Empirical Research Article

Automated Sentiment Analysis in Tourism: Comparison of Approaches

Journal of Travel Research
2018, Vol. 57(8) 1012–1025

DOI: 10.1177/0047287517729757
journals.sagepub.com/home/jtr
SAGE

Andrei P. Kirilenko[1], Svetlana O. Stepchenkova[1], Hany Kim[2], and Xiang (Robert) Li[3]

Abstract

Interest in applying Big Data to tourism is increasing, and automated sentiment analysis has been used to extract public opinion from various sources. This article evaluates the suitability of different types of automated classifiers for applications typical in tourism, hospitality, and marketing studies by comparing their performance to that of human raters. While the commonly used performance indices suggest that on easier-to-classify data sets machine learning methods demonstrate performance comparable to that by human raters, other performance measures such as Cohen's kappa show that the results of machine learning are still inferior to manual processing. On more difficult and noisy data sets, automated analysis has poorer performance than human raters. The article discusses issues pertinent to selection of appropriate sentiment analysis software and offers a word of caution against using automated classifiers uncritically.

Keywords

sentiment analysis, automated classification, social networks, surveys, recommender system

Introduction

The emergence of user-generated content (UGC) on the Internet in the mid-2000s known as Web 2.0 has provided a new source of data for natural and social scientists and industry. Millions of visitors exchange content on popular platforms for mutual benefit, such as social networking (e.g., Facebook), content sharing (e.g., Reddit), blogging (e.g., LiveJournal), micro-blogging (e.g., Twitter), multimedia sharing (e.g., YouTube), location sharing (e.g., FourSquare), review forums (e.g., TripAdvisor), and other sites. A pioneering survey of Internet opinion mining and sentiment analysis approaches by Pang and Lee (2008) found that online opinions and reviews have a significant influence on consumer behavior. For example, consumers were willing to pay 20%–100% more for an item rated 5 stars by fellow consumers in comparison to a 4-star-rated item. Given the relevance of UGC for managerial decisions, the task of extracting specifically defined data from UGC efficiently and with high reliability and accuracy is paramount. However, information specified for extraction is often unstructured. Schmunk et al. (2013, 254) note:

> The arising challenge for tourism managers is to find relevant reviews and analyze them efficiently, which necessitates an automatic extraction of decision-relevant knowledge from . . . UGC with a sufficient quality. Although, many review sites offer scalar ratings, such ratings do not provide any information on specific product characteristics that customers like or don't like. Such information is typically contained within textual reviews and has, thus, to be extracted by techniques from the areas of opinion mining and sentiment analysis.

Thus, data mining techniques directed toward opinion mining and sentiment analysis are gaining importance.

In response to the demand from researchers and businesses, new methods of data storage, retrieval, and analysis referred to as "Big Data analytics" are being developed. The term "Big Data" was famously described by Laney (2001) using three dimensions: volume, variety, and velocity, which are commonly referred to as the "3Vs." The velocity dimension deals with the acquisition and processing of a large volume of data in a short time. This dimension is perhaps less critical for tourism applications, with the exception of a limited set of applications where urgent response is of the utmost importance, such as crisis management. The variety and

[1]Department of Tourism, Recreation & Sport Management, College of Health and Human Performance, University of Florida, Gainesville, FL, USA
[2]Department of Business and Tourism, Mount Saint Vincent University, Halifax, Nova Scotia, Canada
[3]Department of Tourism and Hospitality Management, Temple University, Philadelphia, PA, USA

Corresponding Author:
Andrei P. Kirilenko, Department of Tourism, Recreation & Sport Management, College of Health and Human Performance, University of Florida, 240B Florida Gym, P.O. Box 118208, Gainesville, FL 32611-8208, USA.
Email: andrei.kirilenko@ufl.edu

volume dimensions of data are crucial, however. A typical application may involve the processing of millions of records acquired from social media. Estimates by Cukier (2010) suggest that only 5% of existing data is structured—that is, appearing in the form of spreadsheets or relational databases—while the bulk of the records include unstructured text, imagery, and video data. Because of the unstructured nature of Big Data, traditional sampling and statistical analysis methods are not always suitable, and Big Data analytics unifies a diverse set of methods specifically targeted at finding patterns in the large volumes of data. These methods include predictive analytics, data mining, artificial intelligence, natural language processing, and other tool categories (Russom 2011). This article addresses only a small set of these tools used in connection with sentiment analysis of textual data, which is a task that is frequently encountered in tourism research.

Theoretical Foundations of Automated Sentiment Analysis

"Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes" (Liu, Bing, and Zhang 2012, 215). Sentiment analysis is conceptually grounded in the works of Osgood and his associates on content analysis of people's evaluative judgments and affective responses to stimuli in the texts (Heise 1970). Osgood and colleagues distinguish between three dimensions of meaning: Evaluation, Potency, and Activity (EPA system) (Osgood, Tannenbaum, and Suci 1957). Cognitive appraisals (e.g., good–bad or favorable–unfavorable) lie along the Evaluation dimension of meaning, intensity of these evaluations (e.g., strong–weak or powerful–powerless) constitutes Potency, and such aspects as, for example, fast–slow or active–passive comprise the Activity dimension (Osgood, Tannenbaum, and Suci 1957). The EPA system gave rise to the attitude research with semantic differential measures anchored by contrasting adjectives at each end and numerous studies supported validity of the approach (Heise 1970). However, not only adjectives but also concepts can be placed along the EPA dimensions. For example, concepts such as family, piece, and beauty are at the positive end on the Evaluation dimension while hate and disease are located at the negative pole. The concept of war is at the negative pole of the Evaluation dimension, but high on Potency, and Activity dimensions, while baby in the EPA space is located toward the positive evaluative pole and low potency pole. Research has found the stability of EPA structure across various cultures (Osgood 1964; Jakobovits 1966).

Currently, a large amount of works on sentiment analysis involves determining valence, which can be roughly equated with the evaluative EPA dimension; that is, where the sentiment is located on the positive–negative scale (e.g., Pang and Lee 2008; Liu 2015). There has been growing interest in automated sentiment analysis on detecting both valence and emotions, especially in the texts obtained from social media platforms (Mohammad 2015). However, a number of emotions and affective states can be assigned positive or negative valence as well; for example, joy is considered as carrying positive valence and, thus, indicates positive sentiment while anger is indicative of a negative sentiment. The intensity of the sentiment can be measured by how far from a neutral point on the positive–negative dimension a concept is located. This idea that various concepts, descriptors, and affective states have valence and, thus, can be assigned a score on a positive–negative dimension, lies at the foundation of the automated sentiment analysis (e.g., Pang and Lee 2008).

Methodological Approaches to Automated Sentiment Analysis

The main goal of sentiment analysis is to extract a sentiment expressed in a document toward a certain aspect based on the subjectivity and the linguistic characteristics of the words within an unstructured text (García et al. 2012). In other words, sentiment analysis is a process of computationally identifying and categorizing people's opinions in order to determine the writer's attitude toward a particular issue. The computer-based natural language processing aims to detect sentiment in attitudes to make "an account of emotion that could in principle be used in an Artificial Intelligence (AI) system that would, for example, be able to reason about emotion" (Ortony et al. 1988, 2). For example, Whitelaw et al. (2005) suggested to detect three different groups of attitudes: affect (subjective emotional state, e.g., "gloomy"), appreciation (assessment of the reviewed object, e.g., "inferior"), and social judgment (e.g., "famous"). The majority of practically available packages, however, concentrate on holistic classification of the sentiment expressed in a document, topic, or product that is on a scale ranging from negative to positive.

