Sentiment analysis in tourism: Capitalising on Big Data

Alireza Alaei¹, ², Susanne Becken², Bela Stantic¹
¹School of Information and Communication Technology
²Griffith Institute for Tourism
¹, ²Griffith University, 4222, Australia
{a.alaei, s.becken, b.stantic}@griffith.edu.au

Abstract
Advances in technology have fundamentally changed how information is produced and consumed by all actors involved in tourism. Tourists can now access different sources of information, and they can generate their own content and share their views and experiences. Tourism content shared through social media has become a very influential information source that impacts tourism in many ways. However, the volume of data on the Internet has reached a level that makes manual processing almost impossible, demanding new analytical approaches. Sentiment analysis is rapidly emerging as an automated process of examining semantic relationships and meaning in reviews. In this paper, different approaches to sentiment analysis applied in tourism are reviewed and assessed in terms of the datasets used and performances on key evaluation metrics. The paper concludes by outlining future research avenues to further advance sentiment analysis in tourism as part of a broader Big Data approach.

Keywords: Big Data; Sentiment analysis; Social media; Lexicon; Machine learning
Introduction

The use of Big Data is rapidly entering the domain of tourism research (Fuchs et al., 2014). The four Vs of Big Data, namely volume (scale), variety (different types of data), velocity (high speed, and real time), and veracity (uncertainty, and validity) are particularly relevant in consumer research (IBM, nd), with its increasing need for real-time and customized information. The tourism industry, as an industry where customer experience is crucial for its growth and reputation, has mainly adapted to the evolving technology and the availability of new data sources. Most tourist services are now available on the Internet through online booking websites. In addition, travel is one of the dominant topics on social media, for example on Facebook and Twitter (Neidhardt et al., 2017; Travelmail Reporter, 2013). It is, thus, not surprising that tourism has been recognized as the number one sector in terms of online engagement (Mack et al., 2008).

All Internet-based activities leave a digital footprint. It is timely to examine how tourism researchers are making use of these data, and whether these new types of data form a part of a new research paradigm that entails novel methodologies and has the potential to further advance our theoretical understanding of tourism. To date, online data sources have mainly been used in applied research, whereby advantage was taken of the large and often free-of-charge volumes of data which provide insights into activities of the tourism/travel industry and its customers. Not surprisingly, the focus of previous research was on business strategy development, innovation and product development, and marketing campaigns (Ellion, 2007; Kuttainen et al., 2012; Pan et al., 2007).

In the context of tourism, a service-based industry that relies on positive customer emotions and feedback, the concept of visitor satisfaction is of critical importance. Satisfaction as a theoretical construct has been explored and discussed for a long time, and multiple instruments exist to operationalise and measure it (Wang, 2016). Most rely on
collecting data through surveys. It is well established that survey-based approaches suffer from several shortcomings, including costs and logistics, and potential for multiple bias. Since visitors made a high investment in their travel, their responses to the survey questions may reflect an inherently positive assessment as a result of confirmation bias (Dodds et al., 2015). Interviewer bias and cultural influence in answering particular questions are other known problems of survey-based approaches (Veal, 2006). In addition, questionnaires cover only pre-determined aspects of the destination and, thus, they lack comprehensiveness. On the contrary, the availability of online user-generated content (UGC) and new technologies provided researchers with a new approach that travellers’ perceptions and possibly their level of satisfaction can be approached through ‘sentiment analysis’. Sentiment analysis, in general, aims to determine the overall contextual polarity of a text document, a review, an opinion or an emotion expressed in online UGC, whereby polarity can be positive, neutral or negative. Whilst highly relevant for tourism, sentiment analysis in tourism is only beginning to gain in popularity (Feldman, 2013, Gao et al., 2015, Ribeiro et al., 2016).

The purpose of this paper is to review and critically examine the state-of-the-art sentiment analysis methods in tourism research. To advance this type of analysis for the particular domain of tourism and to understand whether such Big Data-based approaches offer new research pathways, this review asks the following questions:
1. What are the key elements and methods used in sentiment analysis?
2. To what extent has sentiment analysis been applied in tourism and how do different methods perform?
3. Can sentiment analysis, as part of a wider Big Data approach, be a novel way of improving tourism research methods and increase theoretical understanding in tourism?
Background: The digitally supported tourism industry

Technological changes related to the Internet, including smartphones and tablets, have revolutionized the tourism industry from a brick-and-mortar and person-to-person service industry to a heavily digitally supported and omnipresent travel service network. Individual travellers or groups now have much greater control over planning, building and personalizing their trips. They not only interact with a range of platforms and online intermediaries to extend their knowledge in relation to travelling and decision making in tourism, but also associate with other travellers who share their experiences. Travelers have access to online platforms to provide feedback and make recommendations for other travellers (Neidhardt et al., 2017; Yang et al., 2017; Ye et al., 2009). As a result, new Internet technologies have empowered people who previously did not have a voice (Hepburn, 2007). The most successful professional platforms in relation to travel and tourism are TripAdvisor, Expedia, VirtualTourist, and LonelyPlanet (Bjorkelund et al., 2012; Gretzel et al., 2007; Rabanser & Ricci, 2005). TripAdvisor alone counts 350 million unique visitors per month on their website and generates over 320 million reviews that cover accommodations, restaurants, and attractions (TripAdvisor, 2016). Information provided through these independent platforms has been found to be superior and more trustworthy compared with companies’ websites and professional reviews (Akehurst, 2009; Gretzel et al., 2007; Rabanser & Ricci, 2005; Xiang et al., 2009).

In addition to professional systems, online social media, such as Twitter, Instagram, Facebook, FourSquare, Sina Weibo, and GooglePlus, play a significant role in creating electronic word-of-mouth (e-WOM) (Confente, 2015; Garcia-Pablos et al., 2016; Leung et al., 2013; Phillips et al., 2017). Importantly, online social media, travel professional websites and platforms, and blogs present inexpensive means to gather rich, authentic, and unsolicited data on travellers’ opinions. Whilst personal advice often ranks as the most influential source
of pre-trip decision making, the overall credibility of blogs and online social media compared
to that of traditional WOM is relatively high (Akehurst, 2009). Therefore, social media and
blogs nowadays complements opinions attained from relatives, friends, colleagues, and
official sources (Cantallops & Salvi, 2014; Chua & Banerjee, 2013; Filieri et al., 2015;
Hepburn, 2007; Mack et al., 2008).

However, as the number of online information is increasing at an extremely fast pace,
searching, manipulating and aggregating the data to extract relevant and useful insights about
tourists’ attitude, behaviour and experience quality becomes a tedious and time-consuming
task for both, travellers and industry users as well as professional and academic researchers
(Cantallops & Salvi, 2014; Ellion, 2007; Dodd, 2014, Xiang et al., 2015a; Ye et al., 2009). To
analyse large data volumes more effectively, the demand for automatic multi-aspect
algorithmic and machine-operated systems is increasing.

The importance of using social media data and data mining tools and procedures in
tourism was studied in the literature (Dhiratara et al., 2016). Data collection, data cleaning,
mixing process, and then evaluation and understanding of the results are the major steps used
in most of the applications in relation to social media data analysis in tourism (Hippner &
Rentzmann, 2006; Schmunk et al., 2014). Text summarization, and text classification along
with natural language processing (NLP) are earlier technologies used to facilitate information
processing and data analysis (Cantallops & Salvi, 2014; Ghose et al., 2012; Pan et al., 2007;
Stringam & Gerdes, 2010; Xiang et al., 2015a).

