ISSN 1870-4069

Prediction of the Polarity of Opinions in the Domain of Tourism through Machine Learning

Marcos A. Leiva-Vasconcellos, Mireya Tovar-Vidal

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

marcos.leiva@alumno.buap.mx,
mireya.tovar@correo.buap.mx

Abstract. The analysis of the polarity of any type of comments has increased thanks to the development of Web 2.0 where millions of opinions are currently generated by users of various sites, with high information content. Opinion mining focuses on automatically determining the polarity of publications for research and development of real-world applications. This article aims to determine which of the proposed algorithms (Decision Trees, Support Vector Machine and *Naïve Bayes*) are appropriate for predicting the polarity of opinions in the tourism domain, for this a set of opinions (487) about hotels are extracted from TripAdvisor. The experimental results obtained show that Support Vector Machine (SVM) and *Naïve Bayes* are the best classifiers for this type of task with an accuracy of 62% and 61% respectively, a result that will improve by increasing the training set.

Keywords: Opinion mining, natural language processing, machine learning.

1 Introduction

With the creation of Web 2.0, the user went from being a consumer of resources to a content creator, having the possibility of issuing criteria and evaluating the content on the Internet. TripAdvisor was created in 2000 and although it was not intended for users to exchange opinions about the sites visited.

From 2004, consumer comments exceeded professional comments, it became a collection of comments from travelers from around the world where the individual experience of the places visited was exposed, according to a radio interview with Stephen Kaufer, creator from TripAdvisor.

Fig. 1 shows the number of views in millions from 2014 to 2020 from the TripAdvisor.com website, confirming that the website is the largest travel guide in the world. According to [9, 4], 1 in 16 people consult TripAdvisor.com to plan their vacation, which is why the opinions posted on the site are so important to the tourism sector.

According to the World Tourism Organization, tourism is defined as that which includes the activities conducted by people during their trips and stays in places other

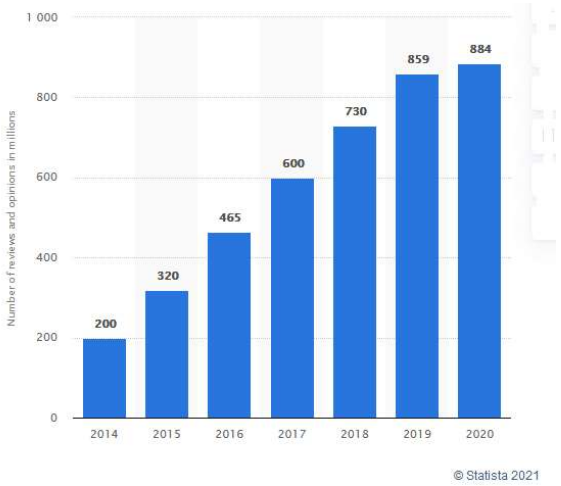

Fig. 1. Number of reviews, expressed in millions, on TripAdvisor. Source: Statista.com.

than their usual environment, for a consecutive period of less than one year for leisure purposes, for business and others [6].

Tourism is one of the fundamental activities in many countries, including Mexico, which represented 8.7% of its GDP (Gross Domestic Product) at the end of December 2019, according to [8], which is why it is constantly receiving feedback to improve the quality of tourism services. It is of vital importance, due to the importance that tourists give it, to consult the opinions that are published by the people of the tourist places to improve their quality.

Sentiment analysis or opinion mining is a field of research within Natural Language Processing that automatically extracts subjective information expressed in a text about a given domain [7]. In this way, the author's attitude about a particular topic can be known, which can normally be positive, neutral, or negative. The interest in opinion mining has increased over time due to the large amount of information that circulates in the networks and that it would be impossible for a person to characterize it [22]. In this sense, the tourism sector can rely on opinion mining to automatically extract the polarity of tourists.