The approaches for automated sentiment analysis draw from two competing ideas: lexical and nonlexical text classification (Thelwall et al. 2010). The lexical approach starts with a set of words, for which a typical sentiment (positive, negative, or neutral) is predefined. This set may be created manually or semiautomatically (Taboada et al. 2011). The sentiment of the entire textual unit under analysis is derived based on the balance of words with known negative and positive sentiment, subjected to linguistic rules (e.g., good–positive; not good–negative; see Miller 1995). The nonlexical approach is based on machine learning, where an algorithm of assigning sentiment is trained on a thematically close text corpus. Multiple machine learning classification approaches can be applied to task, but the Support Vector Machines and Naïve Bayes are the most frequently used (Taboada et al. 2011; for detailed description of the algorithms, see Murphy 2012).

Both approaches require initial manual classification of the words or documents according to the expressed sentiment. In this respect, the lexicon-based approach allows to use the already developed dictionaries specific for the document's language, for example, the widely used SentiWordNet for documents in English (Baccianella, Esuli, and Sebastiani 2010). A generic dictionary, however, does not account for differences in sentiment specific to different domains. For example, in a movie review, movie director's mention tends to coincide with a negative sentiment (Taboada et al. 2011). The details such as domain differences are impossible to include in a lexicon-based approach, but they are automatically accounted for in a machine learning–based algorithm.

Machine learning sentiment classifiers require "training" on the sentiment expressed in manually classified texts from the same domain, as opposed to the lexicon-based classifiers. Consequently, the machine learning classifiers are known for poor performance when used in a domain different from the one they were trained on (Pang and Lee 2008), as the sentiment is expressed differently in different types of content, for example, in blogs and newspapers (Balahur et al. 2013). Therefore, the relative performance of the machine learning and lexicon-based approaches depend on the domain in which they are used.

The machine learning sentiment analysis mainly employs three mathematical approaches to sentiment classification. The Support Vector Machine is a geometric approach that attempts to find the best linear separation between the data expressing positive and negative sentiment. The Naïve Bayes is a probabilistic approach that estimates the probability of a certain sentiment given the document's properties. Finally, the artificial neural network (ANN) and derivative Deep Learning algorithms mimic a biological brain by processing the data through a self-organizing network of "neurons." Generally, the ANN approach is believed to return best results, but is extremely computationally demanding. On the contrary, SVM and Naïve Bayes are popular because of their fast performance.

Before the sentiment analysis of the documents, the algorithm should be trained and then validated (tested) on independent, manually classified samples. For example, during the training phase, the ANN process data records with known sentiment, attempting to improve sentiment by changing the links between its neurons. During the test phase, the trained ANN is offered an independent data set, and the results of sentiment classification are evaluated by comparison with manual processing. The same train and test steps are necessary for all supervised machine learning methods of sentiment analysis. In turn, these three approaches are realized in dozens of software programs, including those available online (e.g., Ribeiro et al. 2015). Moreover, technically inclined researchers may use freely available sentiment analysis computer modules such as the Natural Language Toolkit (http://www.nltk.org/) to build their own text classifiers.

Regardless of which approach is used in the automated algorithm, determining the sentiment of a document is challenging. The challenges include the likely presence of multiple aspects of an issue in a document, ambiguous language, slang, and a sentiment's dependence on context. In fact, IBM suggests adding a new component to the 3Vs of Big Data: veracity (Normandeau 2013). This would address the high amount of subjectivity in Big Data—for example, in determining public sentiments or opinions on social networks. Because of the high amount of subjectivity in analyzing and interpreting the data, agreement between two people classifying the sentiment of the same text (human concordance) will never be perfect. Since the same ambiguity exists for computer sentiment analysis algorithms, a successful automated sentiment analysis algorithm would have an agreement with a human rater similar to that between two human raters. The acceptable "realistic" agreement between human coders measured by an accuracy coefficient may vary between 0.70 and 0.79 (Donkor 2014), as evidenced by, for example, an Amazon MTurks data analysis (Ogneva 2010). Thus, a reasonable target for automated sentiment analysis performance can be defined as approximately 70% agreement (Donkor 2014).

Automated Sentiment Analysis in Tourism

A number of scholars in tourism research have studied the application of automated sentiment analysis to tourism and hospitality-related data. A few have compared the performance of both machine learning and lexicon-based methods. For example, Schmunk et al. (2013) applied four automated sentiment analysis tools (one lexicon-based and three machine learning) to reviews posted by visitors to a ski resort on the websites TripAdvisor.com (127 reviews) and booking.com (81 reviews). They found good performance when either method was applied to separate sentences with approximately 70% accuracy. While this level of agreement seems very encouraging, the investigated resort reviews were overwhelmingly positive,[1] which renders high expected agreement by chance alone. If correction for chance agreement is not applied, the agreement value may be overestimated. It is quite possible that multiple authors would notice higher accuracy for positive sentiment extraction compared to negative sentiment because of this effect and the smaller proportion of negative reviews in a their data (e.g., Kang, Yoo, and Han 2012). We are not aware of tourism-related sentiment analysis reports from machine learning that used more robust rater agreement indices such as Cohen's kappa (Cohen 1960). The performance measures usually reported include accuracy (raw percentage agreement), precision (percentage of identified negatives), recall (percentage of identified positives), and F score, which is a multiplicative combination of precision and recall (Manning and Schütze 1999). Additionally, many authors do not report any performance measure (Table 1).

Table 1. The Methods of Automated Sentiment Analysis Used in Tourism Research.

Reference	Machine Learning Methods	Lexicon-Based Methods	Data	Performance Evaluation
Capriello et al. 2013	No	Developed by the author	Farm tourism reviews: TripAdvisor	None
Garcia-Pablos, Cuadros, and Linaza 2016	SVM+CRF	No	Hotel reviews: Zoover, HolidayCheck	P, R, F
Garcia-Pablos et al. 2016	SVM+CRF	No	Destination reviews: Facebook, FourSquare, Google Place	P, R, F
Kang, Yoo, and Han (2012)	Naïve Bayes	No	Restaurant reviews: web crawling	P, R
Marrese-Taylor et al. 2013; Marrese-Taylor, Velásquez, and Bravo-Marquez 2014	No	Developed by the author	Hotel and restaurant reviews: TripAdvisor	P, R, F
Neidhardt, Rümmele, and Werthner (2016)	No	SentiWordNet	Tour reviews	None
Hu and Chen 2016	No	OpinionFinder	Hotel reviews: TripAdvisor	None
Schmunk et al. 2013	SVM, Naïve Bayes, k-NN	Developed by the author	Hotel reviews: TripAdvisor, Hotels.com	A
Serna, Gerrikagoitia, and Bernabé (2016)	No	WordNet	Easter and summer holiday–related hashtags	None
Shi and Li (2011)	SVM	No	Hotel reviews	P, R, F
Ye, Zhang, and Law 2009	SVM, Naïve Bayes, k-NN	No	US and Europe destination reviews on Yahoo Travel	A, P, R
Zhang et al. 2011	SVM, Naïve Bayes	No	Restaurant reviews: OpenRice.com	A, P, R

Note: Performance evaluation: A = accuracy; P = precision; R = recall; F = F score.

