Research Article

INTERNATIONAL JOURNAL OF ENGINEERING BUSINESS MANAGEMENT

International Journal of Engineering Business Management
Volume 15: 1–16

DOI: 10.1177/18479790231151913
journals.sagepub.com/home/enb
SAGE

The verbalization of numbers: An explainable framework for tourism online reviews

Francesco De Nicolò[1,2,†], Loredana Bellantuono[3,4,†], Dario Borzì[5], Matteo Bregonzio[5], Roberto Cilli[2], Leone De Marco[5], Angela Lombardi[2,4], Ester Pantaleo[2,4], Luca Petruzzellis[2], Ariona Shashaj[6], Sabina Tangaro[7,4], Alfonso Monaco[4], Nicola Amoroso[8,4,‡], and Roberto Bellotti[2,4,‡]

Abstract

Online reviews have been found very useful in decision-making. It is important to design and implement accurate systems to analyze the reviews and, based on textual information, predict their ratings. Given the different sources, languages and evaluating systems, intelligent systems are needed to use textual and numerical reviews to better understand the evaluation of the tourist experience and derive useful information to improve the offer. This paper aims to present an eXplainable Artificial Intelligence framework that contributes to the discussion on numerical and textual evaluations of the hospitality experience. It combines sentiment analysis and machine learning to accurately model and explain the evaluation of the tourist experience. The main findings are that review ratings should be used with caution and accompanied by a sentiment evaluation and explainability plays a central role in identifying which are the key concepts of positive or negative ratings, providing invaluable intelligence about the tourist experience.

Keywords

Tourism intelligence, explainable artificial intelligence, sentiment analysis, machine learning

Date received: 13 October 2022; accepted: 26 December 2022

Introduction

The importance of reviews in consumer decision-making has been widely confirmed by both academic research and practice.[1,2,3,4,5] Different aspects have been considered such as valence, volume, variation, perceived usefulness[6,7,8] as well as their outcomes such as review-based product rankings, trust in online reviews and management responses to consumer reviews.[9,10,11]

Research in tourism has highlighted online reviews as a major driver of brand choice and sales,[12] hotel performance,[13] hotel bookings[14] and destination choice.[15] In particular, their effect on guests' satisfaction[16] has opened the discussion on numerical and textual evaluations of the hospitality experience.[17]

[1]Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Bari, Italy
[2]Dipartimento Interateneo di Fisica, Università degli Studi di Bari Aldo Moro, Bari, Italy
[3]Dipartimento di Scienze Mediche di Base, Neuroscienze e Organi di Senso, Università degli Studi di Bari Aldo Moro, Bari, Italy
[4]Istituto Nazionale di Fisica Nucleare - Sezione di Bari, Bari, Italy
[5]3rdPlace SRL, Milano, Italy
[6]Network Contacts SRL, Molfetta, Italy
[7]Dipartimento di Scienze del Suolo e della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy
[8]Dipartimento di Farmacia e Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Bari, Italy

[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work

Corresponding author:
Luca Petruzzellis, Dipartimento Interateneo di Fisica, Università degli Studi di Bari Aldo Moro, Bari, Italy.
Email: luca.petruzzellis@uniba.it

Numeric characteristics like the number of stars and the number of words included in a text, have been studied in both decision-making[18,19,20] and customer satisfaction research.[21,22] However, the scalar ratings do not provide any information on those characteristics that customers like or do not like, while textual reviews display consumers' preferences and all the nuances of satisfaction, which can be extracted with specific techniques such as opinion mining and sentiment analysis.[23]

Previous research[24,25,26] used a mixed-method approach to analyze the numeric (ratings) and textual (reviews' text) information of online reviews to provide a deeper understanding of such a complex phenomenon. Recent studies[23,27] have investigated the possibility to design and implement accurate systems to analyze the reviews and, based on textual information, predict their ratings. The variety of sources, the nuances of languages and the different evaluating systems (ratings vs verbalization) call for intelligent systems to use textual and numerical reviews to better understand the evaluation of the tourist experience and derive useful information to improve the offer. The volume, subjectivity, and heterogeneity of social web-data require the adoption of specific methods combining Natural Language Processing (NLP) techniques to tokenize customers' reviews and carry out a subsequent sentiment analysis.[28,29] However, the reliability of these approaches is strongly affected by the reliability of the ratings; in fact, misleading data, i.e. reviews with positive evaluations and negative ratings or vice versa, are common due to psychological meachanisms such as social pressure.[30]

Previous research[31] has highlighted the need for a simplification of learning models and an improvement in the speed of analysis of big data. This paper aims to provide a unified framework to analyze the evaluation of the tourist experience and outline its key factors based on both ratings and textual reviews. Specifically, we used a multidisciplinary approach, combining computer science to collect data from an online platform, sentiment analysis to detect the anomalous reviews whose score does not match with the measured sentiment, machine learning to train the classifier and the Shapley paradigm[32,33] to explain the decisions taken by the model.

The theoretical contribution of the study is twofold, first, it contributes to the literature on online reviews by clarifying the impact of the combination of numbers and texts to help understanding and predicting tourist preferences. Second, to the best of our knowledge, it is one of the first studies using a cross-validation framework of the forecast model[33] to avoid biased results based on the particular train-test subdivision of the dataset. Moreover, from a methodological point of view, this paper uses sentiment analysis and classical machine learning methods in a fairly simple combination, obtaining results comparable to those achieved with deep learning models,[34], even though in a binarized-class problem.

The results also offer insights for practitioners and policy makers on how reviews should be analyzed to understand better their customers in order to improve their experiences and what they look for to characterize them.

The remainder of the paper is organized as follows. In the next section the theoretical background is presented analyzing also the different methods, then the proposed framework is discussed outlining the research methodology. Finally, the results of our empirical analysis are presented followed by discussion, future research and limitations.

Theoretical background

Literature has widely investigated the effects of online reviews on tourist experience,[35,36,37,38] analyzing both antecedents (i.e., review extremity, length, readability, and sentiment) and consequences (i.e., bookings, product evaluation, higher prices and room sales).

Since the dual rating system (i.e., a star-numeric value and a textual description) is based on travellers' expectation-experience congruence,[39] the level of matching between experience and expectations, should resonate in the ratings. On the other hand, the textual content describes the individual experience and evaluation in details, which cannot be represented with a single numeric measure like review rating.[40,41]

The online information overload and the decision-making costs push tourists to rely more on review ratings than textual reviews.[13] While the numeric rating has been studied in terms of valence or absence versus presence on websites,[1,42,43] the textual part has been studied with simple definitions and semantic, sentiment and linguistic measures.[44,45]

The evaluation of the experience, the rating

Rating is the reviewer's overall, numeric evaluation of the product and actual experience, which reflects the level of satisfaction with the product.[12] It consists often of a scale of 5, ranging from terrible experience to excellent experience (5 stars). As an indicator of travelers' experience and satisfaction, it represents an important information source to evaluate a hospitality or tourism product in a purchasing decision.[46] Ratings may induce positive judgements about the utility of online reviews,[7] being reviews with extreme ratings (either positive or negative) perceived as more useful and enjoyable than those with moderate ratings.[12,47] showed that moderate reviews (3 stars) were considered as uninformative because they contain ambiguous information, while clearly positive (4–5 stars) or negative (1–2 stars) reviews have clear implications for the purchase decision.

Indeed, numerical ratings are popular among consumers because they can easily be processed and can lessen information asimmetry.[48] Recent literature has demonstrated that customers' ratings influence the information adoption of travelers[49] and their purchase intent;[50] besides, they can improve the performance of restaurants,[51] the sales and

prices of hotel rooms.[52] In particular,[8] showed how specific factors like cleanliness have positive impact while others like price a negative one.

The interactive effect among valence, volume and variance has been also widely documented;[8,53] however, the level of explicability remains questionable. Therefore, our first research question is,

RQ1 – Does the numerical rating fully explain the experience evaluation?

