Hotel Online Reviews: Creating a Multi-Source Aggregated Index

Structured Abstract

Purpose

To develop a model to predict online review ratings from multiple sources, which can be used to detect fraudulent reviews, create proprietary rating indexes, or which can be employed as a measure of selection in recommender systems.

Methodology

This study applies machine learning and natural language processing approaches to combine features derived from the qualitative component of a review with the corresponding quantitative component and, therefore, generate a richer review rating.

Findings

Experiments were performed over a collection of hotel online reviews—written in English, Spanish, and Portuguese—and they show a significant improvement over the previously reported results and demonstrate not only the scientific value of the approach but they also strengthen the value of review prediction applications in the business environment.

Originality/value

This study shows the importance of building predictive models for revenue management and the application of the index generated by the model. It also demonstrates that, although difficult and challenging, it is possible to achieve valuable results in the application of text analysis across multiple languages.

Keywords: natural language processing, machine learning, multi-language, online reviews

Classification: research paper

Introduction

Social reputation is now one of the main aspects that influence a customer's booking decision process (Anderson, 2012; Cantallops and Salvi, 2014; Kwok et al., 2017; Viglia et al., 2016; Zhao et al., 2015). In particular, Internet ubiquity and the ease of use have stimulated the fast growth of user-generated content on social media platforms, especially of service reviews (Duan et al., 2016; Zhao et al., 2015). Online reviews act as a form of electronic word of mouth, which is a technological variation of the traditional word of mouth. The influence of online reviews can affect up to 50% of all hotel booking decisions (Duan et al., 2016).

Social reputation is so important in the hospitality industry that there are now several companies that provide specialized services for collecting, analyzing, and managing hotel online reviews. One service that these companies provide is the collection of reviews from different sources (e.g., Tripadvisor.com, Booking.com, and HolidayCheck, among others) to produce an index that summarizes their ratings. Using the index from one of these companies (ReviewPro), Anderson (2012) demonstrated that a 1% increase in the index rating can lead to a 0.54% increase in hotel occupancy and a 1.42% increase in hotel revenue per available room. Meanwhile, authors such as Kim et al. (2015) and Torres et al. (2015) have revealed that customers are willing to pay more for a room as the review ratings increase. Viglia et al. (2016) demonstrated that a one-point increase in the review score is associated with a 7.5% increase in the occupancy rate.

Although there is an extensive body of knowledge about the impact of online reviews on the hospitality industry (Anderson, 2012; Cantallops and Salvi, 2014; Duan et al., 2016; European Commission, 2014; Kwok et al., 2017; Zhao et al., 2015), most only take advantage of the quantitative components of the reviews (e.g., overall ratings, topics ratings, hotel stars etc.) to measure that impact (Duan et al., 2016; Han et al., 2016; Kwok et al., 2017; Liu et al., 2017). Consequently, most of the previous research employs quantitative methods. In fact, a recent literature review on the subject of online review research, which was performed on seven major hospitality and tourism journals by Kwok et al. (2017), revealed that, from a total of 67 articles published between January 2000 and July 2015, 70% employed quantitative methods, 24% employed qualitative methods, and only 4% employed mixed methods. Yet, as recognized by Noone and McGuire (2014), customers seem to favor the information-rich, textual components of reviews over the quantitative ratings when making value judgments. From the hotel's perspective, the qualitative (textual) component of reviews "can potentially yield insights not indicated in the ratings for how hotels can improve their operations and better meet customer expectations" (Han et al., 2016, p. 4). The evolution of text mining and natural language processing algorithms has facilitated the extraction of information from the textual component of online reviews (McGuire, 2017). Recently, the number of published studies taking advantage of the textual component of reviews has increased, focusing on issues such as identifying relevant topics mentioned in reviews (Calheiros et al., 2017), understanding what satisfied and unsatisfied customers mention (Berezina et al., 2016; Xu et al., 2017), assessing the impact of social media on a hotel's service (Duan et al., 2016), understanding what guests think of hotels (Han et al., 2016; He et al., 2017; Xiang et al., 2015; Xu and Li, 2016), examining the consumers' pre-purchase decisions (Noone and McGuire, 2014), identifying deceptive review comments (Lin et al., 2017), and review opinion classification predictions (Bjørkelund et al., 2012; Salehan and Kim, 2016). Although some of these works resort to sentiment analysis and machine learning, to the extent of the authors knowledge, only three of them use these tools to predict review ratings (i.e., Ganu et al., 2013; Lei and Qian, 2015; López Barbosa et al., 2015).

The prediction of an online review rating based on the quantitative and qualitative components of reviews yields the power to identify discrepancies between these components and it contributes to the detection of fraudulent reviews. It also can be used in the creation of a proprietary rating system that is based on the combination of reviews from different sources, which can then be applied in a competitive set analysis or to rate hotels without a star classification system. It can also be used to build recommender systems to assist users in the selection of a hotel based on review ratings and textual descriptions.

To fill a gap in the research on online review rating predictions, this study applies natural language processing and machine learning to the online reviews of 56 Portuguese hotels (29 city hotels and 27 resort hotels) published on the Booking.com and Tripadvisor.com websites to obtain a prediction model for review ratings. Given that reviews can be written in many different languages, and executing text analysis in multiple languages is a notoriously complex and difficult process, we have restricted our analysis to reviews written in three languages: English, Spanish, and Portuguese. The rationale behind this is that a case study of Portuguese hotels and these languages represent more than 70% of the languages used by the guests of Portuguese hotels (Instituto Nacional de Estatística, 2016). To create this model, this study uses not only existing features (in traditional statistics, these features are known as independent variables) such as hotel type, hotel stars, and review source but also additional features, some of which have been derived from the application of sentiment analysis of the textual component of these reviews.

Due to the cyclical nature of prediction modeling projects, the structure of this study is slightly different from a conventional study. A brief review of the related works is followed by a detailed description of the data and methodology applied in the model. However, because the results need to be assessed in each cycle to determine the next step, the results are presented interleaved with the methodology. These results are then discussed in the conclusion, together with the implications, limitations, and directions for future research.