Besides, sentiment can also be modelled by machines for automation, and integration
across various applications (Choi et al., 2007; Rabanser & Ricci, 2005). Sentiment analysis
basically refers to the use of computational linguistics and natural language processing to
analyse text and identify its subjective information. Whilst research on sentiment analysis
goes back to the 1970, only recently it has received increasing attention from both researchers
and practitioners (Brob, 2013; Pang et al., 2002). The interest is driven by: a) escalation of web- and social media-based information, b) evolution of new technologies, especially machine learning approaches for text analysis, and c) development of new business models and applications that make use of this information. Despite its popularity, sentiment analysis is still in its infancy compared to earlier technologies, such as data mining and text summarization (Pan et al., 2007).

This review argues that sentiment analysis can become an important tool in tourism research. Moreover, it may be an indication of how data-driven research models might be of relevance to tourism research. Whilst this review will not provide a final answer to such challenging questions, it will examine tourism-specific material to further explore whether Big Data is merely a continuation of inductive science (Fricke, 2013) or whether it is the ‘end of theory’ and constitutes a radically new paradigm (Anderson, 2008, Kitchin, 2014). In the meantime, the following postulates are useful:

- The volume of online data relevant to the tourism context is increasing exponentially. Data can be structured, semi-structured, unstructured, textual (in different languages), pictorial, or audio-visual. For example, in the case of online surveys as a source of structured data, the first company in Australia registered in 2001. In 2013, more than 40 companies and 150 market and social research consultancies provide services using online surveys (Dolnicar et al., 2013; Stantic & Pokorný, 2014).

- Online data related to tourism activities are generated at such velocity that they outstrip the potential of traditional (paper and pen) surveys to capture events in real-time in order, for example, to monitor service quality and recovery.
Tourism is part of the ‘experience economy’ and those involved in the travel industry are increasingly seeking to understand the emotional and experiential elements of tourist activities (Ma, Scott, Gao & Ding, 2017).

Online platforms represent a two-way avenue of producing and consuming information, and ‘co-creating experiences’ (Sigala, 2016).

Integration of multiple Big Data sources, e.g., heterogeneous data sources, in the form of structured and unstructured data, such as customer feedback, reservation and booking data, and web search / navigation data, in customer and supplier sides (Höpken et al., 2013, Höpken et al., 2015), may reveal new insights that were not able to be detected with traditional approaches.

In the following, sentiment analysis is reviewed to provide a starting point for future discussions on how Big Data can be used in the tourism context. As an inter-disciplinary research domain, sentiment analysis approaches draw on progress in several areas, including computer science, information technology, and linguistics. Therefore, a brief overview of the technical aspects of sentiment analysis is provided, followed by an assessment of sentiment analysis methods in the tourism domain, including an evaluation of datasets and performances. The paper concludes with recommendations for future research in this area, including an assessment of the potential of using Big Data in the tourism domain.

What is sentiment analysis?

Opinion mining based on sentiment orientation was studied in recent years to understand perceptions and characteristics of population or market groups, and to determine the credibility of content and motivations for posting reviews (Ribeiro et al., 2016). Different sentiment analysis methods were developed in various domains, triggering a small number of
review papers on this topic (Gonçalves et al., 2013; O’Leary, 2011; Ribeiro et al., 2016). None of the reviews to date focus on tourism.

**Overview**

Sentiment analysis, in particular in relation to customer reviews, is built on the premise that information provided through text (e.g., a review) is either subjective (i.e. opinionated) or objective (i.e. factual). Subjective reviews are based on opinions, personal feelings, beliefs, and judgment about entities or events. Objective reviews are based on facts, evidences, and measurable observations (Feldman, 2013). Consumer reviews and social media posts often reflect happiness, frustration, disappointment, delight and other feelings (O’Leary, 2011). Tapping into these large volumes of subjective e-WOM is of great value to tourism organizations and businesses who seek to improve customer management and business profitability (Choi et al., 2007; Kuttainen et al., 2012; Ye et al., 2009).

Methodologically, sentiment analysis represents a polarity classification problem. Considering different numbers of classes, sentiment polarity classification can be conceptualised as binary, ternary or ordinal classification. In a binary classification, we initially assume that a given customer review is subjective. In other words, a binary classification assumes that the given text is predominantly either positive or negative, and then it determines the polarity of the given review as ‘positive’ or ‘negative’. The definition of the two poles of sentiment as positive and negative depends on the particular application and domain. For example, in the context of tourism, ‘positive’ and ‘negative’ may, respectively, refer to “satisfied” and “unsatisfied”, but further research to link sentiment polarity to the theoretical constructs of satisfaction would be required.

Reviews may not always be subjective, therefore, the binary classification needs to be extended to a ternary classification that contains a third, ‘objective’ category. In the ternary classification problem, the classifier implicitly performs a classification to differentiate
between objective and subjective sentences, providing a class-label as ‘positive’, ‘negative’, or ‘neutral’. Neutral polarity is sometimes interpreted as a polarity between positive and negative. The sentiment analysis can also be treated by the means of a cascaded approach, composed of a binary classifier to differentiate between subjective and objective reviews and a binary polarity classifier to further classify subjective reviews into two groups, namely positive or negative. Objective reviews generally do not contain those words that are clearly defined as positive or negative in a dictionary. They may also contain mixed polarities without a clear perspective of direction. In addition to the simple binary and ternary classification, ordinal classification can be performed by the means of a rating scale (e.g., 1 to 5 stars) of the sentiment strength (Brob, 2013).

In sentiment analysis, it is also important to understand what a sentiment relates to. The detection of a target and aspect (i.e. topic detection, Menner et al., 2016), relates to determining the subject of a sentiment expression. Sentence level sentiment analysis supports aspect-based review mining. Based on the level of granularity of analysis, a sentiment aspect may refer to a concrete or tangible entity or to a more abstract topic. A target or an aspect might be referred to either implicitly or explicitly. Reviews with explicit targets or aspects are easier to analyse than those with implicit ones. A hotel review may be composed of different aspects of a hotel, for example, “the size of the bed was small and there was a noisy refrigerator” is a review, which explicitly describes two aspects of a “hotel room” as “small bed” and “noisy”. Whereas in the review “hotel was expensive!”, the word “expensive” is an implicit aspect that refers to the “price” of the hotel. Aurchana et al. (2014) found that extracting both implicit and explicit aspects accurately in reviews results in an increase in the accuracy of sentiment analysis results.

A comprehensive sentiment analysis also includes data on who provided the information and at what point in time. Thus, sentiment analysis of an opinion or review can be technically
formulated by a quintuple \((o, h, t, a, p)\), where \(o\) is an opinion, \(h\) is the opinion holder, \(t\) is the time when the opinion \(o\) is expressed by \(h\), \(a\) is a topical aspect of the opinion \(o\), and \(p\) is the polarity orientation of the opinion \(o\) in relation to aspect \(a\) (Liu, 2010).

Sentiment analysis can be employed at the word, sentence, paragraph and document levels. Relatively less research has focused on sentence level analysis, since it is more challenging to accurately extract polarity from a small number of words compared with paragraphs and documents (Brob, 2013; Choudhury, 2016; Höpken et al., 2016; Schmunk et al., 2014; Ribeiro et al., 2016). For a clear explanation and understanding of the different sentiment analysis methods, the relevant key terms are defined in Table I.

*Table 1 here*

**Sentiment analysis methods**

Sentiment analysis comprises a multi-step process: a) data retrieval, b) data extraction and selection, c) data pre-processing, d) feature extraction, e) topic detection, and f) data mining process (e.g., Hippner & Rentzmann, 2006; Schmunk et al., 2014).

