‘You will like it!’ Using open data to predict tourists’ responses to a tourist attraction

Eleonora Pantano, Constantinos-Vasilios Priporas and Nikolaos Stylos

Abstract

The increasing amount of user-generated content spread via social networking services such as reviews, comments, and past experiences, has made a great deal of information available. Tourists can access this information to support their decision making process. This information is freely accessible online and generates so-called “open data”. While many studies have investigated the effect of online reviews on tourists’ decisions, none have directly investigated the extent to which open data analyses might predict tourists’ response to a certain destination. To this end, our study contributes to the process of predicting tourists’ future preferences via *Mathematica*™, software that analyzes a large set of the open data (i.e. tourists’ reviews) that is freely available on tripadvisor. This is devised by generating the classification function and the best model for predicting the destination tourists would potentially select. The implications for the tourist industry are discussed in terms of research and practice.

Keywords: open data, online reviews, tourism, travel propositions

1. Introduction

The recent advances in digital media technologies and environments, as well as the introduction and acceptance of sophisticated interactive software applications, have driven the digital evolution of marketing in the information society epoch (Garrigos-Simon, Lapiedra Alcamí & Barberá Ribera, 2012; Mekonnen, 2016). Digital social media has played a key role in this recently established sub-field of marketing and its rapid spread has transformed how information is accessed and shared (Di Noia, Mirizzi, Ostuni, Romito, & Zanker, 2012; Pantano, 2014). In particular, the impact of social networking sites (SNSs) on word-of mouth communications and decision making processes has been well reported (Chu & Kim, 2011; Fotiadis & Stylos, 2016; See-To & Ho, 2014). Digital marketers realize that to successfully attract and leverage the interest of SNS users, they need to increase the utility of social networks by offering value added services (Diffley, Kearns, Bennett, & Kawalek, 2011). Thus, SNSs are now expanding their capabilities by offering a diverse portfolio of build-in applications (apps) to meet social media users' needs for novel experiences (Tung, Jai, & Davis Burns, 2014); namely, customized topic-specific virtual spaces to better support user-generated content (UGC) (e.g. Facebook apps, YouTube), including reviews, comments on past experiences and recommendations for future purchases (Turban, King, Lee, Liang & Turban, 2015). As researchers note, online reviews based on SNS users' profiles and established preferences are integral to formulating future preferences and affecting consumer purchases (Baka, 2016; Chevalier & Mayzlin, 2006). The premise is that behavior is influenced not only by individuals' beliefs, feelings, impressions, and behavioral norms, but also by recommendations and prior experiences stemming from the social environment, which in turn produce attitudes and intentions (Cheng & Huang, 2013; Tsai & Bagozzi, 2014; White, 2005). In fact, the more the product online review features available to consumers, the higher the likelihood for sales of related items within the product category (Chevalier & Mayzlin, 2006).

Similarly, in a travel and tourism context, tourists' recommendations via tripadvisor, Yelp etc. influence other travelers' decisions about many different aspects of their trips, e.g. selection of a tourist destination, accommodation and attractions to visit (Hudson, 2014; Pantano & Di Pietro, 2013; Xiang, Magnini, & Fesenmaier, 2015; Filieri, Alguezaui & McLeay, 2015). Notwithstanding the fact that some researchers have indicated that many reviews are fake, or overly positive or negative, consumers perceive online reviews as more trustworthy than content provided by official destination websites (Fotis, Buhalis & Rossides, 2012). Drawing on a huge amount of UGC, marketers make systematic efforts to exploit as much open data as possible to support digital marketing effectiveness. These efforts could potentially improve online sales and the profitability of e-travel services (e.g., accommodation, transportation, restaurants, entertainment, sightseeing and tourism destination information) (Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012; Nguyen & Cao, 2015).

Nevertheless, up until now, most research in UGC and most online reviews have underlined the importance of analyzing ratings to increase the likelihood of travelers' having enjoyable trips (e.g. Fang, Ye, Kucukusta & Law, 2016; Neidhardt, Seyfang, Schuster, & Werthner, 2015; Phillips, Zigan, Silva & Schegg, 2015; Sotiriadis & van Zyl, 2013; Zhang et al., 2016), though only a few studies explicitly examine the impact of reviews on SNS users' future choices (i.e. Ayeh, Au & Law, 2013; Jalilvand, Samiei, Dini & Manzari 2012; Pantano & Di Pietro, 2013; Sparks, Perkins & Buckley, 2013). These studies focus on the readability, credibility and helpfulness of online reviews, however they do not explore the extent to which recommendations maybe perceived as useful to other travelers willing to travel to the same destinations (Schuckert, Liu & Law, 2015). Moreover, they do not investigate ways of using this data to improve traveler review sites' consultation capabilities to the benefit of individuals, hospitality businesses and tourist destinations at large.

Taken together, this study seeks to examine the extent to which open data analysis may apply to the tourists' process of selecting tourist destinations and/or services. In particular, we attempt to predict travelers' attitudes toward a tourist attraction by transforming large amounts of open data into value propositions. In doing so, we implement the random decision forest algorithm approach (Coussement & De Bock, 2013; Xie, Ngai & Ying, 2009) drawing on data available on a popular travelers' review site.

Given that few studies have explored the potential of open data to serve as means of providing opt vacations-related automated database-driven recommendations (Buhalis & Law, 2008; Jannach, Zanker, & Fuchs, 2014; Gretzel, Sigala, Xiang, & Koo, 2015), the objectives of the study are twofold. First, it aims to investigate the potential benefits of using open data sources to form appropriate future travel propositions, thus moving one step forward from the standard method of investigating the influence of perceived value as well as the reliability of online reviews on formulating intentions (Fang, Ye, Kucukusta, & Law, 2016; Korfiatis et al., 2012; Lee, Law, & Murphy, 2011; Liu & Park, 2015; Sparks et al., 2013). Secondly, it seeks to highlight the effectiveness of leveraging a limited bulk of open data, as an alternative to big data sets, in terms of providing useful outputs.

From a theoretical point of view, this study draws attention to the huge potential of using online open data sources to influence tourists' attitudes and behaviors. Practically, we propose a computational tool that can greatly contribute to the effective positioning of hospitality organizations and tourist destinations.