Related Work

In the context of the criteria defined by Surowiecki (2005), online reviewers could be called a "crowd" (group) because they have a diversity of opinion, independence, decentralization, and aggregation. Surowiecki (2005) adds that crowds represent a diverse collection of independent individuals who are better at making certain types of decisions or predictions than its members, or even experts. In particular, customers give more credit to a hotel's online ratings than to their official classification, or stars (Öğüt & Taş 2012). In fact, online review ratings are currently a more significant predictor of hotel performance than traditional customer satisfaction surveys (Woo Gon Kim and Seo Ah Park, 2017).

Technology-based methods are required if we wish to obtain the full potential of reviews, which mostly involve text mining, data mining and big data approaches. As acknowledged by Kwok et al. (2017): "Although much is known about online reviews, the advancement of technology is constantly challenging our current understanding and asking for new insights." These authors recognized the potential of technology, mainly, of big data and data analytics, to better comprehend customers. Magnini et al. (2003) aimed to use technology to extract meaningful patterns and build customer-behavior models to help in decision making, and they proposed five categories of tasks where data mining could be applied in the hospitality industry. One of these categories was the prediction of future value of continuous variables. The authors argued that "with forecasting one can also use data trends to project which hotel amenities are of growing importance to consumers and will be key drivers of the consumer's future perception of value" (Magnini et al., 2003, p. 98). Due to the recent tremendous growth of online reviews, this has never been truer than today.

Online reviews not only influence customer purchase decisions (Cantallops and Salvi, 2014; Kim et al., 2015; Kwok et al., 2017; Zhao et al., 2015) but they also allow hoteliers to exert greater pricing power (Anderson, 2012; Kim et al., 2015). Furthermore, they allow the proprietors to make the management teams accountable for the hotel's reputation (Torres et al., 2015). However, three obstacles must be overcome if we wish to convert online review data into meaningful and actionable information, as follows: the large volume of data, the unstructured nature of the textual component of reviews, and the dynamics of changing information (i.e., new reviews are created every day). The first obstacle incapacitates the manual processing of data, while the second and third obstacles do not guarantee unbiased interpretations (Han et al., 2016; Kwok et al., 2017). Therefore, applying machine learning and natural language processing seems to be the solution to convert the unstructured textual component of reviews into a structured form that could be used to create features to predict review ratings. Sentiment analysis is one of the techniques that is frequently employed to create features based on text. Sentiment analysis, or opinion mining, is the computational study of people's opinions toward entities, individuals, events, topics, and their attributes. Sentiment analysis quantifies opinions according to their valence—that is, positive, negative, or neutral polarity (Liu and Zhang, 2012). It also grants the extraction of features related to the identification of topics, keywords, and concerns (Schuckert et al., 2015a).

The existing literature on the employment of sentiment analysis in the online reviews of products and services is vast (Wang et al., 2016). However, the specific focus of the present study is on the works that have examined the specificities of hotel reviews. In a literature survey related to online reviews in the tourism and hospitality industries, Schuckert et al. (2015a) revealed that in 50 relevant articles published from 2003 to 2014, eight (16%) employed sentiment analysis. However, these and other recent studies that use sentiment analysis to create prediction features employ it to predict the polarity of sentiment and not the rating itself (Berezina et al., 2016; Bjørkelund et al., 2012; Calheiros et al., 2017; Duan et al., 2016; Han et al., 2016; He et al., 2017; Hu and Chen, 2016; Marcheggiani et al., 2014; Markopoulos et al., 2015; Zheng and Ye, 2009). Consequently, most of these studies consider the problem as classification rather than regression, which corresponds to considering the prediction outcome as a class/category/discrete value instead of a continuous value. In other words, these studies manually attribute a polarity to reviews (positive, neutral, or negative) and they then apply machine learning to the textual component of the reviews to predict that polarity. Only Ganu et al. (2013), Lei and Qian (2015), and López Barbosa et al. (2015) use sentiment analysis to predict hospitality reviews' ratings; that is, use the review rating as the expected outcome. However, this expected outcome differs among these studies. As in this study, Ganu et al. (2013) and Lei and Qian (2015) aim to predict review ratings. The first aims to predict restaurant review ratings. The second aims to predict hotel and travel review ratings. In contrast, López Barbosa et al. (2015) aim to predict a hotel's overall ratings and not the review ratings. In common with the present study, all these studies recognize the predictive power of sentiment analysis.

Machine learning is commonly defined as the automatic detection of meaningful patterns in data so that we can make and improve predictions based on that data. Together with the vast amount of data at our disposal, the availability of better and cheaper computational power has recently contributed to the development of new and improved machine learning algorithms and methods. These algorithms and methods allow the exploration of structured and unstructured data, as is the case with online reviews, in ways that were not previously possible. In their literature review, Schuckert et al. (2015a) describe that one of the problems of applying sentiment analysis in hospitality-related studies is that hospitality is a global industry and hotel guests coming from all over the world, which makes it “difficult for opinion mining programs to handle different languages” (Schuckert et al., 2015a, p. 613). Consequently, most of these studies only use English or Chinese reviews. One exception is the work of Markopoulos et al. (2015), who studied reviews written in Greek. Nevertheless, the textual content of online reviews has vast potential in terms of research. As acknowledged by Han et al. (2016, p. 17), “text analysis across multiple languages presents methodological difficulties. However, when those issues are overcome, online reviews will potentially yield insights about cultural effects that can further aid hotel managers in improving their customer experiences.”

Modeling

Hotels are categorized as business, extended-stay, resort, or as a mix of all three, they can also be categorized by size and location (Talluri and Van Ryzin, 2005). However, most studies of online reviews only target city hotels. Therefore, to gain a more general perspective, reviews from two different types of hotels—city and resort—were collected in this study. In total, this study collected review data from 56 hotels in Portugal using two different sources: Booking.com and Tripadvisor.com, which are two of the largest travel websites with hotel reviews. Because of the recognized difficulty in carrying out sentiment analysis in several languages, only reviews written in the top three languages—English, Spanish, and Portuguese—were collected.