Data retrieval requires the identification and definition of the data source, for example, a commercial service provider portal or a social media network. To collect the review data from these sources, a specific web crawling mechanism is necessary to fetch the data and then save them in a database considering the format of data (Menner et al., 2016; Schmunk et al., 2014). After collecting data in a database, the review data needs to be extracted from within a set of heterogeneous data fields. For example, in the case of TripAdvisor data, a review is embedded within a retrieved HTML document, which is composed of different elements, such as footers or headers, tags, and the review text itself (Menner et al., 2016; Schmunk et al., 2014). The review text needs to be extracted using appropriate expressions. Each extracted review contains one or several sentences reflecting the reviewer’s opinion.
Different tasks including splitting a review into sentences, splitting a sentence into words, tokenisation, filtering of stop-words, Part-of-Speech (POS) tagging, stemming and the transformation to lower/upper cases are performed on the reviews in the pre-processing step to prepare them for the next step (i.e. feature extraction) (Schmunk et al., 2014). POS tagging is an important pre-processing task that generally forms a part of sentiment analysis by assigning each word a particular label (e.g., noun, verb, and adjective).

Feature extraction is known as the process of deriving a set of discriminative, informative and non-redundant values to numerically represent a review or text. One of the commonly used feature extraction techniques is based on term occurrences, called term frequency (TF) or term frequency-invers document frequency (TF-IDF). Using the TF feature extraction technique, reviews or sentences are converted into a ‘term document matrix’ (Pang et al., 2002; Hippner & Rentzmann, 2006; Menner et al., 2016).

Topic detection is a multiclass classification problem where a text is classified to an appropriate topic class depending on its content and application. Topic detection research dates back to 1998 where topic identification in the context of broadcast news was studied (Allan et al., 1998). Hu and Liu (2004) later proposed a method to summarize customer reviews based on different product features. Suggested approaches mainly involved word dictionaries, clustering, and similarity measures. Since, the overview of topic detection methods in the literature is out of the scope of this paper, readers are referred to Menner et al. (2016) for an overview.

In the data mining process, different types of sentiment analysis methods can be distinguished in the literature; namely (i) machine learning, (ii) rule/dictionary-based and (iii) hybrid approaches (Feldman, 2013; Ribeiro et al., 2016). Machine learning methods are further categorized into supervised and unsupervised approaches. The dictionary-based
approach also includes a subcategory called semantic-based approach (Tsytsarau & Palpanas, 2012). A detailed description of these five categories is provided in the following.

Supervised machine learning approach

A sentiment analysis method based on supervised machine learning involves creating a model by using annotated data or weakly labelled corpora. In the manually annotation process, for example, “what a wonderful holiday!!” is annotated as a sentence with “positive” sentiment polarity. Weakly labelled data are those data where the class labels were determined heuristically by the machine. For example, user-generated content on review platforms often contains weakly labelled data when reviewers assign categories (e.g., restaurant) and ratings (e.g., stars) to their reviews (Brob, 2013).

Supervised machine learning approaches follow several steps (Figure 1). After applying pre-processing techniques to clean, segment and tokenize the text data, a feature extraction method is applied to characterize the review. Features extracted from the reviews are then fed to a classifier to train the classifier. The trained classifier is finally used to determine the polarity of new text. Support Vector Machine (SVM) and Naïve Bayes are the key machine learning methods used for sentiment analysis in the literature (Brob, 2013; Kang et al., 2012; Markopoulos et al., 2015; Shi & Li, 2011; Shimada et al., 2011; Ye et al., 2009), as they were conventionally designed for two-class classification problems. A SVM is a classifier which uses annotated data for training to obtain an optimal separating hyperplane/line to accurately categorise new samples data into different groups. A Naïve Bayes classifier is a probabilistic classifier, which uses Bayes' theorem in the classifier's decision rule, with an assumption that the features are independent. SVM and Naïve Bayes methods need comparably less annotated data for training the models compared to the neural network approach. Neural network and deep learning models (Irsoy & Cardie, 2014; Socher et al., 2013) and k-nearest neighbour method (Schmunk et al., 2014) were also employed for semantic analysis in the literature.
Unsupervised machine learning approach

Cluster analysis, as an unsupervised machine learning approach, has been used for data mining, pattern recognition, and image analysis. Clustering is the task of grouping a set of data in such a way that items in a cluster are more similar to each other compared to those in other clusters. Clustering techniques, such as k-means (Xiang et al., 2015b), and statistical models based on the probability distribution of reviews in sentiment space (Rossetti et al., 2015) were employed in the literature for sentiment analysis of short text data. In addition, Naïve Bayes models were also adapted in an unsupervised fashion for sentiment analysis (e.g., Shimada et al., 2011).

Dictionary-based approach

As dictionary-, lexicon- and rule-based approaches were used in the literature interchangeably, this review also uses the terms as synonyms. To provide an overview of dictionary-based methods, a complete framework of a common dictionary/rule-based sentiment analysis method is represented in Figure 2. In this approach, the detection of subjectivity versus objectivity can be integrated into the framework or it can be handled by the sentiment polarity detection process itself. Aspect or topic detection can also be included within the framework based on the specific needs of the application. Dictionary-based systems rely on the use of comprehensive sentiment lexicons and sets of fine-tuned rules. A sentiment dictionary can be created either by humans, by machine or by both humans and machine (semi-automatically). For instance, a dictionary may contain words, such as “good”, “nice”, “fantastic”, “bad”, “worse”, and “ugly”, with their associated values of polarity. While creating dictionaries, the polarities are assigned to the words without considering any contextual information.
Different methods were developed for dictionary-based approaches (Bjorkelund et al., 2012; Bucur, 2015; Garcia et al. 2012; Hutto & Gilbert, 2014; Levallois, 2013). SentiWordNet in itself (Bucur, 2015; Garcia et al. 2012), and in combination with a simplified Lesk Algorithm, was also used in sentiment analysis (Bjorkelund et al., 2012). The Lesk algorithm is an algorithm for disambiguating word sense that works based on the hypothesis that words in a given "neighbourhood" have the same topic (Bjorkelund et al., 2012). Valence Aware Dictionary for Sentiment Reasoning (VADER) is a method that has provided promising results on Twitter data (Hutto & Gilbert, 2014). VADER combines a lexicon and a series of intensifiers, punctuation transformation, and emoticons, along with some heuristics to compute sentiment polarity of text. Five general rules that embody grammatical and syntactical conventions for emphasizing sentiment intensity are used for computing the sentiment polarity. The VADER sentiment lexicon is composed of more than 7,000 items, along with their associated sentiment intensity measures, validated by humans and specifically adapted to sentiment in microblog-like contexts, such as Twitter (Hutto & Gilbert, 2014). Umigon is another dictionary-based method, which uses a lexicon with heuristics for sentiment detection in Twitter reviews (Levallois, 2013). It is a fast and scalable method, which can handle negations, elongated words and hashtags. Umigon provides additional semantic features, such as time or subjectivity (Levallois, 2013).

**Semantic approach**

The dictionary-based approach was improved by introducing semantic-based analysis methods (Tsytsarau & Palpanas, 2012). The semantic approach is mainly a rule-based linguistic model to obtain a polarity for each text segment. In this approach a dictionary of domain specific terms and their associated polarity values is required.
Hybrid approach

In hybrid approaches, dictionary and machine learning-based techniques can work in parallel to compute two sentiment polarities. The results obtained from the dictionary and machine learning based methods are then combined to provide a final sentiment polarity. It is also possible to design a sentiment analysis model by incorporating both dictionary and machine learning based methods at different stages of the model (Waldhör et al., 2008; Claster et al., 2010a; Claster et al., 2010b; Kasper & Vela, 2011; Claster et al., 2013; Pappas & Popescu-Belis, 2013; Schmunk et al., 2014; Chiu et al., 2015). Sentiment-aware nearest neighbour model (SANN) is a combination of dictionary- and learning-based approaches that initially classifies text as either a subjective or objective review (Pappas & Popescu-Belis, 2013). If the text is objective, then the task of sentiment analysis is over. However, if the text is subjective, it is then further classified as either positive or negative. For text with zero polarity, the neutral label is assigned (Pappas & Popescu-Belis, 2013).