2. Theoretical Background

2.1. Open data

Open data has been defined by the Open Knowledge Foundation (http://okfn.org/) in 2005 as "data that can be freely used, shared and built on by anyone, anywhere, for any purpose".

Maccani, Donnellan, & Helfert (2015) point out that there are 3 principles behind this definition: (1) availability and access (people can get the data); (2) re-use and redistribution (people can re-use and share the data); (3) universal participation (anyone can use the data). Furthermore, the volume of the information released through open data platforms is huge (Ojha, Jovanovic, & Giunchiglia, 2015; Wu, Liu, Chu, Chu, & Yu, 2014), it is based on a wealth of information and enables enhanced knowledge creation (Theocharis & Tsihritzis, 2013).

Kitchin (2014) asserts that the focus of open data could be any type of socio-economic or business phenomena but that in general, the emphasis to date has been on opening up data that has a high policy and commercial re-use value, such as economic, transport and spatial data. Today, open data are mostly provided by public and services providers (organizations, institutions, and enterprises) while the potential of open data for business development is still mostly unexplored (Pesonen & Lampi, n.d). For example, governments are trying to exploit open data to support the development of better services for citizens (Chan, 2013; Hielkema & Hongisto, 2013). Processing open data is recognized as a potentially powerful alternative to analyzing data collected via surveys (Gurstein, 2011). In specific, the use of open data is being increasingly acknowledged as a means of supporting knowledge management in various contemporary business and technological applications such as, smart cities (Inayatullah, 2011; Ojo, Curry, & Zeleti, 2015). Cities have been the first to be involved in processing open data in various applications (Longhi, Titz & Viallis, 2014), such as the management of their tourist destination products (Buhalis & Amaranggana, 2013; Mariani, Buhalis, Longhi, & Vitouladiti, 2014), recognizing them as a key component of their smart city strategy (Marine-Roid & Clave, 2015).

2.2 Open data in Tourism

Tourism is by nature an industry in which marketing communications strongly depend on data exchange (Mack, Blose & Pan, 2008). In today's rapidly changing world, various forms of data related to tourism activities and services are produced and utilized across a range of online applications (Buhalis & Law, 2008). This is primarily the outcome of the increasing ability to digitize growing volumes of data, and the development of open-sources and open data policies (Sabou, Brașoveanu, & Önder, 2015; Soualah-Alila, Coustaty, Rempulski, & Doucet, 2016). For tourist destinations there are significant opportunities to use open data to develop cultural sights, transportation, marketing and the environment (Wiggins & Crowston, 2011). As people have increasingly focused on the quality of the tourist experience, the demand for open data in tourism and hospitality research has becomes intense (Wu et al., 2014). A growing amount of tourism-related open data is now available on the platform in XML, CSV, or JSON format (Wu et al., 2014). According to Longhi, Titz, and Viallis, (2014) tourism is the first industry to be concerned with open data. Open data can facilitate local authorities in their planning processes (e.g., advertising) and in adapting to the needs of tourists. Mobile technologies have shifted the focus of the tourist industry from a focus on mass tourism related practices to a focus on "one to one" marketing practices (i.e., real mobile, just in time information about attractions, catering facilities and transportation alternatives), resulting in communication plans that could prove much more effective (Longhi, Titz, & Viallis, 2014). In terms of mobile technology, having reliable real time information always available is crucial in terms of enabling the tourists to find their way. The information primarily interests consumers-tourists, based on this information these tourists can find restaurants near their position and can get information on monuments and sightseeing in the areas they are visiting (Longhi, Titz, & Viallis, 2014). Mariani et al. (2014) assert that the development of the disruptive technologies under 'mobiquity'–a new term emerging from the mobility and ubiquity of smartphone market penetration– combined with the free access open data revolution are profoundly changing the whole tourist industry, bringing along new technologies, new knowledge bases, and new roles for the different stakeholders.

Although research has started soliciting new studies on the adoption of new technologies for smart tourism (Gretzel, Sigala, Xiang & Koo, 2015; Rossetti, Stella, & Zanker, 2016), the benefits emerging from the open data use are still under investigated by current literature in tourism (Fermoso, Mateos, Beato, & Berjon, 2015). In fact, there is an increase in online communities that focus on travel discussions (i.e. tripadvisor and social networks like Facebook and Twitter). These new means for tourists to both obtain information and plan travel force tourism managers to create better tailored and more efficient marketing approaches, as well as develop new models for hospitality (Pantano & Di Pietro, 2013). Efficient analyses of the big data sets might support the development of these new approaches (Pantano & Di Blasi, 2015).

As far as tripadvisor is concerned, it can be identified as a significant source of open data given the figures and reviews on attractions/destinations. For example, in 2015, tripadvisor reached 320 million reviews and had 6.2 million opinions on places to stay, to eat and on things to do – including 995,000 hotels and forms of accommodation, 770,000 vacation rentals, 3.8 million restaurants and 625,000 attractions in 125,000 destinations throughout the world (tripadvisor, 2016).

2.3 Vacation decision making

The vacation decision-making process is much more complex than the decision-making process for tangible goods (Park, Nicolau, & Fesenmaier, 2013). The process depends on whether a person goes on a vacation alone, as a couple or as a family with children, and on the planning process (Decrop, 2006). According to Hyde and Decrop (2011), the process also differs for different types of vacation trips (i.e., short, long, annual family vacation) because different trips include different levels of involvement, different time spent planning and a different number of decisions that must be made before travel. Swarbrooke and Hormer (1999) stated that vacation planning is a high-involvement process, because many people spend large amounts of money on an intangible product with a low level of security and great social implications.