We have used Chapman et al.'s (2000) Cross-Industry Standard Process Model for Data Mining (CRISP-DM) to build the prediction models, from data collection and feature selection

Figure 1. CRISP-DM phases reference model (adapted from Chapman et al. (2000))

for dataset creation to model development and evaluation. Due to its practicality and simplicity, CRISP-DM is used in many different fields, including tourism and hospitality (Antonio et al., 2016). As shown in *Figure 1*, CRISP-DM defines six steps that are necessary to build a prediction model, as follows: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This section will describe these steps in further detail, except for the deployment of the models, which is outside of the scope of this study.

Business and Data Understanding

As in any other CRISP-DM predictive model project, predicting hotel online review ratings requires the definition of success criteria and a method to measure them. Therefore, the present study has adopted two standard evaluation measures, mean absolute error (MAE) and root mean square error (RMSE), which are frequently employed in the evaluation of prediction models. MAE is the average of the absolute value of the difference between predicted and actual values, and it can be calculated by the formula:

$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$, where $\hat{y}_i$ is the predicted value and y_i the actual value. RMSE is the square root of the average of the square of the difference between predicted and actual values. RMSE is calculated by the formula $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$. Both MAE and RMSE express average model prediction error in positive units of rating and they are indifferent to the direction of error (either made by fault or excess). While lower values are better, RMSE amplifies and punishes large errors. Which of these metrics is the best to evaluate prediction models has long been a question of debate (Chai and Draxler, 2014). Therefore, our results are shown using both measures, although our discussion is solely based on MAE because of its easier interpretation. In fact, a difference of 12 is twice as bad as a difference of 6. Therefore, it is not the variance of the frequency distribution of error magnitudes but the errors themselves that are important to measure the goodness of fit. As defined by CRISP-DM, our success criteria were determined so that the average error in Booking.com reviews does not exceed 1.5 (because of the normalization process described in the next subsection) and does not exceed 0.8 in Tripadvisor.com review ratings scales; that is, obtaining a MAE value not above 20% of the rating scale amplitude value.

This research focuses on eight hotels, of which four are city hotels that are located in Lisbon, the capital of Portugal, and the remaining four are resort hotels from the Algarve seaside region. To decide on a competitive hotel set to achieve meaningful results, the management team of each of the hotels was asked to identify the top five hotel competitors list (some of which are competitors of more than one of the chosen hotels). The full hotel set distribution, by star classification and hotel type, is shown in *Table 1*.

Table 1. Hotel set summary

Because of the differences between the Booking.com and Tripadvisor.com websites, two customized web content extractors were built (in Microsoft C#) to extract the reviews. One of the differences between these websites, as documented by Hale (2016), is that while Tripadvisor.com has specific URLs for some languages, and has a source-language parameter and machine translation link that allows the identification of the review's original language, Booking.com does not. Consequently, the Booking.com web extractor had to simulate human behavior so that it could avoid machine-translated reviews and reviews not written in the chosen languages. Therefore, after entering in the hotel booking website page, the extractor used a fully automated procedure to click on the "Guests' experiences" link, and then clicked on the "Show me reviews in" the checkbox, and clicked on "English." It then selected the "Sort by" dropdown list and clicked "Date (newer to older)." After reading all of the reviews on page, the web extractor repeated this operation for Spanish and Portuguese languages, and continued for all of the hotels.

The extractors collected reviews between January 2016 and June 2016. During this period, all reviews in English, Spanish, and Portuguese with a publication date between July 2015 and June 2016 were considered. From the total of 23,353 reviews that were initially collected, 23,322 remained after cleaning of duplicates and deletion of reviews that were incorrectly classified in terms of language.[1] A summary of review frequency and distribution by hotel classification, hotel type, and language is presented in *Table 2*.

Table 2. Review frequency and distribution summary

Although reviews on every website have a similar structure, there are some important differences that should be taken into account when combining reviews from multiple sources, as acknowledged by Bjørkelund et al. (2012). For example, both Booking.com and Tripadvisor.com provide an overall rating and a textual component for each review. Booking.com's rating uses a continuous range from 1 to 10, while Tripadvisor.com's uses a discrete range from 1 to 5. There is also a major difference in the textual component: while Booking.com has two text fields, one for positive and one for negative comments, Tripadvisor.com has a single text field. Another important difference is how these websites present ratings: although both sources allow users to assign ratings by topics (cleanliness, location, comfort, etc.), Booking.com presents aggregated results per hotel, while Tripadvisor.com presents results by review.

Metadata and segmentation information from the reviews, such as variables regarding age group, travel reason, or country of the reviewer, could be important in online review research.

However, it is not mandatory for reviewers to fully identify themselves on most social media platforms, allowing them to maintain anonymity (European Commission, 2014). Therefore, even though this metadata and segmentation information could be captured in some of the reviews, because it was not available in all of the reviews, it was not considered to be of good enough quality for this research and consequently the respective data was discarded.

Data Preparation, Model Building, and Evaluation

The development of a prediction model involves selecting, merging, cleaning, encoding, and constructing features from the original dataset. This process is known as "feature engineering" and it may have a significant impact on the model's performance (Kuhn and Johnson, 2013). The resulting features originate a modeling dataset that is used to build and evaluate the prediction model. Developing and finding the best prediction model is an iterative process that often requires going back in the pipeline. For example, the most suitable feature set is often achieved by performing partial experiments, sometimes requiring us to going back to the business understanding stage to derive new features. Because this is an iterative process, the rest of this section will detail all of the steps involved.

