Predictive Accuracy of Sentiment Analytics for Tourism: A Metalearning Perspective on Chinese Travel News

Abstract

Sentiment analytics, as a computational method to extract emotion and detect polarity, has gained increasing attention in tourism research. However, issues regarding how to properly apply sentiment analytics are seldom addressed in the tourism literature. This study addresses such methodological challenges by employing the metalearning perspective to examine the design effects on predictive accuracy using a sentiment analysis experiment for Chinese travel news. Our results reveal strong interactions among key design factors of sentiment analytics on predictive accuracy; accordingly, this study formulates a metalearning framework to improve predictive accuracy for computational tourism research. Our study attempts to highlight and improve the methodological relevance and appropriateness of sentiment analytics for future tourism studies.

Keywords:

Sentiment analytics, Design effects, Predictive accuracy, Metalearning, Chinese travel news

Introduction

With the proliferation of big data, there is an increasing trend for tourism researchers to adopt computational methods in their studies. In particular, sentiment analytics, which works as an effective method to automatically extract public opinions and analyze sentiment polarity from massive textual data (Cambria 2016), has thus piqued researchers' interest (Alaei, Becken, and Stantic 2017; Kirilenko et al. 2017). Traditionally, sentiment has been studied under different variants (e.g., emotion, attitude, and perception) with a focus on interpreting the causal mechanisms related to tourism development (Deery, Jago, and Fredline 2012; Kim, Uysal, and Sirgy 2013; Nunkoo and So 2016; Sharpley 2014; Uysal et al. 2016). Along with massive increases in the volume and types of online data, researchers have applied sentiment analytics to mitigate the effects of human subjectivity and improve the generalizability of scholarly findings (e.g., Duan et al. 2016; Geetha, Singha, and Sinha 2017; Xiang et al. 2017). However, scant research has considered the predictive accuracy of sentiment analytics, and hence issues regarding how to properly apply sentiment analytics are seldom addressed in the tourism literature.

The increasingly computational nature of social science implores researchers to pay greater attention to the predictive accuracy in addition to the interpretation of causal relations (Hofman, Sharma, and Watts 2017). The lack of research on predictive accuracy may lead to false positives in traditional hypothesis testing (Simmons, Nelson, and Simonsohn 2011) and could present a major obstacle when conducting such studies in tourism (Schuckert, Liu, and Law 2015). More importantly, sentiment analytics are technically sophisticated and domain-specific. The method generally combines many design factors from natural language processing (e.g., feature extraction), statistical techniques (e.g., feature weighting), and machine learning (e.g., sentiment classifier) (Alaei, Becken, and Stantic 2017). Hofman et al. (2017) argued that for such computational methods, prediction results depend largely on "researcher degrees of freedom." A recent study by Kirilenko et al. (2017) noted that a specific set of choices pertaining to datasets, classification models, and performance metrics could generate significant differences in predictive accuracy of sentiment analytics.

It is a challenging task to study the predictive accuracy of sentiment analytics for tourism. Such task is more methodological (e.g., choosing design factors) than technical (e.g., improving a new classification algorithm) (Hofman, Sharma, and Watts 2017). Hence, we need an integrative effort to examine how different design factors in sentiment analytics influence predictive accuracy, and how to ensure predictive accuracy of semantic analytics via a systematic approach. Specifically, this paper attempts to address the following research questions:

(1) What are the key design factors of sentiment analytics for tourism?

(2) How do these design factors influence the predictive accuracy of sentiment analytics for tourism?

(3) How can these design factors be systematically incorporated to improve the predictive accuracy of sentiment analytics for tourism?

To answer these questions, we designed an experiment based on the metalearning theory to uncover the black box of sentiment analytics within the context of Chinese travel news. This context was selected for the following reasons: first, news coverage is generally considered a credible source for social realities when examining public opinion, but the massive volume and high variance of news have made this task of social listening extremely time consuming, costly and often manually impossible (Chiu et al. 2015); and second, tourism development in China, including domestic, inbound, and outbound tourism, has experienced exponential growth over the past decades (Jin and Wang 2016). Thus, research on Chinese tourism has drawn increasing attention from academics and practitioners alike.

Metalearning is a concept emerged from educational psychology and has been subsequently applied to machine learning (Smith-Miles 2008). The key idea is to delve into the machine learning process and adjust learning approaches according to a given task (Lemke, Budka, and Gabrys 2015). Developing a holistic awareness of the interactive mechanism in sentiment analytics, rather than simply considering relevant factors individually, should substantially improve the predictive accuracy of sentiment analytics when analyzing Chinese travel news.

This study’s effort towards predictive accuracy of sentiment analytics for tourism aims to raise methodological awareness and improve the methodological relevance and appropriateness of sentiment analytics in future tourism research. Meanwhile, the resulting sentiment classification models may contribute to extensive sentiment analysis of Chinese travel news for policymakers and industry practitioners.

Literature review

Sentiment analytics

Sentiment analytics is the computational study of people's emotions toward objects (Liu 2012). The most popular task of sentiment analytics is analyzing the sentiment polarity of textual data and classifying texts into sentiment categories, such as positive and negative (Pang and Lee 2008; Schuckert, Liu, and Law 2015). To successfully discover, interpret, and communicate extracted opinions and detected polarity, sentiment analytics relies on multi-disciplinary efforts including natural language processing, statistics, and machine learning (Cambria 2016).

Existing approaches to sentiment analytics can be classified into two broad categories: semantic orientation approaches and machine learning approaches (Chiu et al. 2015). Semantic orientation approaches hold that text is classified into affect categories on the basis of the presence of fairly unambiguous affect words, such as "happy", "sad", "afraid", and "bored." Lexicon resources of affect words include the WordNet-Affect, SentiWordNet, and SenticNet (Cambria 2016). Semantic orientation approaches are popular thanks to their accessibility and economy. However, the weaknesses of these approaches include poor affect recognition given complex linguistic rules, and heavy dependence on the depth and breadth of the employed lexicon resources (Cambria 2016). For a domain lacking of such resources, machine learning approaches can mitigate the above limitations.

By feeding a machine learning algorithm a training corpus of affectively annotated texts, machine learning approaches can not only learn the affective polarity of affect keywords (as in semantic orientation approaches) but can also consider the polarity of other arbitrary keywords (like lexical affinity) and word co-occurrence frequencies (Cambria 2016). However, machine learning approaches rely on statistical models that are meaningful when given a sufficiently large text input; therefore, the approaches can achieve better performance on the document or paragraph level compared to smaller text units, such as sentences or clauses.

The domain of sentiment analytics for Chinese travel news has limited lexicon resources of affect words, but a sufficiently large text corpus (Alaei, Becken, and Stantic 2017), therefore this study focused specifically on machine learning approaches.

Sentiment analytics in tourism

Tourism is an ideal application field of semantic analytics. In this field, sentiment was traditionally studied as a synonym for emotion or attitude, underscoring the significance of sentiment study in creating a harmonious host-guest interaction when developing sustainable destinations. Several review papers (Deery, Jago, and Fredline 2012; Nunkoo and So 2016; Sharpley 2014; Uysal et al. 2016) framed the focus of existing research as identifying and interpreting causal relationships around sentiment using typical qualitative (e.g., content analysis) and quantitative methods (e.g., structural equation modeling).

In recent years, a growing body of literature has applied sentiment analytics in the tourism context (Alaei, Becken, and Stantic 2017; Kirilenko et al. 2017). These studies tend to explore sentiment in electronic word-of-mouth (eWOM), such as positive or negative online reviews left by customers for products or services across various domains, including hotels (e.g., Duan et al. 2016; Hu and Chen 2016; Lee, Jeong, and Lee 2017; Wang et al. 2013), restaurants (e.g., Kang, Yoo, and Han 2012; Marrese-Taylor, Velásquez, and Bravo-Marquez 2014; Zhang et al. 2011), and travel destinations (e.g., Capriello et al. 2013; Geetha, Singha, and Sinha 2017; Luo and Zhai 2017; Philander and Zhong 2016; Sanz-Blas and Buzova 2016; Ye, Zhang, and Law 2009). Although we clearly observe increasing interests in applying sentiment analytics to tourism research, few studies have considered the predictive accuracy of sentiment analytics, and many authors have not reported performance measures (Kirilenko et al. 2017).

Discussions surrounding the methodological challenges of computational social science have increased in recent years (Lazer et al. 2009). For example, Ekbia et al. (2015) synthesized several epistemological dilemmas in big data analytics, and Hofman et al. (2017) pointed out the complementarity between predictive accuracy and interpretability for future computational social science. Several tourism researchers have joined in these deliberations. Xiang et al. (2017) identified issues of predictive accuracy related to data quality in social media analytics by incorporating data from three major online platforms (TripAdvisor, Expedia, and Yelp). Schmunk et al. (2013) studied the predictive accuracy of sentiment analytics by comparing four automated sentiment analysis tools (one lexicon-based and three machine learning-based) to reviews posted by visitors of a ski resort on TripAdvisor and Booking.com. More recently, Kirilenko et al. (2017) evaluated the suitability of different types of automated classifiers for tourism by comparing their performance to that of human raters and emphasized the importance of software selection when applying sentiment analytics.