The verbalization of the experience, the textual review

Feelings play a relevant role in people's experiences.[54,55,56] According to recent studies the way people communicate and how words are used reveal how they perceive reality.[57,58] Hence, these communications, the words they use and the way they are used, can unveil deep psychological insights.[55,59] The existence of a strong bond between the way we communicate and our feelings is now acknowledged.[60,61] Besides, it is also known that our communicating style is usually based on unconscious processes; thus, it can accurately unveil how the message content is related to subjective feelings and perceptions of reality.[62] For example, it has been observed that online consumers tend to adopt an abstract language during the first stages of shopping, the contrary is true for the last stages.[58]

Therefore, the varied characteristics of the review text, namely the semantic, sentiment and linguistic characteristics, the review concreteness[63] as well as the length of text have been widely investigated.[2] While semantic features (i.e., words, topics and semantic relationships between linguistic entities) are used to assess information quality,[53] sentiment has been used to to capture consumers' emotions[64] and their positive or negative orientation.[65]

Various frameworks of opinion mining have been used to summarize visitors' opinions and experiences from reviews according to categories and lexicon-based sentiment.[66,67,68] In-depth qualitative analysis and the big data approach have provided additional insights that address a variety of research topics such as popular keywords used,[69] guest satisfaction in restaurants and hotels,[22,26] tourist experience of a specific destination or an event ,[70,71] and attributes that trigger revisit and referral intentions.[72] However, the combination between verbalization and numbers still needs to be better investigated. Therefore, our second research question is,

RQ 2 – Does the combination of wording and rating represent the real sentiment of the tourist?

The explainability of the travel experience, the Shapley value

The different prediction models used to combine the analysis of textual and numerical reviews[7,73,74] have called for different methods to reach the explanability.[75] Explainability assumes a primary importance when examining data characterized by high dimensionality such as the tourists' online reviews.[76] Among others, the Shapley value has become the basis for several methods that attribute the prediction of a machine-learning model on an input to its base features.[77] Shapley values provide a mathematically fair and unique method to attribute the payoff of a cooperative game to the players of the game. The basic idea is to consider the features as collaborative agents whose goal is to reach a decision about the examined classification task.[78]

Previous studies in tourism[79,80] have used the TF-IDF model (Term Frequency – Inverse Document Frequency) in association with Shapley values to extract the important features from reviews. However, these studies have found the importance of elements such as room and service quality[80] or being part of a renowned international hotel chain,[79] mostly influenced by the context of investigation. However, in such a dynamic industry with very individual preferences it would be more helpful to directly examine tourist reviews to determine the key aspects explaining tourists' decisions and ratings without any assumptions on their expectations. Therefore, our third research question is,

RQ 3 – What factors are evaluated in the reviews and predict future choices?

Methodology

The main objective of the paper is to provide a unified framework to evaluate the tourist experience and outline the key elements driving it. In September 2020, a web scraper was used to collect the reviews posted on TripAdvisor regarding the hospitality infrastructures in Puglia, a very popular tourist destination in the South of Italy. Specifically, we drew a total of 13,399 reviews concerning 974 facilities, posted between May 2004 and June 2020 and related to the summer season.

We draw data from TripAdvisor for three main reasons, (i) among the various tourism-dedicated platforms, it is one of the most accessed with more than 860 millions reviews and 8.7 millions opinions, more than five million registered users who visit the platform 30 million times per month on average; (ii) it deals with heterogeneous facilities and tourism services including accommodations, restaurants, experiences, airlines and cruises; (iii) it includes a numerically-based rating system that allows to develop a supervised model. Given its use and content available,

TripAdvisor implies not only high practical significance of the findings but also strong theoretical contributions.

For each entry a text review and six data fields were included,

- **rating**. The numerical score from1 (bad experience) to 5 (excellent experience) that each user gave to the tourist experience;
- **review_id**. A 9-digit numerical code uniquely identifying the review;
- **struc_id**. A 40-digit alpha-numerical code indicating the facility;
- **struc_name**. The name of the facility identified by the struc_id;
- **date**. The date the review was input in the platform;
- **vicinity**. The address of the reviewed facility.

We considered only reviews in English since they reduce the potential bias of language and NLP tools for the pre-processing (removing stop-words, lemmatisation, stemming) of English texts are well consolidated with respect to other languages.[81]

In terms of rating the data are highly unbalanced; more than half of reviews represents an excellent experience (numerical score equal to 5), 27% are given a score equal to 4, 10% are related to a score equal to 3, while less than 10% reviews have a numerical rating of 2 or less. In other terms, considering as *positive* those reviews having a score greater or equal to 3,[7,12,47] the number of positive reviews is much greater than the negative one. Since this rating imbalance is commonly observed in studies dealing with services' reviews,[1] to obtain reliable results we designed an integrated approach that combined NLP techniques to pre-process the data and perform sentiment analysis, a learning framework to assess to which extent online reviews provide a robust base for an accurate prediction and an explainability analysis of the classification model to highlight the key factors driving the tourist experience.

Data have undergone a thorough cleansing process, given that real-world data contain up to 40% of inconsistent data.[82] To meaningfully tokenize the reviews, the case was changed to lower-case, punctuation and stop words were removed, and data were stemmed.[83] Consistently with previous studies[7,12,47] we binarized the ratings so that reviews with rating lower than 3 were considered negative and labeled **0**, while those having a rating higher or equal to 3 were considered positive and labeled **1**.

We then performed a sentiment analysis of the reviews and compared the users' ratings with the measured sentiment. Since it is possible to have a mismatch between the rating and the sentiment of a review,[84,85] we defined contradictory reviews as those reviews belonging to class 0 (negative-rated) but with a positive sentiment and vice-versa. Therefore, we filtered out 1460 reviews, about 9% of the sample.

The cleaning step is important for our framework's reliability despite the fact that these reviews represent 9% of the whole dataset. Since our dataset is highly imbalanced (less than 10% of reviews are negative), we *undersampled* positive reviews to obtain unbiased classification models, that is, we randomly chose a number of positive reviews equal to that of negative ones. Accordingly, we obtained a perfectly balanced dataset containing all the negative reviews and a subsample of positive reviews. Then, we fed this balanced dataset into machine learning algorithms. This approach was repeated 100 times.

After the textual processing analysis, we were able to create the TF-IDF matrix,[86,87] in which each element is the product of two factors, $TF_{(i,j)}$, representing the frequency of the word i in the review j, and $IDF_{(i,j)}$, an importance measure associated to each term,

$$TF_{(i,j)} = \frac{n_{ij}}{|d_j|}; \; IDF_{(i,j)} = \log \frac{|D|}{|d(i)|},$$

with $n_{i,j}$ = the number of occurrences of the word i in the review d_j, D = the set of the reviews and $d(i)$ = the number of elements in D containing the word i at least once. $|.|$ denotes the cardinality.

On the one hand, the first factor emphasizes the high frequency of a word within a review; the more cited a word, the more its importance. On the other, the second term penalizes the high frequency in the whole set of reviews since a word used in all reviews would yield poor discrimination and emphasize the role of the rarely occurring terms.

Then, the data fed the Random Forest (RF) algorithm to measure their informative content.[88] We also compared its results with other state-of-the-art classifiers such as the Gaussian Naive Bayes (GNB), the Support Vector Machine (SVM) and the XGBoost (XGB) classifiers[89,90,91] to check that the informative content evaluated by the machine learning model was independent from the model used. Appendix A contains a concise but complete description of the functioning of all these algorithms.

Their performance was then evaluated with five distinct metrics:[92]

1. Accuracy (acc), defined as the ratio between correctly classified samples and the total number of samples.

$$acc = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP (True Positive) and TN (True Negative) are the correctly classified samples, while FP (False Positive) and FN (False Negative) are the wrongly classified samples.

2. Sensitivity (sens), also called Recall or True Positive Rate, is the ratio of the positive correctly classified samples.

$$sens = \frac{TP}{TP + \mathrm{FP}}$$

3. Specificity (spec), also called True Negative Rate, is the ratio of the negative correctly classified samples.

$$spec = \frac{TN}{TN + FN}$$

4. Area Under the Curve (AUC) refers to the area under the Receiver Operating Characteristic, a curve whose points are represented in terms of Sensitivity and Specificity. It is a measure of how far a model is from being a random guess.

5. F1-score is defined as the harmonic mean of correctly classified samples.

$$F1 = \frac{TP}{TP + \frac{FP+FN}{2}}$$

In order to ensure statistical robustness to our findings, all analyses were carried out 100 times in a 10-fold cross-validation framework.

To explain how the considered models reached a decision with a specific classification score and, therefore, to understand which factors were driving the tourist experience, we adopted the explainability framework based on Shapley values.[93] Accordingly, for each observation we could evaluate how and why the model reached a specific decision. As a consequence, the same feature has a particular Shapley value for all available observations, whose distribution highlighted the level of importance given by the mean absolute Shapley value. We deliberately decided not to perform any feature selection to keep our learning framework as simple as possible. This helped in obtaining an *ex post* feature importance evaluation that further clarified the key aspects of tourist choices.

