IDENTIFYING SENTIMENTS ON HOTELS IN TENERIFE PUBLISHED ON SOCIAL MEDIA BY ENGLISH-SPEAKING TOURISTS

C.A. Martín[(a)], R.M. Aguilar[(b)], J.M. Torres[(c)], A. García-Aguilar[(d)]

[(a),(b),(c),(d)]Department of Computer and Systems Engineering, University of La Laguna, 38200 La Laguna (Tenerife), Spain.

[(a)]carlos.martin.galan@iac.es, [(b)]raguilar@ull.edu.es, [(c)]jmtorres@ull.edu.es, [(d)]albertogaragui@gmail.com

ABSTRACT

This paper describes an algorithm for automatically identifying the sentiments expressed by tourists on eWOM (Electronic Word of Mouth) platforms. Based on reviews published by tourists after staying in hotels, a classifier is trained using an SVM (Support Vector Machine) supervised learning algorithm. We present a use case for this method involving a group of hotels located on the island of Tenerife (Canary Islands).

Keywords: sentiment analysis, social net, machine learning, support vector machine, turism

1. INTRODUCTION

The tourism industry accounts for approximately ten percent of the world's GDP (WTTC, 2015), comprising almost the entirety of the economy in some regions. In Spain, its contribution to the GDP is sixteen percent. In the ranking of the importance of tourism to the national economy, Spain is in seventh place in absolute terms, and in forty-seventh if its size is taken into account (WTTC, 2016).

According to the latest figures, in 2014 the Canary Islands received over eleven and a half million foreign tourists for the first time in its history (11,511,108). This was 8.68% more than in 2013, or an additional 919,848 foreign tourists. The tourism industry is the real engine of growth and development in the Canary Islands, accounting for a high percentage of its GDP. This has a knock-on effect on the remaining industries and services on the islands, especially in the development of trade, transportation, food production and industry.

Tourism also comprises a very important component in creating jobs in the services sector of the archipelago, which encompasses direct employment in the sun and beach sector, as well as workers in activities that support tourism, such as restaurants, hotels, travel agencies, passenger transport, car rental and recreational, cultural and sports activities.

In 2016 Tenerife received over 7 million tourists, most of them from the United Kingdom, which accounted for thirty-one percent of the passengers arriving at Tenerife's airports in 2016. These figures indicate that the opinion that English tourists have of Tenerife is particularly relevant.

The Canary Islands are home to 17% of all hotel rooms available in Spain (1,363,934). These are offered primarily in medium- to high-category establishments, with 56% of the rooms in four-star hotels, 28% in three-star hotels and 10% in five-star hotels (Instituto Canario de Estadística, 2010a; Instituto Nacional de Estadística, 2010).

Future forecasts agree on the need to adapt tourist activity to a) the new requirements and demands of clients, as well as to changing conditions in the industry, such as the emergence of new destinations and markets, namely Asia and Central and Eastern Europe, b) shorter stays and shorter travel distances, c) the financial situation, d) the growing autonomy of users in arranging their travel and the democratization of the supply, e) the effect of new information and communication technologies in the marketing, distribution and communication processes f) the limits on the environmental carrying capacity of the destination, g) the demand for greater corporate social responsibility, and so on (European Travel Commission, 2010).

In light of the financial importance of tourist activity to the Canary Islands, and of the constantly changing scenario, the market survival of our destinations and companies will depend to a large extent on their ability to continue offering experiences, products and services to tourists that are sufficiently appealing and that respond to their changing tastes and demands. As a result, the important role of research in tourism is undeniable as a means of identifying the key socio-economic, sectorial and corporate variables, and for making decisions

ISBN 978-88-97999-85-0; Affenzeller, Bruzzone, Jiménez, Longo and Piera Eds.

that will lead to successfully managing both destinations and companies.

According to Peppers and Rogers (Peppers, 2011), a company that caters to its customers is one that uses information to gain a competitive edge, grow and become profitable. In its most general form, Customer Relationship Management (CRM) may be considered as a group of practices designed to put a company in much closer contact with its customers so it can learn more about them. The broader goal is to increase the value of its customers, and thus of the company. This is why when a management model is based on customer satisfaction, we have to consider four dimensions: customer identification, customer acquisition, customer retention and customer development. This fourth dimension involves finding the real market for a tourism company and product by discovering and learning from its own customers. Focusing on the customer allows developing a product that the customer actually needs. In short, the goal is to have an in-depth knowledge of the customer.