However, it remains unclear how design factors influence the predictive accuracy of sentiment analytics and how to synthesize these design factors to ensure predictive accuracy in tourism research. In the following section, we elaborate on the metalearning theory to identify key design factors and examine their effects on predictive accuracy in the context of Chinese travel news.

Design factors and hypotheses

Educational psychologist John Biggs (1985) explained that metalearning is a higher-level cognitive ability related to “learning about learning”, defined as learners’ awareness and control over their own learning; in other words, metalearning is one’s ability to understand and adapt to the act of learning instead of the learned subject knowledge. Moreover, metalearning is the fundamental capability for people to assess and adapt their learning approach to complete demanding tasks.

Metalearning researchers view the machine learning process for a particular task (e.g., sentiment analytics) as a normal learning process, focusing on the prediction results of the machine. Over the past 20 years, machine learning research has generated myriad approaches, including those related to parameterization, preprocessing, and post-processing. The machine learning process has accordingly become increasingly complex. Researchers are eager to know the preferable approaches for given machine learning tasks rather than selecting appropriate algorithms and parameterization methods through trial and error.

The success of sentiment analytics is generally determined by an appropriate set of design factors including approaches to represent and classify documents (Serrano-Guerrero et al. 2015; Alaei, Becken, and Stantic 2017). Instead of directly identifying the opinions expressed in news articles, metalearning focuses on identifying the key design factors of sentiment analytics, the possible approaches to address these design factors, and the combination of these approaches that performs the best (Lemke, Budka, and Gabrys 2015).

Several key design factors exist in sentiment analytics for Chinese travel news, namely news feature extraction, news feature weighting, and news sentiment classifier.

News feature extraction. The unstructured textual data of news articles must be transformed to structured data prior to processing by sentiment classifiers. The literature has identified three popular approaches to extracting news article features: bag-of-words (BOW)

(Tan, Zhang, and Jin 2008; Wang et al. 2013; Chiu et al. 2015), character unigrams, and character bigrams (Kang, Yoo, and Han 2012; Zhang et al. 2011). BOW is based on the assumption that a document can be represented by a set of words. In BOW, each word in a news article is represented as a separate variable, which allows for simple and efficient sentiment analysis. Unlike English, Chinese uses pictograms/characters with no spaces between words. When using BOW in Chinese news, word segmentation can assist the computer in dividing Chinese characters into words. Character unigrams and bigrams are two widely used forms of character *n*-gram (Zhang et al. 2011). A character *n*-gram is a contiguous sequence of characters with length *n*. A given document is simply segmented into a set of sequences of *n* ordered and adjacent characters. For example, a bigram represents two adjacent characters, and a unigram is a single character. The main advantages of the *n*-gram model include its language independency and simplicity. Word segmentation is not needed in a character *n*-gram.

News feature weighting. Another significant factor in news sentiment classification is feature weighting (Nassirtoussi et al. 2015). The weight of each feature of a news article measures its impact on the final sentiment classification. Prior research has indicated three popular feature weighting approaches: Boolean occurrence (BO) (Wang et al. 2013; Wu, Zheng, and Olson 2014), term frequency (TF) (Ye, Zhang, and Law 2009; Zhang et al. 2011), and term frequency inverse document frequency (TFIDF) (Bravo-Marquez et al. 2014; Chen, Liu, and Chiu 2011). The BO approach is also called Boolean weighting or binary representation; it is a basic technique of weighting features where two values (i.e., 0 and 1) represent a feature’s absence or presence. TF is an extensively used feature weighting approach, which weighs each feature based on the frequency with which a term appears. Lastly, TFIDF is another efficient weighting approach in sentiment classification; it considers terms that appear frequently in all documents because the terms that are frequently present in all documents may contain limited sentiment information for the purpose of sentiment classification (e.g., the term *tourist* in travel news). News corpora can be projected into the feature vector space after the preceding process. The feature vector matrix can then be manipulated by classifiers as a type of structured data.

News sentiment classifier. Most supervised machine learning methods can be used to classify news into positive and negative categories. Naïve Bayes (NB) (Tan, Zhang, and Jin 2008; Ye, Zhang, and Law 2009) and support vector machines (SVM) (Schmunk et al. 2013; Wang et al. 2013) are two classic classifiers for sentiment analytics. Given a feature vector matrix, an NB algorithm computes the posterior probability that a news article belongs to different sentiment polarity classes and assigns it to the class with the highest posterior probability. Given a training news set, an SVM training algorithm builds a model that assigns news articles to one sentiment class or another. The algorithm maps articles represented as points in a high-dimensional feature space by constructing a maximum margin hyperplane, thereby separating sentiment classes. Good separation is achieved by the hyperplane with the greatest distance to the nearest training data points of any class (i.e., functional margin); a considerably large margin means a low generalization error for the classifier. The dimension of feature space is quite large in sentiment classification, so the linear kernel is commonly used because the problem is linearly separable (Xia, Zong, and Li 2011).

According to the metalearning perspective, these design factors and corresponding approaches may interact to influence the predictive accuracy of sentiment analytics. For instance, in terms of the sentiment analytics of Chinese online hotel reviews, Shi and Li (2011) adopted BOW to extract features, TFIDF to weigh features, and SVM to classify the text. In contrast, Ye et al. (2009) employed the TF feature weighting approach given its good performance with English-language online travel destination reviews. Wu et al. (2014) used the BO weighting approach along with BOW and SVM to examine Chinese financial forum posts. However, Zhang et al. (2011) determined that character bigrams, BO, and NB performed better in Cantonese online restaurant reviews. Therefore, we propose the following hypotheses regarding the internal dynamics of sentiment analytics:

Hypothesis 1 (H1): There exists a significant three-way interaction effect among the design factors of feature extraction, feature weighting, and sentiment classifier on the predictive accuracy of sentiment analytics for Chinese travel news.

Hypothesis 2 (H2): There exists a significant interaction effect between the design factors of feature weighting and sentiment classifier on the predictive accuracy of sentiment analytics for Chinese travel news.

Hypothesis 3 (H3): There exists a significant interaction effect between the design factors of feature extraction and sentiment classifier on the predictive accuracy of sentiment analytics for Chinese travel news.

Hypothesis 4 (H4): There exists a significant interaction effect between the design factors of feature extraction and feature weighting on the predictive accuracy of sentiment analytics for Chinese travel news.

The sentiment analytics process used to test these hypotheses, including key design factors and corresponding approaches, is illustrated in Figure 1. The process begins with benchmark building and ends with performance evaluation, including three important design factors regarding the Chinese news sentiment analytics task: news feature extraction, news feature weighting, and news sentiment classifier. In particular, we employed three different news feature extraction approaches and three news feature weighting approaches to project unstructured news textual data into structured vector space. For news sentiment classifier, we selected two commonly used machine learning algorithms, NB and SVM. Accordingly, 18 possible combinations of approaches built a complete sentiment classification model. The research task of improving predicative accuracy of sentiment analytics for Chinese travel news consisted of empirical examination and comparison of these 18 approach combinations to identify the most effective one. To pinpoint the main and interaction effects of these three factors, we established a benchmark corpus and employed a standard accuracy metric with 10-fold cross-validation to evaluate the alternative models' performance. We also fine-tuned the classifier parameters for the 18 sentiment classification models.

=============================

Please insert Figure 1 about here.

=============================

Experiment design

This section elaborates on the metalearning experiment design according to the proposed sentiment analytics process. The process was mainly implemented using RapidMiner Studio 7.3, one of the leading data science platforms, which provides advanced analytics capabilities by combining analysis operators and setting corresponding parameters. Operators are the building blocks of professional data analysis in RapidMiner Studio, including basic data processing and classification. Several Python programs were also developed, thereby facilitating the building of a benchmark corpus and extending the Chinese text processing capability of RapidMiner Studio.

Benchmark building

Given the lack of a publicly available dataset for sentiment analytics of Chinese travel news, we began this experiment by building a benchmark corpus. Specifically, we incorporated the following steps (Xu et al. 2008): searching and crawling, selecting, and annotating.