Among machine learning methods, Support Vector Machine (SVM) and Naïve Bayes are the most popular in tourism-related sentiment analysis. Ye, Zhang, and Law (2009) compared the performance of three machine learning methods, including SVM and Naïve Bayes, for a sample of tourist reviews from seven most popular destinations in the United States and Europe published on the Yahoo Travel portal (currently rebranded as a travel magazine). They reported better performance of the SVM classifier with an accuracy greater than 0.8. Shi and Li (2011) applied the same method and reported precision close to 0.85. In a similar study framework, Zhang et al. (2011) compared the performance of SVM and Naïve Bayes algorithms to extract the sentiment expressed in restaurant reviews in Cantonese language and reported an accuracy higher than 0.9 with superior performance of Naïve Bayes.

A lexicon-based approach has also been used in tourism-related research. For example, Serna, Gerrikagoitia, and Bernabé (2016) used the WordNet (Miller 1995) lexical database to extract emotions from tweets referring to two holiday periods, Easter and summer. Neidhardt, Rümmele, and Werthner (2016) used SentiWordNet (Baccianella, Esuli, and Sebastiani 2010; Esuli and Sebastiani 2006), an opinion-mining extension of WordNet, to mine emotions expressed in an online travel forum. Capriello et al. (2013) compared two lexicon-based sentiment analysis methods with manual analysis of farm tourism experiences in travel reviews—specifically, manual content coding, corpus-based semantic analysis with the computer-assisted text analysis (CATA), and stance-shift analysis with CATA. They reported that the automated analysis yielded superior results compared to manual processing, but no numerical indicators were published.

Marrese-Taylor et al. (2013) and Marrese-Taylor, Velásquez, and Bravo-Marquez (2014) proposed an original framework of opinion mining in tourism to summarize visitors' opinions and experiences from TripAdvisor reviews. The framework classifies the reviews into multiple categories and performs lexicon-based sentiment extraction in each category. Hu and Chen (2016) studied factors affecting the perceived helpfulness of online reviews for hotels in Las Vegas and Los Angeles by computing the sentiment of reviews using the OpinionFinder tool. Finally, García-Pablos et al. (2016) used SVM approach to sentiment analysis in their system for automated processing of customer reviews in the hospitality sector. They evaluated the performance of their tool using a sample of hotel reviews in six European languages and reported accuracy of 0.65 to 0.85, depending on the language; the best rate of correct sentiment detection was achieved for reviews in English. The same tool was applied to analyze reviews of eight destinations worldwide (García-Pablos et al. 2016) to research differences in sentiment toward a destination between speakers of different languages.

While human emotions and affective reasoning are recognized as important factors in consumer decision making (e.g., Davis, Jeff, and Cosenza 1988; Duverger 2013; Li, McCabe, and Song 2017), overall, automated sentiment analysis studies are still rare in tourism and hospitality research (Lu and Stepchenkova 2015). While the numbers have been increasing in recent years, the majority of studies discuss and compare the analysis methodology rather than using sentiment analysis methods for research. Research also tends to concentrate on reviews published on travel-related forums, which tend to be overwhelmingly positive (e.g., Lu and Stepchenkova 2012) and are written following the rules of a language syntax and grammar (particularly English). Both characteristics make this kind of reviews easier to classify compared to texts obtained from social networks like Twitter or from answers to open-ended survey questions, which use slang, abbreviations, and short phrases rather than full sentences. Finally, while some publications do include performance evaluation, robust indices such as Cohen's κ are rarely reported, which makes the results of sentiment analysis hard to assess.

Thus, this article proposes to evaluate the suitability of different types of automated classifiers for applications typical in tourism, hospitality, and marketing studies by comparing their performance to that of human raters. Following Donkor (2014), we assume human concordance as a benchmark for automated classifiers performance. Specifically, the study applies four different types of automated classifiers to three types of UGC data frequently encountered in research, namely, open-ended survey responses, customer reviews on recommender system websites, and messages on Twitter. The compared classifiers that comprise one lexicon-based and three machine learning algorithms, with two algorithms (one machine learning and one lexicon-based) used on an as-is basis. The other two machine learning programs represent customized tools for supervised sentiment analysis, which requires more time and effort and potentially costs on the part of the researchers. The article compares six different performance measures of these automated classifiers across programs and in relation to human raters. Given a high likelihood of increasing number of studies using automated sentiment analysis in the coming years, the article discusses issues pertinent to selection of appropriate sentiment analysis software and offers a word of caution against using automated classifiers uncritically.

Data

To compare the software for automated sentiment analysis, we used three distinct data sets; in this article, they are named SUR, REC, and SND. Each data set represents a random sample of a larger set of the real-world tourism-related data used by the authors in other studies. The SUR data set is a sample of 209 surveys of American pleasure travelers regarding their perception of China as a travel destination (out of a total of 2,232 responses with qualitative data). The REC data set contains 332 TripAdvisor reviews of a historic attraction in the city of St. Augustine, Florida, posted between March 12, 2014, and October 12, 2015, by visitors to the site. Finally, the SND data set is a sample of 200 English-language Twitter messages related to the 2014 Winter Olympics held in Sochi, Russia (total data set size 762,475). Together, the data sets represent three frequently used sources of data in tourism analytics: traditional surveys (SUR), travel recommender systems (REC), and social network data (SND). The size of the sample data sets was determined to represent a realistic training/testing target for a research team, which prevents deteriorated quality of classification because of coders' fatigue (see Neuendorf 2002, 145).

The SUR data set was extracted from data collected via an online panel survey aiming at understanding American leisure travelers' international travel experiences and their perceptions of China and other Asian destinations. The respondents were asked the following open-ended question: "Please describe, in your own words, what would make Mainland China a more appealing international vacation destination for you?" This particular question generated 2,232 valid responses from which a random sample was selected. Typically, a response would be represented by a single sentence with 9.2 words on average. The way the question was formulated made it possible to elicit a large share of responses with negative sentiment, as participants were meant to express their reservations about visiting China. The following 14 topical categories were identified (reported in order of frequency): Political system, Cost, People, Safety, Environment, Culture, Tours, Language, Information, Human rights, Image, Distance, Crowds, and Nothing/Not sure (Stepchenkova, Kirilenko, and Li 2016).

The REC data set was collected from TripAdvisor (www.tripadvisor.com) and concerns visitors' reviews for a historic attraction in St. Augustine, Florida. The city of St. Augustine is one of the oldest in the United States and offers its visitors a walk into the past, with guided tours that include architecture, reenactments, and exhibitions of the colonial history of the city. The attraction site also features a communal area, restaurants, and bars in a pleasant setting. The originally collected data set is a census of reviews for a period spanning a year and a half in 2014–2015 and comprising 500 reviews. A typical review contains a few sentences with 38.9 words on average. Visitors comment on the quality of their experience in terms of excellence and educational value of a tour, historic and authentic atmosphere of the attraction, suitability of the experiences for families and children, particular features of the attraction such as demonstrative activities and self-guided tours, and service amenities (Kim, Babalou, and Stepchenkova 2016).