Review of tourism-focused sentiment analysis

Building on the technical overview of sentiment analysis, this section explores how sentiment analysis has been applied in tourism. Of particular interest is whether tourism related studies are using state-of-the-art methods or whether there are further opportunities to advance the application of sentiment analysis.

Identified studies and datasets used

To identify sentiment analysis studies in tourism, combinations of key words, such as “sentiment analysis of tourism”, “tourism sentiment data”, “sentiment analysis of hotel reviews”, and “sentiment analysis of restaurant reviews” on Google search engine, instead of a specific search within Scopus and Web of Science websites, has been used to broadly
search and retrieve relevant papers published on the Internet. We have further studied recent review articles on sentiment analysis to extract those references that dealt with tourism. As a result, we believe that a critical mass of tourism-related sentiment studies have been identified for this review.

An overview of seminal tourism-related studies and their specific datasets is provided in Table 2. Tourism researchers have typically used two types of online content for their sentiment analysis: reviews of tourism obtained from professional websites (e.g., TripAdvisor, Booking, and Ctrip) and social media posts (e.g., Twitter). Both types of sources usually contain short text. Twitter, for instance, allows tweets of up to 140 characters in length, lending itself to a mostly sentence-level sentiment analysis. Manual and automatic annotation processes were used to label the reviews in order to train and evaluate the sentiment analysis methods. It is also noted that most of the datasets used in the literature relate to hotel accommodation (e.g., Kasper & Vela, 2011 and 2012; Tan & Wu, 2011; Bjorkelund et al., 2012; Gräbner et al., 2012; Bucur 2015; Marrese-Taylora et al., 2013; Markopoulos et al., 2015; Rossetti et al., 2015). A small number of studies focus on restaurants (Ganu et al., 2009; Zhang et al., 2011) and airlines (Misopoulos et al., 2014).

Table 2 here

Both supervised and unsupervised machine learning, dictionary based, semantic and hybrid sentiment analysis approaches were used in the tourism literature. In terms of supervised machine learning approach for sentiment analysis in tourism, SVM (Ganu et al., 2009; Ye et al., 2009; Zheng & Ye, 2009; Shi & Li, 2011; Zhang et al., 2011; Brob, 2013; Markopoulos et al., 2015; Pablos et al., 2015; Schmunk et al., 2014), Naïve Bayes (Schmunk et al., 2014), Conditional Random Fields (CRF) (Pablos et al., 2015), Nearest Neighbour (Schmunk et al., 2014) and entropy based classifiers (Brob, 2013) were employed. Different
types of features, such as term frequency (TF) (Ye et al., 2009), term frequency–inverse
document frequency (TF-IDF) (Shi & Li, 2011), stemmed word (Ganu et al., 2009), bag-of-
words (Markopoulos et al., 2015), information gain (IG) (Zheng & Ye, 2009), n-gram (Brob,
2013; Kang et al., 2012; Markopoulos et al., 2015; Zhang et al., 2011; Pablos et al., 2015)
were proposed to characterise tourism reviews.

An unsupervised machine learning approach based on Naïve Bayes classifier was
implemented by Shimada et al. (2011) to produce a sentiment analysis of tourism data at the
sentence level. The Naïve Bayes sentiment classification approach was trained using
automatically labelled data. Emoticons, such as 😊 and ☹, were used to represent positive and
negative seeds to label data for training instead of words, such as “excellent” and “poor”.
Therefore, reviews that contained a smiley face, for example, were considered as positive and
those with an angry face were classed as negative (Shimada et al., 2011). K-mean clustering
techniques and statistical models based on probability distribution of reviews in sentiment
space (Rossetti et al., 2015, Xiang et al., 2015b) were also employed on tourism data.

Several tourism studies have drawn on dictionary based approaches. Misopoulos et al.
(2014) used a lexicon type method to assess the polarity of Twitter posts relevant to airline
service delivery. The results revealed those aspects of the airline customer service where
customers were dissatisfied, satisfied, or even delighted. The analysis was, however, based on
a limited number of 20 keywords (10 positive and 10 negative), which posed a significant
restriction to the findings from this research. Moreover, negation was not incorporated in the
system to accurately capture the meaning of opinions, such as “not bad”. Another example
can be found in Sharma et al. (2015) who analysed travel reviews at the sentence level using
a lexicon based system. Other dictionary-based analysis focused on hotel and restaurant
customer reviews (Bucur, 2015; Gräbner et al., 2012; Marrese-Taylora et al., 2013; Schmunk
et al., 2014). The dictionary used in Gräbner et al.’s (2012) study was a hotel domain specific
lexicon of semantically relevant words. Previous research established that features with high intensity over different time periods can be useful to detect abnormal changes in hotel reviews, and to analyse the reasons for these changes. Such trends can be particularly useful when visualised (e.g., on a map) to potential customers (Bjorkelund et al., 2012).

Xiang et al. (2015a) proposed semantic approach for text analysis to understand hotel guest experience and their satisfaction. As a part of the system proposed by Kasper and Vela (2011 and 2012), a rule-based linguistic model using semantic information helped to obtain a polarity for each text segment in addition to topic identification. In this approach, a domain specific dictionary was, however, used that makes this system domain dependent.

Finally, several tourism researchers have used hybrid methods. Earlier work by Waldhör and Rind (2008) proposes to combine a linguistic parsing methodology with information and terminology extraction methods in order to determine sentiment polarity of online blog reviews. Using binary choice keywords and a Naïve Bayes algorithm helped measure sentiment polarities of tweets related to different tourist destinations (Claster et al., 2010a; Claster et al., 2010b; Claster et al., 2013). Binary choice keywords are two sets of subjective keywords that constitute antonyms; for example bad versus good. Another sentence level hybrid sentiment analysis system was presented by Kasper & Vela (2011 and 2012) in the context of German language hotel reviews. The study first applied a language filter to select reviews written in German. The filtered texts were then disaggregated into individual sentences, and these sentences were subjected to a polarity classifier and linguistic information extraction process for detecting the respective topics and their polarities. For the information extraction, a dictionary of hotel-specific terms and a sentiment dictionary that associates basic polarity values with these terms were created. The polarity values from the statistical and the linguistic classification were then combined into a joint global polarity value to present the sentiments on the user interface. Sentiment analysis on Chinese e-WOM
was proposed by Chiu et al. (2015). Combining a supervised probabilistic model and a heuristic n-phrase rule was used to effectively obtain customer opinions about hotel attributes. Schmunk et al. (2014) further discussed a system to initially detect the subjectivity of a sentence by a dictionary-based method. Then, the classification of the sentence into positive or negative was performed using bigrams features along with a SVM classifier (Schmunk et al., 2014). However, this type of approach suffers from the promoting of the errors occurred in the first step to the subsequent steps of the system. This drawback may be mitigated by using a backward feedback from the current to the earlier steps.

It is also noted that for subjectivity detection, SVM, Naïve Bayes and k-nearest neighbour methods with regard to machine learning approach, and dictionary-based approach were applied to tourism (Schmunk et al., 2014). In addition, for topic detection, SVM, Naïve Bayes, and k-nearest neighbour methods (as supervised machine learning approach) terminology and dictionary-based approach as well as frequent words, latent-semantic indexing, sequential pattern mining, and cluster analysis (as unsupervised techniques) were applied in the tourism context (Brob, 2013; Markopoulos et al., 2015; Höpken et al. 2016; Schmunk et al., 2014). These methods can detect explicitly mentioned topics in reviews.