Searching and crawling. News coverage spreads through mass media (e.g., newspapers and television) and social media (e.g., blogs and tweets). However, Flanagin and Metzger (2000) determined that newspapers boast the highest information credibility among all sources. Therefore, we searched and selected news articles from WiseNews (http://wisenews.wisers.net), a comprehensive Chinese news database covering news from major newspapers published in the Greater China region, including Mainland China, Hong Kong, Taiwan, and Macau (Chow et al. 2007). The present study focused on travel news articles published in 135 major newspapers in the Greater China region (i.e., 88 in Mainland China, 18 in Hong Kong, 17 in Taiwan, and 12 in Macau) between January and December 2015. We searched for relevant news articles using keywords (i.e., *tourism*, *travel*, *tour*, *tourist*, and *traveler*) and developed a web crawler to collect, parse, reformat, and store news articles from the web page results. In all, we collected 11,321 news articles.

Selecting. To improve the sample's relevance and rigor, we further selected news articles for annotation that met the following criteria: each article should be over 100 words, the typical length of meaningful news articles (Moznette and Rarick 1968); and the news release time should be evenly spread throughout the year. The sample size for a news analysis needs to be manageable. Given that the sample size in recent studies analyzing news articles ranged from 88 (Liu and Pennington-Gray 2015) to 829 (Peel and Steen 2007), we employed RapidMiner Studio to sample 2,000 articles that fulfilled the preceding criteria. Additionally, for this study, the topics of the selected news articles needed to be sufficiently relevant to tourism. For example, after manual review, news articles pertaining mostly to the stock market and advertisements were excluded. This process ultimately resulted in 1,769 eligible articles for sentiment annotation.

Annotating. The sampled news articles were manually annotated to train and test the sentiment classification models. Communication studies have indicated that opinion in most news articles is either explicitly or implicitly embedded (Van De Kauter, Breesch, and Hoste 2015). Therefore, most tourism sentiment analysis studies have classified textual data into two classes (e.g., Duan et al. 2016; Schuckert, Liu, and Law 2015), and we adopted the same strategy. The polarity of positive news was labeled as +1, indicating support for tourism development or other relevant tourism issues. The polarity of negative news was coded as −1, denoting the opposite viewpoint. Overall sentiment polarity depended on the dominant narrative when annotating news with mixed sentiment polarity (i.e., both positive and negative).

Unlike user-generated content, news articles are often written from a third-person point of view and frequently include implicit sentiments. For example, one news report presented information related to an influx of inbound tourists to a destination; however, the focus was on traffic jams, environmental problems, and local residents' inconvenience due to tourists. Therefore, the sentiment polarity of this news piece was annotated as negative even though it included objective narratives.

To minimize annotator bias, we used control coding to estimate whether annotators assigned news sentiment polarities objectively (Stepchenkova and Eales 2011). We employed two native Chinese-speaking annotators who were graduates of a tourism management program to provide independent annotations based on codebook guidelines. Annotation differences were resolved through follow-up discussions between the two annotators. For the control coding process, we measured inter-annotator agreement and obtained a percentage agreement of 91% and kappa score of 0.86, suggesting a high level of agreement between the two annotators (Landis and Koch 1977).

A balanced benchmark dataset with an equal number of positive and negative articles has been widely adopted in sentiment analysis research (Moraes, Valiati, and Gavião Neto 2013; Pang and Lee 2004). Most classification algorithms (e.g., NB and SVM) assume that datasets are evenly distributed among various classes and may ignore minority classes (Lee and Lee 2012). Zhang et al. (2016) stated that for machine learning-based sentiment classification models, a balanced training dataset is preferable for building better classifiers and achieving higher accuracy. A balanced testing dataset also produces more reliable results in the model performance evaluation, which helped to reveal internal dynamics in the present study. We used stratified sampling and randomly selected 1,000 articles (i.e., 500 positive and 500 negative) from the 1,769 annotated samples to build the benchmark corpus. The final corpus included five data fields, namely ID, title, content, publication date, and sentiment polarity.

Key design factor specifications

For the BOW feature extraction approach, we developed a Python program based on Jieba (https://github.com/fxsjy/jieba), a popular Chinese word segmentation tool, along with a manually refined lexicon containing simplified and traditional Chinese characters. After word segmentation, the program removed stop words to filter features that provided limited content information, such as time, numbers, pronouns, prepositions, and conjunctions. The other two feature extraction approaches (i.e., character unigrams and character bigrams) were conducted using RapidMiner Studio's Tokenize Operator. The Process Documents from Data Operator of RapidMiner Studio also provided the three feature weighting approaches (i.e., BO, TF, and TFIDF).

The NB and SVM algorithms were provided by the Naïve Bayes Operator and Support Vector Machine Operator, respectively. We fine-tuned the parameters accordingly because a classification algorithm's parameters may influence its performance (Hagenau, Liebmann, and Neumann 2013). For NB, we employed the Laplace correction to prevent the high influence of zero probabilities. For SVM, we focused on the *C* value, which specifies SVM's cost parameter (i.e., penalty parameter of the error) with a linear kernel function. Parameter tuning and optimal approach selection were based on the performance evaluation that will be discussed later. Apart from the specifications of these key factors, we adopted the default settings in RapidMiner Studio to implement various sentiment classification models for Chinese travel news.

Performance evaluation

To evaluate the sentiment classification models' performance, we needed to split the benchmark corpus into two subsets (i.e., training and testing). The training set, a dataset of known classification, was utilized to train classifiers to formulate complete classification models. The testing set, a dataset of unknown classification, was employed to test the trained classifiers' sentiment classification performance. There are several metrics to evaluate classification performance; the present study specifically adopted the *accuracy* metric, which has been extensively used in sentiment classification (Liu et al. 2013; Ye, Zhang, and Law

2009).

The accuracy metric measures the percentage of correctly classified news articles. It is computed using the following formula (1); Table 1 shows the confusion matrix.

$$\text{Accuracy} = \frac{\text{Number of correctly predicted news articles}}{\text{Number of all news ariticles}} = \frac{A+D}{A+B+C+D} \qquad (1)$$

============================

Please insert Table 1 about here.

============================

To increase the evaluation’s validity and reliability, we used a 10-fold cross-validation in place of a conventional validation estimating performance metric on a single testing set (Nassirtoussi et al. 2015). One round of cross-validation partitioned a data sample into two complementary subsets, which allowed us to analyze one subset and validate the prediction on the other. Multiple rounds of cross-validation were performed using different partitions, and the validation results over the rounds were averaged. The 10-fold cross-validation partitioned the original sample dataset into 10 subsets of equal size. In each round of training and testing, only one subset was used as the testing set; the remaining nine subsets were assembled as the training set. The cross-validation process was repeated 10 times, with each of the 10 subsets being used once as the testing set. The 10 results were then averaged to obtain an aggregate measure to estimate the news sentiment classification model’s performance. The 10-fold cross-validation process is illustrated in Figure 2.

============================

Please insert Figure 2 about here.

============================

Results and discussion

Overall results

To analyze possible interactions among the three key design factors (i.e., news feature extraction, news feature weighting, and news sentiment classifier) on predictive accuracy, this metalearning experiment followed a 3 × 3 × 2 factorial design. Accordingly, we formulated 18 alternative models, which were evaluated using 10-fold cross-validation. Table 2 summarizes the best average accuracy of the 18 news sentiment models across the three factors. The results revealed that BOW–TFIDF–SVM achieved the best performance. The accuracy of this combination peaked at 0.902, which shows promise for the model's application in future Chinese travel news analysis.

==============================

Please insert Table 2 about here.

==============================

Interactions among key design factors

A full factorial three-way ANOVA was employed for hypothesis testing and to identify the main and interaction effects among the three key design factors. The independent variables were news feature extraction (E), news feature weighting (W), and news classifier (C). The dependent variable was accuracy (A). E, W, C, E × W, E × C, W × C, and E × W × C were simultaneously entered into the ANOVA model. The Shapiro–Wilk W test indicated that the normal distribution assumption was satisfied for the independent variables. The significant interaction effects of the dependent variable were followed by an analysis of the simple effects using Tukey's analysis (Tukey 1991). Table 3 summarizes the three-way ANOVA results with a total R^2 of 0.731. These results lend strong support to H1, demonstrating a significant three-way interaction effect among the key design factors (E × W × C) on predictive accuracy [*F*(4,162) = 3.665, *p* = 0.007].

==============================

Please insert Table 3 about here.

==============================

(1) Interactions between Feature Weighting and Classifier

Table 3 shows strong interactions between W and C [*F*(2,162) = 4.711, *p* < 0.001], supporting H2. The pairwise comparisons of overall performance show that the NB classifier using the TF weighting approach (M = 0.843, SD = 0.17) achieved a higher accuracy than that using TFIDF approaches (M = 0.796, SD = 0.17, mean difference = 0.047, *p* = 0.006) and BO (M = 0.774, SD = 0.17, mean difference = 0.069, *p* < 0.001). However, the SVM classifier using the TFIDF weighting approach (M = 0.892, SD = 0.01) obtained higher accuracy compared to the TF (M = 0.871, SD = 0.01, mean difference = 0.022, *p* =0.03) and BO (M = 0.862, SD = 0.01, mean difference = 0.03, *p* = 0.003) approaches.