Prediction modeling requires a two-dimensional dataset that is comprised of rows and columns. Each row represents the unit of analysis while the columns represent attributes, descriptors, or variables (Abbott, 2014). We used the R tool to create this dataset. This process started with the merging of the review data from Booking.com and Tripadvisor.com. Only those features present in both sources were included in this dataset. In addition, two new features were added: review description and normalized rating. In the case of the reviews from Tripadvisor.com, the review description is just the transposition of the text component of the review, while in the reviews from Booking.com this variable is the result of the concatenation of both the positive and the negative text fields. The normalized rating feature was necessary because of the differences of scales between Booking.com and TripAdvisor. This feature was created by applying min-max normalization, which is one of the most common normalization methods to scale variables (Abbott, 2014). The ratings were normalized to a value ranging from 1 to 100 using the following formula: $x' = \frac{(x - \min(x))}{(\max(x) - \min(x)} \times 100$. This scale is typically used in indexes that aggregate ratings from sources, such as the one used by Anderson (2012). Because of the Booking.com rating scale distortion (Mellinas et al., 2016), the minimum rating of Booking.com reviews was considered to be 2.5. *NormalizedRating* was considered to be the outcome (which in traditional statistics is known as the dependent variable). Besides these two new features, the dataset is composed of hotel common ID, hotel type, hotel stars, source, and language.

Figure 2. Initial dataset visualization

Figure 2 depicts this dataset, revealing a good distribution in terms of review ratings per normalized ratings, hotel types, hotels, hotel stars, and languages. The particularities shown in the figure are mostly related to the different rating scales used by Booking.com and Tripadvisor.com. In some levels of ratings, this can lead to a distortion in the proportion of reviews per hotel type, hotel stars, source, and language.

Before applying sentiment analysis to the review descriptions to extract the additional features, the most critical step in the transformation of text from an unstructured to a structured form is text preprocessing (Han et al., 2016). This process allows the retention of relevant information and the removal of irrelevant information. This was accomplished through the application of the following preprocessing steps, some of which are language dependent:

1. Transform all of the review descriptions into lowercase.
2. Normalize the common words that appear in different forms in the three languages; for example, “wi-fi,” “wi fi,” and “wifi” were all considered as the same token.
3. For each language, perform a stemming of common hospitality words such as “rooms”, “restaurants”, “bars”, and others that could be meaningful for data interpretation.
4. For each language, normalize the terms used to write some words or expressions that could be written differently or misspelled; for example, in English, consider “didn’t”, “didnt”, and “did not” as the same token.
5. For each language, normalize the terms used to write important aspects related to the study’s domain of hospitality; for example, in English, “staff” is a common word used to designate hotel staff, but in Portuguese, many words are used: “equipa” (team), “pessoal” (personnel), “funcionários” (employees), or “colaboradores” (collaborators). Other examples related to the origin of guests must also be taken into consideration. Portuguese from Brazil has some differences from European Portuguese, and because Brazil is an important market in Portugal, terms from Brazilian Portuguese like “café da manhã,” “ônibus”, or “metrô” were transformed to their European Portuguese equivalents, respectively, “pequeno-almoço”, “autocarro”, and “metro” (in English, “breakfast”, “bus”, and “metro/underground/subway”).
6. Remove the punctuation, numbers, and stop words.

After preprocessing the text, the data was transformed into a bag-of-words representation, which is one of the most popular methods used in text mining. This method resulted in the creation of two document-term matrices, one by document (review) and the other by sentences (sentences in reviews). Each matrix consists of rows, each of which represents a document, and columns, each of which represents a word frequency (i.e., all words presented in a document). These matrices produced the information to create two new dataset features, namely: number of words per review and number of sentences per review.

Sentiment analysis was then applied to the pre-processed text component of the reviews to obtain additional dataset features. Sentiment strength polarity was taken under two diverse perspectives: per document (full text of review description), in a similar fashion to that reported by Han et al. (2016) and Bjørkelund et al. (2012); and also per sentence, following the work of Duan et al. (2016). While the first allows an understanding of the global opinion of the review, the second relates the opinion to particular aspects.

There are several approaches to the application of sentiment analysis. In this study, we chose to employ an approach based on polarity lexica. This approach relies on dictionaries of opinion words with a polarity classification (Ravi and Ravi, 2015). Therefore, dictionary selection is an important methodological consideration (Han et al., 2016). One relevant aspect for dictionary selection is its adequacy to the domain of the text. Usually, domain-oriented dictionaries produce better results. However, because no dictionaries related to the hospitality industry were found in any of the languages that we studied, dictionaries were chosen based on their structure (i.e., dictionaries had to have the same structure or be easily transformed to a similar structure), completeness (i.e., dictionaries had to have an extensive range of words), and openness (i.e., dictionaries should not be specific to any type of domain). Based on this criteria, the SentiLex-PT 02 sentiment lexicon (Silva et al., 2012) was selected for Portuguese, the ElhPolar dictionary (Saralegi and San Vincente, 2013) was selected for Spanish, and the well-known Opinion Lexicon by Hu and Liu (2004) was selected for English.

The sentiment strength of each sentence was calculated by counting the positive and negative words of each sentence and then applying the formula used by Bjørkelund et al. (2012): $sentiment\ strength = \frac{\sum positive\ words}{\sum positive\ words + \sum negative\ words}$. This formula returns a value between 0 and 1, where 0 is perfectly negative and 1 is perfectly positive. Each review global sentiment strength was then calculated as the average sentiment strength of all of the review sentences. Reviews with no text were considered to be neutral (a sentiment strength of 0.5).

The Microsoft Azure Machine Learning platform was used in the application of the machine learning methods. The following five regression methods were used to later select the one that performed better: Bayesian linear regression, boosted decision tree, decision forest, linear regression, and neural networks. This evaluation was carried out using cross-validation, specifically *k*-fold cross-validation, which is a well-known and widely used model evaluation technique. Although cross-validation can be computationally costly (Smola and Vishwanathan, 2008), it allows for the development of models that are not over fitted and it can be generalized to independent datasets. *k*-fold cross-validation works by randomly partitioning sample data into *k*-sized subsamples. In this study, the dataset was divided into 10 folds—a typical number of chosen folds (Smola and Vishwanathan, 2008). Then, each of the 10 folds was used as a test set and the remaining nine were used as as training data. As is a common practice in cross-validation, MAE and RMSE were calculated for each of the 10 folds, and the global mean (**μ**) and standard deviation (**σ**) were used to assess each method's performance. The results can be seen in *Table 3*.