In fact, selecting the most suitable choice of destination, travel mode and accommodation is a time and effort consuming process (Hsu et al., 2012; Li et al., 2015). This selection may be made on the basis of expectations, preferences, purposes, previous accommodation experience, costs, transport mode, etc. (Li et al., 2015), or even on the basis of others' past experience, word-of-mouth (WOM) and electronic word of mouth or word-of-mouse (eWOM) (Law, Buhalis, & Cobanoglu, 2014). When WOM is mediated through electronic means, internet users pass information to others via social networks, instant messages, news feeds and travel review sites that users can freely access (Liu & Park, 2015). Actually, the influence of the Internet has been significantly transforming the tourist industry in a number of ways: it has become one of the most efficient means of reaching new tourist markets and foster revisiting the same destinations (Pan, Xiang, Law, & Fesenmaier, 2011) and is now the leading information source for tourists due to the many online tourism communities it supports (Pantano & Di Pietro, 2013; Liu & Park, 2015). Although the abundance of online tourists' reviews makes information retrieval easier, it could overexpose tourists to a huge bulk of information thus making it harder for them to select the most useful information (Zhang, Zhang, & Yang, 2016). Individuals are able to process only a part of the available information and only according to certain personal criteria (Johnson, Bellman, & Lohse, 2003; Zhang et al., 2016). This implies that they can only process the information which is included in their selection criteria, which might exclude a large amount of information (Zhang et al., 2016). For this reason, current developments in information and communication technologies are looking at new ways to support consumers in their search for useful information in finalizing holidays planning while avoiding the information overload (McCabe, Li & Chen, 2015; Zhang et al., 2016). The tourist industry has benefited from employing intelligent systems, which are new generation information systems that can provide more applicable and better tailored information, advanced decision support systems, and, ultimately, improved tourism experiences. Examples of intelligent systems employed in this industry are recommender systems, context-aware technologies, autonomous agents searching and mining web resources (Gretzel, 2011). Specifically, recommender systems make use of sophisticated technology that filters out personal information, which could be used for pinpointing interesting items or activities to tourists according to their preferences (Al-Hassan, Lu, & Lu, 2015). Recommender systems' strength relies on their ability to automatically learn tourists' preferences by analyzing their behavioral responses (Batet, Moreno, Sánchez, Isern, & Valls, 2012; Borras, Moreno, & Valls, 2014; Noguera, Barranco, Segura, & Martinez, 2012). Borras et al. (2014) postulate that these systems can facilitate tourists' selection process by dynamically recommending sights of interest based on real time data (i.e. location and context related information). Same authors argue that this setting fosters the development of intelligent autonomous agents, which offer some important benefits through their advanced abilities; first, an enhanced analysis of tourist behavior; second, the ability to provide opt and proactive visit-related recommendations based on automatic learning of tourists preferences and needs. Consequently, smart technologies have the ability to create, develop, manage and deliver intelligent tourism experiences, thus developing an emerging trend in tourism which is characterized by intensive information sharing, relationships building and value co-creation among tourists, tourism managers, organizations, etc. (Prebensen, Kim & Uysal, 2016). In this context, processing and transferring large volume of tourism-relevant data is of upmost importance (Gretzel et al., 2015) and as a consequence, problems related to information overload could be overcome by future progress in technology. Technology would provide a more enhanced service for tourists in terms of ubiquity, connection, context-awareness, and the capacity to process more data (Borras et al., 2014; Gavalas, Konstantopoulos, Mastakas, & Pantziou, 2014).

3. Predictive Models

Open data analyses might support tourism managers in predicting tourists' judgements about a certain tourist attraction. To achieve this goal, it is necessary to introduce predictive models that support information selection within a huge amount of data. A predictive model is a mathematical tool able to produce a mathematical function between a target or "dependent" variable and other features or "independent" variables, aiming at predicting future values of the target variable based on past values of the features, starting from a classification function (Pantano & Di Blasi, 2015). In other words, it allows the prediction of future elements on the basis of past ones. Literature proposes several models in this direction, such as decision trees (DT) (Rokach & Maimon, 2008), regression models (RM) (Freedman, 2005), neural networks (NN) (Rojas, 1996), k-nearest neighbour (k-NN) (Shakhnarovich, Darrell, & Indyk, 2005), support vector machines (SVM) (Cortes & Vapnik, 1995), logistic regression (Issa & Kogan, 2014), Markov series (Ghahramani & Jordam, 1997), and random forest (Prasad, Iverson, & Liaw, 2006; Archer & Kimes, 2008). In particular, DT is usually employed for categorical datasets (e.g. not numerical data); whereas RM, NN, k-NN and SVM are high performing when numerical datasets are available. Pantano and Di Blasi (2015), point out that choosing a computational method for prediction purposes should take into account the analysis of referring context, the nature of data (if numerical, strings, mixed, etc.), and relevant computational cost (incorporating the speed of program execution).

4. Methodology

4.1 Dataset

The present study, which is exploratory in nature, aims to understand the extent to which a tourist would express a positive or a negative judgment about a certain attraction, based on their freely available online profile. To achieve this goal, we used information available on tripadvisor.

tripadvisor provides for each attraction/destination/restaurant/hotels etc. a set of users' reviews marked with stars, from 0 stars (terrible) to 5 stars (excellent). In this case, data collection and analysis focused on those reviews that use only the extremes of the five-point tripadvisor evaluation scale, i.e. 0 and 5 in rating. This approach would facilitate clearer predictions of tourists' future choices. Appendix A provides the technical explanation of the algorithm employed.

Moreover, each reviewer can develop a profile including interest in the following 18 topics (multiple preferences are allowed): foodie, shopping fanatic, history buff, urban explorer, nightlife seeker, peace and quiet seeker, art and architecture lover, vegetarian, thrifty traveller, eco-tourist, backpacker, luxury traveller, beach goer, trendsetter, thrill seeker, family holiday maker, nature lover, behaving like a local. These elements are represented in binary mode: 1 if they indicated the interest in the specific topic, 0 if otherwise.

Five-hundred online reviewers of the Empire State Building were randomly selected from the tripadvisor database, half of the reviewers gave the experience a rating of 5 and the other half gave it a rating of 0. Drawing upon the data available, a database was formed, assigning the value = 1 where a user is interested in a specific topic/characteristic (e.g. foodie), and 0 where otherwise. Thus, a dataset emerged as a result of randomly considering the reviews posted between December 2015 and January 2016.

The key question of the analysis undertaken is: 'Does the user express a positive or negative evaluation?' To facilitate handling of the data, recoding of the variables took place with 0 regarded as negative (0 stars), and 1 as a positive answer (5 stars). From a mathematical point of view, an I set of data is developed comprising a 18-bit string (which means that the set is cardinality of the set S is 18, in other words $n= 2^{18}$). If we consider S= {0,1} the set of possible values of the tourist attraction (which means 1 if they give 5 stars, 0 if they give 0 stars), our function will be:

$$f: I \rightarrow S$$

$$(x_1, x_2, \cdots, x_{18}) \in I \rightarrow s \in S$$

Where x_i, $\forall i = 1, \cdots, 18$.