==============================

Please insert Figure 3 about here.

==============================

These results, and the patterns of mean accuracy depicted in Figure 3, underscore the importance of the weighting approach to classifiers. The TFIDF and BO weighting approaches led to a considerable increase in accuracy under the SVM classifier compared to under the NB classifier; see Figure 3(a). In contrast, the TF weighting approach led to a small increase in accuracy under the SVM classifier compared to under the NB classifier.

Significant three-way interaction effects of extracting, weighting, and classifier approaches were also observed [$F(4,162) = 3.665$, $p = 0.007$]. The results and patterns of mean accuracy depicted in Figure 3(b), (c), and (d) suggest that the interactions of the weighting approach and classifier also varied significantly with extraction approaches. In particular, Figure 3(b) shows that when using BOW as the extraction approach, the BO weighting approach achieved higher accuracy than the TF and TFIDF approaches under the NB classifier. The TFIDF weighting approach led to a significant increase in accuracy under the SVM classifier compared to under the NB classifier and performed best among the three weighting approaches. However, when using the unigram and bigram extraction approaches, the interaction patterns of weighting approaches and classifiers were similar to the overall results; see Figure 3(c) and (d), respectively.

(2) Interactions between Feature Extraction and Classifier

Table 3 indicates a strong interaction between E and C [$F(2,162) = 24.294$, $p < 0.001$], which supports H3. The pairwise comparisons of overall performance show that the NB classifier using the unigrams extraction approach (M = 0.735, SD = 0.013) demonstrated significantly lower accuracy than the BOW (M = 0.837, SD = 0.013, mean difference = −0.101, $p < 0.001$) and bigrams (M = 0.840, SD = 0.013, mean difference = −0.105, $p < 0.001$) approaches. Similarly, the SVM classifier using the unigrams extraction approach (M = 0.859, SD = 0.010) achieved significantly lower accuracy than the BOW (M = 0.887, SD = 0.010, mean difference = −0.027, $p = 0.007$) and bigrams approaches (M = 0.879, SD = 0.010, mean difference = −0.020, p =0.050).

==========================

Please insert Figure 4 about here.

==========================

These results and Figure 4 suggest the importance of the extraction approach to classifiers. The unigrams extraction approach led to a considerable increase in accuracy under the SVM classifier compared to under the NB classifier; see Figure 4(a). In contrast, the BOW and bigrams approaches led to relatively small increases in accuracy under the SVM classifier compared to under the NB classifier. The interaction patterns of extraction approaches and classifiers were similar to the overall results of the three weighting approaches (BO, TF, and TFIDF); see Figure 4(b), (c), and (d), respectively.

(3) Interactions between Feature Extraction and Weighting

The results shown in Table 3 also indicate a strong interaction between E and W [$F(2,162) = 15.818$, $p = 0.001$], providing support to H4. The pairwise comparisons of the overall performance show that the BO weighting approach with BOW (M = 0.854, SD = 0.20) and bigrams extraction approach (M = 0.841, SD = 0.20) achieved significantly higher accuracy than the unigrams approach (M = 0.760, SD = 0.20, mean difference = 0.094 and 0.081, $p < 0.001$). The TF weighting approach with BOW (M = 0.867, SD = 0.12) and bigrams (M = 0.867, SD = 0.12) extraction approaches achieved significantly higher accuracy than the unigrams approach (M = 0.836, SD = 0.12, mean difference = 0.031 and 0.031, p = 0.016). The TFIDF weighting approach with bigrams (M = 0.871, SD = 0.21) and BOW (M = 0.864, SD = 0.21) extraction approaches demonstrated higher accuracy than the unigrams approach (M = 0.796, SD = 0.21, mean difference = 0.075 and 0.068, p = 0.001 and 0.002).

==========================

Please insert Figure 5 about here.

==========================

These results and Figure 5 suggest the influence of the feature extraction approach on the feature weighting approach. Using the TF weighting approach, unigrams extraction led to a considerably larger increase in accuracy over BOW and bigrams compared to the BO approach; see Figure 5(a). However, when using the TFIDF weighting approach, the unigrams extraction approach led to a substantial decrease in accuracy over BOW and bigrams compared with the TF approach.

Significant three-way interaction effects of extracting, weighting, and classifier approaches were observed [$F(4,162) = 3.665$, $p = 0.007$]. Figure 5(b) and (c) suggest that the interactions of the extraction and weighting approaches varied by classifier. Under the NB classifier, the interaction patterns of extraction and weighting approaches were similar to the overall results as shown in Figure 5(b). Yet under the SVM classifier, changes in weighting approaches led to somewhat similar variations in accuracy across the three extraction approaches as depicted in Figure 5(c).

In summary, all hypotheses proposed in this study were supported by the experiment; hence, we can draw the following conclusions. First, the unigrams extraction approach is sensitive to classifiers and to weighting approaches, particularly under the NB classifier. Second, the BOW and bigrams extraction approaches achieve relatively better overall accuracy than unigrams across different weighting approaches and classifiers. Third, the performance of the TFIDF weighting approach is more sensitive to classifiers compared to the BO and TF approaches. Fourth, the TF and TFIDF approaches achieve relatively good performance; they outperform the BO approach across the different extraction approaches. Fifth, the TFIDF performs best among the three weighting approaches under the SVM classifier regardless of extraction approach, whereas the TF approach is well suited to the NB classifier.

Parameter tuning for SVM

We used the preceding analyses to conclude that the combination of BOW–TFIDF–SVM may generate the best classification performance. We further elaborated penalty parameter *C* tuning for SVM. The value of penalty parameter *C* is the most important parameter for a linear kernel (Chen and Wang 2007). Researchers have suggested that a reasonable range of log C is $[-1, 3]$ (Chang and Lin 2011). In addition, selecting a large C value may increase the risk of the model overfitting the training data (Huang and Kecman 2005). Accordingly, we tuned the C value from 0.05 to 10 in 200 step-wise calculations. Given the feature set generated by BOW as the feature extraction approach and TFIDF as the feature weighting approach, accuracy performance with different C values is shown in Figure 6. The accuracy became stable when C was greater than 1, and accuracy peaked at a C value of 5.25.

==============================

Please insert Figure 6 about here.

==============================

A metalearning framework for sentiment analytics

The strong interaction effect among key design factors of sentiment analytics implies that we should consider these factors holistically rather than in a piecemeal manner. Therefore, we formulated our experiment process into a metalearning framework based on the study by Rice (1976). As shown in Figure 7, the framework has four essential components: (1) task space *T* represents the set of instances of sentiment analytics tasks, such as sentiment analytics for Chinese travel news; (2) factor space *F* contains the design factors that influence classification performance; (3) approach space *A* is a set of all considered approaches (e.g., news feature extraction approaches) to address the identified factors; and (4) performance space *P* represents a set of performance metrics, such as predictive accuracy, to measure classification performance.

============================

Please insert Figure 7 about here.

============================

The metalearning framework reflects two important elements of metalearning: awareness and control. Awareness of the classification learning process includes extracting design factors, identifying potential approaches, and defining performance metrics. Control of the classification learning process involves controlling interactions during the factor–approach selection process to maximize sentiment classification performance.

Our experiment provides a precise outline for implementing this framework, and the proposed sentiment analysis model demonstrated better predictive performance than currently available sentiment analysis tools. We also compared our model with two popular sentiment analysis tools: Semantria and ROST Content Mining. Semantria is a popular commercial online sentiment analysis software that performs well with online hotel reviews (Gao, Hao, and Fu 2015; Serrano-Guerrero et al. 2015). ROST Content Mining is a sentiment analysis tool based on a general sentiment lexicon, which has achieved good performance with Chinese online reviews (Luo and Zhai 2017). The accuracies of Semantria and ROST Content Mining with respect to our news corpus are 0.690 and 0.580, significantly lower than the proposed model (0.902). It is also worth noting that the metalearning framework need not be limited to sentiment analytics. Other computational methods may benefit from this framework to improve predictive accuracy when applied in future tourism research.

Conclusion and implications

Computational social science is gaining much attention of tourism researchers. However, a partial hiatus exists between innovative theories and conventional approaches in tourism research practice (Chang, Kauffman, and Kwon 2014; Cohen and Cohen 2012). This study attempted to bridge this methodological hiatus by examining the design effects on predictive accuracy of sentiment analytics for Chinese travel news. In particular, through a metalearning lens, we identified an optimal combination of key design factors inherent to sentiment analytics for Chinese travel news. We evaluated the main and interaction effects of these design factors on the predictive accuracy of sentiment classification using a full-factorial experiment. Results demonstrated significant interactions among the factors of feature extraction, feature weighting, and sentiment classifier in terms of the predictive accuracy of sentiment analytics. Such strong interaction effects among these factors necessitate a holistic metalearning framework for tourism researchers to apply computational methods in future tourism studies.