Results

First, we found that the reviews were asymmetrically distributed in terms of ratings. The positive ratings (i.e., higher than or equal to 3) outnumbered the negative ones (i.e., less than 3) - the positive accounted for the 91% of the entire dataset. Also, the number of reviews showed a steepen continuous growth, which is commonly observed in studies dealing with services' reviews.[94] Since the rating distribution is highly skewed in favor of positive reviews and a different threshold for binarization would have not yielded any significant differences, we adopted undersampling to balance the data to avoid any bias of the learning model.[95,96]

Then, we obtained a 11,848 × 16,898 matrix, with reviewers' evaluation of their experience as target variable. To set up an effective rating forecast model, we studied emotions expressed in the reviews so that we found not only contradictory reviews, but also the emotions that mostly affected our model's performances. In fact, the sentiment analysis highlighted four emotions, Happiness, Sadness, Anger and Surprise. The intensity of the emotions showed that the reviews express happiness more than other emotions, thus confirming a positive experience (Figure 1). Indeed, higher-rating reviews are determined by the happiness of the experience they report.

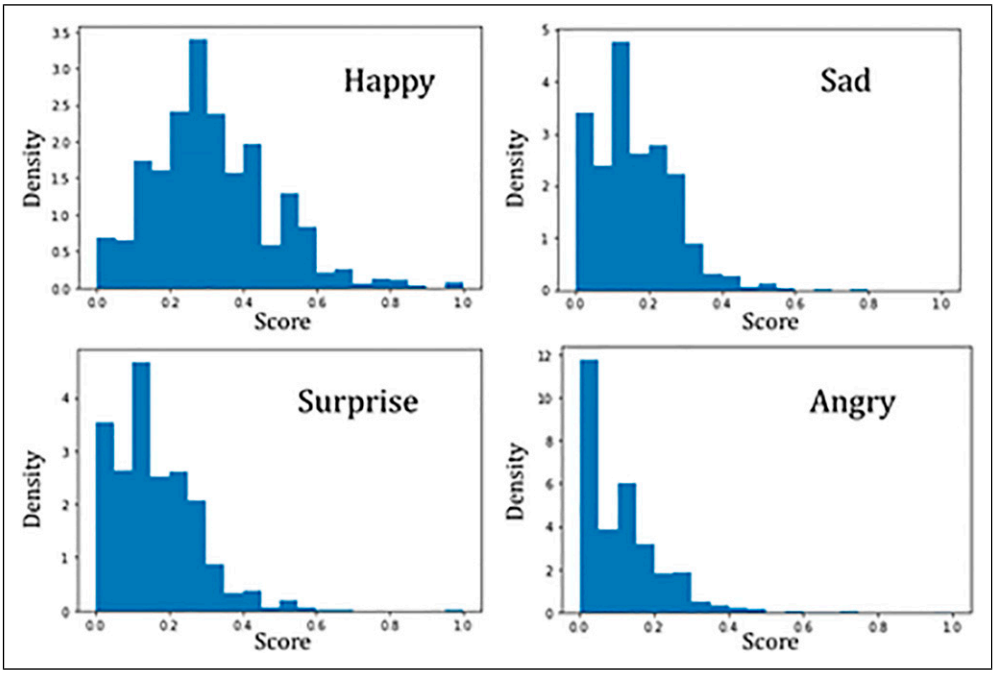

Figure 1. Score density distribution of the emotions. Each review was enriched with four continuous scores (one for every emotion) and the scores were normalized.

These results suggest a general correspondence between the numerical rating and the verbalization of the rating (experience). Also, the number of reviews whose ratings did not match with the expressed sentiment confirmed the effect on the classification performance. We evaluated with a RF model that about 80% of the correctly classified positive reviews (TP) have Happiness as the predominant emotion, while the correctly classified negative reviews (TN) have Sadness as the predominant emotion. On the other hand, FP and FN reviews are more balanced in the presence of Happiness and Sadness (Figure 2), suggesting the mismatch between ratings and verbalization.

Indeed, since TP and TN reviews are more polarized in terms of emotions, the model is able to correctly recognize them in relation to the emotions they express. On the contrary, we found reviews with a predominant happy sentiment in the FN ones as well as reviews with a predominant sad sentiment in the FP ones. These reviews do not express a clear feeling, making the model classify them in a wrong way. Therefore, we decided to exclude them from the analysis. Besides, to ensure that the informative content evaluated by the machine learning model was independent from the model used, we compared the results of RF with other machine learning algorithms. The cross-validation comparison of the models (Figure 3) shows that the RF model resulted the most accurate, although all models reached satisfactory levels of accuracy (Table 1).

Despite this, RF model scored significantly better than the other models in all the measured metrics, as established using a Mann-Whitney statistical test ($p < 0.01$) (Table 1).

The general agreement among the models was confirmed by the Pearson correlation coefficients of the corresponding classification scores; RF and SVM showed the highest correlation (0.908) closely followed by XGB (0.881), while GNB showed the lowest one (0.519). Also, these values greatly improved with respect to the performance obtained on the whole dataset (see Table 2). Consistently with the literature on the impact of noisy data (i.e., the contradictory reviews) on machine learning algorithms[97,98,99] we expected such a decrease in performances.

Therefore, we were able to ensure that this measurement depended only on the informative content provided by the reviews and not from the specific machine learning algorithm adopted.

Finally, to highlight the key factors driving the classification we identified the most important features and their contribution to the classification score based on their intrinsic value through the Shapley values (Figure 4).

The words with the highest occurrence are breakfast, work and staff, which predict the likelihood of a review to be positive, thus influencing the decision. The absolute mean Shapley value, which is a measure of the words' impact on the model, shows that the vast majority of the available terms has a low, if no, impact on the model. In particular, using the elbow-point method we determined 64 important features (Table 3).

As in the example in Figure 5, the positive review is mainly explained by the words visit, breakfast and quiet, while the negative review show a general unsatisfactory situation, mostly affected by specific factors like breakfast, staff and help.

Discussion

The results show a classification framework to evaluate the rating and verbalization of the tourist experience and highlight its determinants to predict future satisfaction from the reviews. The overall pipeline determined by our work consists in the following steps (see Figure 6): (1) A review feeds the Random Forest model; (2) this model accurately

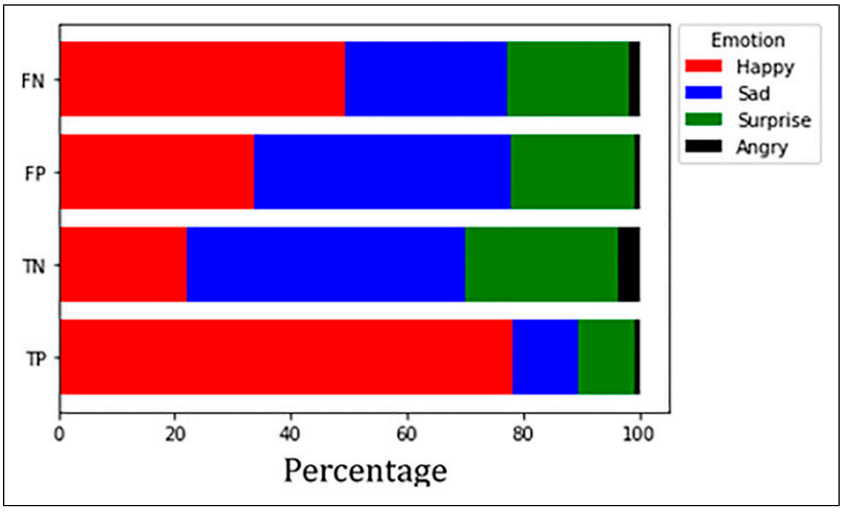

Figure 2. Percentage of emotions in TP, TN, FP, FN reviews determined by the Random Forest model.

