Expert Systems with Applications 38 (2011) 7674–7682

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Sentiment classification of Internet restaurant reviews written in Cantonese

Ziqiong Zhang *, Qiang Ye, Zili Zhang, Yijun Li

*Dept. of Management Science & Engineering, Harbin Institute of Technology, Harbin 150001, China*

ARTICLE INFO

*Keywords:*
Sentiment classification
Online review
Cantonese
Restaurant
Machine learning

ABSTRACT

Cantonese is an important dialect in some regions of Southern China. Local online users often represent their opinions and experiences on the web with written Cantonese. Although the information in those reviews is valuable to potential consumers and sellers, the huge amount of web reviews make it difficult to give an unbiased evaluation to a product and the Cantonese reviews are unintelligible for Mandarin Chinese speakers.

In this paper, standard machine learning techniques naive Bayes and SVM are incorporated into the domain of online Cantonese-written restaurant reviews to automatically classify user reviews as positive or negative. The effects of feature presentations and feature sizes on classification performance are discussed. We find that accuracy is influenced by interaction between the classification models and the feature options. The naive Bayes classifier achieves as well as or better accuracy than SVM. Character-based bigrams are proved better features than unigrams and trigrams in capturing Cantonese sentiment orientation.

## 1. Introduction

The Internet continues to become an essential part of everyday life. people are now able to access not only opinions from family members and friends, but also from strangers located around the world who may have used a particular product, visited a certain destination, or seen a movie. Internet provides a virtual environment for consumers to share their experiences with world-wide travelers via the electronic word-of-mouth (WOM) communication channel (Cheung, Shek, & Sia, 2004). The importance of WOM has been widely documented in the existing literature (Cheung et al., 2004; Goldenberg, Libai, & Muller, 2001). WOM not only strongly influences consumers' decision making process (Goldenberg et al., 2001), but also has important implications for managers to consider their brand building, product development, and quality assurance (Dellarocas, 2003).

As today's consumers are increasingly making their opinions and experiences available online (Horrigan, 2008), there have accumulated a huge amount of consumer reviews for products or service on the Web. When trying to locate user opinions of a product, a general online search will turns up millions of web pages. Getting an overall sense of those reviews can be daunting or time-consuming, however, if only few reviews were read the evaluation would be biased. Sentiment classification aims to address this problem by automatically classifying user reviews into positive or negative opinions.

Review sentiment classification has become one of the foci of recent research endeavors. Many sentiment classification techniques have been developed for English, Japanese, and Mandarin Chinese. But the interest in the sentiment analysis is worldwide to provide support for various NLP applications. Researches on automatic sentiment analysis should be conducted in more new languages such as the Cantonese.

Cantonese is an important dialect spoken in and around the cities of southern China where are typical areas with rapid development in China. In those areas, Cantonese is widely used in social settings and many native Cantonese consumers are not well literate in Mandarin Chinese. Take Hong Kong for example. According to statistics of Hong Kong Census and Statistics Department for 2006 population, Cantonese was the most commonly used language at home for about 91% of the population. Only about 40% of the population claimed to be able to speak Mandarin Chinese,[1] and the percent capable of writing would be less. Those Cantonese-speaking consumers are very likely to express themselves with written Cantonese in informal settings such as Internet forum; however, due to the difference between Cantonese and Mandarin Chinese, Mandarin speakers cannot read the online Cantonese contents (or finds it so difficult that the effort will rapidly be abandoned). Given the importance of written Cantonese (Snow, 2004), innovative techniques that can automatically detect the consumer opinions in Cantonese reviews are urgently required.

* Corresponding author.
*E-mail address:* xiaojia0459@yahoo.com.cn (Z. Zhang).

[1] Hong Kong Census and Statistics Department, http://sc.info.gov.hk/gb/www.censtatd.gov.hk/press_release/other_press_releases/index.jsp?sID=1860&sSUBID=8299&charsetID=1&displayMode=DU.

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.12.147

In this paper, standard machine learning techniques are incorporated into the domain of online Cantonese-written restaurant reviews to automatically classify user reviews as thumbs-up or thumbs-down. Two popular text classification algorithms – naive Bayes and SVM, and six feature presentations concerning *n*-gram presence/frequency are chosen to examine the effects of the classifiers and the feature options on Cantonese sentiment classification. This study seeks empirical answers to the following research questions:

1. Dose the SVM classifier beat naive Bayes regarding Cantonese sentiment-based classification?
2. Are high order *n*-grams better features than unigrams to capture sentiments in the Cantonese text?
3. Is feature presence a better text presentation than feature frequency regarding feature selection and text classification?
4. How dose the size of feature set affect the performance of classifiers?

## 2. Literature review

Sentiment classification aims to automatically classify the text of written reviews from customers into positive or negative opinions. It has emerged as a hot research area. While it is still in a preliminary stage, there has been much work related to various languages, such as English (Liu, Hu, & Cheng, 2005; Pang, Lee, & Vaithyanathan, 2002), Japanese (Fujii & Ishikawa, 2006), Mandarin Chinese (Ku, Liang, & Chen, 2006).

In this paper, we focus our interest on written Cantonese which can be viewed as a written variety of Chinese. In fact, Cantonese is primarily a spoken language. The most important mechanism by which Cantonese is represented in written form is phonetic borrowing, that is, using standard Chinese characters that have similar sound when pronounced to represent the Cantonese words to be written. When sometimes confronting the "sound but no character" problem, Cantonese speakers resorted to the strategy of creating a new character to represent a Cantonese word, so written Cantonese is a mixture of standard Chinese characters and over a thousand extra characters invented specifically for Cantonese (marked Cantonese) (Cheung & Bauer, 2002), which is an unintelligible language for Mandarin speakers.