In this paper, we seek to monitor users' perceptions of hotel establishments in Tenerife through comments made on social media so as to ascertain the level of satisfaction with the tourism destination.

The analysis or tourists comments are part of the sentiment analysis field. According to Liu B. (Liu B, 2012) the purpose of sentiment analysis is to define automatic tools capable of extracting subjective information to create structured knowledge that can be put into operation.

Sentiment analysis is a complex field of research (Sentiment Analysis in Social Networks, 2016), in which great difficulties can appear like context analysis or figurative language devices such as irony and sarcasm. These sort of problems are common when users comments in social networks. Despite this, companies are increasingly mindful of the importance of opinions on their products and services that users publish on social media. Specifically, the appearance of social media has provided us with a new and valuable source of information in the tourism industry for studying the perception of various destinations. This way of learning tourists' opinions offers advantages, since the information provided is spontaneous, natural and not conditioned by how a questionnaire is designed. We can thus learn which factors are truly important at the destination from the tourists' point of view and their opinions of said factors. However, due to the diversity and amount of information, monitoring the social networks in real time is a huge challenge that can only be undertaken through automation. Even so, the problem poses serious difficulties due to the heterogeneity of the information and the added challenge of processing natural language. In this regard, the application of machine learning or network analysis techniques is essential to overcome these barriers.

The goal of our work is to automatically analyze the perception of hotels in Tenerife (Canary Islands) through information provided by English-speaking tourists on social media. To achieve this objective, we propose creating a tool that detects the sentiment expressed in the comments published on the tourism platforms booking.com and tripadvisor.com so as to determine the level of satisfaction with the services received during stays at hotels on the island of Tenerife.

2. METHODS

A text can be automatically classified through machine learning. This entails using a classification algorithm that, based on a set of words, is capable of assigning text to a predefined category. To create this algorithm we need to have a set of data (text) whose category is known (sentiment) that can serve as training data. The next step involves using a classification algorithm to determine the different parameters of the model. In this paper we describe the design for classifying text with a support vector machine algorithm. This process will be divided into the four main steps shown below:

- Definition of the category tree
- Gathering of training data
- Labeling of training data
- Training the classifier

2.1. Definition of the category tree

The first issue to address is what categories or labels we want to assign to the texts. This will depend on the problem to be solved. If the goal is to analyze the sentiment of the text, we could assign the following categories: Positive or Negative. Then, the texts that are entered into the classifier will be labeled as positive ("the hotel was clean") or negative ("the food was bad").

The sentiment analysis can have two possible goals:

- Polarity analysis: to ascertain if a given text has positive, negative or neutral connotations. One of the relevant applications in this area is to analyze user reviews of a product.
- Identification of names: to output a list of the names mentioned in the text, and certain implementations also classify them. For example, one feature of the identification of names is to list the people, organizations or locations mentioned in a text.

ISBN 978-88-97999-85-0; Affenzeller, Bruzzone, Jiménez, Longo and Piera Eds.

Specifically, in this work we will conduct a polarity analysis to determine if, based on their comments, tourists have a good or bad perception of the service received in hotels in Tenerife.

2.2. Gathering of training data

Once the category tree is defined, the next step is to gather training data. This refers to the texts that will be used as training examples and that are representative of the future texts that we want the model to classify automatically. These data are everywhere and can be obtained automatically from the web through APIs and open data sets. If our goal is to learn tourists' perception of the service received at a hotel, we can resort to comments published on eWOM tourism platforms like tripadvisor.com or booking.com.

2.3. Labeling of training data

Once the data are collected, they have to be labeled with the corresponding categories in order to create the set of training data. This is a tedious process because each text has to be labeled using the set of predefined categories. When determining tourist sentiment, each comment used for training has to be identified as positive or negative. The set of training data tells the classifier that for a given input (text), it has to assign a specific output (category). The success of any training algorithm based on supervised learning (it has a data set with its associated classification) depends on the accuracy of how the training data set is initially labeled.

The number of training samples needed depends to a large extent on the specific use case; that is, on the complexity of the problem and on the number of categories to be used. For example, it is not the same to train a sentiment analysis model for comments as it is to train a classifier to identify specific elements in a text. Sentiment analysis is a much harder problem to solve and requires much more training data. Analyzing comments published on eWOM platforms is much more difficult than analyzing well-written and well-structured critiques. In summary, the training data have to be sufficient (at least 3,000 samples for sentiment analysis) and adequately proportioned in terms of the number of positive and negative texts. (Pozzi, 2017)

2.4. Labeling of training data

A support vector machine (SVM) is a supervised learning algorithm; therefore, it has to be trained using a set of properly categorized samples so as to yield a classifier that outputs the category of a new case. The SVM model is supported by the statistical theories proposed by Vapnik (Vapnik,1999).