The function f assigns to a data configuration, a value of 0 or 1. The goal is, thus, to find a function describing these relationships, which are usually described as a rules table (Figure 1).

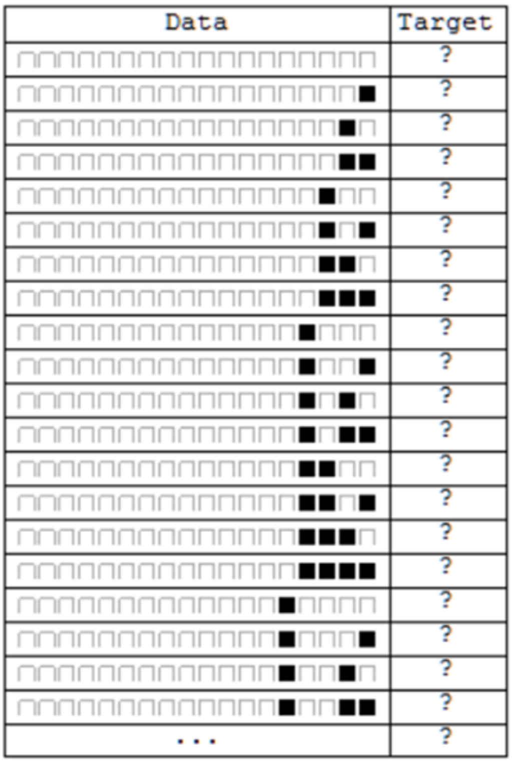

Figure 1: Rules table describing the function f for our problem

The prediction system is designed to deliver outputs; in case the 2^{18} values of the target would be defined. Figure 2 represents how the table of rules appears for our set, just considering a sample of 10 configurations as example (where 0 is white and 1 is black).

Data	Target
■□■■□■□□■□□■□□□□■■	■
■□□■□■□□□□□□■□□■□□	■
□□□□□□□□□□□□□□□□□■	□
■□□□□■□□□□□■■□□■□□	□
■□□□□■□□■□□■■□■■■□	■
□□□□□□□□□□□□□□□□□□	■
□■□■□□□□□□□□□□□□□■	□
□□□□□■□□□□□■□□□□□■	□
■■■■■■□□□□□□■□□■□■	□
■□■■□□■□□□□□□□□□■□	■

Figure 2: Example of rules table describing the function f for our problem for a sample of 10 configurations.

4.2 Procedure

Two hundred and fifty items of review data were used to create the training set and the entire sample and to determine the successful cases for the classifying function. To achieve this goal, we developed and applied the following code (instructions for *Mathematica*^{TM}).

The code provided in Appendix B draws the system input from the excel file and includes some specific columns in our dataset. Then, it builds the set of rules for the training set, linking the input data with the expected results. Then, it builds for x times the prediction function, and compares the results applying the result emerging from the prediction function with the target value of all the data in the sample. In particular, we conducted two different experiments with a different number of classifying functions (50 and 200 respectively).

Experiment 1

We built 50 classifying functions. Figure 3 shows the results by graphically describing the sum of percentage values of the cases in which the system identified the right value of 0 and 1.

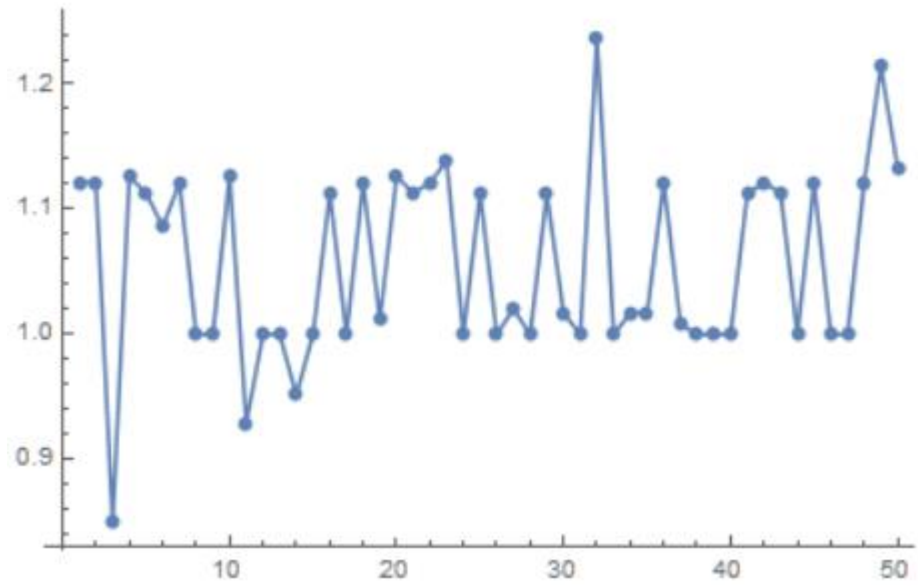

Figure 3: Results of the experiment 1 based on the building of 50 classifying functions.

From the figure, it is possible to note that only in 2 cases the sum overcome the value of 1.2. In particular, in the 32nd case, the percentage of success of identification of the right value of 1 is 0.66, while the percentage of success of 0 is 0.57, which lead to a total value of 1.237. To achieve this result the Random Forest method was selected and employed by *Mathematica* software.

Experiment 2

We built 200 classifying functions. Emerging results are summarized in figure 4.