Table 3. Ten-fold cross-validation results obtained with base features and global sentiment analysis

Bayesian linear regression, decision forest, and neural networks presented promising results, with an MAE (i.e., a mean difference between the predicted ratings and the actual ratings) from 14.4 to 15.1, which are already above the initial objective. Likewise, the low standard deviation values also show that predicted values are closer to the mean, thus revealing a good model performance.

At this point, the features presented in the dataset used to build the model were: *NormalizedRating* (numeric), *HotelType* (categorical), *HotelCommonID* (categorical), *HotelStars* (numeric), *Source* (categorical), *Language* (categorical), *NumberOfWordsPerReview* (numeric), *NumberOfSentencesPerReview* (numeric) and *SentimentStrength* (numeric). The results so far were auspicious but nevertheless another test was performed to determine whether additional features could improve model performance. These features come from two common techniques that are used in prediction modeling research to create features, which are: Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) (Liu and Zhang, 2012). TF is a numerical statistic that represents how frequently each term is used in a document. TF-IDF is also a numerical statistic but it represents the composite weight of each term in a document. The terms could be a single word (also known as a unigram, or *n*-gram) or they could be a contiguous sequence of *n* words in a text. This study presents experiments conducted with unigrams, bigrams, and trigrams (i.e., one, two and three words, respectively) and employed both techniques. TF-IDF with unigrams achieved the best results, which were obtained with a TF-IDF matrix built with words with a minimum of three characters and a maximum of 25. Since each of the three languages has its own vocabulary, the datasets were subdivided by language. Again, cross-validation was employed to evaluate the results (see *Table 4*).

Table 4. Ten-fold cross-validation with TF-IDF unigram features

It is possible to globally verify in Table 4 that Bayesian linear regression presents the best results, both in terms of MAE and RMSE average. Based on these results, optimized models were then built for each language. Each language dataset was split into stratified training and test sets, with 70% and 30% of data, respectively. Bayesian linear regression (BLR) was applied to build the optimized models. The reason behind this choice was that in terms of standard deviation, with the exception of the RMSE of the decision forest for the Portuguese reviews, the BLR method produced the best results. The optimized models using BLR presented a MAE of 13.3 and a RMSE of 17.2 for Portuguese, a MAE of 13.4 and a RMSE of 17.1 for Spanish, and a MAE of 12.5 and a RMSE of 16.3 for English reviews.

Discussion And Conclusions

Conclusions

Although there is substantial literature concerning the study of online reviews, and much of it focuses on the impact that online reviews have on the hospitality industry, researchers have only recently started to take advantage of the power of the automated analysis of the textual component of online reviews. In fact, it is possible to perceive some differences between manually and automated sentiment analysis of reviews. Although Jiang et al. (2010) performed a manual classification of sentiment and found that there was a disconnection between the textual component of the review and the review rating, this present work has demonstrated that the text sentiment strength of a review is associated to the corresponding review rating, similar results were found by He et al. (2017), Han et al. (2016) and Duan et al. (2016), who also employed automated sentiment analysis. As Lei and Qian (2015) revealed, this association, together with service reputation, outperformed traditional recommender systems in the prediction of online review ratings, which is in line with the findings reported in this study. By achieving MAEs of 12.5, 13.4, and 13.4 in English, Spanish, and Portuguese, respectively, this study outperforms the initial objective of having an MAE below 20, which means that the average error between the predicted rating and the actual rating diverges in between 12.5 to 13.4 in a scale from 1 to 100, according to the language.

Another important outcome is the demonstration that rating predictions can be executed in multiple languages. Although most sentiment analysis research is conducted in English and Chinese, and it is difficult to execute text analysis across multiple languages (Han et al., 2016; Schuckert et al., 2015a), this study has shown that good prediction results can be achieved in Portuguese, Spanish, and English. These good results in the prediction of review ratings open the door to many different applications for these models. An obvious application is to predict how a customer would rate quantitatively a hotel stay based on his or her text comments. For example, this could be applied to the standard checkout surveys that hotels ask their customers to complete. However, rating prediction has the potential to be used as a measure and a proxy on its own. Given that the rating prediction result is a measure created from several online review features, including features related to the textual component, one can extrapolate that it offers a more holistic score of the review than the usual "review rating." The prediction result not only has the potential to be used in scientific research as a measure to substitute or complement review ratings but it can also be used in terms of the following business applications:

- *Fraud detection or discrepancy assessment*. As Noone and McGuire (2014, p. 574) recognized "the validity of consumer reviews constitute a legitimate concern for any organization." Substantial differences between the predicted results and the actual review ratings can be used to identify reviews where the actual ratings are outside the rating patterns of similar reviews. These differences could be created by discrepancies between the quantitative and qualitative components of reviews, which could indicate that reviews are fraudulent or focus on topics and aspects that similar reviews do not. For example, instead of requiring the hotel staff to go through each review and pinpoint possible fraudulent reviews, a hotel could run this model periodically to automatically select which reviews could be fraudulent based on the difference between the predicted review rating and the actual rating. Only reviews above or below a pre-defined threshold would be verified by the hotel staff to validate the reason for the discrepancy and understand if it might have been posted fraudulently.

For example, a review on TripAdvisor for one of the 4-star resort hotels presented a discrepancy of -43.4. The user assigned a *NormalizedRating* of 0 (which corresponds to a 1 on the TripAdvisor scale), but the model predicted the value as 43.4. The user wrote (pre-processed text): "somebody needs buy place renovate whole place raze room hideous bathroom dangerous thought going slip bathtub fall elevators dont work times wifi antiquated thing good location view dont bother". Although the reviewer had made a positive statement about the location, he/she is very negative in general. Taking into account that the hotel has a *NormalizedRating* mean of 81.88, then it has to be asked if the problems reported by this user justify this extremely low rating?

Another example can be seen in a review from TripAdvisor for one of the 4-star resort hotels that had a global *NormalizedRating* mean of 63.76 and presented a discrepancy of -39.9. The user gave an overall rating of 1 on the TripAdvisor scale (0 in

NormalizedRating) and wrote the following: "booked hotel july beach thought full confidence following lovely pics indoor pool jacuzzi since removed payed similar hotels booked especially young boy wanted spa facilities please warned still advertising spa hotel yet facilities even though emailed accepting liability nothing can stop dreadful don't go use money better facilities hotels" (pre-processed text). In this case, it is possible to understand that the user was so upset by not being able to use the spa facilities that he/she gave the hotel a very bad rating. Another user reading this review would probably comprehend the reviewer's frustration and not give relevance to the rating. As for the hotel staff, they would probably understand the rating and not consider this to be a fraudulent review.