Studies have been conducted using sentiment analysis to assess opinions about a topic. Since TripAdvisor is the largest site for tourism opinions, it has been used as a data source to develop algorithms that evaluate the opinions of users regarding sites, and thus be able to influence problems in a practical way.

Evaluating sentiment in large volumes of data is not always unambiguous, even when done manually the result may not be the same, depending on the encoder. This article aims to determine the appropriate algorithms for prediction of opinion polarity in the tourism domain using opinion mining. The study is carried out using TripAdvisor as a data source.

This article is structured as follows: Section 2 present the works related. Section 3 shows the proposed solution of the identification of entities. Section 4 shows the results obtained, and in Section 5 the conclusions.

2 Related Works

Opinion mining has been investigated for some years using Natural Language Processing, below are some works related to this study.

In [3] association rules are applied to the database, in this case Twitter, with these rules, opinions can be categorized, and people's feelings identified. In the research work developed by [16], the authors propose a model composed of three phases: Pre-processing, Identification of Aspects and Polarity Identification, obtaining 50% effectiveness in the SemEval 2016 Competition forum. They use the sentiment dictionary SentiWordNet, which is generally so that some words in specific contexts have one polarity and, in another context, a different one, is used for comments in Spanish.

In the research [1] they implement a scheme for the unsupervised detection of the polarity of opinions from new lexical resources SentiWordNet 4.0 and 4.1 obtaining values of accuracy and *F1* of 85% much higher than version 3.0.

The author in [12] proposes a model based on 3 modules: text processing, attribute selection and classification with machine learning, extracting comment data from TripAdvisor, Booking, Expedia and Trivago. These comments are classified into 2 classes (good and bad), the classifiers used were Support Vector Machine (SVM), *Naïve Bayes* and decision trees, the experiments showed that *Naïve Bayes* was the most accurate, although the accuracy levels are accurate for SVM and *Naïve Bayes*.

In [19] the author uses lemmatization and normalization to train the proposed model, which uses SVM, which makes the classification set obtain better results and is done for comments in Spanish.

The authors in [2] extract the keywords from the comment to obtain lists of concepts and keywords through the Microsoft Knowledge Graph. In the research [17] the authors use 3 classifiers to evaluate comments on Twitter, they are: Decision trees, *k nearest* neighbors and *Naïve Bayes*, the best classifier of all was the decision tree.

Sentitext is used in [11], which is a sentiment analysis system, based on domain-independent linguistic knowledge using the Freeling morphological analyzer. It uses comments in Spanish from TripAdvisor and has a success rate close to 90%, although it detects more positive segments than negative ones. This work performs a sentiment analysis classification using the main classifiers according to previous studies to determine which ones provide the best results.

3 Proposed Solution

Next, the proposed solution is described, as well as the methodology to follow for the prediction of the polarity of opinions in the tourism field.

3.1 Description of the Proposed Solution

After reviewing the main trends in sentiment analysis, it can be said that more research is aimed at classifying opinions as positive, neutral or negative.

Most are in the English language and very few in Spanish. The most used classifiers are SVM and *Naïve Bayes*, giving good results in each of them.

