Aslib Journal of Information Management

Evaluating hotels rating prediction based on sentiment analysis services

Rutilio Rodolfo López Barbosa Salvador Sánchez-Alonso Miguel Angel Sicilia-Urban

Evaluating hotels rating prediction based on sentiment analysis services

Rutilio Rodolfo López Barbosa, Salvador Sánchez-Alonso and Miguel Angel Sicilia-Urban
Department of Computer Science, University of Alcala, Alcala de Henares, Spain

Received 13 January 2015
Revised 12 May 2015
Accepted 12 May 2015

Abstract

Purpose – The purpose of this paper is to assess the reliability of numerical ratings of hotels calculated by three sentiment analysis algorithms.

Design/methodology/approach – More than one million reviews and numerical ratings of hotels in seven cities in four countries were extracted from TripAdvisor web site. Reviews were classified as positive or negative using three sentiment analysis tools. The percentage of positive reviews was used to predict numerical ratings that were then compared with actual ratings.

Findings – All tools classified reviews as positive or negative in a way that correlated positively with numerical ratings. More complex algorithms worked better, yet predicted ratings showed reasonable agreement with actual ratings for most cities. Predictions for hotels were less reliable if based on less than 50-60 percent of available reviews.

Practical implications – These results validate that sentiment analysis can be used to transform unstructured qualitative data on user opinion into quantitative ratings. Current tools may be useful for summarizing opinions of user reviews of products and services on web sites that do not require users to post numerical ratings such as traveler forums. This summarizing may be valuable not just to potential users, but also to the service and product providers and offers validation and benchmarking for future improvement of opinion mining and prediction techniques.

Originality/value – This work assesses the correlation between sentiment analysis of hotels' reviews and their actual ratings. The authors also evaluated the reliability of results of sentiment analysis calculated by three different algorithms.

Keywords Sentiment analysis, Consumer-generated content, Intra-class correlation, Opinion mining, TripAdvisor reviews

Paper type Research paper

1. Introduction

Developing an automated method for identifying a customer's feelings or attitude toward people, services or products based on customer-generated content holds tremendous value for researchers, corporations and customers themselves. This is the goal of sentiment analysis, which aims to identify customer opinions or attitudes on the basis of their spoken or written comments. Converting these opinions and attitudes into numbers is a way to rapidly synthesize and analyze customer experiences, allowing customers to make decisions about buying the product or contracting the service, and allowing companies to make decisions about launching new products or redesigning products and services.

The significant impact of online reviews on electronic word of mouth (eWOM) communication is now well established in the literature (Trusov *et al.*, 2009; Zhu and Zhang, 2006). The growing reliance of consumers on online reviews of products and services has led to such an abundance of user-generated content that no potential customer can hope to sift through it all. Sentiment analysis tools, which can condense

Aslib Journal of Information Management
Vol. 67 No. 4, 2015
pp. 392-407
© Emerald Group Publishing Limited
2050-3806
DOI 10.1108/AJIM-01-2015-0004

large amounts of text comments into a few easily digestible numbers, can provide a powerful way to aggregate and summarize the full range of opinions.

The tourism and hospitality sector is an excellent example of an industry in which the success of products and services increasingly depends on large amounts of user-generated content posted on social media sites. Sites such as like.com, Booking.com and HolidayCheck.com allow users not just to rate hotels and their amenities on numerical scales, but also to write their opinions or attitudes as text. However, there are many other web sites, especially traveler forums, such as travelerspoint.com, lonelyplanet.com and losviajeros.com that allow users to share their experiences by writing comment without issuing any numeric rate. Thus, these web sites would improve the user search by posting automatically calculated numeric ratings. The text comments are a powerful form of eWoM and are no less important than the numerical ratings (Gretzel and Yoo, 2008).

Testament to the potential of sentiment analysis for harnessing the power of eWoM is the fact that dozens of commercial and public tools have been developed for this purpose. OpinionFinder was developed by teams of researchers at the University of Pittsburg, Cornell University and the University of Utah comparisons (Wilson *et al.*, 2005a). This algorithm uses a lexicon to identify sentiment expressions based on context (Wilson *et al.*, 2005b). The Recursive Neural Tensor Network (RNTN) tool was developed by researchers at Stanford University; it works by labeling phrases in parse trees of sentences using a data set called Sentiment Treebank, and it functions as a sentiment analysis annotator in the Stanford CoreNLP (Socher *et al.*, 2013). CoreNLP is an integrated suite of natural language processing (NLP) tools for English (Manning *et al.*, 2014). Our group has developed an unsupervised lexicon-induced sentiment analysis tool called SentUAH, which uses the tokenizer, sentence splitter and part-of-speech (POS) tagger from CoreNLP. The tool works in combination with SentiWordNet (Esuli and Sebastiani, 2006; Baccianella *et al.*, 2010) and a naïve Bayesian approach to data mining. The aim of developing this tool was not to improve efficiency but to confront the results of a simple algorithm (naïve Bayes which is well known as efficient for specific cases) with more complex algorithms and assess the reliability of all of them to predict numerical ratings with a large amount of data. We are unaware of studies benchmarking these software tools against a large experimental data set or against one another. Such a study is important for validating sentiment analysis tools in the field, and for guiding the future improvement of these programs.

Therefore we compared the three software tools for their potential ability to predict numerical hotel ratings based on text comments. We examined more than 1 million reviews of 3,535 hotels in seven cities posted on TripAdvisor.com, this was all the English comments, more than 75 percent of the total amount. We used sentiment analysis to classify the comments as positive or negative and thereby generate predicted ratings from those comments. Then we compared the model-generated ratings with the actual ratings. This is the first study to our knowledge that assesses the ability of sentiment analysis to provide a bridge between comments per hotel (qualitative) and ratings (quantitative) from the same source doing this at the review (comment) level instead of at the sentence level.