- *Creation of an aggregated social reputation index*. Due to the multitude of social reputation websites that exist, each with their own rating scales, hotels have a hard time comprehending their global social reputation position at any given moment. Although there are now companies who create aggregated indexes and offer them as a service to hotels (e.g., ReviewPro or TrustYou), these indexes only reflect the quantitative component of the reviews. They do not reflect the aspects related to the textual component or the language of the review and, as expected, customers from different languages rate hotels differently (Hale, 2016; Schuckert et al., 2015b). Therefore, the predicted rating value could be used as a measure to create an aggregate index of social reputation that reflects both the quantitative and the qualitative components of the reviews, including the emotive power of the language in which it was written, instead of using only the more usual quantitative component. This capacity to reflect both the quantitative and qualitative components of reviews can be seen in the examples that we have provided in the previous point, where sometimes the users do not justify their low ratings in the text.

- *Implementation of recommender systems*. In their booking engines, hotel chains, online travel agencies, or meta searchers' websites usually allow the customers to filter their search by the hotel's characteristics, price, location, and social reputation ratings. However, this social reputation rating is usually the average rating of a particular website. If instead of using this social reputation rating, the website used an aggregated rating created with our proposed method, then recommendations of reviews to be read or services/products to buy would also take into account the qualitative component. For example, a hotel chain website could recommend hotels to a user based on the aggregated index of their own hotels, instead of using the rating of a particular website or other criteria. This could be a more transparent and reliable method for users who wish to select hotels by social reputation.

Most high-performance machine learning models are "essentially a black box with a highly complex prediction equation" (Kuhn and Johnson, 2013, p. 221). The models generated in this study are no exception and they cannot be depicted easily. However, by following the steps described in this paper, these models can be replicated and generalized for any type of hotel and in almost any language.

Theoretical Implications

Given the importance of online reviews to the hospitality industry, this study contributes to the research on the impact of social reputation in this industry by confirming Kwok et al.'s (2017) views of the benefits of the application of innovative methods in online reviews research. In this case, the combination of machine learning and natural language processing techniques enables the accurate prediction of online review ratings. First, this work has shown that the quantitative features of the reviews can be complemented with sentiment analysis features, extracted from reviews' textual component, which leads to improved review rating prediction models. Second, this work has shown that additional features based on TF-IDF also contribute to improved prediction models. Third, this study was able to show that online rating predictions are possible today across multiple languages thanks to the existing tools that are available for each language, which is able to overcome the established idea that sentiment analysis is difficult to apply in reviews across multiple languages (Han et al., 2016; Schuckert et al., 2015a).

This study highlights the value of the review rating prediction and it details how it can be employed to build an aggregated index that better reflect reviews from multiple sources. Consequently, in addition to the multiple applications in business environments, this index should also be considered by researchers of online reviews in substitution of, or in complement to the review rating.

Practical Implications

This study shows that is possible to create features based on a review's textual components and that these features are predictors of review ratings. As expected, this study shows an association between a review's quantitative and qualitative components. This means that these features, or even the error in rating predictions, can be used as a measure of discrepancy between a review rating and its textual description. Therefore, this measure can be used to detect and provide an alert for fraudulent reviews or reviews where the components do not match. The hoteliers could then further analyze these reviews and where they consider the review to be fraudulent, they can report it to the source website. Most review websites have procedures to investigate fraudulent reviews and they can remove them if the fraud is proven. By demonstrating that it is possible to predict with good accuracy a review rating, this study has shown that the prediction outcome can also be used as an alternative rating thanks to its association to the review rating. Being able to make use of not only the quantitative component of reviews but also the qualitative component offers the potential to better identify the review's true value. This is in accordance to Noone and McGuire's (2014) finding of customers weighting their decisions based on both components of the review. This alternative rating could be used by hoteliers to pinpoint important reviews, and to analyze the weaknesses and strengths mentioned in those reviews. Because of its completeness, this alternative rating could also be employed with other features in the creation of indexes that aggregate reviews from different sources and as a feature in recommender systems. In addition, this feature could be included in a machine learning model in a website/mobile application booking engine for a hotel chain or in an online travel agency to recommend hotels for the customer.

Limitations and Recommendations for Future Research

Although models specifically built for each source may present better results, this study's aim to create a unique model seems to be on the correct track to build a rating that aggregates reviews from different sources. Our study used data from two different review sources, comprising several different rating scales. However, we have shown that scale could influence the distribution of reviews per hotel type, hotel stars, or review language. Consequently, it is recommended that future research should use additional sources to address this situation.

This study uses MAE and RMSE, which are the two most common metrics for measuring accuracy in regression problems. However, these metrics do not reflect the models' reliability over time. Therefore, it is recommended that future research should assess this reliability with the application of other measures or methods, such as the "walk-forward" approach (Stein, 2007).

Although this study does not perform a semantic characterization of the terms, techniques such as word-sense disambiguation allow for a better comprehension of a word according to the sentence and the context where it is being used. Applying these techniques can lead to improved scores and they are certainly worth considering as a possible next step.

This study's multi-language approach gave raise to limitations that are not common in other online review studies, as follows:

1. Most of the reviews used in this study were written in English, and reviews in Portuguese and Spanish were clearly outnumbered. This unbalanced data may have had an impact on the final results obtained for Portuguese and Spanish reviews.
2. Performing text analysis across multiple languages is a difficult and time-consuming task. Consequently, our study only considered reviews written in the three most frequently used languages: English, Spanish, and Portuguese. Therefore, future research should extend the text analysis to other languages.
3. Because every language has a different degree of expressive power (Ravi and Ravi, 2015) and there were no domain-specific dictionaries with a common degree of polarity, it is possible that differences among the results could be attributed to the dictionaries that we used. Hence, it is recommended that future research should experiment with approaches

free from dictionaries or use domain-specific dictionaries (if available), with polarities standardized across languages.