This type of model can be used for binary or multi-category classifications or, in the most general case, to approximate a regression.

Its performance relies on finding the vectors that define the hyperplanes separating the groups with the maximum margin. To understand this idea, suppose that we want to make predictions about a binary classification. For example, we have the values for a hotel inspection and we want to determine if it increases its star rating or not. We could graphically represent the samples on a Cartesian plane on whose edges we only consider two of the variables included in the inspection (Figure 1). The hotels whose star rating is increased are drawn as circles, and those whose rating is not increased are drawn as squares.

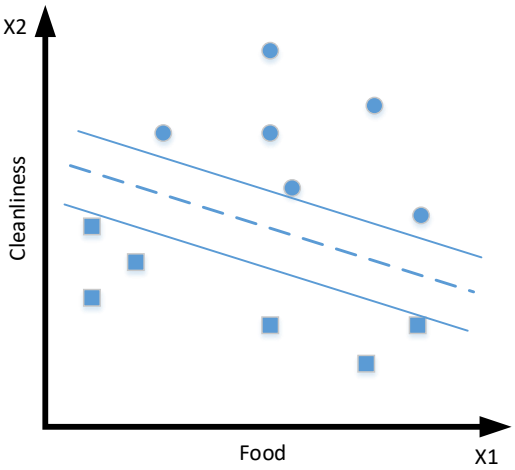

Figure 1: Hotel with increased star rating (circles) and with same star rating (squares)

The SVM model will assign a surface to separate the negative and positive samples by the largest margin possible on either side of the hyperplane (the solid lines in Figure 1). In this case, since it is a two-dimensional problem, the SVM will look for a straight line. In a case with three attributes to classify (e.g. food, cleanliness, staff), it would be a plane. In a case with more than three dimensions, it would be a hyperplane with the appropriate number of variables.

There are numerous methods for determining this hyperplane, such as neural networks and optimization. What sets SVMs apart is that the resulting hyperplane is such that the margin separating the data is as large as possible (dashed line in Figure 1). The further away the hyperplane is from the samples to be classified, the better the result will be because small variations in the variables do not result in being assigned to another group.

Therefore, if we regard each available case (squares and circles) as a vector in space, the SVM algorithm looks for the vectors on which to

support the parallel planes, one toward the positive set and another toward the negative. The hyperplane of maximum separation desired will be the vector drawn exactly in between these vectors.

In the event that a non-linear surface has to be used to separate the sets (Figure 3 left), the SVM transforms the space of attributes into a linear space of higher dimensionality. It does this by mapping each point in the data set by means of a non-linear transformation $\phi(x)$: $R^2 \rightarrow F$, in the so-called higher dimension feature space F, such that each point X_n is projected to a point $\phi(X_n)$. If the transformation is selected correctly, the feature space F will be a higher dimensional space in which there will be a linear relationship between the projected data set and the factors that underlie the variability in the original data set. This process of mapping data in a feature space by means of a non-linear transformation is part of what is known as kernel-based methods. Figure 2 left shows two data sets in R^2. In principle, it would be very complicated to train a classifier that would be able to separate the two sets; however, each point can be mapped using a non-linear transformation $\phi(x)$ into a feature space F with M dimensions. If the non-linear transformation is properly selected, both data sets could be easily separated along a boundary by hyperplane h in F. A schematic illustration of this process is shown in Figure 2.

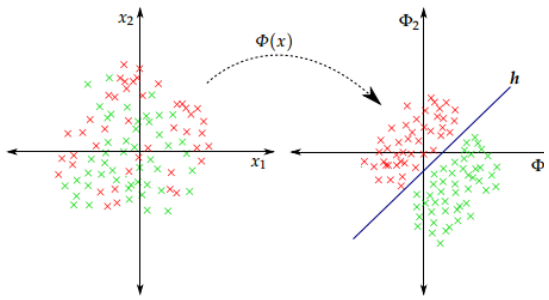

Figure 2: In an SVM, two data sets that are hard to classify in R^2 (left) are mapped via ϕ (x) into the feature space (right), where they can be classified by using hyperplane h as a boundary.

