SVM classifier (%)	Character based N-gram model (%)	χ^2	*P*
1	40	200/197	56.18	65.39	66.21	1.1502	0.5626
2	100	200/197	62.34	72.70	77.14	5.8891	0.0526*
3	160	200/197	76.19	76.25	78.91	0.4860	0.7843
4	240	200/197	77.80	80.64	80.55	0.5819	0.7475
1	320	200/197	78.61	83.66	83.63	2.2066	0.3318
6	400	200/197	79.33	82.84	84.66	1.9586	0.3756
7	500	200/197	78.96	85.54	84.08	3.3476	0.1875
8	600	200/197	79.64	85.99	84.89	3.2993	0.1921
9	700	200/197	79.97	86.96	84.51	3.6132	0.1642
10	794	200/197	80.32	85.20	86.13	2.8595	0.2394

* Significant at a 0.1 level.

perform better in sentiment classifications for all three algorithms for the reviews about travel destinations.

This research has examined the methods to automatically analyze the opinions of reviews on travel blogs about travel destinations. Specifically, this study has demonstrated the promising attempt of incorporating sentiment classification into travel blog analysis. In the context of travel and tourism, recent studies have confirmed that the image of a destination can be directly affected by travel-related websites (Choi et al., 2007; Govers & Go, 2005; Pudliner, 2007). In addition, Choi et al. (2007) advocated that travel blogs will become a popular source for destination information. Pan et al. (2007) further argued that the feedback that is available on travel blogs can be richer in content and more detailed than Likert-scale based questionnaire surveys. Findings of this research would, therefore, make a meaningful contribution to understand travelers' perception of travel-related products and services.

Findings of this study are likely to lead to a new trend of information processing in the tourism industry. This could influence the design of information systems in travel and tourism in different functional areas. As well, empirical results of this research can

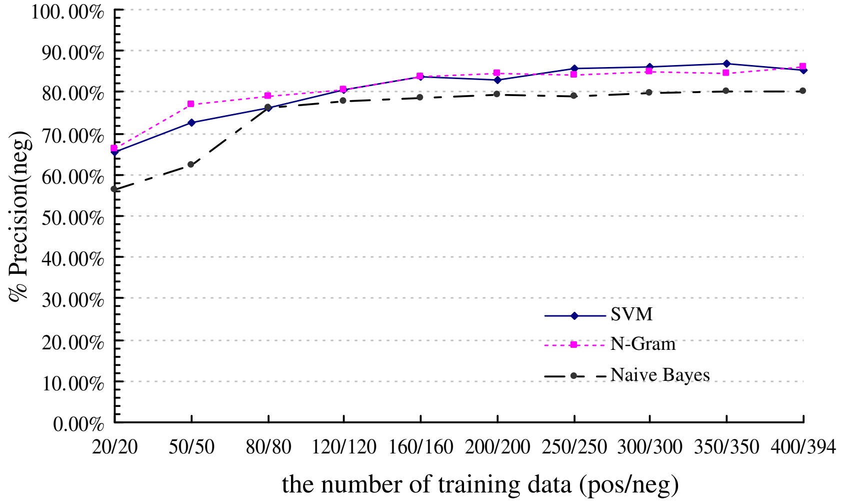

Fig. 5. Diagrammatic presentation of precisions for negative corpus.

Table 7
Recall for positive corpus in testing data set

Round of experiments	Numbers of reviews in training dataset	Numbers of reviews in testing dataset (positive/negative)	Recall on positive corpus				
			Na Bayes model (%)	SVM classifier (%)	Character based N-gram model (%)	χ^2	*P*
1	40	200/197	93.50	68.33	67.33	48.7557	0.0000**
2	100	200/197	85.00	71.83	81.33	11.2879	0.0035**
3	160	200/197	80.00	77.37	81.33	1.2066	0.5470
4	240	200/197	80.50	81.24	82.33	0.2501	0.8824
1	320	200/197	79.00	84.50	84.83	3.0110	0.2219
6	400	200/197	79.33	83.50	85.50	2.7772	0.2494
7	500	200/197	78.50	86.17	84.67	4.7047	0.0951
8	600	200/197	79.50	86.50	85.67	4.3270	0.1149
9	700	200/197	80.17	87.50	85.33	4.2974	0.1166
10	794	200/197	79.67	85.50	87.17	4.6420	0.0982