2. Background

2.1 Sentiment analysis

Sentiment analysis is a subfield of NLP that draws on approaches from information retrieval and computational linguistics to identify opinions expressed in text. It is considered a specific type of text mining (Han *et al.*, 2011), and it has been called opinion mining. While the terms appraisal extraction or review mining have also been applied, they are not always completely accurate (Pang and Lee, 2008). The main goal of sentiment analysis is to identify positive or negative overall attitudes or opinions toward a brand, product or service based on text comments (Liu, 2010; Han *et al.*, 2011).

Several machine learning and data mining algorithms have been used to detect sentiment (Khoo *et al.*, 2012), mood (Mishne, 2005) and sentiment strength (Thelwall *et al.*, 2010). Most of these algorithms are reasonably effective (Wiegand and Klakow, 2010; O'Connor *et al.*, 2010). Even simple algorithms have been shown to work well with large data sets, as in the case of the naïve Bayesian approach (Wu and Kumar, 2008). More complex algorithms are sometimes needed for particular contexts. Pang and Lee (2005) predicted star ratings of movie reviews based on a five-point sentiment scale instead of merely classifying the reviews as positive or negative. They employed a novel similarity measure with a meta-algorithm based on metric labeling and performed several comparisons of pairs of reviews to identify when the first review was less positive than, more positive than or as positive as the second review. Bai (2011) proposed a heuristic search-enhanced Markov blanket model to capture dependencies among words when extracting sentiment from movie reviews; the author found that combining a Markov blanket with Tabu searching to analyze word dependencies together with keywords and high-frequency words can lead to more reliable sentiment detection than using naïve Bayes, support vector machine or maximum entropy approaches.

Other authors have used before the transformation of sentiment analysis results into numbers. Ganu *et al.* (2009) used positive-sentence percentage (PSP) to evaluate ratings in restaurants. PSP has motivated much sentiment analysis work (Pang and Lee, 2005) and is a technique to rate a single comment by computing the percentage of its sentences rated as positive. The present paper is about to rate a hotel by the transformation of the positive percentage of its comments.

The OpinionFinder algorithm identifies sentiment expressions based on context, while the RNTN tool identifies the sentiment in phrases; the latter is used as a sentiment analysis annotator in Stanford CoreNLP. SentiWordNet is a publicly available resource containing words and their associated sentiment scores for positive, negative or objective connotation. SentiWordNet has been incorporated into various algorithms to analyze the polarity of customer reviews, such as film reviews (Ohana and Tierney, 2009) and Amazon product reviews (Hamouda and Rohaim, 2011). SentiWordNet has also been used in conjunction with semi-supervised, machine-learning algorithms for classification. Ye *et al.* (2009) compared the effectiveness of three machine-learning algorithms for classifying online reviews from travel blogs: Support Vector Machine, N-gram and Naïve Bayes with SentiWordNet. Those authors found that in general, the larger the training data set, the better the algorithm performs.

Our group has developed an unsupervised lexicon-induced sentiment analysis tool that works in combination with SentiWordNet and takes a naïve Bayesian approach. It identifies sentiment polarity by using SentWordNet scores to indicate the probability that a given word reflects a positive or negative sentiment.

2.2 Importance and economic impact of online reviews

According to Gretzel and Yoo (2008), 75 percent of travelers visit sites like TripAdvisor.com to read reviews about hotels from other users and to guide decisions to book or hire hospitality services. This is the essence of eWOM: users of products and services affect one another's behavior via informal communication on the internet (Litvin *et al.*, 2008). With eWOM, users can openly share their opinions of brands, products or services

(Jansen *et al.*, 2009). This consumer-generated content can significantly influence product sales (Zhu and Zhang, 2006). Thus, users and sellers alike look to online reviews as an important source of feedback about services and products.

Various studies have demonstrated that potential consumers of tourism services prefer recommendations by other consumers over seller advertising, and that such reviews can be the most influential factor in customer travel decisions (Pan *et al.*, 2007; Gretzel and Yoo, 2008). While user reviews in the form of numbers (ratings) are important, so are text comments (Ghose *et al.*, 2009). In fact, Ghose and Ipeirotis (2011) found that the subjectivity, readability and linguistic correctness of text comments makes a difference to how other users perceive the comment and to how much a particular comment influences sales.

The disadvantage of text comments is that they can quickly accumulate on a web site, such that a user cannot access the full range of comments; priority may be given, for example, to comments that have been posted more recently. The site TripAdvisor.com is home to more than 75 million of registered users (Brown, 2013) and 150 million reviews (TripAdvisor.com Fact Sheet, 2014, www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html); it hosts 260 million unique visitors each month (Kaufer, 2014). Many hotels feature hundreds or thousands of reviews and many cities have hundreds or thousands of hotels. It is impractical for potential consumers to read a large number of reviews to get an overall insight into the hotel's quality of service.

If users and sellers had access to an algorithm to convert a large number of text comments to numerical ratings, it would be much easier to assess the polarity of the comments and gain a clearer global picture of user experience. Sentiment analysis seems well suited to this task, but it has yet to be validated for converting text opinions into numbers.

2.3 Problem statement

Before the arrival of eWOM and related technology, economic and marketing researchers used surveys to gather structured data about consumers' opinions of a product or service. In contrast, user-generated reviews provide opinions usually in an unstructured format in the user's own language, making them richer sources of data but also more difficult to quantify and analyze automatically.

Given the potential of sentiment analysis for allowing large amounts of text data to be converted to a few quantitative measures of sentiment, we undertook the present study to address the following research questions:

RQ1. Does sentiment analysis of hotel reviews correlate with overall ratings on TripAdvisor?

RQ2. Can sentiment analysis of text comments reliably predict overall ratings on TripAdvisor?