The SND data set was extracted from 762,475 messages pertaining to the 2014 Sochi Winter Olympic Games and published on the microblogging platform Twitter from November 2013 to March 2014. For a full data description,

refer to (Kirilenko and Stepchenkova 2017). Twitter limits the length of each message to 140 characters, and the average message length is accordingly 14.7 words. Twitter messages frequently contain URL references to external resources such as photographs, newspaper articles, videos, and messages published on social networks; these references were removed prior to the analysis and were unavailable to both the automated and human raters. The collected tweets discuss not only sport events, athletes, or national teams but also broader issues related to the image of Russia as the hosting country, such as terrorism, infrastructure problems, ecology of the region, politics, and human rights.

All data sets have undergone a thorough "data cleansing" process. It has been estimated that the real-world data contain up to 40% of inconsistent data (Fayyad, Piatetsky-Shapiro, and Uthurusamy 2003). Even the tightly controlled clinical research records contain from 2.3% to 26.9% errors; however, the errors are usually systematic and can be identified and, in some cases, corrected (Goldberg, Niemierko, and Turchin 2008). We followed the established three phases of data-cleansing process: (1) determining error types that are specific to the data sets, (2) identification of error instances, and (3) correction of the found errors (Maletic and Marcus 2009).

Specifically, the process of data cleansing included (1) removing duplicated records, (2) removing corrupted and empty or no-data records (e.g., "nothing," "don't know" in the SUR data set), (3) removing non-Latin letters, (4) removing web links, (5) controlling the records for inconsistencies (e.g., "one" and "1") and abbreviations, (6) controlling for misspellings, and (7) removing non-alphanumeric symbols such as the hash sign and similar technical symbols (but not emoticons). For machine learning algorithms, data were additionally processed: (8) the case was changed to lowercase, (9) punctuation was removed (except emoticons), (10) stop words were removed, (11) data were stemmed, (12) data were tokenized to bigrams, and (13) the most and least frequent tokens were removed.

Methodology

Each data set was analyzed by four automated sentiment classification programs and by two human raters (a faculty and a graduate research assistant proficient in content analysis or two research assistants) working independently, with subsequent cross-comparison of the classification results. The raters classified entire data, hence producing the labeled data sets for all three types of data (survey responses, tweets, and online reviews). The automated classifiers included two popular software packages, *SentiStrength* and *Deeply Moving*, and two programs developed by the first author using the *RapidMiner* visual programing platform:

- SentiStrength (Thelwall et al. 2010) was developed at the University of Wolverhampton, United Kingdom. The program performs a lexicon-based sentiment analysis using a set of words derived from the social network MySpace. This makes the program specifically suitable for rating short informal comments such as those encountered in blogs and on social networks.
- Deeply Moving was developed at Stanford University. The software algorithm uses an Artificial Neural Network (ANN) with two hidden layers (Socher et al. 2013). Sentiment is detected through supervised training. For a ready-to-use application of the program, the developers provide a model pre-trained for the classification of crowdsourced movie reviews. This pre-trained algorithm was used in this study to represent an "off-the-shelf" machine learning sentiment analysis.
- RapidMiner (Hofmann and Klinkenberg 2013) was initially developed in the Technical University of Dortmund under the name YALE. RapidMiner is a programming environment for rapid development of data mining and text mining programs. With RapidMiner, two sentiment detection programs based on two different machine learning approaches were developed, namely, Naïve Bayes and Support Vector Machine (SVM). Both programs follow standard text mining steps (Feldman and Sanger 2007): tokenization, filtering, case change, stemming, and finally a TF-IDF[2] word vector generation to transform textual data to vectors in a multidimensional word space. Sentiment is then detected through supervised training on the respective data sets with model performance estimated through 10-fold cross-validation; that is, the data set was partitioned into 10 subsamples with nine subsamples retained for training and one for testing. The validation is then repeated 10 times by switching the training and testing subsamples (for reference, see Refaeilzadeh, Tang, and Liu 2009).

These four programs were selected to represent choices that researchers face when dealing with sentiment data analysis, primarily in terms of the programs' underlying approach and the amount of effort required on the part of the researcher prior to data analysis. SentiStrength uses a lexicon-based approach, meaning that it requires a dictionary that contains the sentiment of the words included. The document's sentiment is computed from the sentiment of its words extracted from the dictionary and subject to linguistic rules. Compared to other English lexicon-based programs, SentiStrength has an extensive set of word sentiment dictionaries in languages other than English, which enables sentiment analysis in more than 20 different languages. This feature makes SentiStrength very popular among researchers. Deeply Moving, Naïve Bayes, and SVM are machine learning algorithms that extract the document's sentiment by finding a similar document with an already known sentiment.

With respect to effort, SentiStrength and Deeply Moving require no additional information or extensive training to estimate the sentiment of textual data. SentiStrength employs existing knowledge on typical sentiment expressed by words in English. With machine learning algorithms, it is possible and desirable to train the program on data from the domain of interest prior to data analysis, but it seems that applied researchers prefer using pre-trained models. Indeed, it might be less challenging for a researcher to develop a new customized machine learning application than to learn how to train an existing model (cf. Dickinson and Hu 2015). Thus, Deeply Moving represents an "off-the-shelf" machine learning sentiment analysis algorithm. The version used in this study was pre-trained on a set of movie reviews posted on discussion boards. Customized alternatives are represented by Naïve Bayes and SVM and require a manually rated sample of data, which is used to train and evaluate the performance of the model. The trained model is then used to estimate the sentiment of the rest of the data.

The ways to report a document's sentiment scores also differ across programs. SentiStrength returns two values, one for positive and another for negative sentiment. Deeply Moving, Naïve Bayes, and SVM return one overall value, with Deeply Moving using a scale of [–5, 5] and the other two models using [–1, 1]. For all four models, the results are centered on zero (neutral sentiment); the further the values from zero are in either direction, the stronger the sentiment (positive or negative) expressed by the respondent is. To ensure compatibility of the sentiment classifications, the output of all programs was rescaled to a scale of {–1, 0, 1}, where –1 corresponds to predominantly negative, 0 to neutral, and 1 to predominantly positive sentiment. Specifically, (1) the positive and negative sentiments were summarized for SentiStrength and (2) for all classifiers, negative sentiment values were reclassified to –1 and positive to 1. Note that we tested other reclassification mappings that increased the "neutral" category and found that small resultant changes to the overall results did not justify increasing complexity.

The four sentiment analysis instruments were evaluated based on their performance relative to human coding. Human raters were asked to evaluate the sentiment expressed in three data sets using the same {–1, 0, 1} scale, so that each textual record would have two independent human classifications. The raters had prior experience in content analysis, were uniformly trained, and were provided with a list of instructions to classify the textual units as negative, neutral, or positive. The raters had no knowledge of the other rater's classifications. The software algorithms' performance was evaluated against human raters based on the following measures:

- Three measures typically used in machine-based classification: accuracy, precision, and recall
- Cohen's kappa (Cohen 1960), which is a robust agreement measure showing how well the model outperforms random classification, with the score of 1 indicating a perfect agreement and a score of 0 indicating a performance not better than random. Cohen's kappa is a more reliable index than a simple interrater agreement when group sizes differ significantly.
- Correlation using Kendall's τ. The τ is appropriate for use with the ordinal scale variables and reflects the differences between probabilities of correct classification and misclassification.
- The ratio of opposite classifications (e.g., one rater giving positive and the other giving negative classification of the same item) to the total number of items.