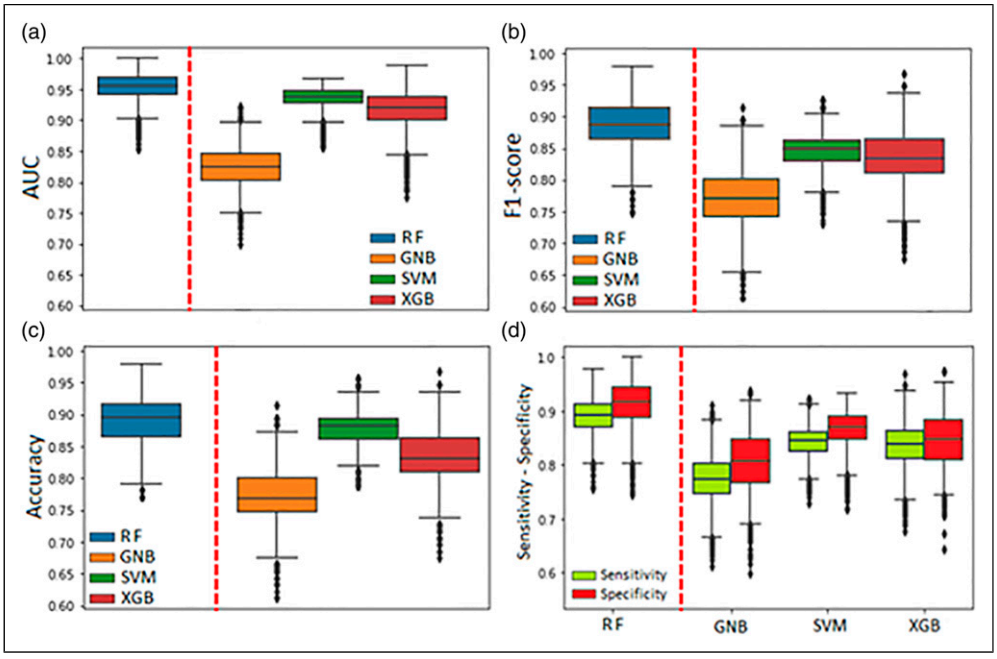

Figure 3. Performance measures for the machine learning models. **a.** AUC; **b.** F1-score; **c.** Accuracy; **d.** Sensitivity and Specificity.100 cross-validation iterations (10-fold) of 100 undersamplings of the original data.

Table 1. Models' performance measures obtained by filtering out contradictory reviews.

Model	acc (%)	AUC (%)	F1-score (%)	sens (%)	spec (%)
RF	89 (82–95)	96 (91–99)	89 (82–95)	89 (82–95)	92 (82–98)
GNB	77 (70–85)	83 (76–85)	77 (68–82)	76 (69–85)	81 (68–91)
SVM	88 (82–91)	94 (91–96)	85 (80–88)	88 (80–86)	87 (76–88)
XGB	83 (76–91)	93 (86–97)	84 (77–90)	88 (76–91)	84 (73–94)

Table 2. Models' performance measures obtained by including contradictory reviews.

Model	acc (%)	AUC (%)	F1-score (%)	sens (%)	spec (%)
RF	62 (58–66)	66 (62–70)	62 (58–66)	62 (58–66)	60 (54–65)
GNB	54 (49–60)	58 (53–62)	54 (50–60)	54 (49–59)	61 (58–64)
SVM	60 (57–64)	60 (58–65)	59 (55–62)	58 (54–63)	61 (58–65)
XGB	58 (55–62)	62 (58–64)	56 (53–61)	60 (57–64)	58 (55–63)

predicts its positive/negative rating based on textual data; (3) explainability algorithms (Shapley values) are determined, in order to highlight the most important words influencing the model's rating.

Basically, we evaluated to which extent online reviews allow a reliable assessment of the tourists' experience and their satisfaction. First, we observed the presence of misleading evaluation among the collected reviews; in many cases the numerical assessment did not match the sentiment expressed. Considering how the contradictory reviews are distributed among positive and negative reviews, we observe that 80% of negative reviews are contradictory. Since all the balanced datasets have twice the number of negative reviews and contains all negative reviews, we fed the machine learning

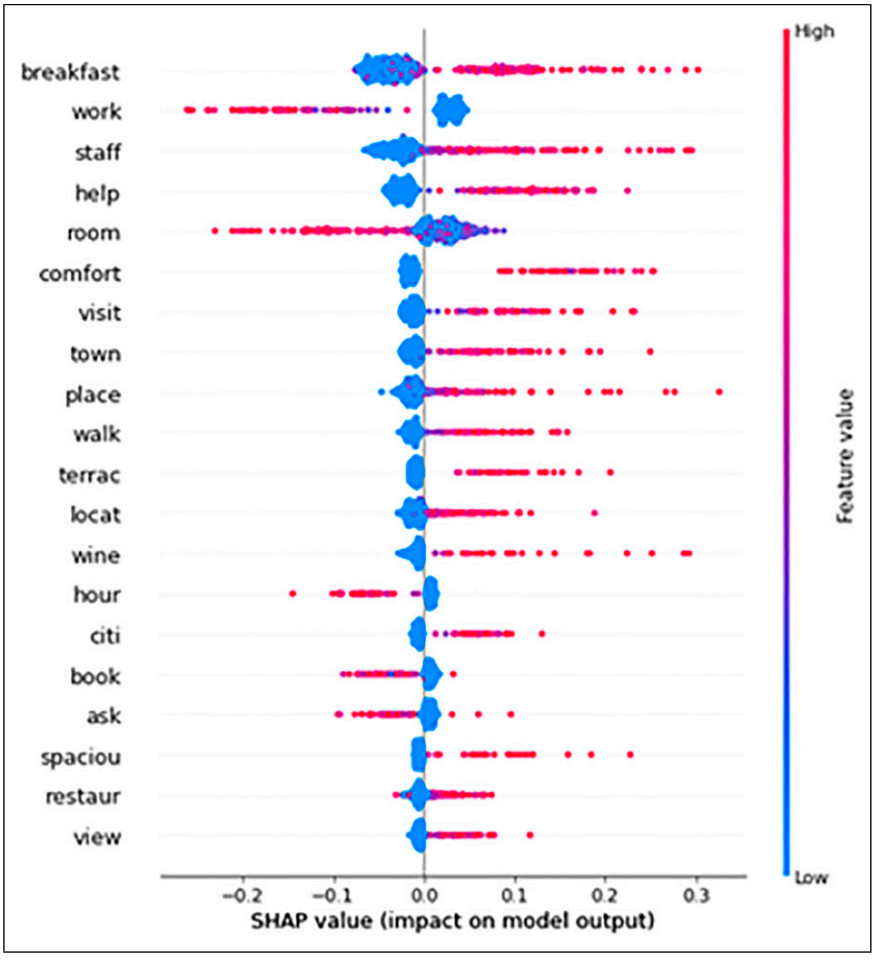

Figure 4. The Shapley values of the first twenty important words. The contributions towards a positive or a negative review are distinguished according to the frequency a word appears within the text (high/low).

Table 3. The most important words in reviews' classification. Words are sorted in ascending order according to their Shapley values.

1	Great	17	Bad	33	Welcom	49	Meat
2	Friendli	18	Work	34	Hotel	50	Dinner
3	Love	19	Recommend	35	Waiter	51	Say
4	Breakfast	20	Best	36	Book	52	Charg
5	Excel	21	Tell	37	Host	53	Walk
6	Town	22	Area	38	Enjoi	54	Air
7	Help	23	Room	39	Terrac	55	Terribl
8	Comfort	24	Amaz	40	Stai	56	Lecc
9	Perfect	25	Water	41	Fantast	56	Atmospher
10	Nice	26	Delici	42	Star	58	Wine
11	Staff	27	Ask	43	Disappoint	59	Free
12	Good	28	Place	44	Rude	60	View
13	Poor	29	Park	45	Worst	61	hour
14	Locat	30	Clean	46	Highli	62	Fresh
15	Beauti	31	Relax	47	Definite	63	Local
16	Wonder	32	Pool	48	Arriv	64	Peopl

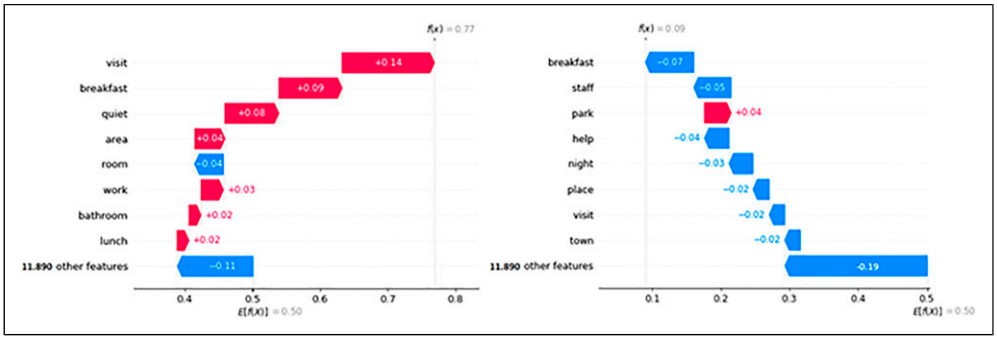

Figure 5. The Shapley values for two correctly classified reviews. Note, On the left a positive-rated review and on the right a negative-rated one.