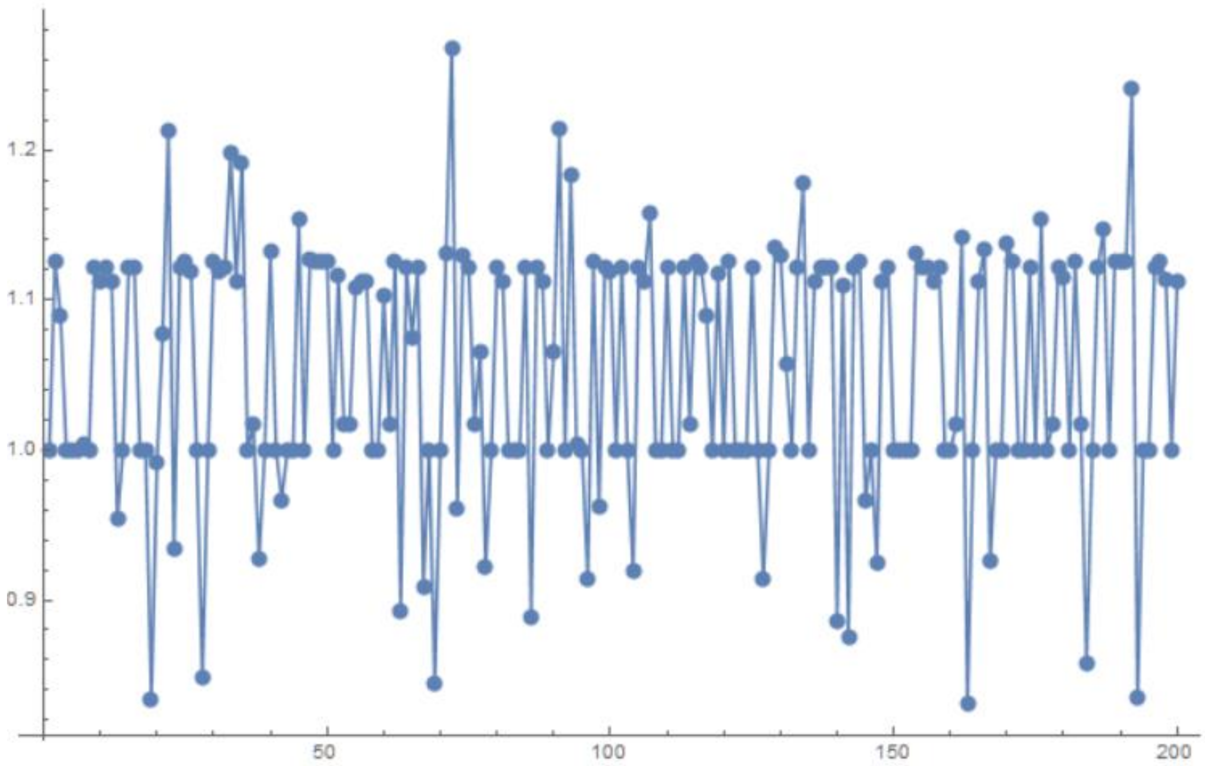

Figure 4: Results of the experiment 2 based on the building of 200 classifying functions.

Like the preceding case, the values rarely exceed the value of 1.2. In contrast to experiment 1, the best prediction appears on the 72nd case, where the percentage of successful identification of 1 is 0.69 and the successful identification of 0 is 0.58, for a global value of 1.268; while the reliability of the of our results is equal to 0.728. Similarly to experiment 1, to achieve this result the Random Forest method was selected and employed by *Mathematica* software. Therefore, it is possible to build a new rules table comparing a random sample of 20 cases and the prediction of our classifying function (Figure 5).

Data	Target	Prevision
■□■■□■□□■□□■□□□□■■	■	■
■□□■□■□□□□□□■□□■□□	■	■
□□□□□□□□□□□□□□□□□■	□	■
■□□□□■□□□□□■■□□■□□	□	□
■□□□□■□□■□□■■□■■■□	■	■
□□□□□□□□□□□□□□□□□□	■	■
□■□■□□□□□□□□□□□□□■	□	■
□□□□□■□□□□□■□□□□□■	□	■
■■■■■■□□□□□□■□□■□■	□	□
■□■■□□■□□□□□□□□□■□	■	■
□□□■□□□□■■□□□□□□□□	■	□
■■■■□□■□■□□□□■■□■■	□	□
□□□□□□■□■□□■□□□□□□	■	□
□□■■□□□□□□□■□□□□□□	■	■
□□□□□□□□□□□□□□■□■□	■	■
□□□□□□□□□□□■□□□□□□	□	■
■□□■□□□□□□□□□□□□□■	□	■
□□□□□□□□□□□□□□□■□□	□	□
■□■□□■□□□■□□□□□□□□	□	□
■□□■□■□□□□□■□□■□□■	■	■

Figure 5: Rules table emerging from the comparison a random sample of 20 cases and the prediction of our classifying function

5. Discussion and Conclusion

Although continuous progress in technology provides new tools to support tourist decision-making (Borras et al., 2014; Gretzel, 2011; Gretzel et al., 2015; Pantano & Di Pietro, 2013), it also creates new challenges due to the large availability of open (free) data and its limited usage by tourist destinations and hospitality managers. In fact, the use of open data for tourism purposes is still limited (Longhi et al., 2014; Soualah-Alila et al., 2016). Tourists are engaged in tourist destinations and are well connected, well informed and active critics (Marchiori & Cantoni, 2015). This research represents one of the first attempts to explore the usage of open data to predict tourists' responses towards a certain destination, in terms of ratings.

Findings show the extent to which our system is able to identify a trend in consumers' appreciation of a certain tourist destination/attraction, by considering a specific attraction reviewed in tripadvisor and a random sample of data consisting of 250 users who considered it terrible (0 stars) and 250 who considered it excellent (5 stars, corresponding to 1 in our data set). Therefore, tourism managers might consider adopting open data analysis to make better predictions about the attractiveness of a certain destination (including hotel, restaurant, monuments, museums, etc.). Moreover, applying this system to a sample at a certain time, and running it again after some changes (i.e. after changing the marketing strategy, renovating the place, adding new services, etc.) could make it possible to evaluate the effectiveness of the adopted strategies. In fact, the use of these analyses would allow managers to better reach the target audience and create tourism products that are better able to meet tourist's needs. This element would provide strong support for better planning and the development of more customized marketing strategies.

Our research shows the extent to which the current increasing and widespread use of online destination reviews (including ranking and ratings) is an opportunity for entrepreneurs, managers and destination marketers to acquire useful insights about the attraction/destination (Marine-Roig & Clave, 2015) .