With the increasing need of information organization and knowledge discovery from text data, many supervised learning algorithms have been used for text document classification. Among these methods, naive Bayes and Support Vector Machines (SVM) are always in the comparison list. Naive Bayes – a generative classifier – is considered a simple but effective classification algorithm (Mitchell, 1997). SVM – a discriminative classifier – is considered the best text classification method to date (Joachims, 1998; Yang & Liu, 1999).

Dave, Lawrence, and Pennock (2003), Pang et al. (2002), Read (2005), Ye, Zhang, and Law (2009) and Yu (2008) all compared naive Bayes and SVM on English web review classification. Pang et al. (2002) used movie review data and simple unigram word features to compare three classification methods naive Bayes, Maximum Entropy and SVM. The results showed that word presence/absence is a better feature representation than word frequency. SVM is worse than naive Bayes using word frequency features and slightly better than naive Bayes using word presence/absence features. The best accuracy achieved is 82.9%, using an SVM trained on unigram features. When applying SVM, naive Bayes and *n*-gram model to the destination reviews, Ye et al. (2009) found that SVM outperforms the other two classifiers. Dave et al. (2003) experimented more feature sets with naive Bayes and SVM, and the two classifiers achieve comparable performance. Yu (2008) also believed that SVM is not a universal winner compared with naive Bayes in literary text classification. Read (2005) studied the dependence of naive Bayes and SVM classification models on domain and time. Again, SVM did not beat naive Bayes in sentiment classification as in topic classification. The above work shows that neither machine learning model consistently outperforms the other, and prompts us to examine the utility of naive Bayes and SVM in Cantonese review classification.

Most of previous work in Mandarin Chinese review classification chooses one classifier to compare the performance of different feature sets. Only a few studies have been conducted on the possible interaction between classification models and the sentiment-based classification problem. Tang, Tan, and Cheng (2007) compared KNN, Winnow, naive Bayes and SVM with web product and service reviews. The results showed that SVM outperforms the other models and Chinese character-based bigrams are better features than unigrams and trigrams.

In addition, decades of work in the domain are at least partially knowledge-based. Some of the work focuses on classifying the semantic orientation of individual words or phrases, using linguistic heuristics or a pre-selected set of seed words (Hatzivassiloglou & McKeown, 1997; Turney & Littman, 2002). The sentiment classification of documents has often involved either the use of manually or semi-manually created polarity lexicons (Das & Chen, 2007; Huettner & Subasic, 2000; Tong, 2001) or rules inspired by cognitive linguistics (M. Hearst & Text-Based Intelligent Systems, 1992). In contract, this paper utilizes completely prior-knowledge-free supervised machine learning methods. We believe an appropriately chosen machine learning model could be able to draw its own conclusions from the distribution of lexical elements in a piece of Cantonese review.

## 3. Methodologies

### 3.1. Data collection

Due to no benchmark data available, we created a corpus of Cantonese-written reviews by retrieving consumer reviews from a Cantonese site OpenRice (URL: http://www.openrice.com). The site allows diners to input text feedback and a three-point satisfaction rating for a restaurant located in Hong Kong. As the majority of OpenRice users are inhabitants of Hong Kong, the feedback are generally written in Cantonese with a few exceptions in English and Mandarin. A crawler was developed by Java to randomly download 1500 positive reviews and 1500 negative reviews.

Two native speakers were trained to label these reviews. Non-Cantonese documents were firstly excluded. The judges were asked to label each review with a categorization tag and an uncertainty degree on a scale of 1–3, with 3 being the most uncertain. We found that, except clearly positive feedback and negative feedback, there are borderline instances in between. We discard a review if (1) it is annotated differently by the two judges; (2) any judge annotates it with 3. This process resulted in 1151 positive and 913 negative reviews. To avoid the skewed class distribution, we randomly selected 900 positive and 900 negative reviews to establish the corpus.

### 3.2. Feature construction and feature selection

One problem with many features is that they may be overly specific. For example, "I like dish A" and "I like dish B" would ideally be grouped as "I like DISH." Substitutions have been used widely to solve these sorts of problems in text mining. As restaurant names and dish names frequently occur in restaurant reviews, we replaced any such tokens with RESTAURANT and DISH.

Once substitution was complete, we combined sets of n adjacent Cantonese characters into *n*-gram features (character unigrams, bigrams and trigrams) to discern sentiment in the text. For example, “難” followed by “食” becomes “難食” in a bigram.

Then each document is represented as a vector with values of feature weights. Let $\{f_1, f_2, \ldots, f_m\}$ be a predefined set of $m$ features that can appear in a document. We investigated two methods of text representation, feature presence and feature frequency. Let $n_i$ be the number of times $f_i$ occurs in document $d$. Each document $d$ is represented by the document vector $d = (n_1, n_2, \ldots, n_m)$, or by setting $n_i$ to 1 if feature $f_i$ appears in $d$.

To reduce computational complexity so as to be adapted in Web applications, we pared down the feature vector by performing feature selection. We calculated information gain (IG) for each $n$-gram. IG is suggested a better term-goodness criterion than document frequency thresholding, mutual information and CHI in Mandarin Chinese sentiment classification (Tang et al., 2007). It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. The IG score of term $k$ is defined as follows:

$$\mathrm{IG}(k) = H(c) - p(k)\cdot H(c|k) - p(\bar{k})\cdot H(c|\bar{k})$$

where $p(k)$ is the fraction of documents when $k$ is present, $p(\bar{k})$ is the fraction of documents when them $k$ is absent, and $H(.)$ is the entropy. The entropy $H(c)$ is estimated as $-\sum_c p(c)\log p(c)$.