3. Methods

We gathered 1,335,781 reviews from every registered hotel in seven cities from TripAdvisor.com, together with overall ratings for hotels (on a scale of 1-5), that is the average of ratings that users are asked to enter in addition to comments. Three of the cities were London, Paris and New York, which are among the cities receiving the largest numbers of tourist visits in the world (Bremner and Grant, 2014; Hedrick-Wong and Choog, 2013). The remaining cities were selected to provide a range of tourist destinations: Las Vegas, Nevada; Anaheim, California; Santa Ana, California; and Alcalá de Henares (Spain).

By accessing TripAdvisor.com we were able to obtain both quantitative and qualitative information from the same web site and for the same user. We applied OpinionFinder, Stanford CoreNLP and our own algorithm (SentUAH) to the comments to determine the percentage of positive reviews for each hotel, and we checked whether these percentages correlated with the actual overall rating for each hotel. Then we converted the percentages into "predicted" overall ratings, which we compared with the actual ratings. Pearson correlation and scatter plots were used to determine correlations between sentiment analysis of comments and overall ratings. Cronbach's α and intra-class correlation (ICC) were used to assess the reliability of predicted overall ratings.

3.1 Data extraction

We developed a crawler based on HtmlUnit API to extract data from TripAdvisor. The crawler navigated through the web site and gathered reviews and overall ratings for every hotel registered for the selected cities. This method can be quite effective for collecting data from dynamically generated web sites (Gerdes and Stringam, 2008).

Data were collected between April 2013 and May 2014. Because some of the tested tools were intended to work only with English-language text, most of the selected cities feature English as the spoken language (Bremner and Grant, 2014; Hedrick-Wong and Choog, 2013). Nevertheless, to test sentiment analysis in a variety of contexts, we also selected cities in countries where the primary language was not English, as well as cities that receive relatively small numbers of tourists. Table I shows the amount and percentage of reviews in English for every selected city.

3.2 Sentiment analysis

We compared three sentiment analysis algorithms. We selected the tools and resources because all of them have been tested for sentiment analysis in different contexts with acceptable efficiency (Socher *et al.*, 2013; O'Connor *et al.*, 2010; He *et al.*, 2008; Wu and Kumar, 2008).

Two of these algorithms are implemented in publicly available tools, while the third is an algorithm that we developed based on unsupervised naïve Bayesian data mining using publicly available resources. OpinionFinder and sentiment annotator in Stanford CoreNLP use complex algorithms and both consider special linguistic characteristics such as negations, intensifications, modalities and comparisons (Wilson *et al.*, 2005a, b; Socher *et al.*, 2013). The third tool is based on a simpler algorithm (Naïve Bayes) combined with sentiment lexicon, and does not purposefully considers negations nor

Table I. Basic information on downloaded hotel reviews for sentiment analysis

City and country	Total no. of reviews	No. of reviews in English	% of reviews in English	No. of hotels in city
London, UK	483,478	372,755	77.10	1,021
New York City, USA	312,307	242,661	77.70	418
Las Vegas, USA	203,208	180,966	89.05	223
Paris, France	294,995	173,026	58.65	1,739
Anaheim, USA	39,698	37,817	95.26	90
Santa Ana, USA	1,576	1,523	96.64	24
Alcala de Henares, Spain	519	123	23.70	20
Total	1,335,781	1,008,871	75.53[a]	3,535

Note: [a]Average of percentages

other complex characteristics for the sake of simplicity. The aim of developing the latter was to confront the results of a simple algorithm with more complex algorithms when analyzing a large amount of data for sentiment and assess the reliability of their results to predict numerical ratings. Because some of the tools were intended to work only in English, we limited our analysis to the reviews in English (1,008,871). The following sections will provide details of the tools and approaches used to analyze sentiment polarity in the reviews. We focussed only on whether the sentiment of the overall comment was positive or negative, not on sentiment strength. This simplification allowed us to validate and compare different sentiment analysis tools as a necessary first step toward more sophisticated studies.

3.2.1 OpinionFinder (OFV2). OpinionFinder comprises several software packages and although originally intended to detect subjectivity (Wilson *et al.*, 2005a), several researchers have used its sentiment detection feature to classify documents according to sentiment (O'Connor *et al.*, 2010; He *et al.*, 2008). The sentiment detection feature (Wilson *et al.*, 2005b) is a modification of a "boosting" machine-learning algorithm and is based on context: for example, whether "love" is an expression of positive opinion or attitude depends on the context. We used version 2.x of this software, which was released at the end of 2013 (OFV2).

Since OpinionFinder identifies every sentence in a document or text and then analyzes sentiment expression for every sentence, each review (r_i) was treated as a collection of sentences $\{s_1, s_2, \ldots, s_n\}$:

$$r_i = \{S_1, S_2, \; \ldots, \; S_n\} \tag{1}$$

Given the set classes C with c_1 (positive) and c_2 (negative):

$$C = \{c_1, c_2\} \tag{2}$$

Expressions of "neutral" sentiment do not give any positive or negative opinion but only express objective facts, so they were not taken into account when classifying sentences and reviews.

Sentence classification: the class of every sentence was calculated according to:

$$C^{*}(S_i) = \underset{j}{argmax} \left(\sum_{k=0}^{n} e_k c_j \right) \tag{3}$$

where *argmax* of *j* is the argument *j* (positive or negative class) for which the sum gets its maximum result, e_k is every sentiment expression and c_j is the corresponding class.

Review classification: the review class was determined using:

$$C^{*}(r_i) = \underset{j}{argmax} \left(\sum_{k=0}^{n} s_k c_j \right) \tag{4}$$

Every sentence was classified as positive or negative based on whether most expressions in it were positive or negative (Equation (3)). Then the overall polarity of the review was determined based on whether most sentences were positive or negative (Equation (4)).