This research has several limitations. First, generalization of our findings is limited to the balanced benchmark dataset and context of this experiment, which might create subtle nuances in practical tourism applications. Future research should attempt to replicate and extend our results using different balanced and unbalanced datasets in various tourism domains. Second, this study focused on machine learning-based sentiment analytics; future work should study the predictive accuracy of other sentiment analytics approaches (Alaei, Becken, and Stantic 2017). Third, this study employed only one popular performance metric to evaluate design factor effects; however, different performance metrics may influence design factor configurations (Kirilenko et al. 2017). Subsequent research should involve relevant comparisons across different performance metrics, including human raters.

Notwithstanding these limitations, this study aimed to make significant methodological contributions to tourism research. Despite the data soundness and practical relevance of computational social research, the inherent complexity of fieldwork is rarely evident in current tourism research. Our research has transparently exemplified the complex internal dynamics in sentiment analytics for tourism. The revealed significant interactions among design factors in terms of predictive accuracy should raise tourism researchers' methodological awareness in the following respects:

1) *Interpretation and prediction*. Historically, researchers have tended to explore tourism phenomena by establishing interpretable causal mechanisms. The increasingly computational nature of social science requires researchers to recognize the importance of predictive accuracy and to effectively integrate interpretation with prediction in the pursuit of tourism knowledge.

2) *Transparency and replicability*. Tourism research involving computational methods, such as sentiment analytics, should carefully select design factors and report design choices to improve the replicability of scientific claims (Hofman, Sharma, and Watts 2017; Simmons, Nelson, and Simonsohn 2011).

3) *Limits to prediction*. To fully understand predictive accuracy, we need not only relevant comparisons across different performance metrics, but also an understanding of the best possible performance. Due to the complexity of human behavior, the predictive accuracy of the proposed model is limited by methodology and theory development and subject to the human nature of the phenomenon being studied (Hofman, Sharma, and Watts 2017). Comparing different approaches to sentiment analytics with human raters (Kirilenko et al. 2017) is a valuable way to contextualize such limitations.

We also proposed a metalearning framework to holistically consider the effects of design factors and improve the predictive accuracy of sentiment analytics. Although these reflections represent a set of empirical inferences related to the current tourism literature, they have the potential to be generalized further to provide guidance on research design and methodological choice.

This study also offers significant practical implications. Although this research was methodological and focused on a specific topic, it could contribute to a new line of tourism research that utilizes unstructured data, potentially advancing our understanding of several key practical issues such as destination image, resident attitude, and tourist experience. Social listening has become a critical task for tourism service providers, and the proposed metalearning framework could improve the predictive accuracy of such tasks. In addition, the best combination of design factors revealed in our study serves as a useful reference for policymakers and industry practitioners when conducting practical sentiment analytics, particularly when discovering sentiment in massive volumes of news articles.

References

Alaei, Ali Reza, Susanne Becken, and Bela Stantic. 2017. “Sentiment Analysis in Tourism: Capitalizing on Big Data.” *Journal of Travel Research*, 0047287517747753. doi:10.1177/0047287517747753.

Biggs, J. B. 1985. “The Role of Metalearning in Study Processes.” *British Journal of Educational Psychology* 55: 185–212.

Bravo-Marquez, Felipe, Marcelo Mendoza, Barbara Poblete, Mendoza Marcelo, and Poblete Barbara. 2014. “Meta-Level Sentiment Models for Big Social Data Analysis.” *Knowledge-Based Systems* 69 (1): 86–99. doi:10.1016/j.knosys.2014.05.016.

Cambria, Erik. 2016. “Affective Computing and Sentiment Analysis.” *IEEE Intelligent Systems* 31 (2): 102–7. doi:10.1109/MIS.2016.31.

Capriello, Antonella, Peyton R. Mason, Boyd Davis, and John C. Crotts. 2013. “Farm Tourism Experiences in Travel Reviews: A Cross-Comparison of Three Alternative Methods for Data Analysis.” *Journal of Business Research*, International Tourism Behavior in Turbulent Times, 66 (6): 778–85. doi:10.1016/j.jbusres.2011.09.018.

Chang, Chih-chung, and Chih-jen Lin. 2011. “LIBSVM: A Library for Support Vector Machines.” *ACM Transactions on Intelligent Systems and Technology* 2: 1–27. doi:Artn 27 10.1145/1961189.1961199.

Chang, Ray M., Robert J. Kauffman, and YoungOk Kwon. 2014. “Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data.” *Decision Support Systems* 63: 67–80. doi:10.1016/j.dss.2013.08.008.

Chen, Kuan-yu, and Cheng-hua Wang. 2007. “A Hybrid SARIMA and Support Vector

Machines in Forecasting the Production Values of the Machinery Industry in Taiwan." *Expert Systems with Applications* 32: 254–64. doi:10.1016/j.eswa.2005.11.027.

Chen, Long-Sheng, Cheng-Hsiang Liu, and Hui-Ju Chiu. 2011. "A Neural Network Based Approach for Sentiment Classification in the Blogosphere." *Journal of Informetrics* 5: 313--322. doi:10.1016/j.joi.2011.01.003.

Chiu, Chaochang, Nan-Hsing Chiu, Re-Jiau Sung, and Pei-Yu Hsieh. 2015. "Opinion Mining of Hotel Customer-Generated Contents in Chinese Weblogs." *Current Issues in Tourism* 18 (5): 477–95. doi:10.1080/13683500.2013.841656.

Chow, C. K. K., S. K. W. Chu, S. H. Ng, C. S. J. Fong, W. Y. Kwan, and A. A. T. Leung. 2007. "WiseNews Database for Primary Four Inquiry-Based Learning Projects." In *Conference on Integrated Learning*, 14–15. Hong Kong: The Hong Kong Institute of Education.

Cohen, Erik, and Scott A. Cohen. 2012. "Current Sociological Theories and Issues in Tourism." *Annals of Tourism Research* 39 (4): 2177–2202. doi:10.1016/j.annals.2012.07.009.

Deery, Margaret, Leo Jago, and Liz Fredline. 2012. "Rethinking Social Impacts of Tourism Research: A New Research Agenda." *Tourism Management* 33 (1): 64–73. doi:10.1016/j.tourman.2011.01.026.

Duan, Wenjing, Yang Yu, Qing Cao, and Stuart Levy. 2016. "Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach." *Cornell Hospitality Quarterly* 57 (3): 282–96.

Ekbia, Hamid, Michael Mattioli, Inna Kouper, G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R. Sugimoto. 2015. “Big Data, Bigger Dilemmas: A Critical Review.” *Journal of the Association for Information Science and Technology* 66 (8): 1523–45. doi:10.1002/asi.23294.

Flanagin, A. J., and M. J. Metzger. 2000. “Perceptions of Internet Information Credibility.” *Journalism & Mass Communication Quarterly* 77: 515–40.

Gao, Shanshan, Jinxing Hao, and Yu Fu. 2015. “The Application and Comparison of Web Services for Sentiment Analysis in Tourism.” In *12th International Conference on Service Systems and Service Management, ICSSSM 2015,*. IEEE. doi:10.1109/ICSSSM.2015.7170341.

Geetha, M., Pratap Singha, and Sumedha Sinha. 2017. “Relationship between Customer Sentiment and Online Customer Ratings for Hotels - An Empirical Analysis.” *Tourism Management* 61: 43–54. doi:10.1016/j.tourman.2016.12.022.

Hagenau, Michael, Michael Liebmann, and Dirk Neumann. 2013. “Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Capturing Features.” *Decision Support Systems* 55: 685–97. doi:10.1016/j.dss.2013.02.006.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. “Prediction and Explanation in Social Systems.” *Science* 355 (6324): 486–88. doi:10.1126/science.aal3856.

Hu, Ya-Han, and Kuanchin Chen. 2016. “Predicting Hotel Review Helpfulness: The Impact of Review Visibility, and Interaction between Hotel Stars and Review Ratings.” *International Journal of Information Management* 36 (6): 929–44. doi:10.1016/j.ijinfomgt.2016.06.003.

Huang, Te Ming, and Vojislav Kecman. 2005. “Gene Extraction for Cancer Diagnosis by

Support Vector Machines--an Improvement.” *Artificial Intelligence in Medicine* 35: 185–94. doi:10.1016/j.artmed.2005.01.006.

Jin, Xin, and Ying Wang. 2016. “Chinese Outbound Tourism Research: A Review.” *Journal of Travel Research* 55 (4): 440–53. doi:10.1177/0047287515608504.

Kang, Hanhoon, Seong Joon Yoo, and Dongil Han. 2012. “Senti-Lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews.” *Expert Systems with Applications* 39 (5): 6000–6010. doi:10.1016/j.eswa.2011.11.107.