References

Abbott, D. (2014), *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*, Wiley, Indianapolis, IN, USA.

Anderson, C.K. (2012), "The impact of social media on lodging performance", *Cornell Hospitality Report*, Vol. 12 No. 15, pp. 4–11.

Antonio, N., Almeida, A. de and Nunes, L. (2016), "Using data science to predict hotel booking cancellations", in Vasant, P. and M, K. (Eds.), *Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry*, Business Science Reference, Hershey, PA, USA, pp. 141–167.

Berezina, K., Bilgihan, A., Cobanoglu, C. and Okumus, F. (2016), "Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews", *Journal of Hospitality Marketing & Management*, Vol. 25 No. 1, pp. 1–24.

Bjørkelund, E., Burnett, T.H. and Nørvag, K. (2012), "A study of opinion mining and visualization of hotel reviews", *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services*, ACM, pp. 229–238.

Calheiros, A.C., Moro, S. and Rita, P. (2017), "Sentiment classification of consumer-generated online reviews using topic modeling", *Journal of Hospitality Marketing & Management*, Vol. 0 No. 0, pp. 1–19.

Cantallops, A.S. and Salvi, F. (2014), "New consumer behavior: A review of research on eWOM and hotels", *International Journal of Hospitality Management*, Vol. 36, pp. 41–51.

Chai, T. and Draxler, R.R. (2014), "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature", *Geoscientific Model Development*, Vol. 7 No. 3, pp. 1247–1250.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000), "CRISP-DM 1.0: Step-by-step data mining guide", *The Modeling Agency*, available at: https://the-modeling-agency.com/crisp-dm.pdf (accessed 10 September 2015).

Duan, W., Yu, Y., Cao, Q. and Levy, S. (2016), “Exploring the impact of social media on hotel service performance: A sentimental analysis approach”, *Cornell Hospitality Quarterly*, Vol. 57 No. 3, pp. 282–296.

European Commission (Ed.). (2014), “Study on online consumer reviews in the hotel sector. Final report.”, European Union, available at: http://rpaltd.co.uk/uploads/report_files/hotel-reviews.pdf (accessed 28 June 2016).

Ganu, G., Kakodkar, Y. and Marian, A. (2013), “Improving the quality of predictions using textual information in online user reviews”, *Information Systems*, Vol. 38, pp. 1–15.

Hale, S.A. (2016), “User reviews and language: How language influences ratings”, *Human-Computer Interaction*, pp. 1208–1214.

Han, H.J., Mankad, S., Gavirneni, N. and Verma, R. (2016), “What guests really think of your hotel: Text analytics of online customer reviews”, *Cornell Hospitality Report*, Vol. 16 No. 2, pp. 3–17.

He, W., Tian, X., Tao, R., Zhang, W., Yan, G. and Akula, V. (2017), “Application of social media analytics: A case of analyzing online hotel reviews”, *Online Information Review*, Vol. 41 No. 7, pp. 921–935.

Hu, M. and Liu, B. (2004), “Mining and summarizing customer reviews”, in Kim, W. and Kohavi, R. (Eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, pp. 168–177.

Hu, Y.-H. and Chen, K. (2016), “Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings”, *International Journal of Information Management*, Vol. 36 No. 6, pp. 929–944.

Instituto Nacional de Estatística. (2016), “Tourism statistics-2015”, *Estatísticas Do Turismo*, July, available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=265858123&PUBLICACOEStema=55581&PUBLICACOESmodo=2 (accessed 1 September 2016).

Jiang, J., Gretzel, U. and Law, R. (2010), "Do negative experiences always lead to dissatisfaction? — Testing attribution theory in the context of online travel reviews", in Gretzel, U., Law, R. and Fuchs, M. (Eds.), *Information and Communication Technologies in Tourism 2010*, Springer, Vienna, pp. 297–308.

Kim, W.G., Lim, H. and Brymer, R.A. (2015), "The effectiveness of managing social media on hotel performance", *International Journal of Hospitality Management*, Vol. 44, pp. 165–171.

Kuhn, M. and Johnson, K. (2013), *Applied Predictive Modeling*, Springer New York, New York, NY, available at: (accessed 7 June 2015).

Kwok, L., Xie, K.L. and Richards, T. (2017), "Thematic framework of online review research: A systematic analysis of contemporary literature on seven major hospitality and tourism journals", *International Journal of Contemporary Hospitality Management*, Vol. 29 No. 1, pp. 307–354.

Lei, X. and Qian, X. (2015), "Rating prediction via exploring service reputation", *Proceedings of the 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, pp. 1–6.

Lin, C.H., Hsu, P.Y., Cheng, M.S., Lei, H.T. and Hsu, M.C. (2017), "Identifying deceptive review comments with rumor and lie theories", *Lecture Notes in Computer Science*, Vol. 10386 LNCS, presented at the 8th International Conference in Swarm Intelligence, ISCO 2017, Springer Verlag, pp. 412–420.

Liu, B. and Zhang, L. (2012), "A survey of opinion mining and sentiment analysis", in Aggarwal, C.C. and Zhai, C.X. (Eds.), *Mining Text Data*, Springer, pp. 415–463.

Liu, Y., Teichert, T., Rossi, M., Li, H. and Hu, F. (2017), "Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews", *Tourism Management*, Vol. 59, pp. 554–563.

López Barbosa, R.R., Sánchez-Alonso, S. and Sicilia-Urban, M.A. (2015), "Evaluating hotels rating prediction based on sentiment analysis services", *Aslib Journal of Information Management*, Vol. 67 No. 4, pp. 392–407.

Magnini, V.P., Honeycutt Jr., E.D. and Hodge, S.K. (2003), "Data mining for hotel firms: Use and limitations", *The Cornell Hotel and Restaurant Administration Quarterly*, Vol. 44 No. 2, pp. 94–105.