3.2.2 Stanford CoreNLP (RNTN). Stanford CoreNLP is a set of NLP tools that provide model files for analysis of English texts. Every integrated tool is intended for a specific NLP task; these tools include a POS tagger, named entity recognizer, parser, co-reference resolution system and sentiment analysis. Every tool can easily be activated as an annotator. Every annotator is related to a group of annotations, which are simply the data generated for each tool.

A sentiment annotator has recently been integrated into the CoreNLP; it is based on RNTN technology, which determines the sentiment of a sentence based on word composition in phrases (Socher *et al.*, 2013). This annotator is a supervised multiclass classifier that assigns every sentence to one of five classes: very positive, positive, neutral, negative and very negative. The algorithm performs this classification by identifying specific *n*-grams with different intensities of positive or negative sentiment. The sentence: “Your cat is beautiful,” is classified as “very positive” and carries the following annotation:

((your_0 cat_0)_0((is_0 beautiful_++)_++._0)_+)_++

where “_++,” “_+,” “_0,” “_−” and “_−−” are the tags for very positive, positive, neutral, negative and very negative, respectively. Figure 1 shows the tree resulting from the sentence classification.

For our experiment, we used the sentiment annotator of CoreNLP API to analyze hotel reviews.

Sentence classification: in order to compare results across all three tools in our analysis, we transformed this five-class classification into a three-class classification. Thus, we classified sentences as positive, negative or neutral by considering the intensifier “very” simply as an indicator of double strength. Thus, one “very positive” sentence was weighted as two “positive” sentences, and the same was done for “very negative” sentences.

Review classification: we calculated the polarity of reviews by summarizing the polarity of the sentences within it and then assigning the predominant polarity class to the entire review. We used the same Equation (4) as in OpinionFinder.

3.2.3 SentUAH. We used SentiWordNet combined with a naïve Bayesian approach to develop our classifier. SentiWordNet is an opinion lexicon derived from WordNet where every term is associated with three numerical scores. The terms are classified into four categories: adverbs, adjectives, nouns and verbs. The scores indicate the extent to which each term is positive, negative or objective. Esuli and Sebastiani (2006) and Baccianella *et al.* (2010) evaluated the effectiveness of SentiWordNet and found it to be adequate for opinion mining. Ohana and Tierney (2009) and Hamouda and Rohaim (2011) further showed that SentiWordNet can function reliably in sentiment classification tasks. Saggion and Funk (2010) classified sentences as positive based on the number of times a term in SentiWordNet was more positive than negative and vice versa.

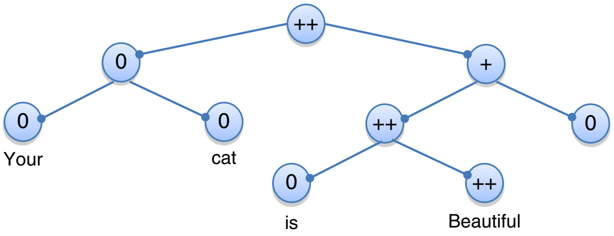

Figure 1.
Example of text classification using RNTN

We adopted the method proposed by Hamouda and Rohaim (2011) to summarize SentiWordNet term scores in lieu of the less effective term-counting method. We interpreted scores as probabilities that terms belonged to a positive or negative class (Esuli and Sebastiani, 2006), and we developed a sentiment analysis tool that determines polarity of sentences based on a naïve Bayesian approach rather than based on simple counts of terms with higher scores. The words in SentiWordNet may have diverse senses and belong to different synsets with different positive, negative and objective scores. Instead of performing word sense disambiguation, we determined the scores for each word by calculating the average of the scores of its entries according to their corresponding categories (adjectives, adverbs, nouns or verbs).

Sentence classification: we identified polarity expressions for each sentence with the help of the CoreINLP POS tagger, after which we calculated scores for SentiWordNet. The sentiment polarity (class) of a sentence $C^{*}(s_i)$ was calculated using:

$$C^{*}(s_i) = \underset{j}{argmax}\left(\Pi_{k=1}^{n} p\left(t_k \middle| c_j\right)\right) \tag{5}$$

where p is the probability that the term t_k belongs to class c_j.

Review classification: we used Equation (4) to determine the polarity of reviews based on the predominant sentence class.

Figure 2 depicts the entire review classification process. SentUAH performs sentence splitting, tokenizing and POS tagging on original reviews. Then, every adjective, adverb, noun and verb is annotated with the corresponding calculated score (sentiment identification), after which sentences and reviews are classified for sentiment polarity as described above and stored in the database. Both the original reviews with their corresponding actual ratings and the classified reviews are stored in the same database.

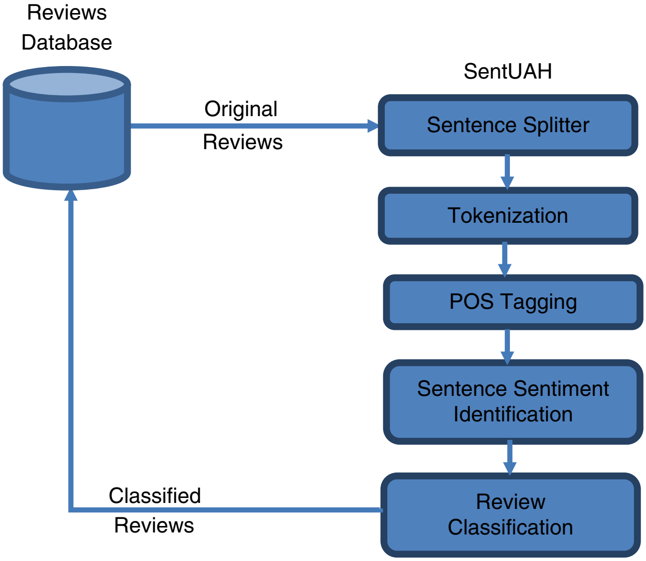

Figure 2. Sentiment classification of reviews using SentUAH

3.3 Statistical analysis

Pearson linear correlation was used to determine the existence and strength of association between a hotel's user-assigned ratings and percentages of reviews classified as positive by each sentiment analysis algorithm.