The human raters' performance was evaluated with the same measures applied to their individual classification vectors.

Results

The performance of automated classifiers varied significantly across the data sets (Tables 2 and 3). Sentiment classifications were generally similar between humans and automated classifiers for lengthier textual units written in conventional English, which are common in reviews on TripAdvisor and other recommender system platforms. All classifications (with Deeply Moving standing somewhat aside) returned similar numbers for positively, negatively, and neutrally rated items, with the great majority of reviews rated positively (Table 2). In addition, Naïve Bayes and SVM had similar performance to those achieved by human raters based on three indices commonly reported in data-mining literature. Accuracy varied between 0.85 and 0.89, recall between 0.34 and 0.47, and precision between 0.33 and 0.52 (Table 3). The other two algorithms, however, demonstrated markedly poorer performance (Table 2) and significantly overestimated the number of negative reviews (Table 2). While the "black-box" algorithm of Deeply Moving does not allow tracing the origin of the discrepancies, SentiStrength allows a deeper analysis of its output. Upon thorough analysis of the TripAdvisor data set, we found that many of the misclassified reviews contained references to a highly praised tour guide, whose name was mentioned in 29% of the reviews in total. The spelling correction feature of SentiStrength interpreted the name of the guide as a misspelled word in its dictionary, which incidentally had a moderately high negative sentiment (rated –3 on a scale of –5 for extremely negative to +5 for extremely positive). Replacing the guide's name with the neutral "Smith" in all textual units vastly improved SentiStrength's performance, which became similar to that of the other classifiers.

Classification of the SUR data presented difficulty for both human raters and automated classifiers. Again, the performance of Naïve Bayes and SVM classifiers (accuracy 0.60–0.74) was similar to that of human raters (accuracy 0.73), while the performance of the unsupervised algorithms

Table 2. Percentage of the Textual Units in the TripAdvisor, China Survey, and Twitter Sochi Olympics Databases Classified as Negative, Neutral, and Positive by Two Human Raters and Four Automated Classifiers (SentiStrength, Deeply Moving, Naïve Bayes, and SVM).

	Rater 1	Rater 2	Raters Mean	SentiStrength	Deeply Moving	Naïve Bayes	SVM
REC							
Negative	5	7	8	12[a]	20	2	2
Neutral	8	4	3	29[a]	22	1	2
Positive	88	89	88	58[a]	58	97	96
SUR							
Negative	52	70	62	25	56	58	59
Neutral	21	11	19	52	25	12	23
Positive	27	19	19	23	18	30	17
SND							
Negative	27	28	27	20	87	31	20
Neutral	36	32	32	54	2	31	35
Positive	38	41	41	26	12	38	44

[a]Poor performance of the lexicon-based SentiStrength classifier was due to misclassification of the name of a highly popular tour guide. After correction, the numbers become 6%, 13%, and 80% (see the text for clarification).

Table 3. Comparison of Agreement between Two Human (Raters 1 and 2) and Four Machine-Based Sentiment Classifications of the Textual Units in the TripAdvisor, China Survey, and Twitter Sochi Olympics Databases.

	Rater 1 vs	SentiStrength* vs		Deeply Moving vs		Naïve Bayes vs		SVM vs	
	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
REC									
A	0.89	0.57	0.57	0.65	0.61	0.87	0.87	**0.87**	**0.88**
R	0.47	0.37	0.37	**0.65**	**0.50**	0.34	0.36	0.47	0.42
P	0.45	0.38	0.35	0.45	0.40	0.33	0.35	**0.52**	**0.42**
κ	0.46	0.07	0.06	**0.23**	**0.17**	0.07	0.10	0.21	0.17
τ	0.77	0.16	0.10	**0.38**	**0.28**	0.13	0.18	0.21	0.18
O	0.02	0.11	0.12	0.14	0.14	0.09	0.05	**0.08**	**0.05**
SUR									
A	0.73	0.55	0.56	0.53	0.42	**0.62**	**0.74**	0.60	0.74
R	0.64	0.49	0.45	0.58	0.54	**0.57**	**0.55**	0.55	0.50
P	0.76	0.52	0.45	**0.61**	**0.52**	0.56	0.56	0.58	0.52
κ	0.52	0.25	0.19	0.33	0.20	**0.38**	**0.40**	0.35	0.37
τ	0.57	0.27	0.27	0.45	0.36	**0.46**	**0.49**	0.40	0.45
O	0.11	0.16	0.14	**0.07**	**0.11**	0.15	0.12	0.15	0.10
SND									
A	0.75	**0.60**	**0.54**	0.39	0.39	0.55	0.54	0.51	0.50
R	0.75	**0.60**	**0.54**	0.44	0.42	0.55	0.53	0.50	0.49
P	0.75	**0.64**	**0.58**	0.62	0.51	0.56	0.53	0.50	0.52
κ	0.61	**0.39**	**0.30**	0.14	0.13	0.32	0.30	0.25	0.23
τ	0.60	**0.45**	**0.23**	0.34	0.02	0.34	0.25	0.26	0.16
O	0.08	**0.06**	**0.06**	0.27	0.30	0.14	0.16	0.15	0.16

The indices are as follows: A = accuracy; R = recall; P = precision; κ = Kohen's κ; τ = Kendall's τ; O = relative number of opposite classifications. The best performing algorithm has numbers in bold for each of the metrics.
*Poor performance of the lexicon-based SentiStrength classifier on TripAdvisor was due to mis-classification of the name of a highly popular tour guide. After correction, agreement significantly improves and the agreement indices become 0.79, 0.45, 0.43, 0.23, 0.34, and 0.05 (mean between two raters; see the text for clarification).

of Deeply Moving and SentiStrength was inferior (accuracy 0.42–0.53). Additionally, SentiStrength seriously underestimated the number of negative reviews (25% vs. 52% and 70% estimated by human raters).

The REC textual units had predominantly positive emotions and SUR units had predominantly negative ones, while the SND data were balanced with approximately the same number of Twitter messages classified as positive, negative,

and neutral by human raters (Table 2). Similar to SUR and REC data sets, the supervised classification algorithms returned close proportions of positive, negative, and neutral classifications in comparison to human raters, while SentiStrength overestimated the number of neutral tweets and Deeply Moving grossly overestimated the number of negative tweets. In terms of accuracy, however, SentiStrength outperformed all other classifiers, which should not come as a surprise since its dictionary contains emoticons and abbreviations commonly used in social media (Thelwall et al. 2010). In this example, human raters outperformed all automated classifiers.

Overall, based on the accuracy, precision, and recall metrics, both SVM and Naïve Bayes classification models exhibited performance similar to or only slightly worse than classification by human raters, while the performance of pretrained models was usually inferior. Other metrics, however, hint at a more complex picture. The number of opposite classifications represents the percentage of items classified as positive by one classifier and negative by the other one. Using this metric, both Naïve Bayes and SVM demonstrated performance similar to that of human raters on REC and SUR data, while SentiStrength's performance was excellent on the SND data set and, after the aforementioned data correction, on REC data as well. In contrast, Deeply Moving returned up to 30% opposite classification on Twitter data.