3. RESULTS

The goal of this work is to create an automatic classifier that can determine the sentiment that tourists who visit the island of Tenerife have regarding the hotels where they stayed. This sentiment is taken from the comments left on eWOM platforms like booking.com and tripadvisor.com. (García-Pablos, 2016)

As noted in the above section, the first step in creating a classifier is to obtain a set of previously classified comments to use as a training sample for the classifier that will be used later. In order to train the classifier properly, we need to find 1500 positive comment samples and 1500 negative comment samples. The first part of the work involved finding said samples. We decided to take the training samples from the British version of the booking.com website, where comments on hotels feature the following characteristics:

- They are scored numerically between 0 and 10 (decimals allowed).
- In some case, information is available on the source and date of the comment.
- The person making the comment has the option of giving the comment a title with a positive and negative aspect from his review.

With these characteristics, and after doing a mass data extraction, we can set up the training sample as follows:

- Take the reviews with a score above 6 and label them as "good", using as training text the concatenation of the title and the positive part of the comment.
- Take the reviews with a score below 5 and label them as "bad", using as training text the concatenation of the title and the negative part of the comment

From the resulting samples, we select 1500 good ones and 1500 bas ones distributed uniformly along the range of scores. An example of the samples used to train the algorithm is shown in Figure 3.

Once the training samples are obtained, we generate an SVM training algorithm using Monkeylearn (infrastructure and software service for analyzing text) (MonkeyLearn, 2017). The following parameters are specified to define the algorithm:

1. Language: must match that in the samples (English in this case).
2. Normalize weights: we enable the normalization to inform the classifier that it must take into account the number of samples for each category when defining their probability. In our case, since the number of positive and negative samples is balanced, it is not enabled.
3. N-gram: defines whether the features used to characterize the text will be:
   - Unigrams or words (n-gram size = 1)
   - Bigrams or terms consisting of two words (n-gram size = 2)
   - Trigrams or terms consisting of up to three words (n-gram size = 3)

The classifications are based on words (unigrams).

4. Stemming: this setting specifies if words have to be derived. The derivation process transforms words into their root version, meaning that inflections and derivatives are grouped. For example, the words "fishes", "fishing" and "fisherman" are all transformed into the root word "fish". This will generally help the classifier to generalize and improve its classification. However, in the case of sentiments based on reviews written in