Cronbach's α was used to measure the internal consistency between published ratings and ratings calculated on the basis of positive percentages of reviews. It is frequently used to measure inter-rater reliability. We also assessed reliability using ICC, which we calculated using the "two-way random" approach since each tool is considered a rater classifying every review and we evaluated only reviews in English. For both Cronbach's α and ICC, a coefficient > 0.9 was considered excellent; 0.81-0.9, good; 0.71-0.8, acceptable; 0.61-0.7, uncertain; 0.51-0.6, poor; and < 0.5, unacceptable (George and Mallery, 2003).

4. Results and discussions

The three sentiment analysis tools were applied to the extracted reviews. The summarized results per hotel were compared with actual overall ratings.

When users submit a review to TripAdvisor.com, they are asked to indicate a rating for the hotel on a five-point scale: 1 (terrible), 2 (poor), 3 (average), 4 (very good) and 5 (excellent). The overall rating for a specific hotel is a simple average of ratings from all users. The web site presents these ratings as multiples of 0.5.

The percentage of reviews classified as positive or negative was calculated over the sum of reviews classified as positive or negative. In other words, we ignored neutral reviews because they express objective information, such as "The hotel was located 5 km from downtown."

4.1 Correlation between overall ratings and percentages of reviews classified as positive

We calculated correlations between the overall rating and the percentage of reviews classified as positive by each sentiment analysis tool. The two types of data correlated positively for all three algorithms (Table II).

OFV2 showed the best correlation for most cities, followed by RNTN. The behavior of naïve Bayesian analysis in our SentUAH model was still acceptable for most cities, even with those for which much smaller proportions of reviews were available. Figures 3-5 compare the correlations obtained with the three algorithms for the three most visited cities in our data set. These figures confirm the trend in Table II that OFV2 and RNTN gave better correlations than a naïve Bayesian approach like that in SentUAH and show that more elaborated algorithms considering linguistic issues such as negations,

Table II. Correlations between overall rating and percentage of reviews classified as positive by each tool

		Correlation	
City	OFV2	RNTN	SentUAH
Alcala de Henares	0.496	0.204	0.604
Paris	0.680	0.688	0.526
Las Vegas	0.716	0.691	0.536
Santa Ana	0.672	0.634	0.473
Anaheim	0.737	0.681	0.675
New York City	0.752	0.741	0.683
London	0.814	0.829	0.672

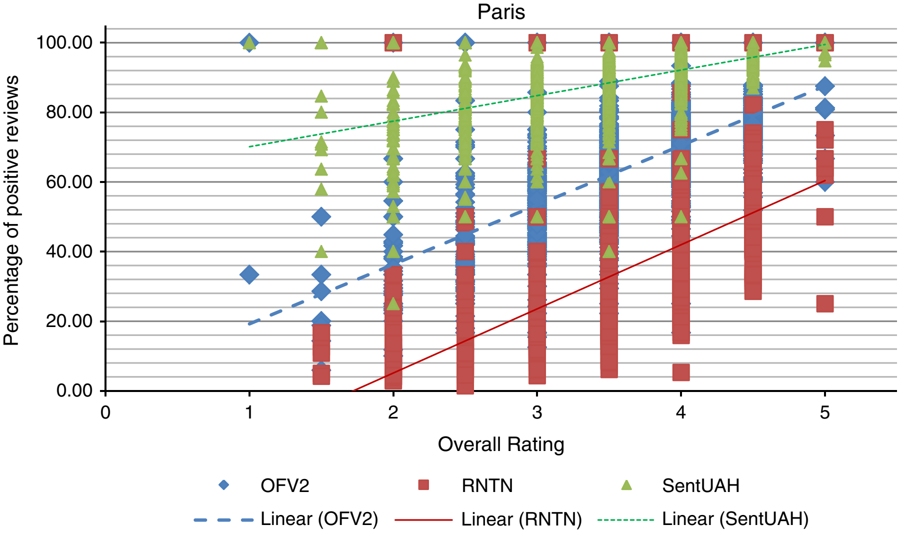

Figure 3. Correlation between overall rating and percent of positive reviews for Paris

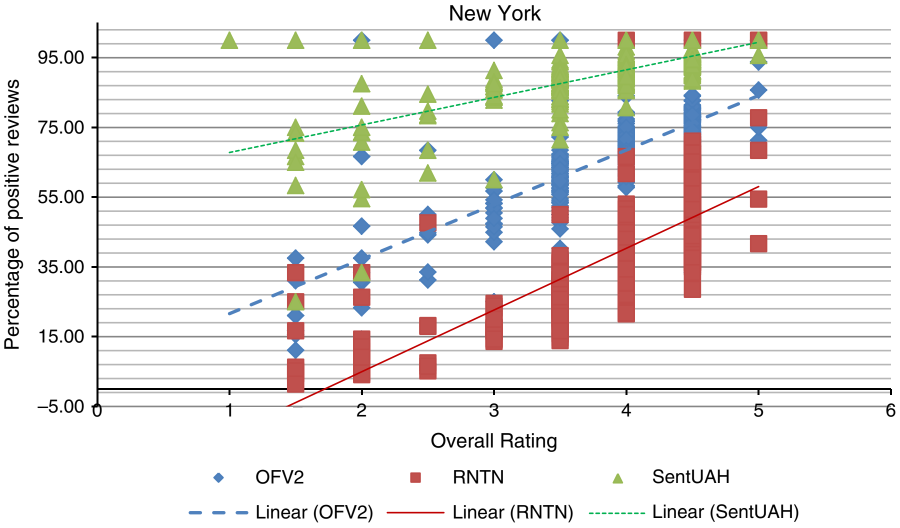

Figure 4. Correlation between overall rating and percent of positive reviews for New York

intensifications, modalities and comparisons work in fact more efficiently than the Naïve Bayesian approach with a big amount of reviews.