Weka IG attribute selection model was employed. The $n$-grams were sorted in descending order by their IG scores and the terms with IG scores greater than 0.001 were collected. This process resulted in about 1600 unigrams, 4300 bigrams and 3450 trigrams. We take the top ranked $n$-grams as features in the experiments and look at the effect of feature set size on classification performance. Tables 1 and 2 show the top $n$-grams selected from the binary documents and the term-frequency-based documents, respectively.

### 3.3. Classifier

One primary concern of ours in this work is to examine whether it suffices to treat Cantonese review classification simply as a case of text classification, or whether special classification models need to be developed. We experimented with two standard algorithms, naive Bayes and support vector machines. The philosophies behind these two algorithms are quite different, but each has been shown to be effective in previous text categorization studies.

#### 3.3.1. Naive Bayes

Naive Bayes classifiers are commonly-used in text categorization. The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naive part of such a model is the assumption of item independence.

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

**Table 1**
Top $n$-gram features selected from binary documents.

| | Unigrams | Bigrams | Trigrams |
|---|---|---|---|
| 1 | 滑 | 好味 | 唔會再 |
| 2 | 味 | 好香 | 好好味 |
| 3 | 差 | 好好 | DISH同DISH |
| 4 | 香 | 唔錯 | 唔新鮮 |
| 5 | 濃 | DISH 好 | 都好好 |
| 6 | 正 | 幾好 | DISH好香 |
| 7 | 脆 | 唔掂 | DISH好好 |
| 8 | 甜 | 點知 | 知所謂 |
| 9 | 錯 | 難食 | 不知所 |

**Table 2**
Top $n$-gram features selected from term-frequency-based documents.

| | Unigrams | Bigrams | Trigrams |
|---|---|---|---|
| 1 | 好 | 好味 | 唔知點 |
| 2 | 味 | 好香 | DISH係有 |
| 3 | 差 | 好好 | 話蘇絲 |
| 4 | 滑 | 唔錯 | 之極之 |
| 5 | 香 | DISH 好 | 佢個同 |
| 6 | 濃 | 幾好 | 野只是 |
| 7 | 正 | 唔掂 | ! 跟住 |
| 8 | 脆 | 態度 | 得幾件 |
| 9 | 甜 | DISH同 | 好食囉 |

$$c^*(d_i) = \arg\max_j p(c_j)p(d|c_j) = \arg\max_j p(c_j)\prod_{t=1}^{|V|} p(w_t|c_j)^{n(w_t,d_i)}$$

where, $p(c_j) = \frac{|c_j|}{\sum_{r=1}^{|C|}|c_r|}$, $n(w_t, d_i)$ is the number of occurrences of $w_t$ in $d_i$, and $p(w_t|c_j)$ is estimated from training documents with known category, using maximum likelihood estimation with a Laplacean prior:

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

#### 3.3.2. SVM

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization (Joachims, 1998; Yang & Liu 1999). SVMs seek a hyperplane represented by vector $\vec{w}$ that separates the positive and negative training vectors of documents with maximum margin (see Fig. 1).

Findings in this hyperplane can then be translated into a constrained optimization problem. Let $y_i$ equals +1 (−1) if document $\vec{di}$ is in class + (−). The solution can be written as:

$$\vec{w} = \sum_{i=1}^{n} a_i^* \, y_i \, \vec{d}_i \quad a_i \geqslant 0 \tag{1}$$

where $a_i^*$ are obtained by solving a dual optimization problem. Eq. (1) shows that the resulting weight vector of the hyperplane is constructed as a linear combination of $\vec{d}_i$. Only those examples contribute, for which the coefficient $a_i$ is greater than zero. Those vectors

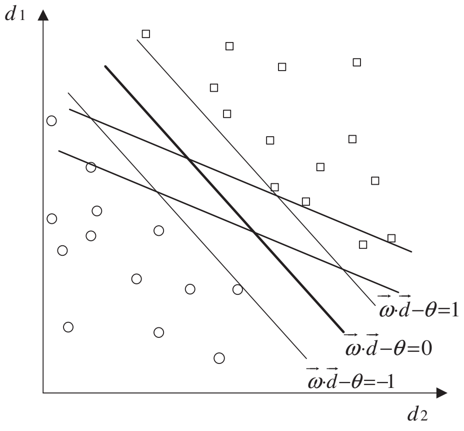
d1
d2
$\vec{\omega}\cdot\vec{d}-\theta=1$
$\vec{\omega}\cdot\vec{d}-\theta=0$
$\vec{\omega}\cdot\vec{d}-\theta=-1$

**Fig. 1.** Maximum margin classifier of SVM.

are called *support vectors*, since they are the only document vectors contributing to $\vec{w}$.

## 4. Performance measures

The category assignments of a polarity classifier can be evaluated using a two-way contingency table (Table 3) which has four cells, where

- cell $a$ counts the documents correctly assigned to positive reviews;
- cell $b$ counts the documents incorrectly assigned to positive reviews;
- cell $c$ counts the documents incorrectly assigned to negative reviews;
- cell $d$ counts the documents correctly assigned to negative reviews.