Kim, Kyungmi, Muzaffer Uysal, and M. Joseph Sirgy. 2013. “How Does Tourism in a Community Impact the Quality of Life of Community Residents?” *Tourism Management* 36 (June): 527–40. doi:10.1016/j.tourman.2012.09.005.

Kirilenko, Andrei P., Svetlana O. Stepchenkova, Hany Kim, and Xiang (Robert) Li. 2017. “Automated Sentiment Analysis in Tourism: Comparison of Approaches.” *Journal of Travel Research*, 0047287517729757. doi:10.1177/0047287517729757.

Landis, J. R., and G. G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33: 159–74. doi:10.2307/2529310.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. “Computational Social Science.” *Science* 323 (5915): 721–23. doi:10.1126/science.1167742.

Lee, Chou-Yuan, and Zne-Jung Lee. 2012. “A Novel Algorithm Applied to Classify Unbalanced Data.” *Applied Soft Computing* 12 (8): 2481–85. doi:10.1016/j.asoc.2012.03.051.

Lee, Minwoo, Miyoung Jeong, and Jongseo Lee. 2017. “Roles of Negative Emotions in

Customers’ Perceived Helpfulness of Hotel Reviews on a User-Generated Review Website.” *International Journal of Contemporary Hospitality Management* 29 (2): 762–83. doi:10.1108/IJCHM-10-2015-0626.

Lemke, Christiane, Marcin Budka, and Bogdan Gabrys. 2015. “Metalearning: A Survey of Trends and Technologies.” *Artificial Intelligence Review* 44 (1): 117–30. doi:10.1007/s10462-013-9406-y.

Liu, Bing. 2012. “Sentiment Analysis and Opinion Mining.” *Synthesis Lectures on Human Language Technologies* 5 (1): 1–167. doi:10.2200/S00416ED1V01Y201204HLT016.

Liu, Bingjie, and Lori Pennington-Gray. 2015. “Bed Bugs Bite the Hospitality Industry? A Framing Analysis of Bed Bug News Coverage.” *Tourism Management* 48: 33–42. doi:10.1016/j.tourman.2014.10.020.

Liu, Shaowu, Rob Law, Jia Rong, Gang Li, and John Hall. 2013. “Analyzing Changes in Hotel Customers’ Expectations by Trip Mode.” *International Journal of Hospitality Management* 34 (1): 359–71. doi:10.1016/j.ijhm.2012.11.011.

Luo, Qiuju, and Xueting Zhai. 2017. “‘I Will Never Go to Hong Kong Again!’ How the Secondary Crisis Communication of ‘Occupy Central’ on Weibo Shifted to a Tourism Boycott.” *Tourism Management* 62: 159–72. doi:10.1016/j.tourman.2017.04.007.

Marrese-Taylor, Edison, Juan D. Velásquez, and Felipe Bravo-Marquez. 2014. “A Novel Deterministic Approach for Aspect-Based Opinion Mining in Tourism Products Reviews.” *Expert Systems with Applications* 41 (17): 7764–75. doi:10.1016/j.eswa.2014.05.045.

Moraes, Rodrigo, João Francisco Valiati, and Wilson P. Gavião Neto. 2013. “Document-

Level Sentiment Classification: An Empirical Comparison between SVM and ANN." *Expert Systems with Applications* 40 (2): 621–33. doi:10.1016/j.eswa.2012.07.059.

Moznette, James, and Galen Rarick. 1968. "Which Are More Readable: Editorials or News Stories?" *Journalism Quarterly* 45 (June): 319–21. doi:10.1177/107769906804500214.

Nassirtoussi, A. K., S. Aghabozorgi, Y. W. Teh, and D. C. L. Ngo. 2015. "Text Mining of News-Headlines for FOREX Market Prediction: A Multi-Layer Dimension Reduction Algorithm with Semantics and Sentiment." *Expert Systems with Applications* 42 (January): 306–24. doi:10.1016/j.eswa.2014.08.004.

Nunkoo, Robin, and Kevin Kam Fung So. 2016. "Residents' Support for Tourism: Testing Alternative Structural Models." *Journal of Travel Research* 55 (7): 847–61. doi:10.1177/0047287515592972.

Pang, Bo, and Lillian Lee. 2004. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271–271. doi:10.3115/1218955.1218990.

———. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2 (1–2): 1–135.

Peel, Victoria, and Adam Steen. 2007. "Victims, Hooligans and Cash-Cows: Media Representations of the International Backpacker in Australia." *Tourism Management* 28 (4): 1057–67. doi:10.1016/j.tourman.2006.08.012.

Philander, Kahlil, and Yun Ying Zhong. 2016. "Twitter Sentiment Analysis: Capturing

Sentiment from Integrated Resort Tweets." *International Journal of Hospitality Management* 55: 16–24. doi:10.1016/j.ijhm.2016.02.001.

Rice, John R. 1976. "The Algorithm Selection Problem." *Advances in Computers* 15: 65–118. doi:10.1016/S0065-2458(08)60520-3.

Sanz-Blas, Silvia, and Daniela Buzova. 2016. "Guided Tour Influence on Cruise Tourist Experience in a Port of Call: An EWOM and Questionnaire-Based Approach." *International Journal of Tourism Research* 18 (6): 558–66. doi:10.1002/jtr.2073.

Schmunk, Sergej, Wolfram Höpken, Matthias Fuchs, and Maria Lexhagen. 2013. "Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC." In *Information and Communication Technologies in Tourism 2014*, 253–65. Springer, Cham. doi:10.1007/978-3-319-03973-2_19.

Schuckert, Markus, Xianwei Liu, and Rob Law. 2015. "Hospitality and Tourism Online Reviews: Recent Trends and Future Directions." *Journal of Travel & Tourism Marketing* 32 (5): 608–21. doi:10.1080/10548408.2014.933154.

Serrano-Guerrero, Jesus, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. 2015. "Sentiment Analysis: A Review and Comparative Analysis of Web Services." *Information Sciences* 311: 18–38. doi:10.1016/j.ins.2015.03.040.

Sharpley, Richard. 2014. "Host Perceptions of Tourism: A Review of the Research." *Tourism Management* 42 (June): 37–49. doi:10.1016/j.tourman.2013.10.007.

Shi, Han Xiao, and Xiao Jun Li. 2011. "A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning." In *Proceedings of International Conference on Machine Learning and Cybernetics*, 950–54. Guilin, China.

doi:10.1109/ICMLC.2011.6016866.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.

Smith-Miles, Kate A. 2008. "Cross-Disciplinary Perspectives on Meta-Learning for Algorithm Selection." *ACM Computing Surveys* 41 (1): 6:1–6:25. doi:10.1145/1456650.1456656.

Stepchenkova, Svetlana, and James S. Eales. 2011. "Destination Image as Quantified Media Messages: The Effect of News on Tourism Demand." *Journal of Travel Research* 50 (2): 198–212. doi:10.1177/0047287510362780.

Tan, Songbo, Jin Zhang, and Zhang Jin. 2008. "An Empirical Study of Sentiment Analysis for Chinese Documents." *Expert Systems with Applications* 34: 2622–29. doi:10.1016/j.eswa.2007.05.028.

Tukey, John W. 1991. "The Philosophy of Multiple Comparisons." *Statistical Science* 6: 100–116.

Uysal, Muzaffer, M. Joseph Sirgy, Eunju Woo, and Hyelin (Lina) Kim. 2016. "Quality of Life (QOL) and Well-Being Research in Tourism." *Tourism Management* 53: 244–61. doi:10.1016/j.tourman.2015.07.013.

Van De Kauter, Marjan, Diane Breesch, and Véronique V. ronique Hoste. 2015. "Fine-Grained Analysis of Explicit and Implicit Sentiment in Financial News Articles." *Expert Systems with Applications* 42: 4999–5010. doi:10.1016/j.eswa.2015.02.007.

Wang, Hongwei, Pei Yin, Jiani Yao, and James N. K. Liu. 2013. “Text Feature Selection for Sentiment Classification of Chinese Online Reviews.” *Journal of Experimental & Theoretical Artificial Intelligence* 25 (4): 425–39. doi:10.1080/0952813X.2012.721139.

Wu, Desheng Dash, Lijuan Zheng, and David L. Olson. 2014. “A Decision Support Approach for Online Stock Forum Sentiment Analysis.” *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44: 1077–87. doi:10.1109/TSMC.2013.2295353.

Xia, Rui, Chengqing Zong, and Shoushan Li. 2011. “Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification.” *Information Sciences* 181: 1138–52. doi:10.1016/j.ins.2010.11.023.

Xiang, Zheng, Qianzhou Du, Yufeng Ma, and Weiguo Fan. 2017. “A Comparative Analysis of Major Online Review Platforms : Implications for Social Media Analytics in Hospitality and Tourism.” *Tourism Management* 58: 51–65. doi:10.1016/j.tourman.2016.10.001.