** Significant at a 0.01 level.

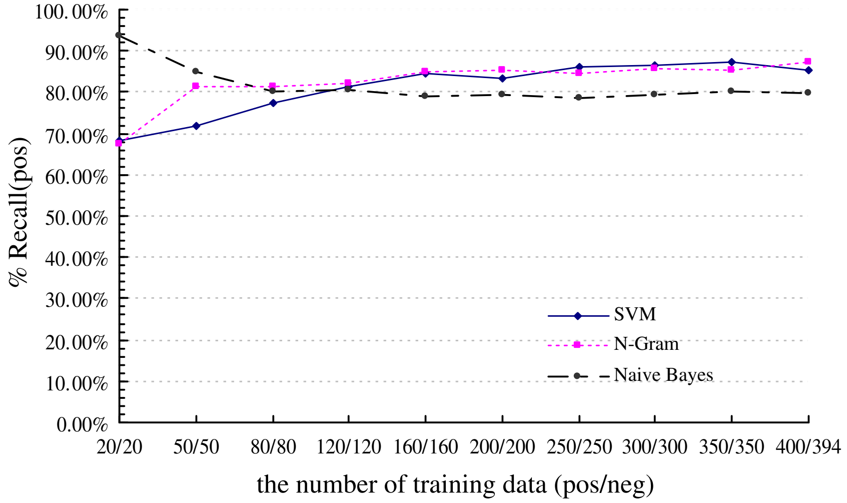

Fig. 6. Diagrammatic presentation of recalls for positive corpus.

provide insights to study how search engines can process information of reviews about travel destinations.

At present, tourists are largely Internet users. They like to conduct an online search about the information of their preferred destinations prior to making any travel decisions. Thus, the opinions in the reviews of other tourists often have a strong influence on their final decisions. This research has investigated the methods that can conduct automatic analysis of the sentiment attitude of

Table 8
Recall for negative corpus in testing data set

Round of experiments	Numbers of reviews in training dataset	Numbers of reviews in testing dataset (positive/negative)	Recall on negative corpus				
			Na Bayes model (%)	SVM classifier (%)	Character based N-gram model (%)	χ^2	*P*
1	40	200/197	8.51	60.74	64.97	157.9230	0.0000**
2	100	200/197	25.21	76.14	63.96	112.6430	0.0000**
3	160	200/197	64.97	74.55	70.90	4.7540	0.0928
4	240	200/197	69.37	80.63	74.28	7.2025	0.0273
1	320	200/197	78.34	80.54	78.68	0.3354	0.8456
6	400	200/197	80.54	80.88	81.22	0.0292	0.9855
7	500	200/197	81.90	83.08	82.23	0.1011	0.9507
8	600	200/197	81.39	84.09	81.73	0.5912	0.7441
9	700	200/197	80.37	84.60	81.22	1.3412	0.5114
10	794	200/197	84.26	84.77	80.88	1.2680	0.5305

** Significant at a 0.01 level.

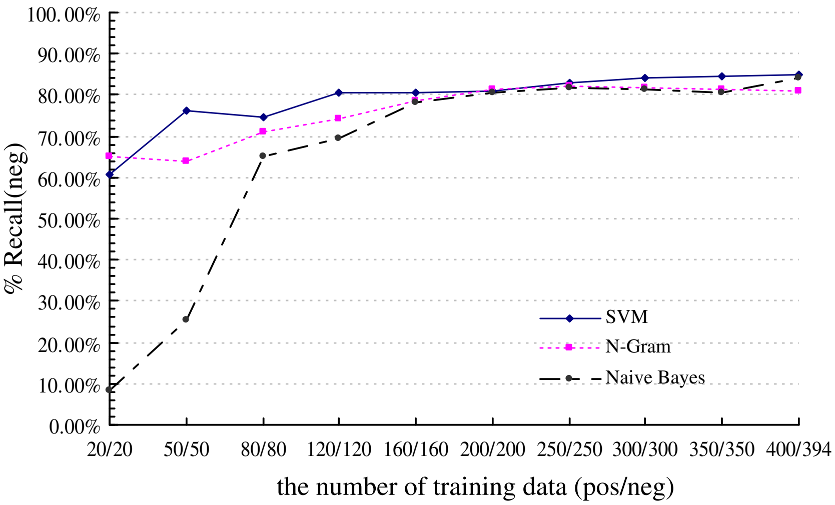

Fig. 7. Diagrammatic presentation of recalls for negative corpus.

the reviews on travel destinations. The research is expected to help both potential visitors and the tourism industry to extract the valuable values from these reviews efficiently. Some potential applications include extracting opinions or reviews from travel forums efficiently, and integrating automatic review mining with search engines to provide useful information of search results for the opinions about certain travel destinations.

A natural extension of this research is to expand the number of destinations. Although this study has analyzed the data for popular destinations in Western countries, the applicability of the presented classification methods in other destinations remain unknown and is thus worthwhile for future investigation. Another direction for future research is the examination of readership for the travel blogs. Lastly, since consumers change their perception frequently it would be interesting to do a longitudinal study to compare and contrast findings between different time periods.

Acknowledgements

This study was partially supported by Research Funding of Hong Kong Polytechnic University (G-YX93), Internet Research Center of China and NSFC grant (70771032, 70501009).