The performance indices discussed so far show that the trained automated classifiers demonstrate performance that is similar or moderately reduced compared to human raters, but other indices show that this success may be misjudged. For example, Cohen's κ, which compares classifier's performance to random classification, is universally lower for automated classifiers in comparison to human performance. For SentiStrength's classification of the REC data set, κ=0.06–0.07, that is, no better than random (although after the name replacement, it improved to κ=0.23). This is still worse than the human raters' performance, however (κ=0.46). Similarly, Kendall's τ shows reduced correlation between the outcomes of automated classifiers and human raters as compared to the correlation between the human raters (Table 3). To summarize, the supervised machine learning algorithms Naïve Bayes and SVM showed the best performance on TripAdvisor (REC) and China (SUR) data sets, while lexicon-based SentiStrength algorithms demonstrated the best performance on the Twitter (SND) data set.

Discussion and Implications

Current interest in Big Data and the data patterns contained in millions of records of unstructured data is supported by new methods of data analysis. One of the important research venues in tourism applications is sentiment analysis. The sentiment classification software automatically extracts the sentiment expressed in texts using either lexicon-based or machine learning approaches. Ideally, the machine learning algorithms should be trained on samples from the studied population of texts, and the lexicon-based approaches should use the lexicon from the document's domain. The associated costs, however, render this task unfeasible for most projects, leading practitioners to use off-the-shelf software.

The problem with using the off-the-shelf classifiers is illustrated in Figure 1. The columns represent results of sentiment analysis of four short sentences by 18 different publicly available sentiment analysis packages (Ribeiro et al. 2015). No shading represents correct sentiment recognition, and dark shades represent incorrect. The first column demonstrates that the positive sentiment in a simple sentence "I really love pizza," is correctly recognized by all classifiers with an obvious exception of Emoticons, which was designed for emoticon recognition only. A slightly more complex sentence structure, "I don't think I love pizza," however, confused all but two classifiers; the majority of classifiers returned obvious misclassifications. Similarly, only five classifiers correctly classified a simple sentence, "I h8 pizza," which uses "h8" in place of "hate," as used on Twitter for example. Finally, only five classifiers were able to recognize an emoticon in the sentence "Pizza :-)." Importantly, no classifier was able to correctly recognize the emotions in all four of the sentences.

There are two fundamentally different approaches in automated sentiment analysis and multiple software packages that use these approaches, as illustrated by Figure 1. Nonetheless, there is little indication to aid tourism researchers in selecting the most appropriate approach and software for their data. We studied the performance of four methods for automated sentiment analysis that use three different approaches and represent currently available choices. The methods were applied to three different sets of travel-related data extracted from open-ended survey question, a recommender system, and a social network. In this section, we outline the issues related to selection between the approaches that we found as an outcome of this study.

From the perspective of performance measurement, we suggest using the human–human versus human–computer interrater agreement as a criterion for evaluating the performance of automated classifiers. Although few authors report the agreement between human raters on machine-classified data, it presents an important baseline for understanding machine performance indices. Donkor (2014) suggests accepting an accuracy relative to human raters of around 0.7 as a benchmark for automated sentiment analysis, and we concur. This accuracy is attainable by automated classifiers, as we have demonstrated (Table 3, REC example). Using this benchmark, the performance of the automated classifiers is typically high. For example, reported accuracy from different articles listed in Table 1 is in the range of 0.65–0.80. Similarly, based on the commonly used performance indices, we found that the best machine learning algorithms have performance similar to human raters.

Classifier	I really love pizza *very positive*	Pizza :-) *positive*	I don't think I love pizza *negative*	I h8 #pizza *very negative*
AFINN	1	0	1	0
Emolex	1	0	1	0
Emoticons	0	1	0	0
EmoticonDS	1	1	1	-1
Happiness Index	1	1	1	1
Opinion Finder	1	0	0	0
NRC Hashtag	1	1	-1	-1
Opinion Lexicon	1	0	1	0
Emolex	1	0	1	0
SANN	1	0	0	0
Senticnet	1	0	1	0
SASA	1	-1	1	0
Sentistrength	1	1	1	-1
SentiWordNet	1	0	1	0
SO-CAL	1	0	1	0
Deep Learning	1	0	0	0
Umigon	1	1	1	-1
Vader	1	0	-1	-1

Figure 1. Differences in sentiment detection with 18 different classifiers (for description, see Ribeiro et al. 2015). Values –1, 0, and 1 designate negative, neutral, and positive emotions, respectively. No shading designates correct emotion recognition, light gray an acceptable recognition, and dark gray incorrect.

However, we also found that other indices such as Cohen's κ show that the automated sentiment analysis is still inferior to the manual method. For example, the SVM processing of REC data returns very high raw agreement (0.87 and 0.88 for comparisons with Rater 1 and Rater 2), but Cohen's κ values are quite small (0.21 and 0.17, respectively, as compared to 0.46 for human raters). We also found that while the accuracy of human raters is homogenous across three data sets, other agreement measures vary substantially—for example, SUR data received 11% opposite sentiment classifications as compared to only 2% for REC data. The leading reason for the high number of opposite classifications for the China data is that the human raters disagreed in the negative emotion detection, which was manifested in significant differences in the percentage of responses classified as having negative sentiments (52% and 70%). While a substantial number of indices is available, selection of the ones suitable for classifier performance evaluation essentially depends on data. For example, precision makes an emphasis on correct selection of the positive units of analysis, while recall emphasizes correct selection of the negative units. Finally, extending Donkor's (2014) suggestion, the same values of a software–human interrater agreement index could be considered satisfactory for a harder-to-classify data set but inadequate for other data.

Despite the less-than-ideal performance of automated classifiers, our results show the feasibility of such approaches for sentiment analysis of tourism-related data. The potential of automatic classifiers to process large volumes of data efficiently means that the overall classification results will be acceptable as long as (1) the outcomes are based on a large amount of data, leading to error reduction with averaging, and (2) classification results of the most frequent words or textual units are validated.

We found several overarching issues that are important for automated sentiment analysis of real-world data. First, the automated classifiers have to be trained on a sample representative of the classified data. This requirement seems more demanding for the machine learning algorithms, which rely more on style consistency of the analyzed documents as compared to the lexicon-based classifiers, which rely on consistency of language semantics and syntax. Indeed, the machine learning tool Deeply Moving exhibited a very poor performance compared to all other classifiers on SND data, which was very dissimilar to movie reviews on which Deeply Moving was trained, but had an adequate performance of REC data. In contrast, Naïve Bayes and SVM algorithms trained on data from the same domain consistently showed good performances. The time-consuming and costly process of the algorithm training is not always feasible, however. These costs are not just limited to manual classification of a sample of the text, but also include learning the specifics of the tool, required trained data set format, data cleansing, and switching between the training, testing, and prediction modes. It should be noted that selection of the representative sample by itself is a nontrivial task for Big Data, since random sample tends to overrepresent frequent patterns. For example, for social media data, random sampling overrepresents top posters such as mass media online outlets and bots (bots are the programs simulating human posters).