Next we compared our measured correlations with previously reported accuracy results obtained with each of the three algorithms (O'Connor *et al.*, 2010; Socher *et al.*, 2013; Hamouda and Rohaim, 2011) (Table III).

The reported algorithms' accuracy from previous research is the result of the analysis of each of those algorithms with different data sources. Most of the data used in those research studies is in fact very similar to the data in our experiments (RNTN: movies reviews; OpinionFinder: Twitter and polls; Naïve Bayes with sentiment lexicon:

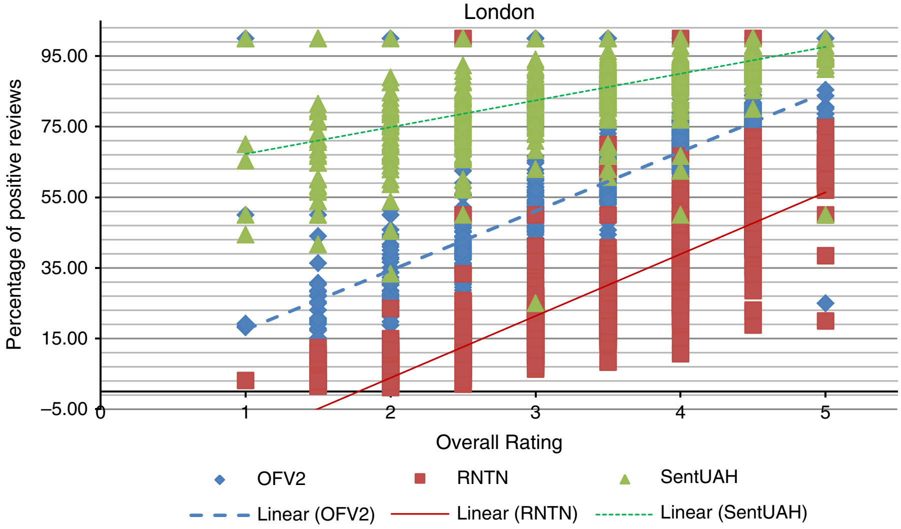

Figure 5. Correlation between overall rating and percent of positive reviews for London

Table III. Comparison of previously reported accuracy and average correlation for each sentiment analysis tool in the present study

Algorithm	Reported accuracy	Range	Correlations in the present study Average for all cities	Average for cities where > 50% of reviews are in English
OFV2	73.9-80.0	0.496-0.814	0.695	0.729
RNTN	85.4-87.6	0.204-0.829	0.638	0.711
Naïve Bayesian (SentUAH)	64.95-68.63	0.473-0.683	0.596	0.594

products reviews). The subjacent idea of comparing reported accuracy from previous research with the resulting correlations of present experiment is that algorithms with better accuracy should ideally generate better correlations. Nevertheless, RNTN which reported the best accuracy shows lower correlations than OFV2. SentUAH was consistent with previous reports of accuracy, especially for cities for which greater than 50 percent of reviews were available. Even when the latter algorithm (SentUAH) is the lowest – compared to the others – it is still acceptably reliable.

4.2 Prediction of overall ratings

Considering that TripAdvisor overall ratings are expressed in the range of 1-5 in multiples of 0.5, we divided the percentage of reviews classified as positive by 20 to predict the overall rating, expressing it in multiples of 0.5, and always rounding it up to the nearest upper 0.5 fraction. Thus, percentages between 0 and 10 were converted to a rating of 0.5; percentages between 11 and 20, to a rating of 1; and so on. In this way, a positive percentage of 70-80, meaning that 20-30 percent of users expressed negative opinions, was transformed into a rating of 4. Cronbach's α was used to determine how close the calculated overall ratings matched with the actual overall ratings on the web site (Table IV). Similar results were obtained based on the ICC (data not shown). The results of the SentUAH algorithm for the Alcala de Henares City are better when compared to OFV2 and RNTN. This could lead us to infer that the Naïve Bayes method can still work reliably in cases with smaller amounts of data, although this result could also be circumstantial. To find out the reason, we carried out a more in depth analysis with those four cities in our experiment with a smaller number of reviews. The analysis showed an increase on the correlation coefficient for the SentUAH algorithm but only for those hotels with a smaller amount of reviews. Unfortunately, this behavior is not consistent so further research selecting more cities with same percentage of English reviews is necessary to effectively assess the SentUAH's performance for those cases.

All three tools gave acceptable or nearly acceptable Cronbach α values for most cities, and as observed in Table II, OFV2 and RNTN showed greater reliability than SentUAH. These data, together with those described above, suggest that while simple naïve Bayesian method of sentiment analysis shows potential, algorithms based on "boosting" machine learning and recursive neural tensor networks perform better, especially when greater than 50 percent of a hotel's reviews are included in the analysis. Best results were obtained in our study when at least 60 percent of available reviews were analyzed. Thus simpler algorithms like the naïve Bayesian approach can provide acceptable results, but they require substantial improvement. Regardless of the algorithm used, training the system with as high a proportion of available data will be crucial for making reliable predictions, consistent with the findings of Ye *et al.* (2009).

5. Conclusions and future work

This paper describes using sentiment analysis to predict overall hotel ratings from text comments. First we compared three different algorithms and showed that all three classified text comments in a way that correlated positively with actual ratings, validating for the first time that sentiment analysis of text can reliably generate quantitative data on user opinions and attitudes. When the predictions of the three algorithms were compared with actual ratings using Cronbach's α coefficient and ICC, all three models showed acceptable or nearly acceptable reliability for most cities examined. The more complicated algorithms based on "boosting" machine learning and recursive neural tensor networks performed substantially better than the relatively simple Naïve Bayesian algorithm.