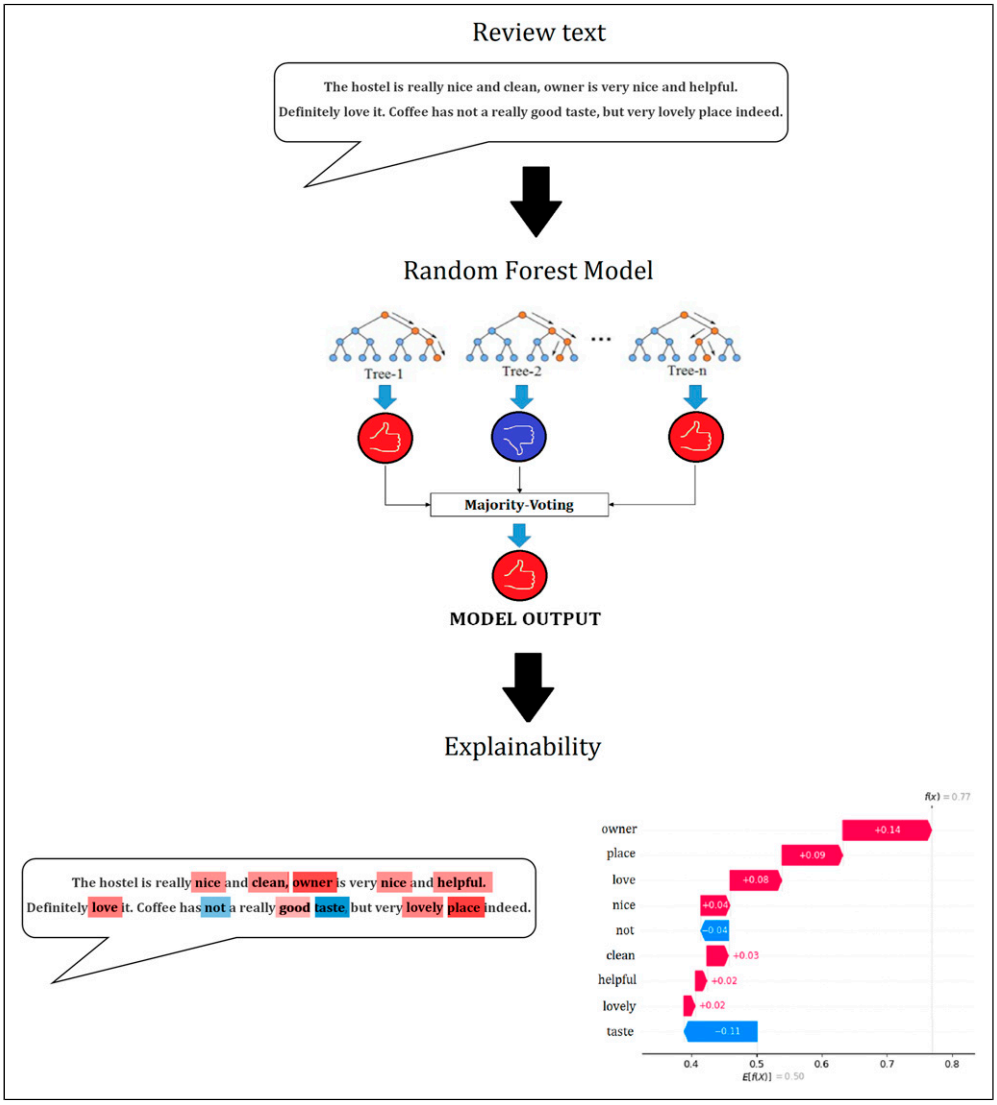

Figure 6. The overall pipeline determined by our work. (1) A review feeds the Random Forest model; (2) this model accurately which predicts its positive/negtive rating based on textual data; (3) explainability algorithms (Shapley values) are determined, in order to highlight the most important words influencing the model's rating.

algorithms with datasets having at least 40% of error (or *noise*) level. Previous studies on the sensitivity of machine learning algorithms to noisy data show that models' accuracy decays almost linearly with the noise level: 40% of error level in data reduces by 30%–40% a model's accuracy.[97,98,99]

This contributes to the literature on text mining,[58] highlighting that text can predict future behavior since the word usage and writing styles are indicative of some stable inner traits as well as more transient states, which affect people's behaviors. The systematic relationship between the words people use and emotional states[100] derives from the human tendency to tell stories and express internal thoughts and emotions through these stories, which are essentially made possible by language. Therefore, even if the content might be similar across different individuals, the manner in which they convey that content differs.

Moreover, text can provide insight into a person's attitudes toward or relationships with other attitude objects—whether that person liked a hotel room or some services. Consistently with previous studies, we were able to cross-validate different models and evaluate their performance. In particular, we demonstrated that the use of sentiment analysis is fundamental to accurately predict if the tourist is going to assign a positive or negative evaluation just based on the textual review.

Second, in terms of methodology, since the main difficulty involved with using text for predictions is that text can generate several features (words) that are all potential predictors for the outcome of interest, the method used showed that the classification performance was robust independently on the adopted model or the specific considered metrics, despite the higher performance of the RF model compared to the others. Also, we found a strong agreement with the predictions of the other models, especially SVM and XGB; thus implying that the explainability analysis is independent on the particular considered model.

Contrary to previous research,[33] we used a cross-validated framework in order to get more robust results. In fact, our results are independent from the train-test subdivision, thus avoiding biased results and inaccurate conclusions and reaching findings comparable with those obtained using state-of-the-art deep learning methods.[34]

In addition, through the Shapley paradigm we explained the RF decisions. On the one hand, the findings underlined that the most important words are related to places, meals and staff, and in particular the word breakfast. This connotates the typical tourist offer and can be explained by both the most common type of hospitality structures, namely bed-and-breakfast, and the connection with food, one of the most importante elements of a tourist experience. On the other hand, the Shapley values also highlighted how feature values affect the classification score. This helps in characterizing the experience and predicting the satisfaction (positive or negative evaluation).

The results have also strong managerial implications in the way the tourist offer can be improved through the creation of personalized services on the basis of the reviews and the features that contribute to living a memorable experience. Understanding the actual behaviors of reviewers through such behavioral-tracking data set can reveal many valuable insights for business improvement and marketing effectiveness. Since consumers' preferences can be dynamic and expensive to monitor, advances in technology can help not only in reducing the cost of collecting and mining data in an efficient and non-intrusive manner but also in providing more useful information to better target the offerings.

This paper proposes an accurate workflow to examine online reviews and exploit their informative content in order to provide valuable insight to tourism stakeholders and policy-makers guide. For the sake of simplicity, we did not consider possible stratifying variables such as nationality or age. However, it is reasonable to assume that these factors can affect the judgements in that, for example, expectations and needs of a teenager are necessarily different from those of a family with children. Future studies will be devoted to enlarge the examined geographical area and take into account potentially confounding factors, such as age or nationality.

Also, while in this paper we considered an *ex post* feature importance analysis based on Shapley values, it would be possible to include a feature importance step in the learning phase to reduce the amount of words to consider and simplify the analysis. The design and implementation of dedicated strategies to maximize and exploit the informative content provided by online reviews would deserve further investigations.

Although the main aim of this paper is to analyze tourists' reviews and give useful insights to tourism stakeholders, the proposed pipeline could be easily implemented in all those fields where textual data are collected/used. By analyzing products and events' reviews this model helps highlighting those aspects that mostly influence reviewers' feelings. In fact, the main components of our workflow are: (1) *Review scraping,* obtained by using packages that are freely available to every programming language development environment.[101] These packages can be used to scrape almost all social media platforms (like Booking.com, Twitter, Facebook, Amazon) and obtain the desired textual information. In this paper we used Python programming language and related packages; (2) *NLP techniques and Sentiment Analysis* that have reached optimal performances in the analysis of textual data and in extracting useful features describing the meaning of texts. These techniques are all encapsulated in *NLTK* and *SpaCy* Python packages;[102] (3) *Machine Learning and Explainability algorithms* that are widely used in various fields such

as wildfire preventions,[103,104] medicine,[105,106,107] drug discovery.[108] All the machine learning algorithms used in this paper are implemented using the *scikit-learn* Python package, which is one of the well-known and mostly-used package used in machine learning application and research.[109]

Conclusions

On platforms like Trip Advisor tourists ensure a continuous flow of information about their experiences and level of satisfaction. In this paper, we proposed an accurate workflow to examine online reviews and exploit this informative content to provide valuable insight to tourism stakeholders and policy-makers guide. On the one hand, we evaluated to which extent online reviews allow a reliable assessment of the tourists' experience and their satisfaction. We demonstrated that the use of sentiment analysis is fundamental to accurately predict if the tourist is going to assign a positive or negative evaluation just based on the textual review. Although all the scripts and packages used in our analysis are applied to analyze tourists' reviews, our pipeline can be used as it is in various contexts, simply changing the scraping package: for example, products' reviews (e.g., analysis of Amazon reviews) or Twitter events logs can be obtained. Accordingly, our pipeline has the potential of being proposed as a general framework to be used to extract useful insights from textual data.