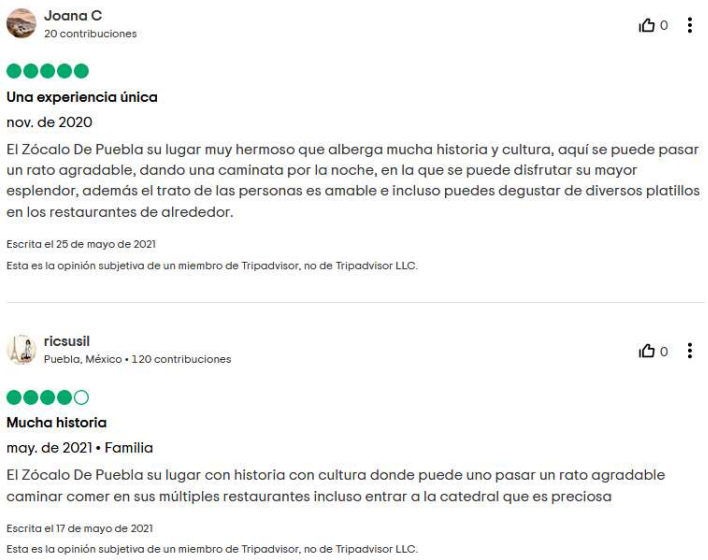

Fig. 2. Reviews on TripAdvisor.com. Source: TripAdvisor.com

There are not many studies of natural language processing using TripAdvisor, most of the research focuses on Twitter and Facebook. The consulted investigations do not consider the date that the opinion was made to verify if the comment was improved over time.

Our proposal consists of 5 types of polarities like the rate on TripAdvisor and are described below:

1. Very Negative: It refers to negative opinions, but with emphasis, it uses expressions such as: very bad, worse, etc.
2. Negative: Use negative expressions, for example: bad, expensive, etc.
3. Regular: They are regular expressions such as: we went to the hotel, room on the 4th floor, etc.
4. Positive: Expressions with an emphasis on positive issues: good hotel, nice pool, etc.
5. Very Positive: They are positive expressions with emphasis: very good, very nice, etc.

To carry out the program, it was determined to use Python due to its robustness in terms of natural language processing through the NLTK (*Natural Language Toolkit*) library, which, although it was not initially designed for the Spanish language, can use the corpus in Spanish as es cess_ esp[10], which has 500,000 words and 610 files[15]. The data structure to be used will be trees, graphs and json mainly for the representation of information.

Reviews are sourced from TripAdvisor.com in Spanish language; Fig. 2 shows the form in which the comments are expressed, according to [20] and [21], users sometimes

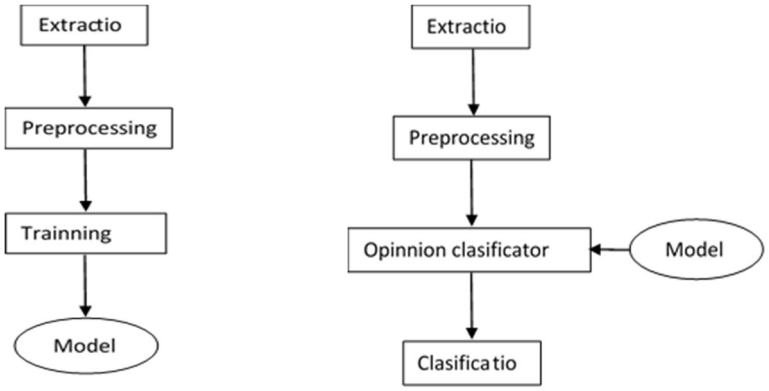

Fig. 3. Methodology for the classification of opinions a) Training b) Classification with the generated model.

evaluate in one way and the opinions expressed contradict the evaluation given, so it is not feasible to take only the numerical evaluation of each user. *Web scraping* is used to obtain the comments of a certain place since TripAdvisor does not have an API to obtain the data.

3.2 Proposed Methodology

The methodology to be used for polarity prediction is described below. Fig. 3 a) shows the methodology for training the classifiers, returning the model to be used in the classification and figure 3 b) shows the methodology for performing the classification according to the model obtained in the training. Next, the stages in each of the steps will be described.

3.2.1 Extraction

The data extraction will be done from the TripAdvisor.com page using web scraping, through a script implemented in Python and the data will be saved in csv files.

3.2.2 Preprocessing

In this stage, a set of techniques will be applied to obtain better results in the later stages, the Python NLTK (Natural Language Toolkit) library is used. Characters that are not letters will be removed first, such as: punctuation marks, numbers and characters that do not belong to the Spanish alphabet. In addition, pieces of the text that may interfere with the analysis of the text will be removed; this is domain dependent. In this phase, all words are converted to lowercase. Finally, the noise will be eliminated, which consists of getting rid of stopwords (words like *the, the, are*, etc.).