Xu, Ruifeng, Yunqing Xia, Kam Fai Wong, and Wenjie Li. 2008. “Opinion Annotation in On-Line Chinese Product Reviews.” In *International Conference on Language Resources and Evaluation, LREC 2008*, 1625–32. Marrakech, Morocco.

Ye, Qiang, Ziqiong Zhang, and Rob Law. 2009. “Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches.” *Expert Systems with Applications* 36: 6527–35. doi:10.1016/j.eswa.2008.07.035.

Zhang, Yan, Yong Zhang, Jennifer Xu, Chunxiao Xing, and Hsinchun Chen. 2016. “Sentiment Analysis on Chinese Health Forums: A Preliminary Study of Different

Language Models." In *International Conference on Smart Health*, 68–81. Springer, Cham. doi:10.1007/978-3-319-29175-8_7.

Zhang, Ziqiong, Qiang Ye, Zili Zhang, and Yijun Li. 2011. "Sentiment Classification of Internet Restaurant Reviews Written in Cantonese." *Expert Systems with Applications* 38: 7674–82. doi:10.1016/j.eswa.2010.12.147.

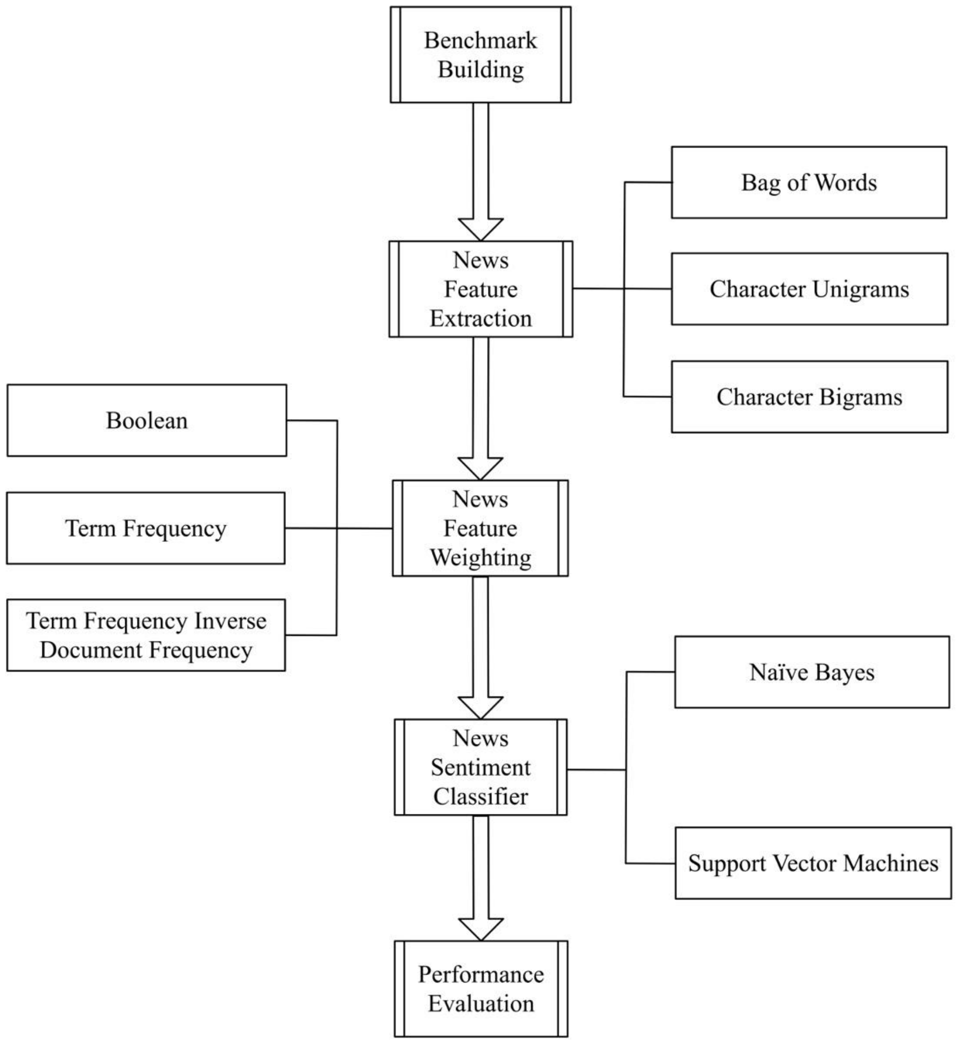


Figure 1. Sentiment analytics process for Chinese travel news.

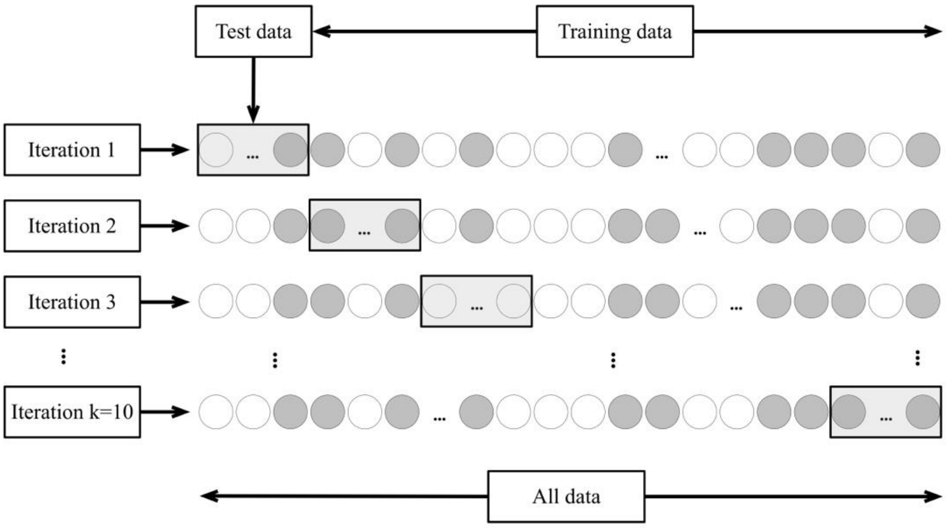
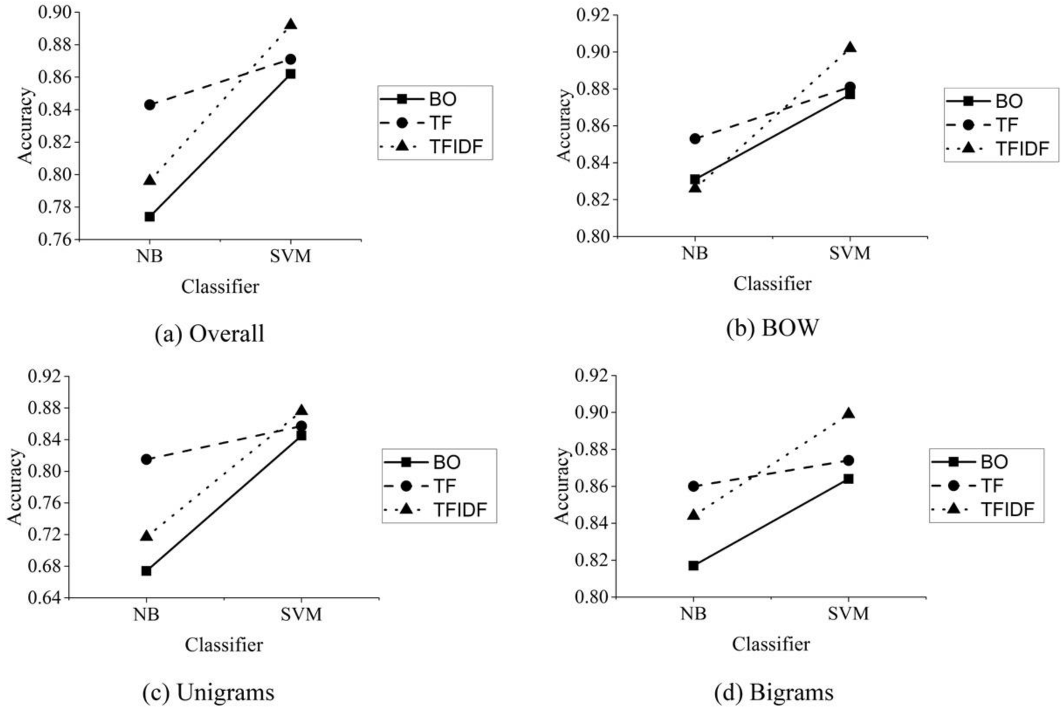


Figure 2. Illustration of the k-fold cross-validation.