Appendix A

Appendix A – Machine Learning models

Random Forest (RF)

Random Forest can be used for both regression and classification problems. In this work, we use it to classify a review as positive or negative based on its textual data, represented in the corresponding column of the TF-IDF matrix. Random Forest is a generalization of decision trees.[88]

In fact, Random Forest is an ensemble learning method that works by constructing a multitude of decision trees at training time. In particular, every tree is trained on a *bootstraped* sample of training data (i.e. sampling with replacement from training data) and each tree uses a random subset of predictors to take decisions, in order to overcome the presence of strong predictors. The output of the random forest is the class selected by most trees (*majority vote* rule). Taking decisions based on an ensemble of trees greatly improves the performance of a single decision tree.[110]

Figure A1 best summarizes how a RF works in classification settings.

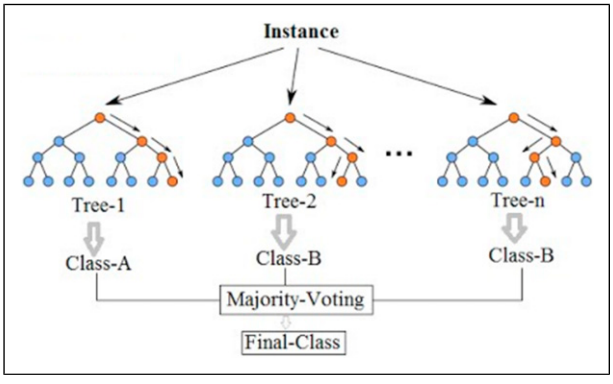

Figure A1. How a Random Forest determines its output from the trees in its ensemble.

Gaussian Naïve Bayes classifier (GNB)

GNB classifier is based on Bayes' Theorem of probability theory and on some strong assumptions about the independence of input variables in determining the probability of an item to belong to an output class.[89] In particular, if an instance determined by N input variables, $(x_1, \ldots, x_N)$, should be assigned to one of K classes, $(C_1, \ldots, C_K)$, GNB aims at calculating the corresponding conditional probabilities $p(C_i|x_1, \ldots, x_N)$, $\forall i \in \{1, \ldots, K\}$. In order to determine these probabilities, GNB refers to Bayes' Theorem

$$p(C_i|x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N|C_i)p(C_i)}{p(x_1, \ldots, x_N)}$$

where: $p(C_i)$ is called *prior* probability; $p(x_1, \ldots, x_N|C_i)$ is denoted as *likelihood* distribution; $p(x_1, \ldots, x_N)$ is referred to as *evidence* distribution.

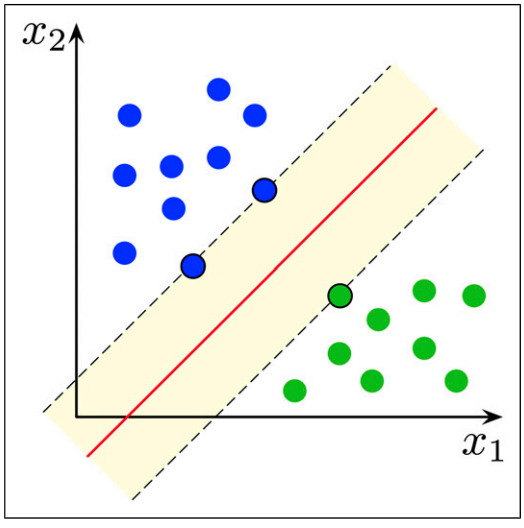

Figure A2. The result of the SVM algorithm applied to a dataset with two input features (x_1 and x_2), for the sake of clarity. The two classes are reported in blue and green directly as colors of the data points. The maximum margin hyperplane is reported in red.

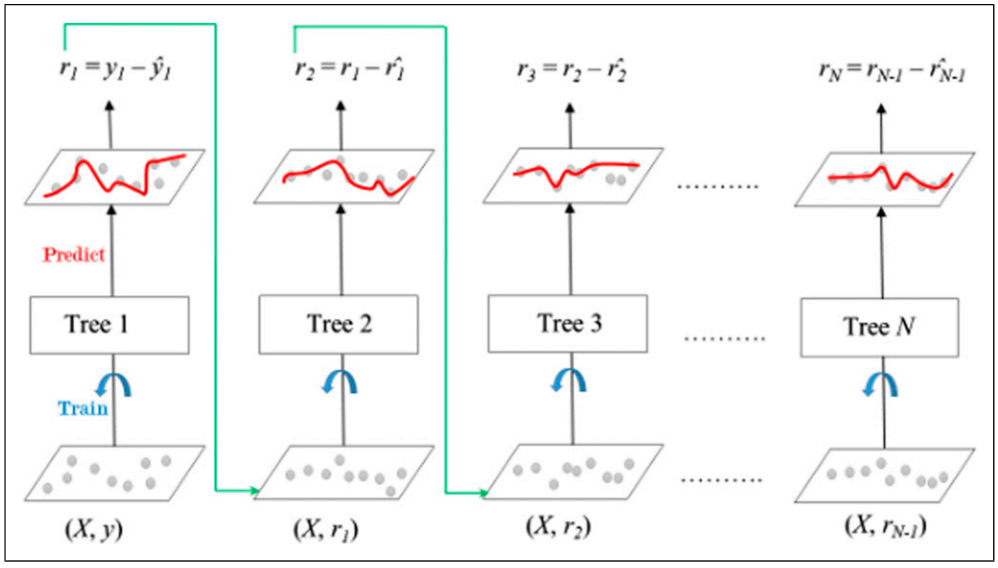

Figure A3. Graphical representation of the XGB algorithm.

GNB assumes that likelihood distributions are Gaussian, whose parameters should be estimated in the training phase of the model.

Since $p(x_1, \ldots, x_N|C_i)p(C_i) = p(x_1, \ldots, x_N, C_i)$, then applying simple probability rules and considering the hypothesis of mutual independence of the N input variables, it can be written

$$p(C_i|x_1, \ldots, x_N) = \frac{p(C_i)}{p(x_1, \ldots, x_N)} \prod_{j=1}^{N} p(x_j|C_i)$$

Then the GNB will assign a class $\hat{y}$ to every item if $\hat{y}$ has the greatest conditional probability. In mathematical terms

$$\hat{y} = \max_{i \in \{1, \ldots, K\}} p(C_i) \prod_{j=1}^{N} p(x_j|C_i)$$

Support Vector Machine (SVM)

SVM can be used for both regression and binary classification problems and it is based on finding, in the feature-space, the best hyperplane subdividing training points (i.e. data) of one class from those belonging to the other one.[90] In particular, consider a training dataset of N items and with M input features. These items may be represented as $(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_N}, y_N)$, where $\overrightarrow{x_i}$ is the M -dimensional vector of input variables of i-th data item and y_i the corresponding binary label (0 or 1, for example). They may be considered as geometrical points in the M -dimensional feature space. The target of SVM algorithm is to find the *maximum margin* hyperplane: the hyperplane which is defined so that the distance between the hyperplane and the nearest point from either group is maximized. Figure A2 clearly explains the result of the SVM algorithm in a dataset with two input-features.

eXtreme Gradient Boosting (XGB)

XGB classifier, like Random Forest, is a model in the form of an ensemble of decision trees (also called *weak learners*), but, differently from RF, it is built in an iterative fashion and learns slowly.[91] In fact, trees in RF are trained on different bootstraped samples taken fron the training dataset, independently of each other; XGB, instead, does not involve bootstrap sampling but every tree is grown using information from previously grown trees, being fit on on a modified version of the training dataset. In particular, the main idea underpinning XGB is that, given the current model, we fit a decision tree to the residuals from the current model. Then, we add this new decision tree into the current model in order to update the residuals. By iteratively fitting trees to the residuals, we improve the current model in areas where it does not perform well. The functioning of XGB is exemplified in Figure A3

Acknowledgements

Research and results presented in this paper have been realized as part of the project "C-BAS - Customer Behaviour Analysis System", funded by POR PUGLIA FESR 2014-2020.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this

article: Regione Puglia; C-BAS - POR PUGLIA FESR 2014-2020.