Our results indicate that, in response to *RQ1*, sentiment analysis of hotel reviews does indeed correlate with overall ratings on TripAdvisor, even when a fairly simple algorithm is used. Our results further indicate that, in response to *RQ2*, sentiment analysis of text comments can be used to predict global numerical ratings. Predictions appear to be more reliable when complex algorithms and at least 50-60 percent of available reviews are used. We conclude that currently available

			Cronbach's α	
City name	% of reviews in English	OFV2	RNTN	SentUAH
Alcala de Henares	23.7	0.279	0.289	0.704
Paris	58.7	0.705	0.758	0.623
Las Vegas	89.1	0.821	0.830	0.750
Santa Ana	96.6	0.861	0.839	0.649
Anaheim	95.3	0.844	0.791	0.770
New York City	77.7	0.833	0.817	0.655
London	77.1	0.870	0.884	0.724
Average		0.745	0.744	0.696

Table IV. Cronbach's α coefficients to measure agreement between actual overall ratings and ratings generated by each of the sentiment analysis tools

sentiment analysis tools are indeed reliable for predicting hotel ratings. Under our conditions, average reliability was 74 percent over a range of cities involving 3,500 hotels.

A good amount of web sites allow travelers to share their experiences but not all of them allow users to assign numeric rating to hotels, especially traveler forums which can improve the users search for recommendations by implementing an automatic rating prediction.

This research supports the growing focus on applying sentiment analysis to hospitality sites and blogs (Choi *et al.*, 2007; Pan *et al.*, 2007) in order to give users and sellers alike the ability to analyze large data sets of opinions and attitudes. Future research should extend the present study to analyze what amenities or features of a product or service are more likely to lead to positive user reviews. Though studies have examined what services generate more opinions, few studies have looked at what services generate more positive opinions. Sentiment analysis is well suited to this task. For example, it should be possible to download hotel reviews and summarize positive and negative comments toward specific hotel services (reception, laundry, room service, cleanliness) and amenities (restaurant, internet, pool). The work of Barreda and Bilgihan (2013) is an important step in this direction; those authors analyzed hotel reviews and identified specific aspects of hotels that were more likely to generate positive or negative comments in users.

This research would be easier if a lexicon containing terms most commonly used in hospitality reviews were available. Building such a lexicon will require caution, since as Ye *et al.* (2009) point out, words such as "unpredictable" may have negative meaning for a tourism-related product or service, but not necessarily for an adventure-tourism experience.

References

Baccianella, S., Esuli, A. and Sebastiani, F. (2010), "SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", *Proceedings of the Language Resources and Evaluation International Conference, European Language Resources Association*, *Valetta*, pp. 2200-2204.

Bai, X. (2011), "Predicting consumer sentiments from online text", *Decision Support Systems*, Vol. 50 No. 4, pp. 732-742.

Barreda, A. and Bilgihan, A. (2013), "An analysis of user-generated content for hotel experiences", *Journal of Hospitality and Tourism Technology*, Vol. 4 No. 3, pp. 263-280.

Bremner, C. and Grant, M. (2014), "Top 100 city destinations ranking – analyst insight from Euromonitor International", available at: http://blog.euromonitor.com/2014/01/euromonitor-internationals-top-city-destinations-ranking.html (accessed June 14, 2014).

Brown, A. (2013), "Here's why we believe TripAdvisor's user base will continue to climb", Trefis.com to Forbes, available at: www.forbes.com/sites/greatspeculations/2013-/03/08/heres-why-we-believe-tripadvisors-user-base-will-continue-to-climb/ (accessed July 5, 2014).

Choi, S., Lehto, X.Y. and Morrison, A.M. (2007), "Destination image representation on the web: content analysis of macau travel related websites", *Tourism Management*, Vol. 28 No. 1, pp. 118-129.

Esuli, A. and Sebastiani, F. (2006), "SentiWordNet: a publicly available lexical resource for opinion mining", *Proceedings of the Language Resources and Evaluation International Conference, European Language Resources association*, *Genoa*, pp. 417-422.

Ganu, G., Elhadad, N. and Marian, A. (2009), "Beyond the stars: improving rating predictions using review text content", *Proceedings of the Web and Databases 12th International Workshop, Providence, RI, June 28.*

George, D. and Mallery, P. (2003), *SPSS for Windows Step by Step: A Simple Study Guide and Reference, 17.0 Update*, Pearson Education.

Gerdes, J. and Stringam, B.B. (2008), "Addressing researchers' quest for hospitality data: mechanism for collecting data from web resources", *Tourism Analysis*, Vol. 13 No. 2, pp. 309-315.

Ghose, A. and Ipeirotis, P.G. (2011), "Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics", *IEEE Transactions On Knowledge and Data Engineering*, Vol. 23 No. 10, pp. 1498-1512.

Ghose, A., Ipeirotis, P. and Li, B. (2009), "The economic impact of user-generated content on the internet: combining text mining with demand estimation in the hotel industry", *Proceedings of the Information Systems and Economics 20th Workshop*, *Phoenix, AZ*, pp. 14-15.

Gretzel, U. and Yoo, K.H. (2008), "Use and impact of online travel reviews", in O'Connor, P., Höpken, W. and Gretzel, U. (Eds), *Proceedings of the Information and Communication Technologies in Tourism 2008 International Conference in Innsbruck, Austria*, *Springer, New York, NY*, pp. 35-46.

Hamouda, A. and Rohaim, M. (2011), "Reviews classification using sentiwordnet lexicon", *The Online Journal on Computer Science and Information Technology*, Vol. 2, No. 1, pp. 120-123.

Han, J., Kamber, M. and Pei, J. (2011), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, Waltham, MA.

He, B., Macdonald, C. and Ounis, I. (2008), "Ranking opinionated blog posts using OpinionFinder", *Proceedings of the Research and Development in Information Retrieval, 31st Annual International ACM SIGIR Conference in Singapore*, pp. 727-728.