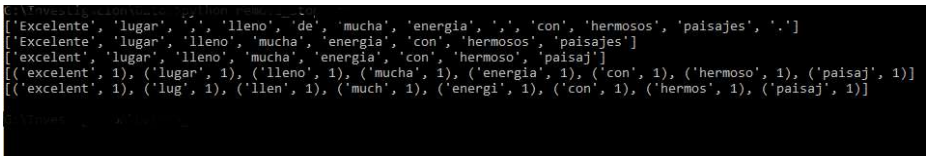

Fig. 4. Pre-processing using the Python NLTK library.

For example, if you have the opinion: "Excellent place, full of energy, with beautiful landscapes. "When applying these steps, the text would read as follows: "Excellent place full of energy with beautiful landscapes". Then it would go to the tokenization process, which consists of separating the words from the text and building a vector composed of each word.

The last method of pre-processing is called lemmatization or stemming, which reduces the original word in its root part, making subsequent classification easier. Continuing with the previous example, after applying tokenization and lemmatization, the result would be as follows: [('excellent', 1), ('lug', 1), ('llen', 1), ('much', 1), ('energi', 1), ('con', 1), ('hermos', 1), ('landscape', 1)], in Fig. 4 shows the whole process using the aforementioned library with *Snowball Stemmer*.

3.2.3 Training

For this phase, *TF-IDF* are used, which are: Term Frequency and Inverse Frequency of Documents and with this it will convert the document feature vectors using TfidfVectorizer. The vectorized document is used for SVM [18], *Naïve Bayes* [13] and Decision Tree [14] training to generate a model, which will be loaded into the classification.

3.2.4 Opinion Classifier

There are various investigations about classifiers for opinion mining, in most investigations machine learning techniques are used, being SVM and *Naïve Bayes* the most popular, of these two techniques SVM is the one that presents greater certainty according to [5]. Due to the aforementioned, it is determined to use SVM and *Naïve Bayes* as opinion classifiers, although it will be trained and classified using Decision Trees to verify its results.

3.3 Evaluation of the Proposal

To measure the performance of a classifier, several terms have been defined that are described below [1]:

Accuracy : *is* the proportion of the total number of predictions that were correct as shown in equation 1:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}. \quad (1)$$

Precision: is the proportion of predicted cases that were positive as shown in equation 2:

$$P = \frac{TP}{TP+FP}. \quad (2)$$

Recall: is the proportion of positive cases that were correctly identified as shown in equation 3:

$$R = \frac{TP}{TP+FN}. \quad (3)$$

F_1: is the harmonic mean that combines the precision and accuracy values as shown in equation 4:

$$F_1 = \frac{2*P*R}{P+R}. \quad (4)$$

Being:

- *TP* (*True Positive*) True Positive: Number of cases that the test declares positive and that are truly positive.
- *TN* (*True Negative)* True Negative: Number of cases that the test declares negative and that are negative.
- *FP* (False *Positive*) False Positive: Number of cases that the test declares positive and that are negative.
- *FN* (False *Negative*) False Negative: Number of cases that the test declares negative and that are positive.

4 Experimental Results

This section describes the data used for the analysis of the opinions and the results corresponding to each class with the SVM, *Naïve Bayes* and Decision Trees algorithms.

4.1 Dataset

The data used is extracted from the TripAdvisor.com page by web scraping 45 hotels in Puebla. Table 1 shows the names of the hotels.

Table 2 shows the total opinions classified by classes from 1 (very negative) to 5 (very positive), out of a total of 487 opinions retrieved from the aforementioned hotels.