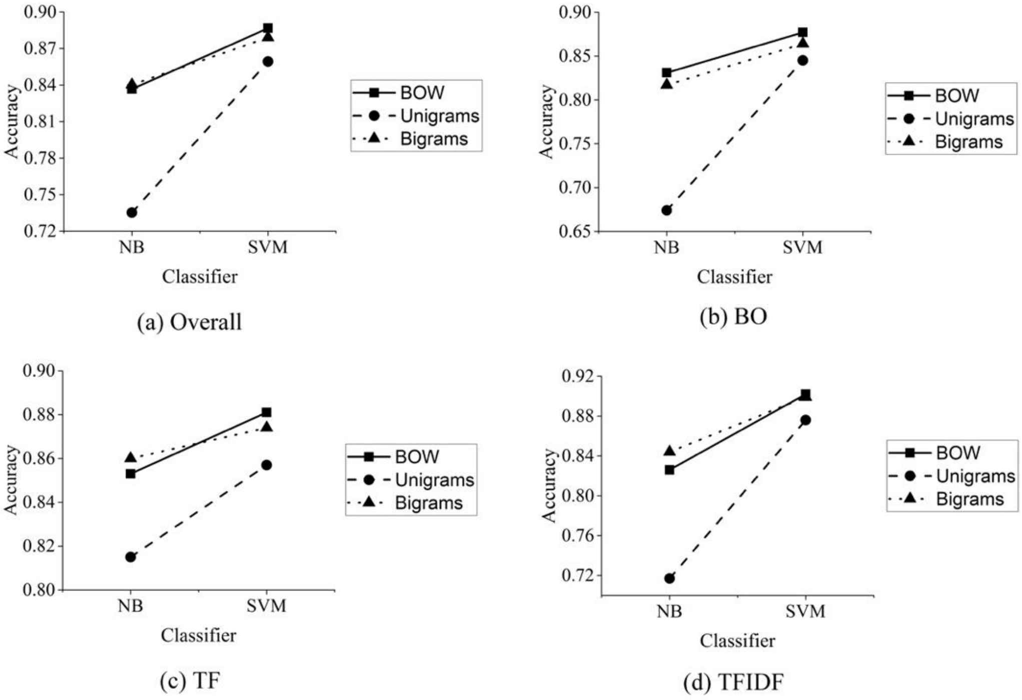
(a) Overall

(b) BOW

(c) Unigrams

(d) Bigrams

Figure 3. Interactions between design factors of feature weighting and classifier.



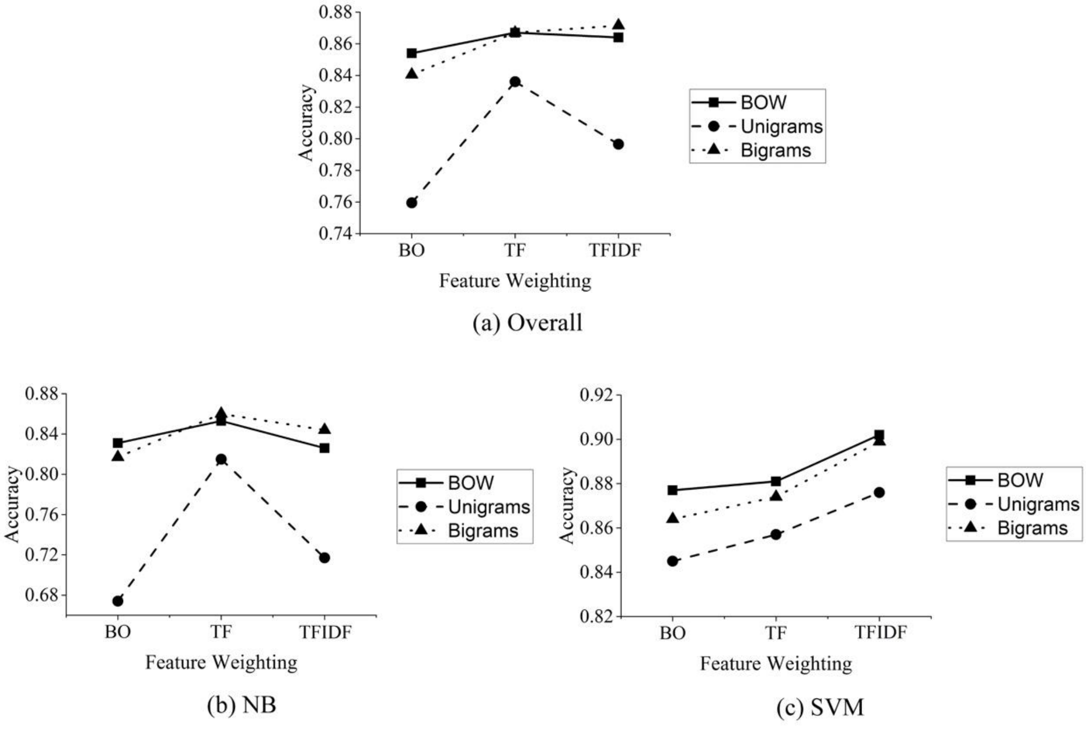
(a) Overall

(b) BO

(c) TF

(d) TFIDF

Figure 4. Interactions between design factors of feature extraction and classifier.



(a) Overall

(b) NB

(c) SVM

Figure 5. Interactions between design factors of feature extraction and weighting.

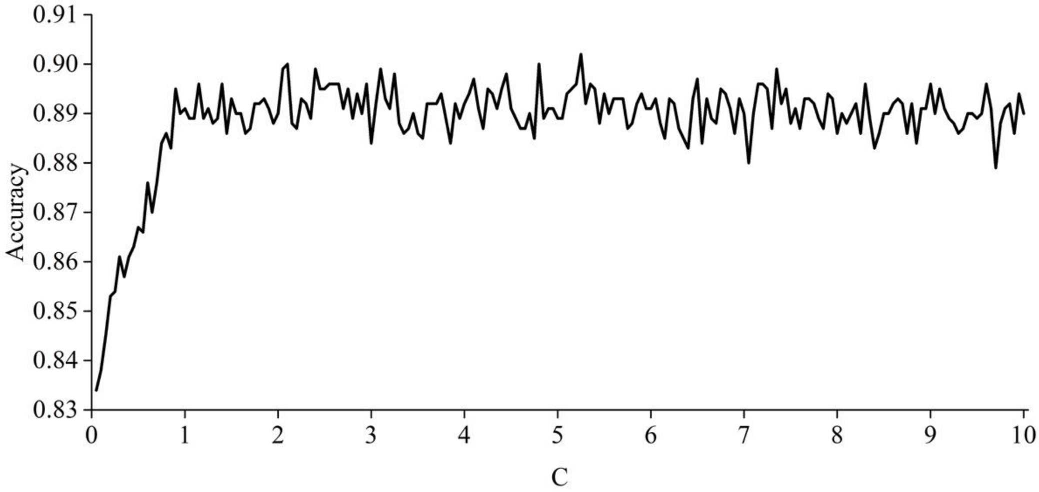


Figure 6. Parameter tuning for the SVM classifier.

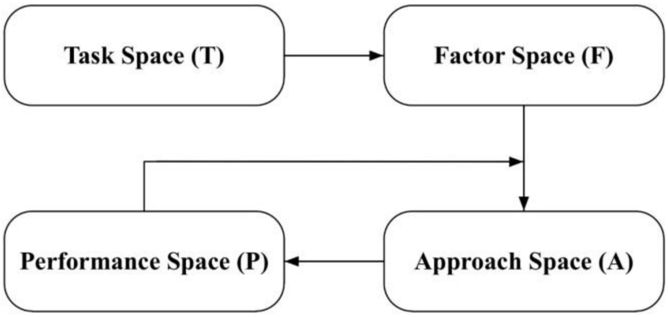


Figure 7. A metalearning framework for sentiment analytics.

Table 1. Classification Confusion Matrix.

	Positive (actual)	Negative (actual)
Positive (predicted)	A	B
Negative (predicted)	C	D

Table 2. Best Average Accuracy of News Sentiment Classification Models across Design Factors.

Variable		Accuracy		
		NB	SVM	*n*
BOW	BO	0.831	0.877	10
	TF	0.853	0.881	10
	TFIDF	0.826	0.902	10
Unigrams	BO	0.674	0.845	10
	TF	0.815	0.857	10
	TFIDF	0.717	0.876	10
Bigrams	BO	0.817	0.864	10
	TF	0.860	0.874	10
	TFIDF	0.844	0.899	10

Note: *n* represents the fold number.

Table 3. Summary of the Three-way ANOVA Analysis.

Variables		Accuracy (A)		
		F measure	Degree of freedom (df)	*p*
Main effects	E	60.542	2,162	< 0.001
	W	17.583	2,162	< 0.001
	C	170.533	1,162	< 0.001
Interaction effects	E × W	15.818	4,162	0.001
	E × C	24.294	2,162	< 0.001
	W × C	4.711	2,162	< 0.001
	E × W × C	3.665	4,162	0.007