Marcheggiani, D., Täckström, O., Esuli, A. and Sebastiani, F. (2014), "Hierarchical multi-label conditional random fields for aspect-oriented opinion mining", *Advances in Information Retrieval*, presented at the European Conference on Information Retrieval, Springer, Cham, pp. 273–285.

Markopoulos, G., Mikros, G., Iliadi, A. and Liontos, M. (2015), "Sentiment analysis of hotel reviews in Greek: A comparison of unigram features", in Katsoni, V. (Ed.), *Cultural Tourism in a Digital Era*, Springer International Publishing, Cham, Switerland, pp. 373–383.

McGuire, K.A. (2017), *The Analytic Hospitality Executive: Implementing Data Analytics in Hotels and Casinos*, John Wiley & Sons, Inc, Hoboken, New Jersey.

Mellinas, J.P., María-Dolores, S.-M.M. and García, J.J.B. (2016), "Effects of the Booking.com scoring system", *Tourism Management*, Vol. 57, pp. 80–83.

Noone, B.M. and McGuire, K.A. (2014), "Effects of price and user-generated content on consumers' prepurchase evaluations of variably priced services", *Journal of Hospitality & Tourism Research*, Vol. 38 No. 4, pp. 562–581.

Öğüt, H. and Onur Taş, B.K.O. (2012), "The influence of internet customer reviews on the online sales and prices in hotel industry", *The Service Industries Journal*, Vol. 32 No. 2, pp. 197–214.

Ravi, K. and Ravi, V. (2015), "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", *Knowledge-Based Systems*, Vol. 89, pp. 14–46.

Salehan, M. and Kim, D.J. (2016), "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics", *Decision Support Systems*, Vol. 81, pp. 30–40.

Saralegi, X. and San Vincente, I. (2013), "Elhuyar at TASS 2013", in Esteban, A.D., Loinaz, I.A. and Román, J.V. (Eds.), *Proceedings of "XXIX Congreso de La Sociedad Española de Procesamiento de Lenguaje Natural"*, presented at the XXIX Congreso de la Sociedad Española de

Procesamiento de Lenguaje Natural, Madrid, Spain: El Congreso Español de Informática, pp. 143–150.

Schuckert, M., Liu, X. and Law, R. (2015a), "Hospitality and Tourism Online Reviews: Recent Trends and Future Directions", *Journal of Travel & Tourism Marketing*, Vol. 32 No. 5, pp. 608–621.

Schuckert, M., Liu, X. and Law, R. (2015b), "A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently", *International Journal of Hospitality Management*, Vol. 48, pp. 143–149.

Silva, M.J., Carvalho, P. and Sarmento, L. (2012), "Building a sentiment lexicon for social judgement mining", in Caseli, H., Villavicencio, A., Teixeira, A. and Perdigão, F. (Eds.), *Computational Processing of the Portuguese Language*, Springer Berlin Heidelberg, New York, NY, pp. 218–228.

Smola, A. and Vishwanathan, S.V.N. (2008), *Introduction to Machine Learning*, Cambridge University Press, Cambridge, UK.

Stein, R.M. (2007), "Benchmarking default prediction models: pitfalls and remedies in model validation", *Journal of Risk Model Validation*, Vol. 1 No. 1, pp. 77–113.

Surowiecki, J. (2005), *The Wisdom of Crowds*, Reprint edition., Anchor, New York, NY.

Talluri, K.T. and Van Ryzin, G.J. (2005), *The Theory and Practice of Revenue Management*, Vol. 68, Springer, available at: http://books.google.es/books?hl=pt-PT&lr=&id=GhdDbEM-_5oC&oi=fnd&pg=PR5&dq=Revenue+Management&ots=wG7RZyy2LU&sig=O7lVKcmTAU0REJIqASqvGQ6e-dI (accessed 23 December 2013).

Torres, E.N., Singh, D. and Robertson-Ring, A. (2015), "Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry", *International Journal of Hospitality Management*, Vol. 50, pp. 77–83.

Viglia, G., Minazzi, R. and Buhalis, D. (2016), "The influence of e-word-of-mouth on hotel occupancy rate", *International Journal of Contemporary Hospitality Management*, Vol. 28 No. 9, pp. 2035–2051.

Wang, B., Huang, Y. and Li, X. (2016), “Combining review text content and reviewer-Item rating matrix to predict review Rating”, *Computational Intelligence and Neuroscience, Article ID 5968705*, Vol. 2016, available at:https://doi.org/10.1155/2016/5968705.

Woo Gon Kim and Seo Ah Park. (2017), “Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance?”, *International Journal of Contemporary Hospitality Management*, Vol. 29 No. 2, pp. 784–802.

Xiang, Z., Schwartz, Z., Gerdes, J.H.J. and Uysal, M. (2015), “What can big data and text analytics tell us about hotel guest experience and satisfaction?”, *International Journal of Hospitality Management*, Vol. 44, pp. 120–130.

Xu, X. and Li, Y. (2016), “The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach”, *International Journal of Hospitality Management*, Vol. 55, pp. 57–69.

Xu, X., Wang, X., Li, Y. and Haghighi, M. (2017), “Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors”, *International Journal of Information Management*, Vol. 37 No. 6, pp. 673–683.

Zhao, X. (Roy), Wang, L., Guo, X. and Law, R. (2015), “The influence of online reviews to online hotel booking intentions”, *International Journal of Contemporary Hospitality Management*, Vol. 27 No. 6, pp. 1343–1364.

Zheng, W. and Ye, Q. (2009), “Sentiment classification of Chinese traveler reviews by support vector machine algorithm”, *Proceedings of the 3rd International Conference on Intelligent Information Technology Application*, IEEE Press, Piscataway, NJ, pp. 335–338.

[1] Although the data that we employed in this study is publicly available at Booking’s and TripAdvisor’s websites, sharing it publicly might be considered an infringement of the use of the data. In addition, because hotel names and hotel staff names are sometimes referred in the textual component of reviews, accessing this data allows the identification of the studied hotels. For these reasons, this data is not shared publicly

in this paper. However, the authors make themselves available to share this data for research proposes, including reproducibility, improvement, or expansion of this study. For those interested, please email one of the authors to obtain the link for both source datasets.