References

Alias-I (2006). LingPipe natural language toolkit. <http://www.alias_i.com/lingpipe> Accessed 02.10.07.

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research, 1*(1), 5–17.

Beineke, P., Hastie, T., & Vaithyanathan, S. (2004). The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL) 2004* (pp. 263–270). Association for Computational Linguistics.

Carpenter, B. (2005). Scaling high-order character language models to gigabytes. In *Proceedings of the 2005 association for computational linguistics software workshop* (pp. 1–14). Ann Arbor: ACL.

Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In R. R. Sprague (Ed.), *Proceedings of the 38th Hawaii international conference on system sciences, Big Island Hawaii* (pp. 1–9). IEEE.

Cheung, C. M. Y., Shek, S. P. W., & Sia, C. L. (2004). Virtual community of consumers: Why people are willing to contribute? In *Proceedings of the 8th Pacific-Asia conference on information systems* (pp. 2100–2107).

Choi, S., Lehto, X. Y., & Morrison, A. M. (2007). Destination image representation on the web: Content analysis of Macau travel related websites. *Tourism Management, 28*, 118–129.

Conrad, J. G., & Schilder, F. (2007). Opinion mining in legal blogs. In A. Gardner (Ed.), *Proceedings of the 11th international conference on Artificial intelligence and law, Stanford, California* (pp. 231–236). ACM Press.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In G. Hencsey & B. White (Eds.), *Proceeding of 12th international conference on World Wide Web* (pp. 519–528). Budapest, Hungary: ACM Press.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424.

Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning, 29*(2-3), 103–130.

DoubleClick Inc. (2005). Search before the purchase: Understanding buyer search activity as it builds to online purchase. <http://www.doubleclick.com> Accessed 01.11.06.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). LargeScale sentiment analysis for news and blogs. In N. Glance & N. Nicolov (Eds.), *International conference on weblogs and social media (ICWSM'2007) Boulder, Colorado, USA*. ICWSM.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., et al. (2005). The firm's management of social interactions. *Marketing Letters, 16*(3), 415–428.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters, 12*(3), 211–223.

Govers, R., & Go, F. (2005). Projected destination image online: Website content analysis of pictures and text. *Information Technology and Tourism, 7*, 73–89.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of 10th European Conference on Machine Learning (ECML-98), Chemnitz, Germany, 1998* (pp. 137–142). Springer.

Klein, L. R. (1998). Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of Business Research, 41*(3), 195–203.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 2*(12), 1137–1143.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of 10th European conference on machine learning (ECML-98), Chemnitz, Germany, 1998* (pp. 4–15). Springer.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In A. Ellis & T. Hagino (Eds.), *Proceedings of the 14th international World Wide Web conference (WWW-2005), Chiba, Japan* (pp. 10–14). ACM Press.

Lo, A., Cheung, C., & Law, R. (2002). Information search behavior of Hong Kong's in-bound travelers – A comparison of business and leisure travelers. *Journal of Travel and Tourism Marketing, 13*(3), 61–81.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization* (pp. 41–48). AAAI Press.

Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). Mining product reputations on the web. In O. R. Zaiane (Ed.), *Proceeding of K.D.D. Edmonton, Alberta* (pp. 1–8). ACM Press.

Okanohara, D., & Tsujii, J. (2005). Assigning polarity scores to reviews using machine learning techniques. *Lecture Notes in Computer Science (LNAI), 3651*, 314–325.

Pan, B., MacLaurin, T., & Crotts, J. (2007). Travel blogs and the implications for destination marketing. *Journal of Travel Research, 46*, 35–45.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In P. Isabelle (Ed.), *Proceeding of 2002 conference on empirical methods in natural language, Philadelphia, US* (pp. 79–86). Association for Computational Linguistics.

Pudliner, B. A. (2007). Alternative literature and tourist experience: Travel and tourist weblogs. *Journal of Tourism and Cultural Change, 5*(1), 46–58.

Stokes, D., & Lomax, W. (2002). Taking control of word of mouth marketing: The case of an entrepreneurial hotelier. *Journal of Small Business and Enterprise Development, 9*(4), 349–357.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In P. Isabelle (Ed.), *Proceeding of association for computational linguistics 40th anniversary meeting, Philadelphia, PA, USA* (pp. 417–424). ACL.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315–346.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Yang, Y. M., & Liu, X. (1999). A re-examination of text categorization methods. In F. Gey, M. Hearst, & R. Tong (Eds.), *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, California, United States* (pp. 42–49). ACM Press.

Zhu, F., & Zhang, X. (2006). The influence of online consumer reviews on the demand for experience goods: The case of video games. In *Twenty-seventh international conference on information systems (ICIS), Milwaukee, 2006*. AIS Press.