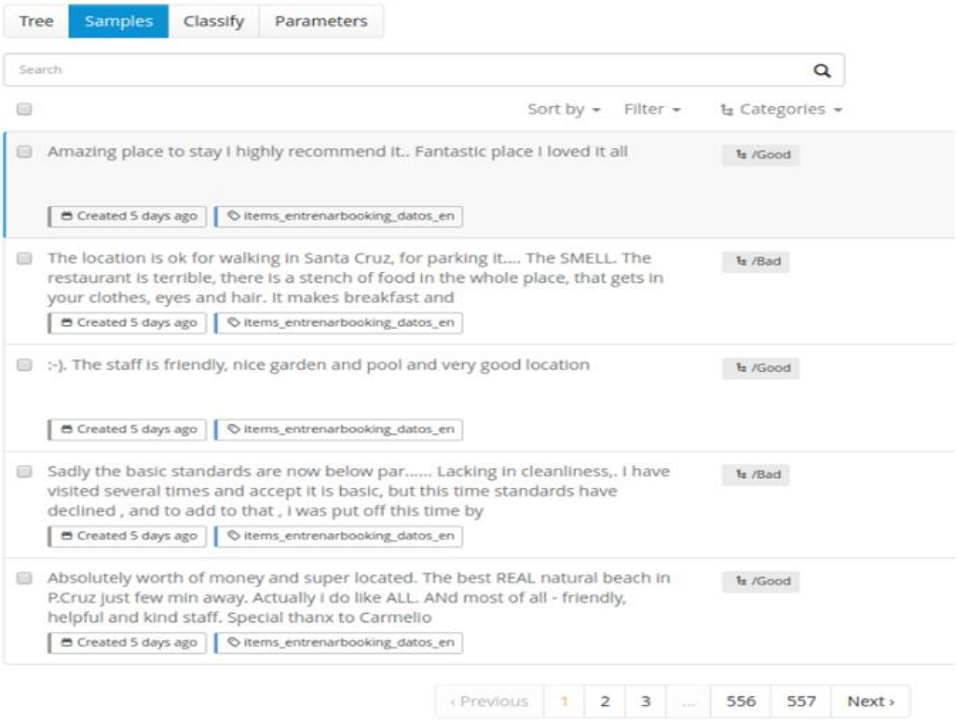

Figure 3: Example of some of the training samples used

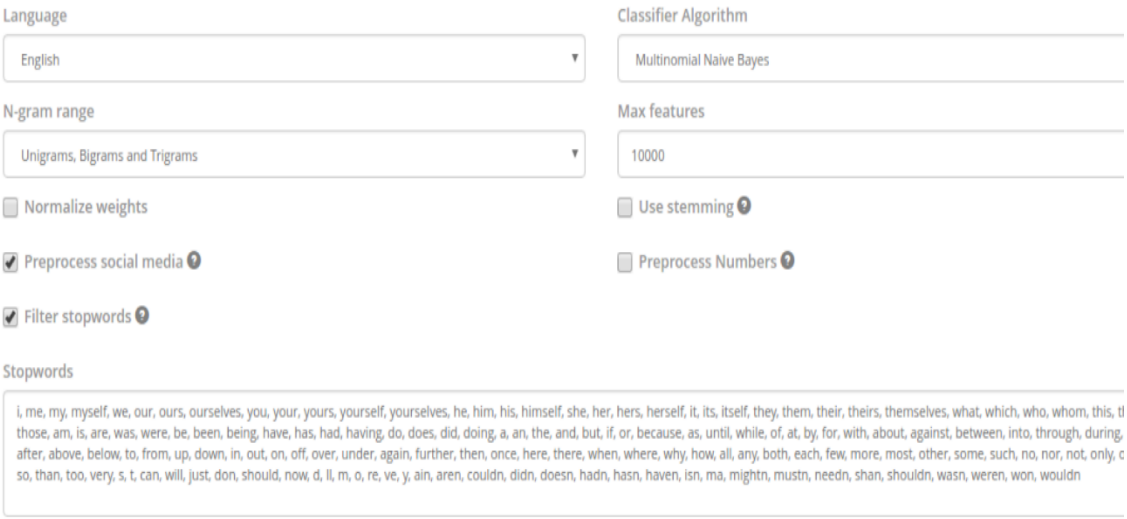

Figure 4 – Configuration parameters for training the SVM

digital platforms, the setting is better without this transformation.

5. Filter inactive words: this active setting filters stopwords. These are high-frequency words that do not normally contribute as classification characteristics, such as articles, conjunctions, etc.
6. Pre-processing of social media: expressions used in conversations on social media are transformed into a word. For example, :-) is transformed into "happy".

The training of the classifier yielded the results shown in Figure 5. The method was calculated to be 95% accurate in identifying good sentiments and 91% for the bad. We can see that the words that are repeated most often in good sentiments are "beautiful" and "amazing", whereas in the bad it is "rude", "dirty" and "cockroaches".

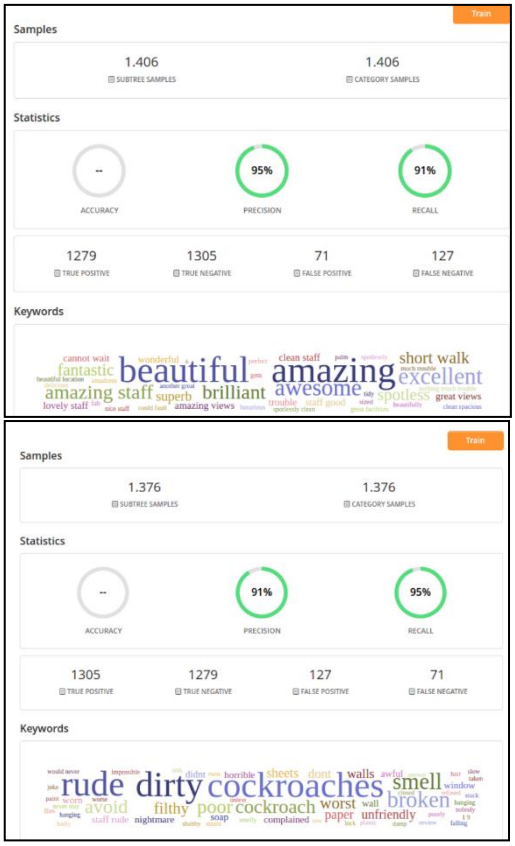

Figure 5 – Performance of the training process for good sentiments and bad sentiments.