In summary, a relatively broad range of studies exist in the domain of tourism, mainly in relation to hotels and accommodation. Most studies have used data written in English for sentiment analysis, but few used reviews written in Chinese, Spanish and German (Zheng & Ye, 2009; Tan & Wu, 2011; Garcia et al, 2012; Kasper & Vela, 2011; Zhang et al., 2011). Furthermore, our review revealed that most tourism sentiment analyses are based on a machine learning approach, although a considerable number of studies have also used a dictionary-based approach. The main advantage of the latter is that there is no need for annotated text corpora for training sentiment extraction models. Moreover, creating a lexicon is a one-time effort and can be used forever and often across different domains. The more
sophisticated hybrid approaches have only been used in a few instances, indicating future research opportunities. In most cases, publicly available dictionaries were used or adapted to the tourism context. Domain-specific lexica are rarely used, thus, compromising the quality of the sentiment analysis. A possible way forward is to use an aspect-dictionary based method first to initially determine a review aspect (e.g., food quality in a restaurant), followed by a machine learning method to obtain the sentiment polarity of the review. This could begin by using weakly labelled data for an initial training of a model and complete the task by using manually annotated data to obtain a refined model.

**Evaluation Metrics**

As mentioned earlier, most sentiment analysis methods provide either a 2-class (positive and negative) or a 3-class (positive, neutral and negative) classification. It is important to evaluate and quantify the performances of different methods. A clean and unambiguous way to present the prediction results of a classifier is to use a confusion matrix, which is also called contingency table (see Table 3 for a 3-class problem). Each letter in Table 3 denotes the number of review instances, which belong to the original class provided by the annotation process and are anticipated as predicted class obtained from a classifier, where class labels are positive (\(\text{Pos}\)), neutral (\(\text{Neu}\)) and negative (\(\text{Neg}\)).

![Table 3 here](https://example.com/table3.png)

The Accuracy \((A)\) is one of the evaluation metrics commonly used in the literature (Ribeiro et al., 2016). It is simply the number of correct predictions of sentiment made, divided by the total number of predictions made. The accuracy measures how accurate the method is in its prediction of the correct output. The metric \(A\), as shown in Formula (1),
assumes that every correct classification of the input reviews independent of the class label has an equal weight.

\[ A = \frac{a + e + i}{a + b + c + d + e + f + g + h + i} \]  

Formula (1)

Precision, Recall, and F1-measure are the other three evaluation metrics frequently used for evaluating the results of sentiment analyses (Brob, 2013; Markopoulos et al., 2015; Ribeiro et al., 2016). Considering a sample sentiment analysis system of three classes, and using the definitions provided in Table 3, the Precision \((P)\) of a class, for example positive \(\text{‘}Pos\text{’}\), is defined as the ratio of the number of instances correctly classified as the class \(\text{‘}Pos\text{’}\) relative to the total number of instances predicted as the class \(\text{‘}Pos\text{’}\). The Recall \((R)\) of a class, for example \(\text{‘}Pos\text{’}\), is then defined as the ratio of the number of instances correctly classified as the class \(\text{‘}Pos\text{’}\) with respect to the total number of instances, which actually should be classified as the class \(\text{‘}Pos\text{’}\). The \(F1\) measure is a weighted harmonic mean of both, the Precision and Recall. The described metrics for the 3-class problem can easily be adapted for the 2-class problem by removing the Neutral column and row from Table 3. Based on the above-mentioned definitions, the \(P\), \(R\) and \(F1\) measures of the \text{‘}Pos\text{’} class are computed as follows:

\[ P(\text{Pos}) = \frac{a}{a + d + g} \]  

Formula (2)

\[ R(\text{Pos}) = \frac{a}{a + b + c} \]  

Formula (3)

\[ F1(\text{Pos}) = \frac{2 \times P(\text{Pos}) \times R(\text{Pos})}{P(\text{Pos}) + R(\text{Pos})} \]  

Formula (4)

**Performance of sentiment analyses in tourism studies**

The evaluation analyses, as introduced earlier, were performed on available tourism datasets (Table 4). The results indicate that the majority of tourism-related sentiment analysis studies used a binary polarity classification (e.g., Zheng & Ye, 2009; Ye et al., 2009; Gindl et al., 2010; Shimada et al., 2011; Zhang et al., 2011; Bjorkelund et al., 2012; Kang et al.,
Some studies followed a slightly different approach, whereby the sentiment analysis was divided into two subtasks, namely: (i) classifying sentences into objective and subjective sentences, and (ii) then determining the polarity of the subjective sentences (Marrese-Taylor et al., 2013; Riloff & Wiebe, 2003). Furthermore, a few studies followed an approach that involved determining polarity of sentences by using multiclass classifiers in a single step; that is they relied on a 3-class classifier (e.g., Ganu et al., 2009; Gräbner et al., 2012; Brob, 2013).

It is worth mentioning that the results are not directly comparable, as the sizes of the databases, the number of classes and the types of data are quite different for the evaluation of each method. There is some indication, however, that better results in terms of Accuracy and F-measure were obtained when only two classes (Positive, Negative) were used in the experimentations. To have a fair comparison of the results obtained using Twitter reviews, three methods (Levallois, 2013; Pappas & Popescu-Belis, 2013; Hutto & Gilbert, 2014) along with a publicly available Twitter based dataset for products (Sanders, 2011) were further considered for experimentation. The results are presented in Table 5. The method proposed by Levallois, (2013) has provided the best results in the binary classification problem. However, the VADER (Hutto and Gilbert, 2014) delivered the best results in the multi-class classification case. As these methods have provided reasonably good performances on Twitter data, and their lexicons also contain tourism related words and emoticons, they are applicable to the tourism domain.
Synthesis of the tourism-specific results

Based on the above insights, it can be noted that the majority of tourism sentiment analyses used a machine learning approach, often trained with small annotated datasets, as this process needs considerable human resources. Future more sophisticated sentiment analyses could draw on machine learning approaches using larger annotated datasets in combination with weakly annotated data to learn more complex rules making the use of potential correlations within data. Furthermore, future studies using lexicon-based methods could improve their performance by further adapting the sentiment lexicon to the tourism domain.

From the results reported in the literature, we noted that most sentiment analysis methods perform better in classifying positive sentences than negative or neutral sentences. One reason might be the existence of a larger number of positive texts and lexicons’ bias towards positivity, as human language is inherently biased towards positivity (Dodds et al., 2015). Moreover, analysing the negation in reviews is semantically a complex task. Related to the issue of bias, the review shows that the overall prediction performance of the methods can still be improved in both 2-class and 3-class sentiment analysis, but particularly in 3-class approaches. It appears that neutral reviews are difficult to detect in most of the sentiment analysis methods (Ribeiro et al., 2016). To address the above issues, tourism-specific analysis might benefit from transferring insights generated from other domains, for example, sentiment analysis of movies, products or advertisement. Should new areas of tourism be explored, beyond the current focus on hotels and restaurants, then new data sources and domain specific lexicons need to be considered.

Assuming further refinements, it is important to investigate whether these types of large scale sentiment analyses might impact on the tourism and travel industry in that they define new forms of customer feedback and service standards that are possibly tailored to specific
market segments. Approaches that take a broader destination-based perspective would be necessary. Such analyses would have to seek to understand the significant societal implications of social media, and Big Data beyond tourism (Ahlqvist et al., 2010).

Recognising the embeddedness of social media in people’s lives and behaviours will also help the tourism industry to develop better systems for product development and delivery, market research, and risk management, to name a few.