The performance measures recall, precision and accuracy are defined and computed from the contingency table as:

$$\begin{aligned}&\text{Accuracy} = (a+d)/(a+b+c+d),\\&\text{Recall(pos)} = a/a+c, \quad \text{Precision(pos)} = a/a+b,\\&\text{Recall(neg)} = d/b+d, \quad \text{Precision(neg)} = d/c+d.\end{aligned}$$

## 5. Results and discussion

Three-fold cross-validation was performed for the experiments reported in this study. The experiments used our own implementation of a naive Bayes classifier and Chang and Lin's (2001) LIBSVM implementation of a Support Vector Machine classifier with all parameters set to their default values. We ran each classifier with various-sized feature sets to examine the effects of feature size on sentiment classification performance. Figs. 2–13 display the average three-fold cross-validation results, and Pre, Rec and Acc denote precision, recall and accuracy, respectively.

### 5.1. Naive Bayes experiments

Fig. 2 shows the results for average accuracy, precision and recall as we varied the number of binary unigram features. The highest average accuracy is 93.17% with 400 features. Fig. 3 shows the results with frequency-based unigram features and the average accuracy peaks at 88.83% with 150 features.

Figs. 4 and 5 display the results with the varying number of bigrams. Binary features achieve accuracy above 95% when 650–1750 features are used with a peak at 95.67%. Frequency features achieve accuracy around 94% when the feature set size is 400–1400, with the highest 94.83% at 1300 features.

Figs. 6 and 7 show that the best accuracies with binary trigrams and frequency-based trigrams are 95.33% (1700 features) and 94.17% (1200 features), respectively.

### 5.2. SVM experiments

For SVM, Fig. 8 shows that binary unigrams yield all the five measures close to 90% based on the feature size considerations, while Fig. 9 shows that frequency-based unigrams yield accuracy between 77% and 87%, higher recall of positive reviews, and higher precision of negative reviews.

**Table 3**
Contingency table for performance evaluation.

| | Actual positive reviews | Actual negative reviews |
|---|---|---|
| Assigned positive | $a$ | $b$ |
| Assigned negative | $c$ | $d$ |

Figs. 10 and 11 show that binary bigrams and frequency bigrams achieve the highest accuracy of 90.67% (around 2000 features) and 94.83% (1950 features). Figs. 12 and 13 show that binary trigrams and frequency trigrams achieve the highest accuracy of 82.50% (2250 features) and 90.17% (2550 features). The performance of frequency features is more unstable when the feature size is small.

Table 4 summarizes the best accuracy with every feature presentation in our experiments. We draw the following observations:

(1) As Figs. 2–13 show, the accuracy of high order $n$-gram peaks with a larger feature size than that of unigram. The reason could simply be that individual terms in $n$-gram representation have, on the average, a lower frequency of appearance than character terms. With the number of features increasing, accuracy starts to decline due to overfitting, and other factors such as the lower quality of the additional features could also play a role in the decline.

(2) Table 4 shows that naive Bayes achieves the best accuracy by accounting only for feature presence, not feature frequency. In contact, SVM yields better performance with unigram presence/absence, bigram frequency and trigram frequency as features. This result is some contradictory to the conclusion drawn by Pang et al. (2002). Moreover, Yu (2008) found that SVM achieves the best accuracy with the normalized word frequency representation and naive Bayes achieves the best accuracy with the word frequency representation rather than Boolean features in novel sentiment classification. Thus, different document property probably requires different types of feature presentation and models of text classification.

(3) Table 4 show that, the best accuracy achieved by naive Bayes with each of the six feature representations is not less than SVM, and especially the naive Bayes classifier outperforms SVM when the feature size is small. The good performance of naive Bayes is surprising because its assumption of conditional independence is violated in our experiments. In fact, there are strong relationships among polarity features extracted from reviews. Sentiment classification perhaps is one of the domains containing clear feature dependences, where naive Bayes often performs surprisingly well (Domingos & Pazzani, 1997). But further empirical and theoretical study is still required to understand the relation between the sentiment classification task and the behavior of naive Bayes.

(4) For both naive Bayes and SVM, bigrams outperform unigrams and trigrams in Cantonese review classification. We examined a number of test documents and found the feature vectors corresponding to some of trigram documents (particularly the short ones) have many zeroes in them. This suggests that the main problem with the trigram model is likely to be data sparseness, and unigrams are not enough to capture the sentiment in Cantonese text. Bigrams have also been proved better features than unigrams and trigrams in Mandarin Chinese sentiment classification (Tang et al., 2007).

(5) More negative reviews are wrongly recognized as positive reviews. Figs. 2–13 show that both classifiers achieve higher recall of positive reviews and higher precision of negative reviews. We examined the selected features and found that there are more positive-sentiment-bearing instances of the highly scored items. For example, the second column in Table 1 show that the top nine features include five obvi-