4.2 Data Set

With the SVM and *Naïve Bayes algorithms* presented in section 3 and adding decision trees, the training and classification are carried out. Table 3 shows the results of Accuracy, Precision, Recall and F_1 for each of the algorithms. The training was carried out with cross validation, in which 75% of the cases were taken for training and the remaining 25% of tests. The Baseline was created with a Random Forest Classifier using a random class to the testing set.

Table 1. Selected Hotels for Feedback.

Casona Maria Boutique Hotel	New Suite Seville	Hotel Posada Cuetzalan
La Quinta by Wyndham Puebla Palmas Angelopolis	Hotel Royal 500	Hotel Panamerican
Hotel Casona Poblana	Moctezuma Luxury Boutique Hotel B&B	San Jose House of Prayer Hotel
Palm House Hotel	Hotel San Miguel	Palace Hotel
Loa_Inn_Puebla	Meson San Sabastian	Hotel Royalty Center
Hotel_Gilfer	Hotel Leones	eight 30
Hotel_One_Puebla_FINSA	Hotel Real Santander	Isabel Hotel
Blue Talavera Hotel	Hotel Las Iglesias	Sonata Hotel & Residences
Fiesta Inn Puebla Finsa	OYO Hotel Del Paseo	Puebla de Antano
Hotel Puebla Plaza	Puebla American Party	Crowne Plaza Puebla
Hotel Star Express	Diana Hotel	Suites La Concordia
El Capricho Boutique Hotel	poblana square	Portal Hotel
Hotel Hacienda Del Molino	Meson Sacristy of the Company	Palacio Julio Hotel
Hotel La Quinta	Hotel Casona San Antonio	Hotel Gilfer
Hotel Plaza Zacapoaxtla	Quinta Esencia Hotel Boutique	Loa Inn Juarez

Table 2. Distribution of opinions by classes (1 to 5).

	1	**2**	**3**	**4**	**5**
Opinions	30	13	42	129	273

Table 3. Results of the experiment using the 3 algorithms mentioned.

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1_**
Naïve Bayes	61.48	1.0	0.61	0.76
SVM	62.30	0.88	0.62	0.71
Decision tree	48.36	0.47	0.48	0.47
Baseline	22.13	0.89	0.22	0.34

5 Conclusions and Future Work

This article presents automatic classification algorithms that allow identifying the polarity of the opinions extracted from TripAdvisor. The polarity is distributed from 1 (very negative) to 5 (very positive).

Even though the training set is very small, the experimental results show that the SVM method achieves good results, obtaining 62%, as well as *Naïve Bayes,* which has 61% accuracy, all the algorithms are above the baseline created. We can see that the decision tree algorithm does not have a good performance (48%) as reflected in the studies consulted.

It has been shown that the SVM and *Naïve Bayes algorithms* are appropriate for polarity prediction in this domain, so it is proposed as future work to enrich the data set. In addition to implementing other artificial intelligence techniques such as Artificial Neural Network to improve the performance.