An important implication of our findings is that destination marketers can evaluate tourists' responses to a certain destination in advance, and can potentially influence the final destination choice by improving marketing strategies accordingly. Destinations might use these analyses to predict the weaknesses or strengths of their image based on the analysis of tourists' open data, which can be freely and quickly accessed online.

Despite the new perspective provided by the present paper, there are some limitations that should be taken into account. The first one relates to the reliability of the adopted proposed framework; a large number of estimates (262,144) were identified, although the training session ran on only 250 items of review data (<0.1%). Hence, the quality of results may be sensitive to the size of the initial data set. Consequently, this proposed framework should be effectively tested using a larger sample of reviews, focusing on the development ad hoc programs for rapidly collecting and converting open data. Second, this study focused on one tourism attraction located at a specific tourist destination to serve as a case for driving data analysis. Thus, future research could test the applicability of data analysis and compare outputs from attractions and tourist destinations across the globe. Third, data analysis was based on the reviews posted within a timeframe of two months. A study incorporating data from a longer time period could improve the predictability of the proposed technique. Moreover, a longitudinal study could offer a better validation of the travel propositions made over time.

References

Al-Hassan, M., Lu, H., & Lu, J. (2015). A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system. *Decision Support Systems*, *72*, 97-109.

Archer K. J., & Kimes R.V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*(4), pp. 2249-2260.

Ayeh, J. K., Au, N., & Law, R. (2013). Predicting the intention to use consumer-generated media for travel planning. *Tourism Management*, *35*, 132-143.

Baka, V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management*, *53*, 148-162.

Batet, M., Moreno, A., Sánchez, D., Isern, D., & Valls, A. (2012). Turist@: Agent-based personalised recommendation of tourist activities. *Expert Systems with Applications*, *39*(8), 7319-7329.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. (2013). Digital business strategy: toward a next generation of insights. *MIS Quarterly*, *37*(2), 471-482.

Borras, J., Moreno, A., & Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, *41*(15), 7370-7389.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory*, (pp.144-152). ACM.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet - The state of eTourism research. *Tourism Management*, *29*(4), 609-623.

Buhalis, D., & Amaranggana, A. (2013). Smart tourism destinations. In *Information and communication technologies in tourism 2014* (pp. 553-564). Springer International Publishing.

Chan, C.M.L. (2013). From open data to open innovation strategie: creating e-services using open government data. In Proceedings of the 46th annual Hawaii international conference on system sciences, pp. 1890-1899.

Cheng, H. H., & Huang, S. W. (2013). Exploring antecedents and consequence of online group-buying intention: An extended perspective on theory of planned behavior. *International Journal of Information Management*, *33*(1), 185-198.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345-354.

Chu, S. C., & Kim, Y. (2011). Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International journal of Advertising*, *30*(1), 47-75.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*(9), 1629-1636.

Diffley, S., Kearns, J., Bennett, W., & Kawalek, P. (2011). Consumer behaviour in social networking sites: implications for marketers. *Irish Journal of Management*, *30*(2), 47-65.

Di Noia, T., Mirizzi, R., Ostuni, V. C., Romito, D., & Zanker, M. (2012). Linked open data to support content-based recommender systems. In *Proceedings of the 8th international conference on semantic systems* (pp. 1-8). ACM.

Duan, K. B., & Keerthi, S. S. (2005). Which is the best multiclass SVM method? An empirical study. Multiple classifier systems. *Lecture Notes on Computer Science*, 3541, 278–285.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics. *Tourism Management*, *52*, 498-506.

Fermoso, A.M., Mateos, M., Beato, M.E., & Berjon, R. (2015). Open linked data and mobile devices as e-tourism tools. A practical approach to collaborative e-learning. *Computers in Human Behavior, 51*, 618–626.

Filieri, R., Alguezaui, S., & McLeay, F. (2015). Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. *Tourism Management*, *51*, 174-185.

Fotiadis, A. K., & Stylos, N. (2016). The effects of online social networking on retail consumer dynamics in the attractions industry: The case of 'E-da'theme park, Taiwan. *Technological Forecasting and Social Change*, http://dx.doi.org/10.1016/j.techfore.2016.06.033.

Fotis, J., Buhalis, D., & Rossides, N. (2012). Social media use and impact during the holiday travel planning process. In M. Fuchs, F. Ricci, & L. Cantoni (Eds.), *Information and communication technologies in tourism 2012* (pp. 13-24). Vienna: Springer.

Freedman, D. A. (2005). *Statistical models: Theory and practice*. New York: Cambridge University Press.

Gavalas, D., Konstantopoulos, C., Mastakas, K., & Pantziou, G. (2014). Mobile recommender systems in tourism. *Journal of Network and Computer Applications*, *39*, 319-333.

Garrigos-Simon, F. J., Lapiedra Alcamí, R., & Barberá Ribera, T. (2012). Social networks and Web 3.0: their impact on the management and marketing of organizations. *Management Decision*, *50*(10), 1880-1890.

Ghahramani Z., & Jordan M.I. (1997). Factorial hidden Markov models. *Machine Learning, 29* (2), 245-273.

Gretzel, U. (2011). Intelligent systems in tourism: a social science perspective. Annals of *Tourism Research, 38*(3), 757-779.

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets*, *25*(3), 179-188.

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone?. *First Monday*, *16*(2). Retrieved form http://firstmonday.org/htbin/cgiwrap/ bin/ojs/index.php/ fm/article/view/3316/2764

Hielkama, H., & Hongisto, P. (2013). Developing the Helsinki smart city: The role of competitions for open data applications. *Journal of the Knowledge Economy*, *4*(2), 190-204.

Hsu, F.-M., Lin, Y.-T., & Ho T.K. (2012). Design and implementation of an intelligent recommendation system for tourist attractions: the integration of EBM model, Bayesian network and Google Maps. *Expert Systems with Applications*, *39*(3), 3257-3264.

Hudson, S. (2014). Challenges of tourism marketing in the digital, global economy. In S. McCabe (Ed.), *The Routledge handbook of tourism marketing* (pp.475-490). London: Routledge.

Inayatullah, S. (2011). City futures in transformation: Emerging issues and case studies. *Futures, 43*(7), 654–661.

Hussein, I., & Kogan, A. (2014). A predictive ordered logistic regression model as a tool for quality review of control risk assessments. *Journal of Information Systems*, *28*(2,) 209-229.