Second, for the pre-trained algorithms like Deeply Moving, domain linguistic differences reduce performance even when the analyzed data are similar to the training data. The Deeply Moving classifier, originally trained on movie

review data, demonstrated moderate performance when applied to the TripAdvisor data, even though both data sets are similarly structured user rating systems in common English language. Linguistic differences become a much higher barrier to automated classification when the data from social networks are being analyzed. Indeed, Deeply Moving demonstrated performance no better than random when applied to the data extracted from Twitter messages. The same data set analyzed with SentiStrength, which includes emoticons (pictorial representations of facial expressions) in its dictionary, returned results comparable with those obtained from human raters.

Third, the possibility of intentional or unintentional misspellings in the analyzed texts contributes to poor performance of pre-trained classifiers. This problem is more characteristic of data from social networks with many misspelled and abbreviated words and expressions. The tools equipped with automated spelling correction such as SentiStrength partially alleviate this problem but introduce another one when spelling correction is done with errors.

Fourth, the data distribution may amplify classification errors. The distribution of words in a natural language corpus follows Zipf's law (1935, 2016); that is, when words are ranked according the frequency of their appearance in texts, each word is used roughly twice as frequently as the next-ranked word. Inherently, misclassification of a single high-frequency word disproportionally affects classification results for the entire data set. For example, the spelling-correcting algorithm of SentiStrength misunderstood a person's name as a misspelled word that carries negative emotions, leading to substantially less positive evaluation of the entire historic St. Augustine attraction: 58% of positive reviews compared to 80% by SentiStrength after error correction and 88%–89% by human raters (Table 2). On social networks, an additional problem is connected with reposting the messages originating from bloggers with high popularity. For example, in our other research using SentiStrength, a Twitter message by the famous figure skater Evgeni Plushenko, "…Mao [Asada, Japanese figure skater] … You're real fighter!!", was misclassified due to the negative connotation of the word "fighter." This message was subsequently re-tweeted 1,600 times in a single day.

Finally, it is important to note that additional information in the form of knowledge of the context is available to human raters as opposed to automated classifiers. While the golden standard of content analysis calls for raters who are not familiar with the research (Krippendorff 2004), this requirement is hard to implement in practice. For instance, in the SUR data, the questionnaire targeted the barriers preventing the respondents from travelling to China. Because of that, negative sentiment prevailed, which was detected by both human raters, who were familiar with the survey question. This provided them with additional information compared with the automated classifiers.

One source of advancement in sentiment analysis is improved Natural Language Processing methods. Another exciting possibility is developments of combined learned-based and lexicon-based methods (Tan, Wang, and Cheng 2008). Nevertheless, no improvements in the algorithms in mathematical linguistics can replace the knowledge of the field of application. We suggest that the future of Big Data mining in tourism research depends on developing machine learning instruments pre-trained on data specific for different tourism-oriented data domains, in different languages. That will include social network data (e.g., Chinese Sina Weibo), travel forums (e.g., Russian forum.awd.ru), recommender systems (e.g., TripAdvisor), and many others.

Another potential approach is "averaging to the mean" the results of multiple classifiers, as suggested by Ribeiro et al. (2015). The averaging, however, discards information that is highly useful to estimate results uncertainty. In natural sciences, there is a trend of treating the differences between the results provided by various models as a measure of uncertainty generated by the differences in models' physics. For example, see discussion of uncertainty in future climate projections by Knutti and Sedláček (2013). Following new developments in treating the uncertainty in natural sciences, we suggest that as methodology matures, the best practice will be to employ multiple tools to estimate the sentiment and then to treat the differences in results to measure the corridor of possible sentiment estimates.

In conclusion, our research has revealed considerable differences in the outcomes of automated classifiers. Hence, comparisons among several software programs available to researchers and human coders should be carried out to determine the most suitable instrument for the data. The results seem to be more accurate if a machine learning algorithm is selected and the algorithm is trained on a text corpus similar to the data being used in the study. Ideally, each time classification is performed, the algorithm should be trained on a representative data sample for machine learning classifiers, and the classifier dictionary should be recalibrated for the lexicon-based classifiers. However, we suggest that a lexicon-based approach such as SentiStrength is preferable to an off-the-shelf machine learning algorithm if no training/calibrating is feasible because of time or cost constraints.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. See reviews at https://www.TripAdvisor.com/Hotel_Review-g670155-d1157031-Reviews-Hotel_Aregarden-Are_Jamtland_

County_Jamtland_and_Harjedalen.html#REVIEWS

2. TF-IDF (term frequency–inverse document frequency) measures the importance of a word for a document in a collection of documents. Generally, "important" words are specific to the documents; that is, they are frequent in a specific document, but rare in the entire collection of the documents.

References

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." *LREC* 10:2200–4.

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. "Sentiment Analysis in the News." *arXiv Preprint arXiv:1309.6202*.

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* 66 (6): 778–85.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46.

Cukier, Kenneth. 2010. "Data, Data Everywhere: A Special Report on Managing Information." *The Economist*. http://www.economist.com/node/15557443.

Davis, Duane, Jeff Allen, and Robert M. Cosenza. 1988. "Segmenting Local Residents by Their Attitudes, Interests, and Opinions toward Tourism." *Journal of Travel Research* 27 (2): 2–8.

Dickinson, Brian, and Wei Hu. 2015. "Sentiment Analysis of Investor Opinions on Twitter." *Social Networking* 4 (3): 62.

Donkor, Ben. 2014. "Sentiment Analysis: Why It's Never 100% Accurate," November. http://brnrd.me/sentiment-analysis-never-accurate.

Duverger, Philippe. 2013. "Curvilinear Effects of User-Generated Content on Hotels' Market Share: A Dynamic Panel-Data Analysis." *Journal of Travel Research* 52 (4): 465–78.

Esuli, Andrea, and Fabrizio Sebastiani. 2006. "Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining." *Proceedings of LREC* 6:417–22.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Ramasamy Uthurusamy. 2003. "Summary from the KDD-03 Panel: Data Mining: The Next 10 Years." *ACM Sigkdd Explorations Newsletter* 5 (2): 191–96.

Feldman, Ronen, and James Sanger. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

García, Aitor, Sean Gaines, Maria Teresa Linaza, and others. 2012. "A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain." *Expert Systems with Applications* 39 (10): 9166–80.

García-Pablos, Aitor, Montse Cuadros, and Maria Teresa Linaza. 2016. "Automatic Analysis of Textual Hotel Reviews." *Information Technology & Tourism* 16 (1): 45–69.

García-Pablos, Aitor, Angelica Lo Duca, Montse Cuadros, María Teresa Linaza, and Andrea Marchetti. 2016. "Correlating Languages and Sentiment Analysis on the Basis of Text-Based Reviews." In *Information and Communication Technologies in Tourism 2016*, edited by Alessandro Inversini and Roland Schegg, 565–77. Cham, Switzerland: Springer.

Goldberg, Saveli, Andrzej Niemierko, and Alexander Turchin. 2008. "Analysis of Data Errors in Clinical Research Databases." *AMIA Annual Symposium Proceedings*, 242–46.

Heise, David R. 1970. "The Semantic Differential and Attitude Research," in *Attitude Measurement*, edited by Gene F. Summers, 235–53. Chicago: Rand McNally.

Hofmann, Markus, and Ralf Klinkenberg. 2013. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Boca Raton, FL: CRC Press.

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.

Jakobovits, Leon A. 1966. "Comparative Psycholinguistics in the Study of Cultures." *International Journal of Psychology* 1 (1): 15–37.

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): 6000–10.