Advancing sentiment analysis, both conceptually and practically, means to focus analyses on specific targets or aspects mentioned in text. Target-specific polarity detection is a key challenge in the field of sentiment analysis, as the sentiment polarity of words and phrases may depend on the aspect. For example, considering the adjective word “small”, in the case of “small room” can be interpreted as negative, but in relation to a “handbag” it might be seen as positive. Further research on the relations between targets and expressions, and implications for sentiment, is necessary. Targets can be further defined through their aspects, and relations among aspects of a target can be modelled using ontology learning techniques (Maedche & Staab, 2001). However, depending on the application and the rules of grouping, hierarchical relationships between aspects and target can be different. Creating an appropriate taxonomy for aspects related to a target helps to determine more precise aspect-oriented sentiment analysis. One important problem in aspect-oriented sentiment analysis is to discover implicit aspects. Consider the following two reviews, for example: “our luggage was delivered very quickly”, and “it took an hour time to receive our luggage!!!”. The first example includes a subjective assessment (“quickly”), the second example merely states a fact (an hour time = late delivery). To provide a negative evaluation of the luggage delivery process in the second example, common sense knowledge is required to interpret that an hour is not acceptable for luggage delivery. In the literature, this form of implicit sentiment is
referred to as objective polar utterance (Fang et al., 2016), evaluative fact (Gräbner et al., 2012), or it is denoted as a polar fact (Leung et al., 2013).

In relation to the features, different types of features, such as lexical features (e.g., N-grams and Bag of Opinions), knowledge-based features (e.g., sentiment lexicons), and linguistic features (e.g., lemmatization, and syntax), and sentiment shifter features (e.g., negation, intensification, and neutralization) have frequently been considered for sentiment analysis in other domains. These features can also be employed on tourism related data to study their performance and applicability. Regarding the role of features in sentiment analysis, it is noted that sentiment shifters and negations most probably modify the sentiment polarity of an individual expression, a sentence, or even a whole document. The word order, contextual and dependency structure of individual phrases may also affect the polarity of a sentence. Features derived based on a sentiment lexicon improve sentiment analysis results. It has further been shown that the use of a simple Bag of Words (BoW) representation for sentiment analysis provides less favourable results compared to traditional topic classification, as in sentiment analysis, semantic information needs also to be modelled by BoW, which is a difficult task. Considering higher order N-grams and complex linguistic features are helpful for polarity classification and can improve the results significantly. However, the use of higher order N-grams and complex linguistic features is beneficial when large corpora are available for training the models. When using smaller corpora, a feature selection step is necessary to obtain satisfactory sentiment analysis results.

In relation to the kind of data used for sentiment analysis in tourism, we noted that most of the travel agencies and hotel booking service providers employ scalar ratings to rate users’ reviews, for example, scores between 1 and 5 stars. Such scores alone cannot help managers or service providers understand what the issues are and where improvements are necessary.
However, from an analytical perspective, the user-provided scores function as weakly labelled data, and can improve the classification accuracy, as well as help to verify polarity.

Concluding remarks and future directions

Whilst compelling in theory, in practice, the task of extracting and processing increasingly high velocity and large volumes of data has become very complex and made it necessary to develop automated machine-based approaches. Various methods exist to extract sentiment from online text, and these have been reviewed in this paper, both from a general and a tourism specific perspective. Due to the difficulty of detecting and finding implicit aspects in reviews, aspect-oriented sentiment analysis remains challenging. In relation to aspect-oriented sentiment analysis, future research requires close collaborations between domain experts (i.e. tourism researchers), information technology and NLP scientists to initially create and make publicly available some specific dictionaries for topics / aspects as well as annotated review databases related to industries involved in tourism. This will firstly help to design a more sophisticated aspect-oriented sentiment analysis model to better deal with the problem of implicit aspect detection in reviews. Secondly, it will enhance the research in the tourism domain by developing new hypotheses, for example, understanding the relation between satisfaction and sentiment (Xiang et al., 2015a) and then estimating tourist satisfaction by analysing aspect-oriented sentiment of text.

Moreover, using Big Data and deep learning approaches can help tourism research to discover dynamics based on large interconnected sets of data and getting more insight from different aspects of Big Data. Tourism research may further move into a new area, where theory driven approaches and data driven practices can support each other to understand or explain phenomena as well as to realize new dimensions in theories. This review article
concludes by suggesting that tourist sentiment analysis is the tip of an iceberg towards a new research paradigm for tourism. Sentiment analysis is only the beginning of more complex approaches using ‘Big Data’. In particular, the integration of several types of data has great potential for generating future insights at scales not seen before. Combining sentiment scores with other data, such as information on transport, weather, the environment, special events, crises, and other destination components may give rise to finding patterns that could not been seen and understood before. For example, the relative importance of weather conditions on visitor satisfaction, moderated by any other potential factors captured through various datasets, could be investigated. Adding other non-traditional data, such as imagery shared through Twitter or other social media, video footage (including security cameras), and electronic transaction footprints, enhanced by deep learning and object recognition technology can provide valuable information and reveal interesting insight that could not be hypothesised to those involved in tourism research and practice.

References


http://www.sananalytics.com/lab/twitter-sentiment/


Travelmail Reporter. 2013. “Travel is the main theme among Facebook posts - and we can't stay offline while away.” Daily Mail, 3 December 2013. Available at (29/01/16) http://www.dailymail.co.uk/travel/article-2517369/Travel-dominant-theme-Facebook-postings-stay-offline-holiday.html


### Table 1 Key terms and definitions

<table>
<thead>
<tr>
<th>Key term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
<td>Every topic or target (see below) in sentiment analysis has different features and characteristics. For example in tourism-related text, ‘restaurant’ as a potential target has various aspects, such as the food and atmosphere, ambiance, cleanliness, price, and location.</td>
</tr>
<tr>
<td>Bag of Words (BoW)</td>
<td>The BoW is a feature extraction method where the frequency of occurrence of each word in a given text/review, disregarding word order and grammar rules in the text, is used as a feature.</td>
</tr>
<tr>
<td>Classification and classifier</td>
<td>In machine learning, classification is the procedure that helps identify to which set of predefined groups a new sample belongs to. The model, which is called classifier, needs to initially 'learn' based on a training set of data that contains instances of text (or individual words) that are representative of a particular group. Once trained, the classifier can then perform the classification task on a new sample.</td>
</tr>
<tr>
<td>Confusion matrix</td>
<td>It is a table used to describe the performance of a classifier on a set of test data for which the true labelled are known.</td>
</tr>
<tr>
<td>Experimental analysis</td>
<td>To evaluate the performance of an algorithmic model, a set of tests/experiment is performed using training and testing data. Considering the results obtained from the test data, evaluation metrics are also computed. This process is called experimental analysis.</td>
</tr>
<tr>
<td>Feature extraction</td>
<td>Feature extraction is the process of building or deriving a set of discriminative, informative and non-redundant values from a set of data, which eventually facilitates the learning process.</td>
</tr>
<tr>
<td>Information Gain (IG)</td>
<td>IG is a feature selection strategy, which uses more important features or more discriminative features for the classification purposes.</td>
</tr>
<tr>
<td>Maximum entropy</td>
<td>Maximum entropy is a classifier, which mainly relies on the concepts of data uniformity and entropy. In the maximum entropy classifier, it is assumed that the probability distribution of the prior data that best represents the current state of data/knowledge should have the largest entropy.</td>
</tr>
<tr>
<td>N-gram</td>
<td>An N-gram is an adjacent order of N items in a given text (review) or speech. In a text (review) the items can be letters or words.</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>Naive Bayes classifier is a probabilistic classifier which works based on a strong assumption that features are all independence.</td>
</tr>
<tr>
<td>K-Nearest Neighbour (K-NN)</td>
<td>K-NN is an instance-based and non-parametric classifier used for classification, where K denotes the K closest training samples. The K-NN algorithm is one of the simplest machine learning algorithms.</td>
</tr>
<tr>
<td>Part-of-Speech (POS)</td>
<td>POS is a category of words (lexical items) which have similar grammatical properties (syntax, morphology) in English. Noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, and sometimes numeral, article or determiner are commonly listed English parts of speech.</td>
</tr>
<tr>
<td>Polarity</td>
<td>In sentiment analysis, the main problem is to determine to which extent a review is positive or negative. The positivity and negativity of reviews are two main poles of human feeling. Therefore, a review generally belongs to either positive or negative polarity.</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td>SVM is a supervised machine learning algorithm, which uses a separating hyperplane/line to categorise a given data. The hyperplane/line needs to be trained using labelled data in such a way that optimally segregates the data.</td>
</tr>
<tr>
<td>Conditional Random Fields (CRF)</td>
<td>CRF is a discriminative undirected probabilistic model which is especially used in NLP to parse a sequential data or predict sequences of class labels for sequences of input samples.</td>
</tr>
<tr>
<td>Target</td>
<td>In sentiment analysis, the topic (or particular subject of text) against which the analysis is performed is known as target. In tourism context, e.g., restaurants or hotels are targets.</td>
</tr>
<tr>
<td>Term Frequency (TF)</td>
<td>TF is the number of times an item (letters or words) occurs in a review.</td>
</tr>
<tr>
<td>Term Frequency–Inverse Document Frequency (TF-IDF)</td>
<td>TF–IDF is the product of TF and IDF. The IDF is a measure to show whether a term is common or rare across all reviews.</td>
</tr>
<tr>
<td>Unigram</td>
<td>Unigram is a special case of N-gram (defined above) where N=1.</td>
</tr>
<tr>
<td>Weakly labelled data</td>
<td>Data with the class labels determined heuristically by machine and not manually by human beings (such as star rating).</td>
</tr>
</tbody>
</table>
Table 2: A brief overview of the methods and datasets previously used for sentiment analysis in the domain of tourism