Jalilvand, M. R., & Samiei, N., Dini, B., & Manzari, P. Y. (2012). Examining the structural relationships of electronic word of mouth, destination image, tourist attitude toward destination and travel intention: An integrated approach. *Journal of Destination Marketing & Management*, *1*(1), 134-143.

Jannach, D., Zanker, M., & Fuchs, M. (2014). Leveraging multi-criteria customer feedback for satisfaction analysis and improved recommendations. *Information Technology & Tourism*, *14*(2), 119-149.

Johnson, E.J., Bellman, S., & Lohse, G.L. (2003). Cognitive lock-in and the power law of practice. *Journal of Marketing*, *67*(2), 62-75.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications, 11*(3), 205-217.

Law, R., Buhalis, D., & Cobanoglu, C. (2014). Progress on information and communication technologies in hospitality and tourism. *International Journal of Contemporary Hospitality Management*, *26*(5), 727-750.

Lee, H. A., Law, R., & Murphy, J. (2011). Helpful reviewers in TripAdvisor, an online travel community. *Journal of Travel & Tourism Marketing*, *28*(7), 675-688.

Li, G., Law, R., Vu, H.Q., Rong, J., & Zhao, X. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management*, *46*, 311-321.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, *47*, 140-151.

Longhi, C., Titz, J. B., & Viallis, L. (2014). Open data: Challenges and opportunities for the tourism industry. In M. Mariani, R. Baggio, D. Buhalis, C. Longhi (Eds.), *Tourism management, marketing, and development*. Volume I: The importance of networks and ICTs. (pp. 57-76). New York: Palgrave Macmillan.

Maccani, G., Donnellan, B., & Helfert, M. (2015). Open data diffusion for service Innovation: An inductive case study on cultural open data services. *PACIS 2015 Proceedings*, Paper 173. Retrieved from http://aisel.aisnet.org/pacis2015/173.

Mack, R. W., Blose, J. E., & Pan, B. (2008). Believe it or not: Credibility of blogs in tourism. *Journal of Vacation Marketing*, *14*(2), 133-144.

Marchiori, E., & Cantoni, L. (2015). The role of prior experience in the perception of a tourism destination in user-generated content. *Journal of Destination Marketing & Management*, *4*(3), 194-201.

Mariani, M. M., Buhalis, D., Longhi, C., & Vitouladiti, O. (2014). Managing change in tourism destinations: Key issues and current trends. *Journal of Destination Marketing & Management*, *2*(4), 269-272.

Marine-Roig, E., & Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management*, *4*(3), 162-172.

McCabe, S., Li, C., & Chen, Z. (2015). Time for a radical reappraisal of tourist decision making? Toward a new conceptual model. *Journal of Travel Research,* doi: 0047287515592973.

Mekonnen, A. (2016). Digital marketing strategy for affinity marketing: Utilising the new marketing arena. In Ozuem, W., & Bowen, G (Eds.), *Competitive social media marketing strategies* (pp. 1-19). Hershey PA: Business Science Reference-IGI Global.

Neidhardt, J., Seyfang, L., Schuster, R., & Werthner, H. (2015). A picture-based approach to recommender systems. *Information Technology & Tourism*, *15*(1), 49-69.

Nguyen, H.T.H., & Cao, J. (2015). Trustworthy answers for top-k queries on uncertain Big Data in decision making. *Information Sciences*, *318*, 73-90.

Noguera, J.M., Barranco, M., Segura, R. & Martinez, L. (2012). A mobile 3D-GIS hybrid recommender system for tourism. *Information Sciences*, *215*, 37-52.

Ojha, S. R., Jovanovic, M., & Giunchiglia, F. (2015). Entity-centric visualization of open data. In Human-Computer Interaction–INTERACT 2015 (pp. 149-166). Springer International Publishing.

Ojo, A., Curry, E., & Zeleti, F. A. (2015). A tale of open data innovations in five smart cities. In *System sciences (HICSS), 2015 48th Hawaii international conference on systems science* (pp. 2326-2335). IEEE.

Pan, B., Xiang Z., Law, R., & Fesenmaier, D.R. (2011). The dymanics of search engine marketing for tourist destinations. *Journal of Travel Research*, *50*(4), 365-377.

Pantano, E, & Di Blasi, G. (2015). Big data analysis solutions for supporting tourist's decision making. In K. Andriotis (Ed.), *Proceedings of the international conference on tourism (ICOT2015),* London, UK, 24-27 June (pp.440-452). International Association for Tourism Policy.

Pantano, E. (2014). Innovation drivers in retail industry. *International Journal of Information Management*, *34,* 344-350.

Pantano, E., & Di Pietro, L. (2013). From e-tourism to f-tourism: emerging issues from negative tourists' online reviews. *Journal of Hospitality and Tourism Technology, 4*(3), 211-217.

Park, S., Nicolau, J. L., & Fesenmaier, D. R. (2013). Assessing advertising in a hierarchical decision model. *Annals of Tourism Research, 40*, 260-282.

Pesonen, J., & Lampi, M. (n.d). Utilizing open data in tourism.

Phillips, P., Zigan, K., Silva, M. M. S., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management*, *50*, 130-141.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9(2),* 181-199.

Prebensen, N. K., Kim, H. L., & Uysal, M. (2016). Cocreation as moderator between the experience value and satisfaction relationship. *Journal of Travel Research, 55(7),* 934-945.

Rojas, R. (1996). *Neural networks - A systematic introduction*. Berlin: Springer-Verlag.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: theory and applications*. Singapore: World Scientific Pub Co Inc.

Rossetti, M., Stella, F., & Zanker, M. (2016). Analyzing user reviews in tourism with topic models. *Information Technology & Tourism*, *16*(1), 5-21.

Sabou, M., Brașoveanu, A. M., & Önder, I. (2015). Linked data for cross-domain decision-making in tourism. In *Information and communication technologies in tourism 2015* (pp. 197-210). Springer International Publishing.

See-To, E. W., & Ho, K. K. (2014). Value co-creation and purchase intention in social network sites: The role of electronic Word-of-Mouth and trust–A theoretical analysis. *Computers in Human Behavior*, *31*, 182-189.