ORCID iD

Francesco De Nicolo https://orcid.org/0000-0003-3036-8108

References

1. Chevalier JA and Mayzlin D. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* 2006; 43(3): 345–354.
2. Mudambi SM and Schuff D. Research note: what makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly* 2010; 34(1): 185–200.
3. Winer RS. New Communications Approaches in Marketing: Issues and Research Directions. *Journal of Interactive Marketing* 2009; 23(2): 108–117.
4. Ye Q, Law R, Gu B, et al. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior* 2011; 27(2): 634–639.
5. Zhu F and Zhang X. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing* 2010; 74(2): 133–148.
6. Duverger P. Curvilinear Effects of User-Generated Content on Hotels' Market Share. *Journal of Travel Research* 2013; 52(4): 465–478.
7. Liu Z and Park S. What Makes a Useful Online Review? Implication for Travel Product Websites. *Tourism Management* 2015; 47: 140–151.
8. Xie K, Zhang Z and Zhang Z. The Business Value of Online Consumer Reviews and Management Response to Hotel Performance. *International Journal of Hospitality Management* 2014; 43: 1–12.
9. Ayeh JK, Au N and Law R. "Do We Believe in TripAdvisor?" Examining Credibility Perceptions and Online Travelers' Attitude toward Using User-Generated Content. *Journal of Travel Research* 2013; 52(4): 437–452.
10. Ghose A, Ipeirotis PG and Li B. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 2012; 31(3): 493–520.
11. Liu B, Kim H and Pennington-Gray L. Responding to the bed bug crisis in social media. *International Journal of Hospitality Management* 2015; 47: 76–84.
12. Park S and Nicolau JL. Asymmetric effects of online consumer reviews. *Annals of Tourism Research* 2015; 50: 67–83.
13. Yang Y, Park S and Hu X. Electronic word of mouth and hotel performance: A meta-analysis. *Tourism Management* 2018; 67: 248–260.
14. Sparks BA and Browning V. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management* 2011; 32(6): 1310–1323.
15. Toral SL, Martínez-Torres MR and Gonzalez-Rodriguez MR. Identification of the unique attributes of tourist destinations from online reviews. *Journal of Travel Research* 2018; 57(7): 908–919.
16. Mauri A and Minazzi R. Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management* 2013; 34: 99–107.
17. O'Connor P. Managing a hotel's image on tripadvisor. *J Hosp Mark Manag* 2010; 19(7): 754–772.
18. Casaló LV, Flavián C, Guinalíu M, et al. Avoiding the dark side of positive online consumer reviews: Enhancing reviews' usefulness for high risk-averse travelers. *Journal of Business Research* 2015; 68(9): 1829–1835.
19. Pan Y and Zhang JQ. Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews. *Journal of Retailing* 2011; 87(4): 598–612.
20. Zhao X, Wang L, Guo X, et al. The influence of online reviews to online hotel booking intentions. *International Journal of Contemporary Hospitality Management* 2015; 27(6): 1343–1364.
21. Briggs S, Sutherland J and Drummond S. Are hotels serving quality? An exploratory study of service quality in the Scottish hotel sector. *Tourism Management* 2007; 28(4): 1006–1019.
22. Xiang Z, Schwartz Z, Gerdes JH, et al. What Can Big Data and Text Analytics Tell Us about Hotel Guest Experience and Satisfaction? *International Journal of Hospitality Management* 2015; 44: 120–130.
23. Kirilenko AP, Stepchenkova SO, Kim H, et al. Automated Sentiment Analysis in Tourism: Comparison of Approaches. *Journal of Travel Research* 2018; 57(8): 1012–1025.
24. Kwok L. Exploratory-triangulation design in mixed methods studies: A case of examining graduating seniors who meet hospitality recruiters' selection criteria. *Tourism and Hospitality Research* 2012; 12(3): 125–138.
25. Lu W and Stepchenkova S. Ecotourism experiences reported online: Classification of satisfaction attributes. *Tourism Management* 2012; 33(3): 702–712.
26. Wang S and Hung K. Customer perceptions of critical success factors for guest houses. *International Journal of Hospitality Management* 2015; 48: 92–101.
27. Jain PK, Yekun EA and Pamulaal R. Consumer recommendation prediction in online reviews using Cuckoo optimized machine learning models. *Comput Electr Eng* 2021: 95.
28. Bansal B and Srivastava S. Hybrid attribute based sentiment classification of online reviews for consumer intelligence. *Applied Intelligence* 2019; 49(1): 137–149.
29. Singh NK, Tomar DS and Sangaiah AK. Sentiment analysis: a review and comparative analysis over social media. *Journal of Ambient Intelligence and Humanized Computing* 2020; 11(1): 97–117.

30. Ho YC, Wu J and Tan Y. Disconfirmation effect on online rating behavior: A structural model. *Information Systems Research* 2017; 28(3): 626–642.
31. Wedel M and Kannan PK. Marketing analytics for data-rich environments. *Journal of Marketing* 2016; 80(6): 97–121.
32. Belle V and Papantonis I. Principles and practice of explainable machine learning. *Front Big Data* 2021; 4: 1–25.
33. Truc HL, Arcodia C, Abreu Novais M, et al. Proposing a systematic approach for integrating traditional research methods into machine learning in text analytics in tourism and hospitality. *Curr Issues Tour* 2021; 24(12): 1640–1655.
34. Zheng T, Wu F, Law R, et al. Identifying unreliable online hospitality reviews with biased user-given ratings, A deep learning forecasting approach. *Int J Hosp Manag* 2021: 92.
35. Filieri R, Raguseo E and Vitari C. Extremely negative ratings and online consumer review helpfulness: the moderating role of product quality signals. *Journal of Travel Research* 2021; 60(4): 699–717.
36. Lee M, Jeong M and Lee J. Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website. *International Journal of Contemporary Hospitality Management* 2017; 29(2): 762–783.
37. Wang X, Tang RT and Kim E. More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness? *International Journal of Hospitality Management* 2019; 77: 438–447.
38. Zhou Y, Yang S, Li Y, et al. Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. *Inf Processi Manag* 2021; 57: 102179.
39. Leung D, Lee HA and Law R. Adopting Web 2.0 technologies on chain and independent hotel websites: A case study of hotels in Hong Kong. In: Law R, Fuchs M and Ricci F (eds). *Information and communication technologies in tourism*. 1st ed.. Berlin: Springer, 2011, pp. 229–240.
40. Fang B, Ye Q, Kucukusta D, et al. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management* 2016; 52: 498–506.
41. Mudambi SM and Schuff D. Research Note: what makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly* 2010; 34(1): 185–200.
42. Floh A, Koller M and Zauner A. Taking a Deeper Look at Online Reviews, The Asymmetric Effect of Valence Intensity on Shopping Behaviour. *J Mark Manag* 2013; 29(5/6): 646–670.
43. Tang T, Fang E and Wang F. Is neutral really neutral? the effects of neutral user-generated content on product sales. *J Mark* 2014; 41(78): 41–58.
44. Kuan KKY, Hui KL, Hui P, et al. What Makes a Review Voted? An Empirical Investigation of Review Voting in Online Review Systems. *Journal of the Association for Information Systems* 2015; 16(1): 48–71.
45. Zhang X, Yu Y, Li H, et al. Sentimental interplay between structured and unstructured user-generated contents. *Online Information Review* 2016; 40(1): 119–145.
46. Radojevic T, Stanisic N and Stanic N. Ensuring positive feedback: Factors that influence customer satisfaction in the contemporary hospitality industry. *Tourism Management* 2015; 51: 13–21.
47. Forman C, Ghose A and Wiesenfeld B. Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. *Information Systems Research* 2008; 19(3): 291–313.
48. Viglia G, Minazzi R and Buhalis D. The influence of e-word-of-mouth on hotel occupancy rate. *International Journal of Contemporary Hospitality Management* 2016; 28(9): 2035–2051.
49. Filieri R and McLeay F. E-WOM and Accommodation. *Journal of Travel Research* 2014; 53(1): 44–57.
50. Filieri R. What makes online reviews helpful? a diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research* 2015; 68(6): 1261–1270.
51. Kim WG, Li JJ and Brymer RA. The impact of social media reviews on restaurant performance: The moderating role of excellence certificate. *International Journal of Hospitality Management* 2016; 55: 41–51.
52. Wang N, Liang H, Zhong W, et al. Resource Structuring or Capability Building? An Empirical Study of the Business Value of Information Technology. *Journal of Management Information Systems* 2012; 29(2): 325–367.
53. Kim W, Lim H and Brymer R. The Effectiveness of Managing Social Media on Hotel Performance. *International Journal of Hospitality Management* 2015; 44: 165–171.
54. Kubler RV, Colicev A and Pauwels KH Social media's impact on the consumer mindset, when to use which sentiment extraction tool? *J Interact Mark* 2020; 50: 136–155.
55. Netzer O, Lemaire A and Herzenstein M. When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. *Journal of Marketing Research* 2019; 56(6): 960–980.
56. Ziemer KS and Korkmaz G. Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis. *Computers in Human Behavior* 2017; 76: 122–127.
57. Aleti T, Pallant JJ, Tuan A, et al. Tweeting with the Stars: Automated Text Analysis of the Effect of Celebrity Social Media Communications on Consumer Word of Mouth. *Journal of Interactive Marketing* 2019; 48: 17–32.
58. Berger J, Humphreys A, Ludwig S, et al. Uniting the tribes, using text for marketing insight. *J Mark* 2021; 84(1): 1–25.
59. Humphreys A and Wang RJH. Automated text analysis for consumer research. *Journal of Consumer Research* 2018; 44(6): 1274–1306.