<table>
<thead>
<tr>
<th>Method</th>
<th>Type of approach</th>
<th>Source of data</th>
<th>Language</th>
<th>Type of reviews</th>
<th>Annotation type</th>
<th>No. of reviews</th>
<th>No. of annotated data</th>
<th>No. of positive reviews</th>
<th>No. of negative reviews</th>
<th>No. of neutral reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bjorkelund et al. (2012)</td>
<td>Machine learning</td>
<td>Tripadvisor.com,</td>
<td>English</td>
<td>Hotel</td>
<td>Automatic</td>
<td>794,962</td>
<td>794,962</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Markopoulos et al. (2015)</td>
<td>Machine learning</td>
<td>Tripadvisor.com,</td>
<td>Greek</td>
<td>Hotel</td>
<td>Semi-Automatic</td>
<td>1,200</td>
<td>1,200</td>
<td>900</td>
<td>900</td>
<td>0</td>
</tr>
<tr>
<td>Pablos et al. (2015)</td>
<td>Machine learning</td>
<td>Tripadvisor.com,</td>
<td>6 languages</td>
<td>Hotel</td>
<td>Manual</td>
<td>60,648</td>
<td>60,648</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Xiang et al. (2015b)</td>
<td>Machine learning</td>
<td>Expedia.com</td>
<td>English</td>
<td>Hotel</td>
<td>Automatic</td>
<td>1,800</td>
<td>1,800</td>
<td>900</td>
<td>900</td>
<td>0</td>
</tr>
<tr>
<td>Markopoulos et al. (2015)</td>
<td>Machine learning</td>
<td>Tripadvisor.com</td>
<td>English</td>
<td>Travel</td>
<td>Manual</td>
<td>1,191</td>
<td>1,191</td>
<td>600</td>
<td>591</td>
<td>0</td>
</tr>
<tr>
<td>Ye et al. (2009)</td>
<td>Machine learning</td>
<td>Travel, yahoo.com</td>
<td>English</td>
<td>Travel</td>
<td>Manual</td>
<td>52,264</td>
<td>3,400</td>
<td>1,904</td>
<td>612</td>
<td>1,884</td>
</tr>
<tr>
<td>Kang et al. (2012)</td>
<td>Machine learning</td>
<td>Restaurant websites</td>
<td>English</td>
<td>Restaurant</td>
<td>Manual</td>
<td>70,000</td>
<td>11,400</td>
<td>5,700</td>
<td>5,700</td>
<td>0</td>
</tr>
<tr>
<td>Zhang et al. (2011)</td>
<td>Machine learning</td>
<td>OpenRice.com</td>
<td>Cantonese (Chinese)</td>
<td>Restaurant</td>
<td>Manual</td>
<td>1,800</td>
<td>1,800</td>
<td>900</td>
<td>900</td>
<td>0</td>
</tr>
<tr>
<td>Rossetti et al. (2015)</td>
<td>Machine learning</td>
<td>YELP, Tripadvisor.com</td>
<td>English</td>
<td>Hotel and Restaurant</td>
<td>Automatic / Manual</td>
<td>12,342</td>
<td>3,733, 12,342, 200,000 / 116</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Shimada et al. (2011)</td>
<td>Machine learning</td>
<td>Twitter</td>
<td>English</td>
<td>Tourism</td>
<td>Automatic / Manual</td>
<td>10,000,000</td>
<td>3,733, 12,342, 200,000 / 116</td>
<td>100,000 / 64</td>
<td>100,000 / 52</td>
<td>0</td>
</tr>
<tr>
<td>Misopoulos et al. (2014)</td>
<td>Lexicon-based</td>
<td>Twitter</td>
<td>English</td>
<td>Airlines</td>
<td>Automatic / Manual</td>
<td>67,953</td>
<td>67,953, 1,587</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Gräbner et al. (2012)</td>
<td>Lexicon-based</td>
<td>Tripadvisor.com</td>
<td>English</td>
<td>Hotel</td>
<td>Automatic / Manual</td>
<td>80,000</td>
<td>80,000</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Garcia et al. (2012)</td>
<td>Lexicon-based</td>
<td>Tripadvisor.com</td>
<td>Spanish</td>
<td>Hotel and Restaurant</td>
<td>Automatic / Manual</td>
<td>1,994</td>
<td>1,994 / 40</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Bucur (2015)</td>
<td>Lexicon-based</td>
<td>Tripadvisor.com</td>
<td>English</td>
<td>Hotel</td>
<td>Manual</td>
<td>3,000</td>
<td>3,000</td>
<td>1,500</td>
<td>1,500</td>
<td>0</td>
</tr>
<tr>
<td>Marrese-Taylora et al. (2013)</td>
<td>Lexicon-based</td>
<td>Tripadvisor.com</td>
<td>English</td>
<td>Hotel and Restaurant</td>
<td>Manual</td>
<td>200</td>
<td>200</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Tan &amp; Wu (2011)</td>
<td>Lexicon-based</td>
<td>Tripadvisor.com</td>
<td>Chinese</td>
<td>Hotel</td>
<td>Manual</td>
<td>4,000</td>
<td>4,000</td>
<td>2,000</td>
<td>2,000</td>
<td>0</td>
</tr>
<tr>
<td>Xiang et al. (2015a)</td>
<td>Semantic approach</td>
<td>Tripadvisor.com</td>
<td>English</td>
<td>Hotel</td>
<td>Manual</td>
<td>60,648</td>
<td>60,648</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Kasper &amp; Vela (2011 and 2012)</td>
<td>Hybrid</td>
<td>Tripadvisor.com</td>
<td>German</td>
<td>Hotel</td>
<td>Manual</td>
<td>4,792</td>
<td>4,792</td>
<td>2,240</td>
<td>1,183</td>
<td>938</td>
</tr>
<tr>
<td>Chiu et al. (2015)</td>
<td>Hybrid</td>
<td>Wretch and Yahoo Blogs</td>
<td>Chinese</td>
<td>Hotel</td>
<td>Manual</td>
<td>2,147</td>
<td>2,147</td>
<td>1,899</td>
<td>248</td>
<td>0</td>
</tr>
<tr>
<td>Schmunk et al. (2014)</td>
<td>Hybrid</td>
<td>Tripadvisor.com,</td>
<td>English</td>
<td>Hotel</td>
<td>Manual</td>
<td>1,516</td>
<td>1,516</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Claster et al. (2010a), Claster et al. (2010b), Claster et al. (2013)</td>
<td>Hybrid</td>
<td>Twitter</td>
<td>English</td>
<td>Tourism</td>
<td>Automatic / Manual</td>
<td>70,570,800</td>
<td>200</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
</tbody>
</table>
Table 3 Confusion matrix of the results obtained for a general 3-class classification problem