Shakhnarovich, G., Darrell, T. & Indyk, P. (2005). *Nearest-neighbor methods in learning and vision.* Cambridge, MA: MIT Press.

Soualah-Alila, F., Coustaty, M., Rempulski, N., & Doucet, A. (2016). Data tourism: designing an architecture to process tourism data. In *Information and communication technologies in tourism 2016* (pp. 751-763). Springer International Publishing

Sotiriadis, M.D., & van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: the use of twitter by tourists. *Electronic Commerce Research*, *13*, 103-124.

Sparks, B. A., Perkins, H. E., & Buckley, R. (2013). Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behaviour. *Tourism Management*, *39*, 1-9.

Theocharis, S. A., & Tsihritzis, G. A. (2013). Open data for e-government the Greek case. *The fourth international conference on information, intelligence, systems and applications* (ISA), pp. 1-6.

Tripadvisor . (2016). Tripadvisor annual report for 2015, http://ir.tripadvisor.com/annuals.cfm

Tsai, H. T., & Bagozzi, R. P. (2014). Contribution behavior in virtual communities: Cognitive, emotional, and social influences. *MIS Quarterly*, *38*(1), 143-163.

Tung, T., Jai, T. M., & Davis Burns, L. (2014). Attributes of apparel tablet catalogs: value proposition comparisons. *Journal of Fashion Marketing and Management*, *18*(3), 321-337.

Turban, E., King, D., Lee, J. K., Liang, T. P., & Turban, D. C. (2015). Social commerce: foundations, social marketing, and advertising. In *Electronic commerce: A managerial and social networks perspective* (pp. 309-365). Springer International Publishing.

Xiang, Z., Magnini, V.P., & Fesenmaier, D.R. (2015). Information technology and consumer behaviour in travel and tourism: Insights from travel planning using the internet. *Journal of Retailing and Consumer Services, 22*, 244-249.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), 5445-5449.

Wang, D., Li, X., & Li, Y. (2013). China's "smart tourism destination" initiative: A taste of the service-dominant logic. *Journal of Destination Marketing & Management*, *2*(2), 59-61.

White, C. J. (2005). Culture, emotions and behavioural intentions: implications for tourism research and practice. *Current Issues in T tourism*, *8*(6), 510-531.

Wiggins, A., & Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. *Proceedings of the 44th Hawaii international conference on system sciences.*

Williams, A.M., & Balaz, V. (2015). Tourism risk and uncertainty: theoretical reflections. *Journal of Travel Research*, *54*(3), 271-287.

Wolfram, S. (2015). An elementary introduction to the Wolfram language. Wolfram Media Inc.

Wu, C. T., Liu, S. C., Chu, C. F., Chu, Y. P., & Yu, S. S. (2014). A study of open data for tourism service. *International Journal of Electronic Business Management*, *12*(3), 214-221.

Zhang Z., Zhang X., & Yang Y. (2016). The power of expert identity: how website-recognized expert reviews influence travelers' online rating behavior. *Tourism Management*, *55*, 15-24.

Appendix A

The algorithm employed can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Formally, it builds a hyper-plane, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training-data point of the true and false classes (Figure 6).

Given some training data set D:

$$D = \left\{ \left(\vec{x}_i, y_i \right) \mid \vec{x}_i \in \Re^n, y_i \in \left\{ \text{true, false} \right\} \right\};$$

Therefore, the algorithm finds the maximum-margin hyper-plane dividing the points exhibiting y_i=true from those having y_i=false; any hyper-plane can be written as the set of points $\vec{x}_i$ satisfying the following equation (see again Figure 1):

$$\vec{w} \cdot \vec{x} - b = 0;$$

In case the training data are linearly separable, then two hyper-planes can be selected in a way that data are separated, thus maximizing the distances between the points.

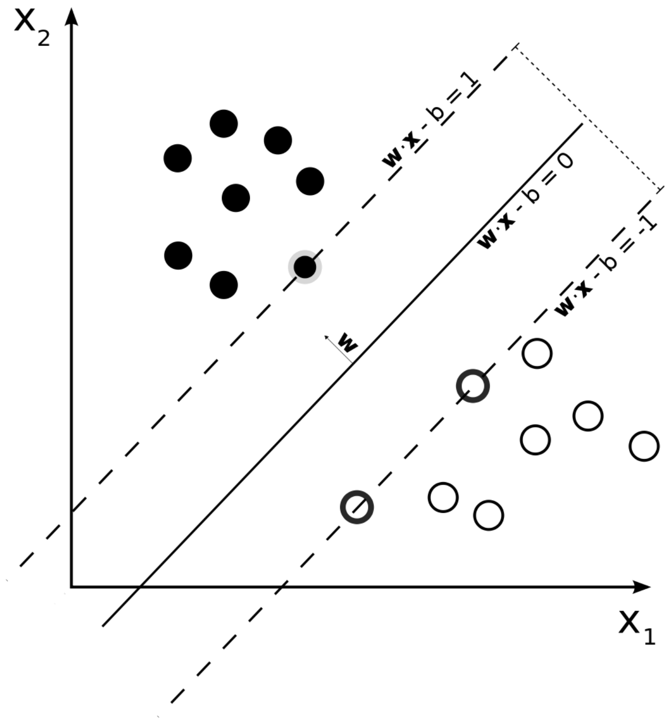

Figure 6: Hyper-plane building among data

Source: Pantano & Di Blasi, (2015, p. 445).

Appendix B

Code used in *Mathematica* software

```
xx3=Table[xx[[1,a]][[9;;26]],{a,2,Length[xx[[1]]]}];

xxa=Table[xx3[[a]]->1,{a,125}];xxb=Table[xx3[[a]]->0,{a,254,385}];xxc=Union[xxa,xxb];

ccy=Table[c=

Classify[xxc];f1=Table[c[xx3[[a]]],{a,253}];f2=Table[c[xx3[[a]]],{a,254,501}];{c,N[Total[f

1]/Length[f1]],N[(Length[f2]-Total[f2])/Length[f2]],N[Total[f1]/Length[f1]]+N[(Length[f2]-

Total[f2])/Length[f2]]},{i,200}];Table[ccy[[a,4]],{a,200}]
```