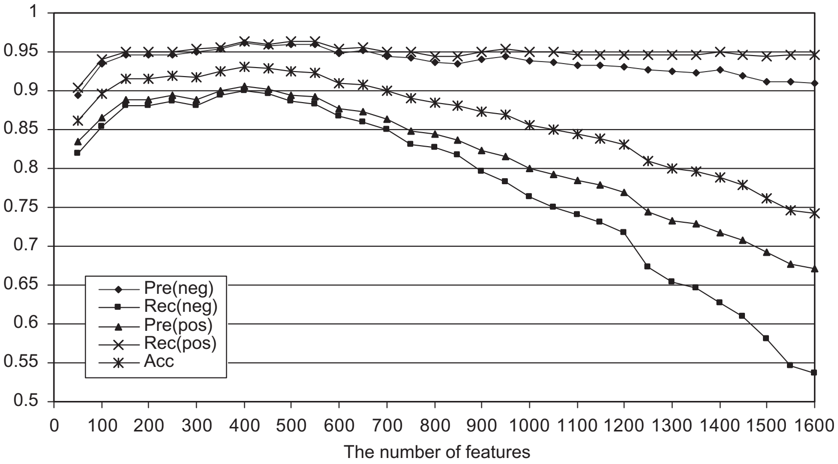

**Fig. 2.** Effect of feature size with binary unigram feature.

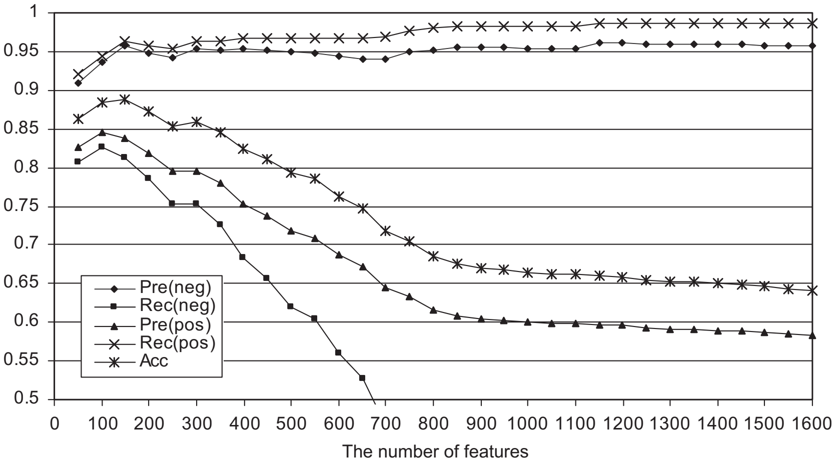

**Fig. 3.** Effect of feature size with frequency-based unigram feature.

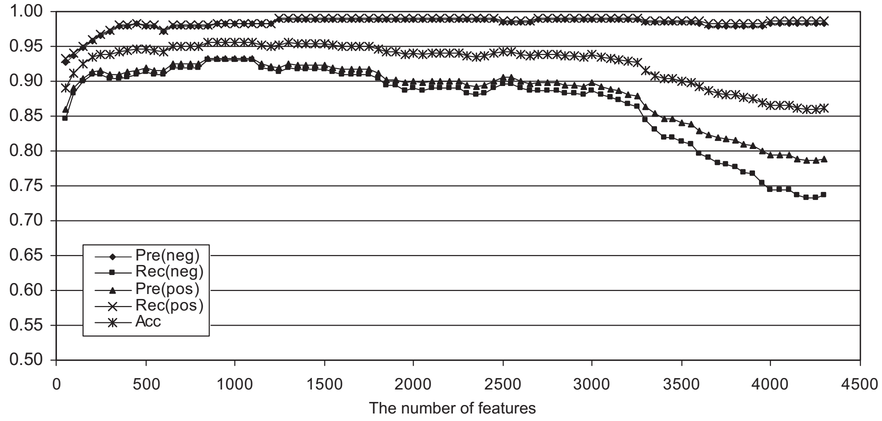

**Fig. 4.** Effect of feature size with binary bigram feature.

ously positive instances 好味 (delicious), 好香 (smell great), 好好 (very good), 唔錯 (good) and 幾好 (very good)) in comparison with two obviously negative instances 唔掂 (not worth) and 難食 (taste terrible). Then an examination of some text reviews revealed that positive Cantonese reviews have more frequent and discriminative items in common, while negative opinions are often expressed with different words depending on the situation. This makes so strong association between the positive-sentiment features and the positive category that not only most positive reviews but a few negative reviews are classified as positive.

## 6. Conclusion

This paper has shown that machine learning techniques perform quite well in the domain of Cantonese review classification. Despite