60. Hirsh JB and Peterson JB. Personality and language use in self-narratives. *Journal of Research in Personality* 2009; 43(3): 524–527.
61. Kosinski M, Stillwell D and Graepel T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 2013; 110(15): 5802–5805.
62. Ludwig S, de Ruyter K, Friedman M, et al. More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. *Journal of Marketing* 2013; 77: 87–103.
63. Shin S, Chung N, Xiang Z, et al. Assessing the Impact of Textual Content Concreteness on Helpfulness in Online Travel Reviews. *Journal of Travel Research* 2019; 58(4): 579–593.
64. Ren G and Hong T. Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Information Processing & Management* 2019; 56: 1425–1438.
65. González-Rodríguez MR, Martínez-Torres MR and Toral S. Post-visit and Pre-visit Tourist Destination Image through eWOM Sentiment Analysis and Perceived Helpfulness. *International Journal of Contemporary Hospitality Management* 2016; 28(11): 2609–2627.
66. Baccianella S, Esuli A and Sebastiani F. Multi-facet Rating of Product Reviews. In: Proceedings of 31st European Conference on Information Retrieval. Toulouse, France, 6 April – 9, 2008.
67. Hu YH and Chen K. Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management* 2016; 36(6): 929–944.
68. Marrese-Taylor E, Velásquez JD and Bravo-Marquez F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications* 2014; 41(17): 7764–7775.
69. Pang Y, Hao Q, Yuan Y, et al. Summarizing Tourist Destinations by Mining User-Generated Travelogues and Photos. *Computer Vision and Image Understanding* 2011; 115(3): 352–363.
70. Fu X, Tanyatanaboon M and Lehto XY. Conceptualizing transformative guest experience at retreat centers. *International Journal of Hospitality Management* 2015; 49: 83–92.
71. Wu MY, Wall G and Pearce PL. Shopping experiences: International tourists in Beijing's Silk Market. *Tourism Management* 2014; 41: 96–106.
72. Zhang JJ and Mao Z. Image of All Hotel Scales on Travel Blogs: Its Impact on Customer Loyalty. *Journal of Hospitality Marketing & Management* 2012; 21(2): 113–131.
73. Ghose A and Ipeirotis PG. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering* 2011; 23(10): 1498–1512.
74. Liu Y, Huang X, An A, et al. Modeling and predicting the helpfulness of online reviews. In: Proceedings of 8th IEEE international conference on data mining. Pisa, Italy, 15-19 December 2008.
75. Burkart N and Huber MF. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 2021; 70: 245–317.
76. Sharma R, Arpit K and Chuah C. Turning the blackbox in a glassbox, an explainable machine learning approach for understanding hospitality customer. *Int J Inf Manag Data Ins* 2021; 1(2): 100050.
77. Merrick L and Taly A. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In: Proceedings of 4th International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Dublin, Ireland, 25-28 August 2020.
78. Miguéns J, Baggio R and Costa C. Social media and tourism destinations, TripAdvisor case study. *Adv Tour Res* 2008; 26(28): 1–6
79. de la Peña MR, Núñez-Serrano JA, Turrión J, et al. Are innovations relevant for consumers in the hospitality industry? A hedonic approach for Cuban hotels. *Tourism Management* 2016; 55: 184–196.
80. Li G, Law R, Vu HQ, et al. Discovering the hotel selection preferences of Hong Kong inbound travelers using the Choquet Integral. *Tourism Management* 2013; 36: 321–330.
81. Chowdary K. *Fundamentals of artificial intelligence*. 1st ed.. New Delhi: Springer, 2020.
82. Fayyad UM, Piatetsky-Shapiro G and Uthurusamy R. Summary from the KDD-03 panel. *ACM SIGKDD Explorations Newsletter* 2003; 5(2): 191–196.
83. Sarica S and Luo J. Stopwords in technical language processing. *PLoS One* 2021; 16(8): e0254937.
84. Feldman R. Techniques and applications for sentiment analysis. *Communications of the ACM* 2013; 56(4): 82–89.
85. Hu M and Liu B. Mining and summarizing customer reviews. In: Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 22 August – 25 August 2004, pp. 168–177.
86. Aizawa A. An information-theoretic perspective of tf-idf measures. *Information Processing & Management* 2003; 39(1): 45–65.
87. Qaiser S and Ali R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications* 2018; 181(1): 25–29.
88. Breiman L, Friedman JH, Olshen RA, et al. *Classification and regression trees*. 1st ed. Abingdon: Routledge, 2017.
89. Paul A, Mukherjee DP, Das P, et al. Improved Random Forest for Classification. *IEEE Transactions on Image Processing* 2018; 27(8): 4012–4024.
90. Rameshbhai CJ and Paulose J. Opinion mining on newspaper headlines using SVM and NLP. *International Journal*

of Electrical and Computer Engineering (IJECE) 2019; 9(3): 2152–2163.
91. Ting SL, Ip WH and Tsang AH. Is Naive Bayes a good classifier for document classification? *Int J Soft Eng Applic* 2011; 5(3): 37–46.
92. Tharwat A. Classification assessment methods. *Applied Computing and Informatics* 2021; 17(1): 168–192.
93. Sundararajan M and Amir N, ICML. The many Shapley values for model explanation. In: Proceedings of 37th International Conference on Machine Learning. In: Virtual Event, 13-18 July 2020, pp. 9269–9278.
94. Mariani MM, Borghi M and Gretzel U. Online reviews: Differences by submission device. *Tourism Management* 2019; 70: 295–298.
95. Ganganwar V. An overview of classification algorithms for imbalanced datasets. *Int J Emerg Tech Adv Eng* 2012; 2(4): 42–47.
96. Blagus R and Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013; 14(106): 106–116.
97. Sáez JA, Luengo J and Herrera F. Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure. *Neurocomputing* 2016; 176: 26–35.
98. Zhu X and Wu X. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review* 2004; 22(3): 177–210.
99. Nazari Z, Nazari M, Sayed M, et al. Evaluation of class noise impact on performance of machine learning algorithms. *Int J Comput Sci Netw Secur* 2018; 18: 149.
100. Tausczik YR and Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 2010; 29(1): 24–54.
101. Dewi LC, Meiliana A and Chandra A. Social media web scraping using social media developers API and regex. *Procedia Computer Science* 2019; 157: 444–449.
102. Rathee N, Joshi N and Kaur J. Sentiment analysis using machine learning techniques on Python. In: Proceedings of the 2nd IEEE International Conference on Intelligent Computing and Control Systems. Madurai, India, 15 June 2019.
103. Iban MC and Sekertekin A. Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin provinces, Turkey. *Ecological Informatics* 2022; 69: 101647.
104. Cilli R, Elia M, D'Este M, et al. Explainable artificial intelligence (XAI) detects wildfire occurrence in the Mediterranean countries of Southern Europe. *Scientific Reports* 2022; 12(1): 16349–16411.
105. Amoroso N, Pomarico D, Fanizzi A, et al. A roadmap towards breast cancer therapies supported by explainable artificial intelligence. *Applied Sciences* 2021; 11(11): 4881.
106. Lombardi A, Diacono D, Amorosoal N at, et al. Explainable deep learning for personalized age prediction with brain morphology. *Front Neurosci* 2021; 15: 578.
107. Tjoa E and Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems* 2020; PP(11): 4793–4813.
108. Jiménez-Luna J, Grisoni F and Schneider G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* 2020; 2(10): 573–584.
109. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
110. Ho TK. A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. *Pattern Analysis & Applications* 2002; 5(2): 102–112.