References

1. Amores-Fernández, M., Arco, L., Borroto, C.: Unsupervised opinion polarity detection based on new lexical resources. Computación y Sistemas, vol. 20, no. 2, pp. 263–177 (2016) doi: 10.13053/cys-20-2-2318
2. Chen, W., Xu, Z., Zheng, X., Yu, Q., Luo, Y.: Research on sentiment classification of online travel review text. Applied Sciences, vol. 10, no. 15 (2020) doi: 10.3390/ app10155275
3. Díaz-García, J. Á., Ruiz, M. D., Martín-Bautista, M. J.: Minería de opinión no supervisada en Twitter. In: XVIII Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2018) Granada, Spain, pp. 1023-1028 (2018)
4. Filieri, R., Acikgoz, F., Ndou, V., Dwivedi, Y.: Is TripAdvisor still relevant? The influence of review credibility, review usefulness, and ease of use on consumers' continuance intention. International Journal of Contemporary Hospitality Management, vol. 33, no. 1, (2020) doi: 10.1108/IJCHM-05-2020-0402
5. Flores, L., Guadalupe, I., Peña-Álvarez, E. P.: Aprendizaje automático para la optimización de procesos de marketing digital en el sector turístico. Universidad Tecnológica de Perú (2020)
6. Ghanem, J.: Conceptualizing "the Tourist": A critical review of UNWTO definition. Master Thesis, Universitat de Girona (2017)
7. Hariguna, T., Sukmana, H. T., Kim, J.: Survey opinion using sentiment analysis. Journal of Applied Data Sciences, vol. 1, no. 1, pp. 35–40 (2020) doi: 10.47738/jads.v1i1.10
8. Instituto nacional de estadística y geografía (INEGI): Turismo. Sistema de Cuentas Nacionales de México (2021) https://www.inegi.org.mx/temas/turismosat/#Informacion _general
9. Kinstler, L.: How TripAdvisor changed travel. The Guardian, London (2018) https://www.theguardian.com/news/2018/aug/17/how-tripadvisor-changed-travel
10. Martí, M., Taulé, M., Recasens, M.: AnCora: Multilevel annotated corpora for Catalan and Spanish. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (2008)
11. Moreno-Ortiz, A., Pineda-Castillo, F., Hidalgo-García, R.: Análisis de valoraciones de usuario de hoteles con Sentitext: Un sistema de análisis de sentimiento independiente del dominio. Procesamiento del Lenguaje Natural, no. 45, pp. 31–39 (2010)
12. Mostafa, L.: Machine learning-based sentiment analysis for analyzing the traveler's reviews on Egyptian hotels. In: Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), pp. 405–413 (2020) doi: 10.1007/978-3-030-44289-7_38
13. Murphy, K. P.: Naive Bayes classifiers. University of British Columbia (2006)
14. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., Brown, S. D: An introduction to decision tree modeling. Journal of Chemometrics, vol. 18, pp. 275–285 (2004) doi: 10.10 02/cem.873
15. Navas-Loro, M., Rodríguez-Doncel, V.: Spanish corpora for sentiment analysis: A survey. Language Resources and Evaluation, vol. 54, pp. 303–340 (2020) doi: 10.1007/s10579-019-09470-8
16. Rosales-Quiroga, M. A., Vilariño-Ayala, D., Pinto, D., Tovar, M., Beltrán, B.: Análisis de sentimientos basado en aspectos: un modelo para identificar la polaridad de críticas de usuarios. Research in Computing Science, vol. 115, pp. 171-180 (2016)
17. Shoeb, M., Ahmed, J.: Sentiment analysis and classification of tweets using data mining. International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 12 (2017)
18. Suthaharan, S.: Support vector machine. Machine learning models and algorithms for big data classification, Springer, pp. 207-235 (2016) doi: 10.1007/978-1-4899-7641-3

19. Ticona-Nina, R.: Minería de opiniones basado en aprendizaje supervisado en la evaluación de destinos turísticos de la región de Puno. Universidad Peruana Unión (2019)
20. Valdivia, A., Luzón, M. V., Herrera, F.: Sentiment analysis in TripAdvisor. IEEE Intelligent Systems, vol. 32, no. 4, pp. 72–77 (2017) doi: 10.1109/MIS.2017.3121555
21. Valdivia, A., Luzón, M. V, Herrera, F.: Sentiment analysis on TripAdvisor: Are there inconsistencies in user reviews? In: International Conference on Hybrid Artificial Intelligence Systems, vol. 10334, pp. 15–25 (2017) doi: 10.1007/978-3-319-59650-1_2
22. Vazquez, K. L., Tovar, M., Vilariño, D., Beltrán, B.: Un algoritmo para detectar la polaridad de opiniones en los dominios de laptops y restaurantes. Research in Computing Science, vol. 128, pp. 91–98 (2016) doi: 10.13053/rcs-128-1-8