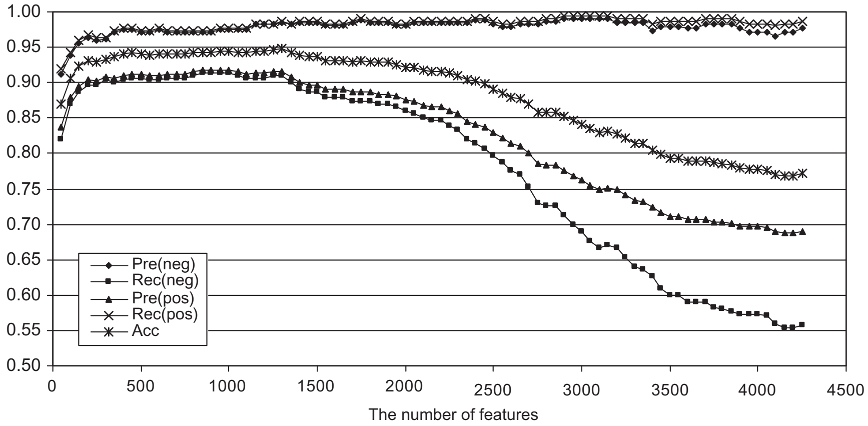

**Fig. 5.** Effect of feature size with frequency-based bigram feature.

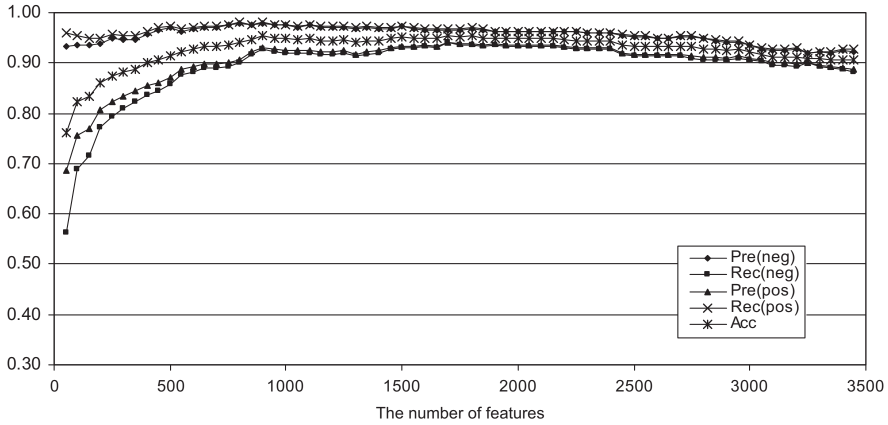

**Fig. 6.** Effect of feature size with binary trigram feature.

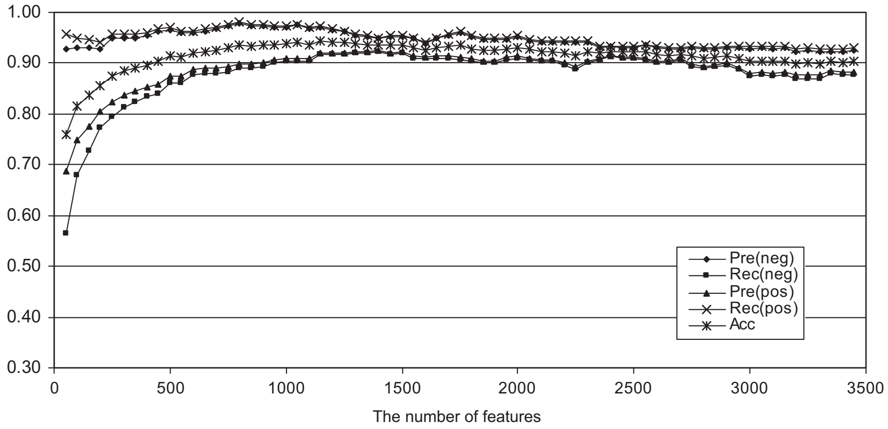

**Fig. 7.** Effect of feature size with frequency-based trigram feature.

its unrealistic independence assumption, the naive Bayes classifier surprisingly achieves comparable, or better performance than SVM. Interactions between classification methods and feature presentation options are observed, and bigram frequency is proved the effective feature in capturing sentiments in the Cantonese text. In addition, we look at the effects of feature set size on the classification performance. As the feature size increases, the accuracies of both classifiers reach their peaks and then decline due to overfitting, with the SVM classifier less sensitive to the number of features.

Although written Cantonese is still a minority written language, accounting for only a small percentage of the total text written and published in China, it is also clear that written Cantonese's percentage of the "market share" has increased rather dramatically over the last decades. Cantonese as a language of identity for cultural communities has come to extend to the use of Cantonese in spoken as well as written form (Snow, 2004), and people within the same community are more readily to share knowledge and opinions with each other.

In the context of travel and hospitality, resent studies have confirmed that consumer' travel-related purchase decision can be directly influenced by online user-generated contents (Gretzel & Yoo, 2008; Pan, MacLaurin, & Crotts, 2007). According to a survey