Kim, Hany, Vahideh Babalou, and Svetlana Stepchenkova. 2016. *Rebranding the Colonial Quarter, St. Augustine, Florida*. Technical Report. Eric Friedheim Tourism Institute, University of Florida.

Kirilenko, Andrei P., and Svetlana O. Stepchenkova. 2017. "Sochi 2014 Olympics on Twitter: Perspectives of hosts and guests." *Tourism Management* 63:54–65.

Knutti, Reto, and Jan Sedláček. 2013. "Robustness and Uncertainties in the New CMIP5 Climate Model Projections." *Nature Climate Change* 3 (4): 369–73.

Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Laney, Doug. 2001. "3D Data Management: Controlling Data Volume, Velocity and Variety." *META Group Research Note* 6:70.

Li, Chunxiao, McCabe, Scott, and Song Haiyan. 2017. "Tourist Choice Processing: Evaluating Decision Rules and Methods of Their Measurement." *Journal of Travel Research* 56 (6): 699–711.

Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. New York: Cambridge University Press.

Liu, Bing, and Lei Zhang. 2012. "A Survey of Opinion Mining and Sentiment Analysis." In *Mining Text Data*, edited by Charu C. Aggarwal and ChengXiang Zhai, 415–63. New York: Springer.

Lu, Weilin, and Svetlana Stepchenkova. 2012. "Ecotourism Experiences Reported Online: Classification of Satisfaction Attributes." *Tourism Management* 33 (3): 702–12.

Lu, Weilin, and Svetlana Stepchenkova. 2015. "User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software." *Journal of Hospitality Marketing & Management* 24 (2): 119–54.

Maletic, Jonathan I., and Andrian Marcus. 2009. "Data Cleansing: A Prelude to Knowledge Discovery." In *Data Mining and Knowledge Discovery Handbook*, edited by Oded Maimon and Lior Rokach, 19–32. New York: Springer.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, vol. 999. Cambridge, MA: MIT Press. http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2000.26.2.277.

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.

Marrese-Taylor, Edison, Juan D. Velásquez, Felipe Bravo-Marquez, and Yutaka Matsuo. 2013. "Identifying Customer Preferences about Tourism Products Using an Aspect-Based Opinion Mining Approach." *Procedia Computer Science* 22:182–91.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41.

Mohammad, Saif M. 2015. "Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text." *Emotion Measurement*, edited by Herbert L. Meiselman, 201–38. Duxford, UK: Elsevier.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Neidhardt, Julia, Nataliia Rümmele, and Hannes Werthner. 2016. "Can We Predict Your Sentiments by Listening to Your Peers?" In *Information and Communication Technologies in Tourism 2016*, edited by Alessandro Inversini and Roland Schegg, 593–603. Cham, Switzerland: Springer.

Neuendorf, K. A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.

Normandeau, Kevin. 2013. "Beyond Volume, Variety and Velocity Is the Issue of Big Data Veracity." *Inside Big Data*. http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.

Ortony, Andrew, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Ogneva, Maria. 2010. "How Companies Can Use Sentiment Analysis to Improve Their Business." *Mashable*. http://mashable.com/2010/04/19/sentiment-analysis/.

Osgood, Charles E. 1964. "Semantic Differential Technique in the Comparative Study of Cultures." *American Anthropologist* 66 (3): 171–200.

Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. 1978. "The Measurement of Meaning. 1957." Urbana: University of Illinois Press.

Pang, Bo, and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2 (1/2): 1–135.

Refaeilzadeh, Payam, Lei Tang, and Huan Liu. 2009. "Cross-Validation." In *Encyclopedia of Database Systems*, edited by M. Tamer Ozsu and Ling Liu, 532–38. New York: Springer.

Ribeiro, Filipe Nunes, Matheus Araújo, Pollyanna Gonçalves, Fabrıcio Benevenuto, and Marcos André Gonçalves. 2015. "A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods." *arXiv Preprint arXiv:1512.01818*.

Russom, Philip. 2011. "Big Data Analytics." *TDWI Best Practices Report, Fourth Quarter*, 1–35.

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*, edited by Zheng Xiang and Iis Tussyadiah, 253–65. Cham, Switzerland: Springer.

Serna, Ainhoa, Jon Kepa Gerrikagoitia, and Unai Bernabé. 2016. "Discovery and Classification of the Underlying Emotions in the User Generated Content (UGC)." In *Information and Communication Technologies in Tourism 2016*, edited by Alessandro Inversini and Roland Schegg, 225–37. Cham, Switzerland: Springer.

Shi, Han-Xiao, and Xiao-Jun Li. 2011. "A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning." In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, 3:950–54.

Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1631–42.

Stepchenkova, Svetlana, Andrei Kirilenko, and Xiang Li. 2016. "Content and Sentiment Analyses of Travel Barriers to China." In *Proceedings of the 4th Interdisciplinary Tourism Research Conference*, 486–91; May 24–29, Bodrum, Turkey. http://www.anatoliajournal.com/kongre_arsivi/graduate/graduate_programme/2016.pdf.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37 (2): 267–307.

Tan, Songbo, Yuefen Wang, and Xueqi Cheng. 2008. "Combining Learn-Based and Lexicon-Based Techniques for Sentiment Detection without Using Labeled Examples." In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 743–44. ACM.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. "Sentiment Strength Detection in Short Informal Text." *Journal of the American Society for Information Science and Technology* 61 (12): 2544–58.

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. "Using Appraisal Groups for Sentiment Analysis." In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 625–31. ACM.

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 (3): 6527–35.

Zhang, Ziqiong, Qiang Ye, Zili Zhang, and Yijun Li. 2011. "Sentiment Classification of Internet Restaurant Reviews Written in Cantonese." *Expert Systems with Applications* 38 (6): 7674–82.

Zipf, George Kingsley. 1935. *The Psychobiology off Language*. Boston, MA: Houghton Mifflin.

Zipf, George Kingsley. 2016. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Vantaa, Finland: Ravenio Books.

Author Biographies

Andrei P. Kirilenko, is an associate professor in the department of Tourism, Recreation, and Sports Management at the University of Florida. He received his first degree in Applied Mathematics, a Ph.D. in Computer Science, and held positions at the Center for Ecology & Forest Productivity in Russia, European Forest Institute in Finland, US, Environmental Protection Agency laboratory in

Oregon, Purdue University, and the University of North Dakota. His research interests include Big Data analysis, tourism analytics, climate change impacts and sustainability issues, especially the water and food security.

Svetlana Stepchenkova, is an associate professor at the Dept. of Tourism, Recreation & Sport Management, University of Florida. The area of her research interests is destination marketing and branding, media and user-generated communications in tourism, and social studies methodology.

Hany Kim, PhD, is an assistant professor in the Department of Business Administration and Tourism and Hospitality Management at the Mount Saint Vincent University. She received her PhD in Tourism from University of Florida. Her current research focuses on destination marketing and branding, tourist perception and cross-cultural issues in media and user-generated media.

Xiang (Robert) Li, PhD, is a professor and Washburn Senior Research Fellow of the Department of Tourism and Hospitality Management, Temple University. Robert's research mainly focuses on destination marketing and tourist behavior, with special emphasis on international destination branding, customer loyalty, and tourism in Asia. Robert's research findings have appeared in numerous top-tier tourism, business, leisure, and hospitality journals.