<table>
<thead>
<tr>
<th>(Predicted)</th>
<th>'Pos'</th>
<th>'Neu'</th>
<th>'Neg'</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Pos'</td>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>'Neu'</td>
<td>d</td>
<td>e</td>
<td>f</td>
</tr>
<tr>
<td>'Neg'</td>
<td>g</td>
<td>h</td>
<td>i</td>
</tr>
</tbody>
</table>
Table 4 The sentiment analysis results obtained from different methods in the domain of tourism industry

| Method                        | Feature          | Classifier                  | Dataset                                      | No. of Annotated reviews | No of classes | A   | P   | R   | F-measure |  |
|-------------------------------|------------------|-----------------------------|----------------------------------------------|--------------------------|---------------|-----|-----|-----|-----------|-
| **Binary classification**     |                  |                             |                                              |                          |               |     |     |     |           |-
| Kasper & Vela (2011)          | N-gram           | Statistical classifier      | Hotel reviews                                | 4,792                    | 2             | 0.82| -   | -   | 0.80      |-
| Bjorkelund et al. (2012)      | N-gram           | Dynamic Language Model Classifier | Hotel reviews from tripadvisor.com              | 501,083                  | 2             | 0.90| -   | -   | -         |-
| Gindl et al. (2010)           | Stemmed words    | Naive Bayes                 | Travel reviews                               | 1,800                    | 2             | -   | 0.81| 0.78| 0.78      |-
| Ye et al. (2009)              | TF               | Naive Bayes                 | Travel reviews                               | 1,191                    | 2             | 0.807| 0.82| 0.82| -         |-
| Ye et al. (2009)              | TF               | SVM                         | Travel reviews                               | 1,191                    | 2             | 0.851| 0.851| 0.851| -         |-
| Zheng & Ye (2009)             |                  | SVM                         | Hotel reviews                                | 479                      | 2             | 0.912| 0.912| 0.901| -         |-
| Markopoulos et al. (2015)     | Unigram          | SVM                         | Hotel reviews                                | 1,800                    | 2             | 0.718| 0.65| 1   | 0.79      |-
| Bjorkelund et al. (2012)      | N-gram           | Dynamic Language Model Classifier | Hotel reviews from booking.com              | 293,879                  | 2             | 0.66 | -   | -   | -         |-
| Shimada et al. (2011)         | Unigram          | Naive Bayes                 | Tourism information                          | 116                      | 2             | 0.92 | -   | -   | -         |-
| Kang et al. (2012)            | N-gram           | Naive Bayes                 | Restaurant reviews                           | 11,400                   | 2             | -   | 0.737| 0.728| -         |-
| Zhang et al. (2011)           | N-gram           | SVM                         | Restaurant reviews                           | 1,800                    | 2             | 0.948| 0.948| 0.948| -         |-
| Zhang et al. (2011)           | N-gram           | Naive Bayes                 | Restaurant reviews                           | 1,800                    | 2             | 0.957| 0.957| 0.957| -         |-
| Chiu et al. (2015)            | N-gram           | SVM, Statistical classifier | Hotel reviews                                | 442                      | 2             | 0.89 | 0.91| 0.89| -         |-
| **Two-step classification**   |                  |                             |                                              |                          |               |     |     |     |           |-
| Marrese-Taylor et al. (2013)  | Lexicon          | Lexicon-based method        | Hotel and Restaurant                         | 200                      | 2/3           | 0.90| 0.93| 0.92| -         |-
| **Multi-class classification**|                  |                             |                                              |                          |               |     |     |     |           |-
| Kasper & Vela (2012)          | N-gram           | Statistical classifier      | Hotel reviews                                | 4,792                    | 3             | 0.81 | -   | -   | -         |-
| Schmunk et al. (2014)         | Bigrams          | SVM + Lexicon               | Hotel reviews                                | 1,516                    | 3             | 0.768| -   | -   | -         |-
| Pablos et al. (2015)          | Unigram          | SVM + CRF                   | Hotel reviews                                | 1,200                    | 3             | 0.76 | 0.49| 0.59| 0.59      |-
| Brob (2013)                   | Unigram          | SVM                         | Hotel reviews                                | 310                      | 3             | -   | 0.67 | 0.66| 0.68      |-
| Gräbner et al. (2012)         | Lexicon          | Lexicon-based method        | Hotel reviews                                | 80,000                   | 3             | 0.68 | 0.57| 0.62| 0.62      |-
| Ganu et al. (2009)            | Stemmed words    | SVM                         | Restaurant reviews                           | 3,400                    | 4             | 0.81 | 0.51| 0.45| 0.48      |-
| Bucur (2015)                  | Lexicon          | Lexicon-based method        | Hotel reviews                                | 3,000                    | 3             | 0.72 | 0.737| 0.856| 0.792     |-
| Garcia et al. (2012)          | Lexicon          | Lexicon-based method        | Hotel and Restaurant                         | 1,994/40                 | 3             | 0.80 | -   | -   | -         |-

39
Table 5 Comparison of the sentiment analysis results obtained from different methods on the Sanders’s Twitter dataset (Sanders, 2011)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Feature</th>
<th>Classifier</th>
<th>No of classes</th>
<th>A</th>
<th>P</th>
<th>R</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary classification</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pappas &amp; Popescu-Belis (2013)</td>
<td>POS + Lexicon</td>
<td>Lexicon + Nearest Neighbour model</td>
<td>2</td>
<td>0.70</td>
<td>0.72</td>
<td>0.71</td>
<td>0.715</td>
</tr>
<tr>
<td>Levallois (2013)</td>
<td>Lexicon</td>
<td>Lexicon-based method</td>
<td>2</td>
<td>0.82</td>
<td>0.83</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>Hutto &amp; Gilbert (2014)</td>
<td>Lexicon</td>
<td>Lexicon-based method</td>
<td>2</td>
<td>0.77</td>
<td>0.79</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>Two-step classification</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pappas &amp; Popescu-Belis (2013)</td>
<td>POS + Lexicon</td>
<td>Lexicon + Nearest Neighbour model</td>
<td>3</td>
<td>0.55</td>
<td>0.46</td>
<td>0.39</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>Multi-class classification</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levallois (2013)</td>
<td>Lexicon</td>
<td>Lexicon-based method</td>
<td>3</td>
<td>0.66</td>
<td>0.58</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Hutto &amp; Gilbert (2014)</td>
<td>Lexicon</td>
<td>Lexicon-based method</td>
<td>3</td>
<td>0.60</td>
<td>0.50</td>
<td>0.69</td>
<td>0.58</td>
</tr>
</tbody>
</table>
Figure 1. An overview of a machine learning based sentiment analysis system.
Figure 2. A general framework of the dictionary/rule-based sentiment analysis system. Dotted boxes indicate that these steps are optional or dependent on the particular model and application.