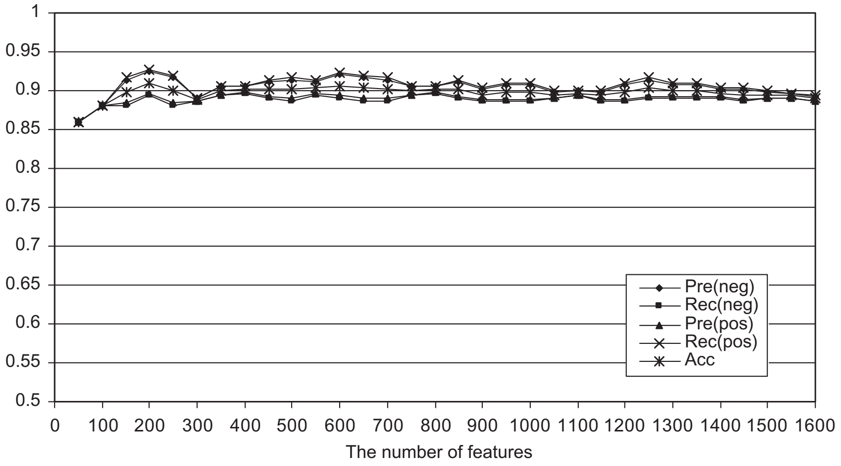

**Fig. 8.** Effect of feature size with binary unigram feature.

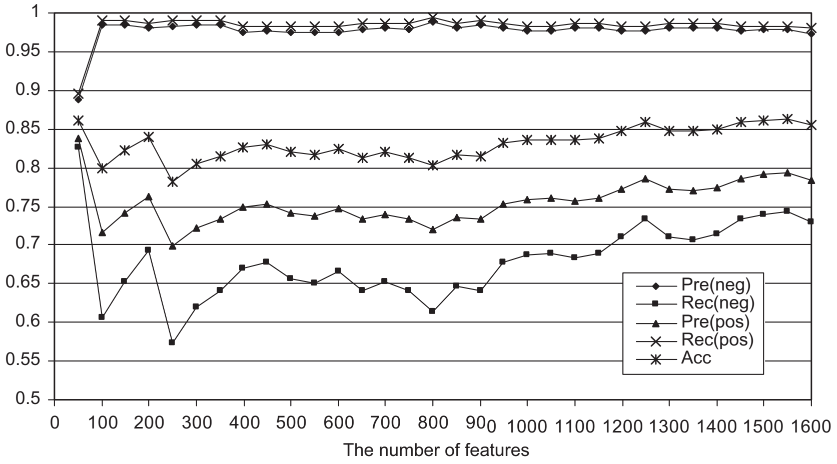

**Fig. 9.** Effect of feature size with frequency-based unigram feature.

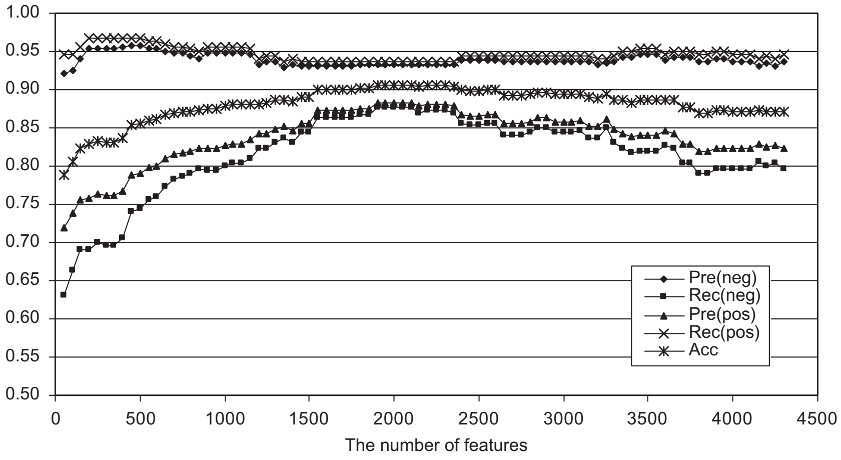

**Fig. 10.** Effect of feature size with binary bigram feature.

of more than 2000 American Internet users,[2] respectively, 79%, 87% and 84% reported that online reviews had a significant influence on their restaurants, hotels and travel services purchasing, and they were willing to pay at least 20 percent more for an Excellent (5 star rating) than for a Good (4 star rating). Findings of this research are expected to make a contribution to understand Cantonese-speaking consumers' perception of travel-related products and services.

Empirical results of this study are likely to lead to a new trend of Cantonese information processing. This could provide insights into how web search engines can process information of Cantonese

[2] ComScore/the Kelsey group, *Online consumer-generated reviews have significant impact on offline purchase behavior*, Press Release, http://www.comscore.com/press/release.asp?press=1928, November 2007.