We monitored a group of hotels in a city on the island of Tenerife; specifically, we analyzed 13,541 comments, yielding the results shown in Figure 6. The X axis shows the actual rating of the reviewer, while the Y axis shows the predictions of the classifier. Note that better results are obtained by classifying negative comments (1-2), than positive (4-5), and as expected better predictions at the extremes.

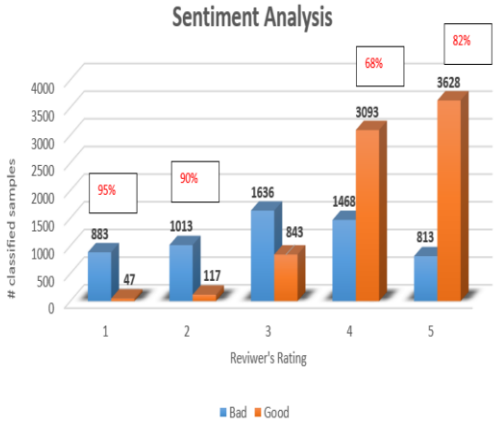

Figure 6 – Results of the classification of 13541 comments on hotels.

4. CONCLUSIONS

This paper clearly shows the need to monitor customer perception in order to maintain a competitive edge. In the case of tourism, the main driver of the economy in the Canary Islands, it is a key component. In this area, however, in which there are millions of users, finding information on their perceptions of the service received is a cumbersome task. Moreover, the use of questionnaires does not provide accurate information on how customers feel about the service they received. As a result, in this paper we propose using the reviews that tourists publish on eWOM platforms like tripadvisor.com and booking.com.

In order to identify the sentiment of a comment, we used an SVM training algorithm that, according to the bibliography, exhibits very good efficiency indices with classification problems, (Xiang'zi, 2017). In this paper we used the online service Monkeylearn to create the classifier.

In the final step, we used this classifier to determine the sentiment expressed for a group of hotels on the island of Tenerife in the Canary Islands. In this use case we can see how the classifier obtains better results by evaluating potentially negative comments and, as expected, shows a better behavior with less neutral comments. In future work we plan to create a classifier that can identify what tourists are discussing so that we can determine what aspects they like and thus what to promote, as well as what elements are associated with negative comments so that they can be improved.

We also plan to conduct a more exhaustive analysis of the results in order to be able to monitor sentiments based on the tourist's place of origin, the length of their stay on the island and other factors.

ACKNOWLEDGMENTS

This work is being supported by the project #2016TUR15 “VITUIN: Vigilancia Turística Inteligente de Tenerife en Redes Sociales “ from research funds of the Fundación CajaCanarias

REFERENCES

García-Pablos A., Lo Duca A., Cuadros M., Linaza M.T., and Marchetti A. 2016, “Correlating Languages and Sentiment Analysis on the Basis of Text-based Reviews,” in Information and Communication Technologies in Tourism 2016, pp. 565–577.

MonkeyLearn Inc, ©2017. All rights reserved. http://monkeylearn.com [accessed 15 May, 2017]

Peppers D. and Rogers M., 2011, Managing Customer Relationships: A Strategic Framework (Second Edition), Wiley.

Pozzi F.A., Fersini E., Messina E., and Liu B., 2017,, “Chapter 1 - Challenges of Sentiment Analysis in Social Networks: An Overview,” in Sentiment Analysis in Social Networks, Eds. Boston: Morgan Kaufmann, pp. 1–11.

Vapnik Vladimir N., 1999, An Overview of Statistical Learning Theory, IEEE Transactions On Neural Networks, Vol. 10, No. 5

WTTC, 2015, Econocmic Impact Analysis 2015, World Travel & Tourism Council. Available from: http://sete.gr/media/2614/150430-economic-impact-2015.pdf [accessed 15 May, 2017]

WTTC, 2016, Econocmic Impact 2016 Spain, World Travel & Tourism Council. Available from: https://www.wttc.org/-/media/files/reports/economic%20impact%20research/countries%202016/spain2016.pdf [accessed 15 May, 2017]

Xiang'zi Leng, Jinhua Wang, Haibo Ji, Qin'geng Wang, Huiming Li, Xin Qian, Fengying Li, Meng Yang, 2017, Prediction of size-fractionated airborne particle-bound metals using MLR, BP-ANN and SVM analyses, Chemosphere, Volume 180, pp: 513-522

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, & Bing Liu. (2016).Sentiment analysis in social networks (1st ed.). US: Morgan Kaufmann Publishers Inc. Retrieved from http://lib.myilibrary.com?ID=961656