Hedrick-Wong, Y. and Choog, D. (2013), "Top 20 global destination cities in 2013, MasterCard WorldWide Insights", available at: http://insights.mastercard.com/position-papers/top-20-global-destination-cities-in-2013/ (accessed June 27, 2014).

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009), "Twitter power: tweets as electronic word of mouth", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 11, pp. 2169-2188.

Kaufer, S. (2014), "TripAdvisor fact sheet", available at: www.tripadvisor.com/-PressCenter-c4-Fact_Sheet.html (accessed July 2014)

Khoo, C.S.G., Nourbakhsh, A. and Na, J.C. (2012), "Sentiment analysis of online news text: a case study of appraisal theory", *Online Information Review*, Vol. 36 No. 6, pp. 858-878.

Litvin, S.W., Goldsmith, R.E. and Pan, B. (2008), "Electronic word-of-mouth in hospitality and tourism management", *Tourism Management*, Vol. 29 No. 3, pp. 458-468.

Liu, B. (2010), "Sentiment analysis and subjectivity", in Indurkhya, N., Damerau, F.J. (Eds), *Handbook of Natural Language Processing*, Vol. 2, Taylor and Francis Group, Boca Raton, FL, pp. 627-666.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. and McClosky, D. (2014), "The stanford CoreNLP natural language processing toolkit", *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mishne, G. (2005), "Experiments with mood classification in blog posts", *Proceedings of Stylistic Analysis of Text for Information Access ACM SIGIR 2005 Workshop in Salvador, Bahia*, *August 19, Vol. 19*, *Citeseer*.

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010), "From tweets to polls: linking text sentiment to public opinion time series", *Proceedings of the International AAAI Conference on Weblogs and Social Media*, *Vol. 11*, *Washington, DC*, pp. 122-129.

Ohana, B. and Tierney, B. (2009), "Sentiment classification of reviews using SentiWordNet", *Proceedings of the 9th IT & T Conference, Dublin Institute of Technology*, *Dublin*, p. 13.

Pang, B. and Lee, L. (2005), "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales", *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, pp. 115-124.

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2 Nos 1-2, pp. 1-135.

Pan, B., MacLaurin, T. and Crotts, J.C. (2007), "Travel blogs and the implications for destination marketing", *Journal of Travel Research*, Vol. 46 No. 1, pp. 35-45.

Saggion, H. and Funk, A. (2010), "Interpreting SentiWordNet for opinion classification", *Proceedings of the Language Resources and Evaluation International Conference Valetta, European Language Resources Association*, *Valetta*, pp. 1129-1133.

Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C. (2013), "Recursive deep models for semantic compositionality over a Sentiment Treebank", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1631-1642.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010), "Sentiment strength detection in short informal text", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 12, pp. 2544-2558.

Trusov, M., Bucklin, R.E. and Pauwels, K. (2009), "Effects of words-of-mouth versus traditional marketing: findings from an internet social networking site", *Journal of Marketing*, Vol. 73 No. 5, pp. 90-102.

Wiegand, M. and Klakow, D. (2010), "Bootstrapping supervised machine-learning polarity classifiers with rule-based classification", *Proceeding of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA*, *Lisbon*, pp. 59-66.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y. and Patwardhan, S. (2005a), "OpinionFinder: A system for subjectivity analysis", *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing on Interactive Demonstrations*, *Association for Computational Linguistics*, *Stroudsburg, PA*, pp. 34-35.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005b), "Recognizing contextual polarity in phrase-level sentiment analysis", *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing on Interactive Demonstrations*, *Association for Computational Linguistics*, *Stroudsburg, PA*, pp. 347-354.

Wu, X. and Kumar, V. (2008), "Top 10 algorithms in data mining", *Knowledge and Information Systems*, Vol. 14 No. 1, pp. 1-37.

Ye, Q., Zhang, Z. and Law, R. (2009), "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 6527-6535.

Zhu, F. and Zhang, X. (2006), "The influence of online consumer reviews on the demand for experience goods: the case of video games", *Proceedings of 27th International Conference on Information Systems*, *Milwaukee, WI*, pp. 367-382.

About the authors

Rutilio Rodolfo López Barbosa is a Full-Time Professor and Researcher for the University of Colima, México. As part of the faculty of Accounting and Businesses, he has participated on several research projects related with Information Systems and Education. He is currently PhD Candidate for the Engineering of Information and Knowledge program at the University of Alcala, Spain. His areas of interest are software development, information retrieval, data mining, machine learning and internet technology. Professor Rutilio Rodolfo López Barbosa is the corresponding author and can be contacted at: rutiliol@hotmail.com

Dr Salvador Sánchez-Alonso is an Associate Professor and a Senior Member of the Information Engineering group, a research unit dependent of the Computer Science Department of the University of Alcala, Spain. He earned a PhD in Computer Science at the Polytechnic University of Madrid in 2005 with a research on learning object metadata design for better machine "understandability," and finished a degree on Library Science on 2011. He has participated or coordinated in several EU-funded projects in the last five years on the topics of learning object repositories and metadata, remarkably LUISA, Organic.Edunet, VOA3R, Organic. Lingua and agINFRA just to name a few. Author of more than 20 high impact factor publications in the last ten years, his current research interests include technology enhanced learning, learning object repositories and Semantic Web.

Dr Miguel Angel Sicilia-Urban is currently Full Professor at the Computer Science Department of the University of Alcala and the head of the Information Engineering research unit, where he leads several European and national research projects in the topics of learning technology and Semantic Web. He is the Editor-in-Chief of the *International Journal of Metadata, Semantics and Ontologies*, published by Inderscience, and serves as Member of Editorial Board of many other scientific journals in the area of Semantic Web, computational intelligence and information systems. He has been active in metadata research and was awarded the 2006 Cyc prize for the best research paper.