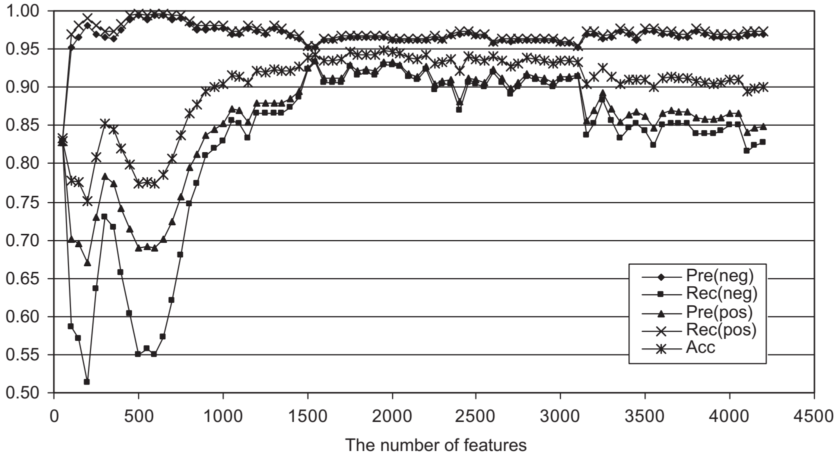

**Fig. 11.** Effect of feature size with frequency-based bigram feature.

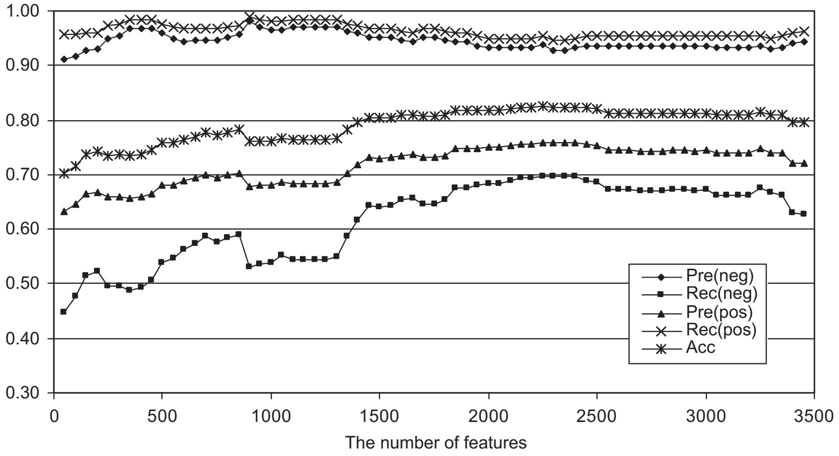

**Fig. 12.** Effect of feature size with binary trigram feature.

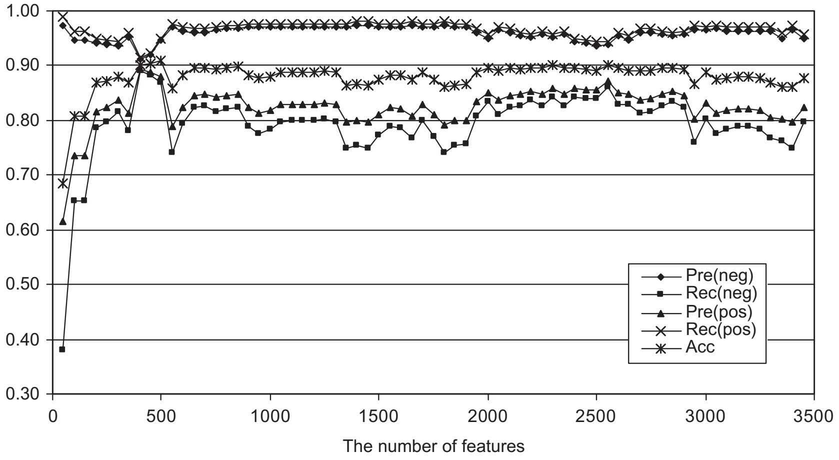

**Fig. 13.** Effect of feature size with frequency-based trigram feature.

reviews, and help the design of travel and hospitality information systems in different functional areas. At present, travelers are largely Internet users. They like to consult an online search about the information of their preferred destinations when making travel planning and the opinions of peer consumers often have a strong influence on their decisions. Since this study has investigated the methods that can conduct automatic analysis of the sentiment attitude in Cantonese reviews, it is expected to help consumers, and the travel and hospitality industry to obtain valuable information from a large amount of online Cantonese-written reviews. Some potential applications include extracting opinions from travel forums efficiently and integrating automatic review mining technolo-

**Table 4**
The peak of every accuracy curve and the corresponding number of features.

| | Unigram | Unigram_freq | Bigram | Bigram_freq | Trigram | Trigram_freq |
|---|---|---|---|---|---|---|
| NB | 93.17 (400) | 88.83 (150) | **95.67** (900–1100) | 94.83 (1300) | 95.33 (1700, 1800) | 94.17 (1200) |
| SVM | 90.67 (600) | 86.33 (1550) | 90.67 (1900–2100) | **94.83** (1950) | 82.50 (2250) | 90.17 (2550) |

*Note:* Boldface denotes the best performance for a given setting (row).

gies with search engines to provide summarized information of search results for users.

## Acknowledgments

This study was partially funded by National Science Foundation of China (70971033, 70890082) and NCET-08-0172.

## References

Chang, C.-C., Lin, C.-J. (2001). *LIBSVM: A library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Cheung, K., & Bauer, R. (2002). *The representation of Cantonese with Chinese characters. Journal of Chinese Linguistics, Monograph Series number 18*. Berkeley: University of California.

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).

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375–1388.

Dave, K., Lawrence, S., Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: *Proceedings of 12th international conference on World Wide Web* (pp. 519–528).

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. (1997). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning, 29*(2–3), 103–130.

Fujii, A., Ishikawa, T. (2006). A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In: *Proceedings of COLING-ACL workshop on sentiment and subjectivity in text* (pp. 15–22).

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.

Gretzel, U., & Yoo, K. (2008). *Use and impact of online travel reviews. Information and Communication Technologies in Tourism*. Wien/New York: Springer-Verlag. pp. 35–46.

Hatzivassiloglou, V., McKeown, K. (1997). Predicting the semantic orientation of adjectives. In: *Proceedings of the 35th annual meeting of the association for computational linguistics* (pp. 174–181).

Hearst, M. (1992). *Direction-based text interpretation as an information access refinement. Text-Based Intelligent Systems*. Lawrence Erlbaum Associates.

Horrigan, J. A. (2008). Online shopping, pew Internet & American life project report.

Huettner, A., Subasic, P. (2000). Fuzzy typing for document management. In: *ACL 2000 companion volume: Tutorial abstracts and demonstration notes* (pp. 26–27).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In: *Proceedings of the 10th European conference on machine learning* (pp. 137–142).

Ku, L.-W., Liang, Y.-T., Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In: *AAAI Symposium on Computational Approaches to Analysing Weblogs* (pp. 100–107).

Liu, B., Hu, M., Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In: *Proceedings of the 14th international World Wide Web conference* (pp. 10–14).

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

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Pan, B., MacLaurin, T., & Crotts, J. (2007). Travel blogs and the implications for destination marketing. *Journal of Travel Research, 46*(1), 35–45.

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the 2002 conference on empirical methods in natural language processing* (pp. 79–86).

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: *Proceedings of the ACL student research workshop* (pp. 43–48).

Snow, D. (2004). *Cantonese as written language: The growth of a written Chinese vernacular*. Hong Kong University Press (in Chinese).

Tang, H., Tan, S., & Cheng, X. (2007). Research on sentiment classification of Chinese reviews based on supervised machine learning techniques. *Journal of Chinese Information Processing, 21*(6), 88–108 (in Chinese).

Tong, R., (2001). An operational system for detecting and tracking opinions in online discussion. *Workshop note, SIGIR 2001 Workshop on Operational Text Classification*.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 417–424).

Yang, Y. M., Liu, X. (1999). A re-examination of text categorization methods. In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49).

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527–6535.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing, 23*(3), 327–343.