www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 14(3), Sep-Dec 2023, 53-60. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2023.v14i3.407

Improving Prediction of Polarity in Tourism Domain using Convolutional Neural Network

Marcos A. Leiva Vasconcellos, Mireya Tovar Vidal
[1] Benemérita Universidad Autónoma de Puebla
marcos.leiva@alumno.buap.mx**,** mireya.tovar@correo.buap.mx

Abstract. The analysis of the polarity of various types of comments has been enhanced by the development of Web 2.0, where millions of user-generated opinions on different sites now offer a wealth of information. Opinion mining focuses on automatically determining the polarity of posts for research and the development of real-world applications. This article aims to determine which of the proposed algorithms (Decision Trees, Support Vector Machine, Naïve Bayes, and Convolutional Neural Network) are most suitable for predicting the polarity of opinions in the tourism domain. For this purpose, a set of opinions (30,210) from the REST-MEX 2022 Sentiment Analysis competition, pertaining to hotels, attractions, and restaurants, is used. The experimental results obtained show that the Convolutional Neural Network (CNN) is the best classifier, achieving an accuracy of 98.83%, followed by the Support Vector Machine (SVM) and Naïve Bayes with accuracies of 71% and 70% respectively. The worst performance was observed with Decision Trees, which achieved 62% accuracy.
Keywords: Opinion mining, natural language processing, machine learning, sentiment analysis.

Article Info
Received Sep 26, 2023
Accepted Dec 11, 2023

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 of TripAdvisor.

Figure 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 Kinstler (2018) and Filieri, Acikgoz, Ndou, & Dwivedi (2020), 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 than their usual environment, for a consecutive period of less than one year for leisure purposes, for business and others (Ghanem, 2017). 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 INEGI (n.d.), which is why it is constantly receiving feedback to improve the quality of tourism services. It is of vital importance, due to the interest 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 (Hariguna, Sukmana, & Kim, 2020). 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 (Vazquez, Tovar, Vilariño, & Beltrán, 2016). In this sense, the tourism sector can rely on opinion mining to automatically extract the polarity of tourists.

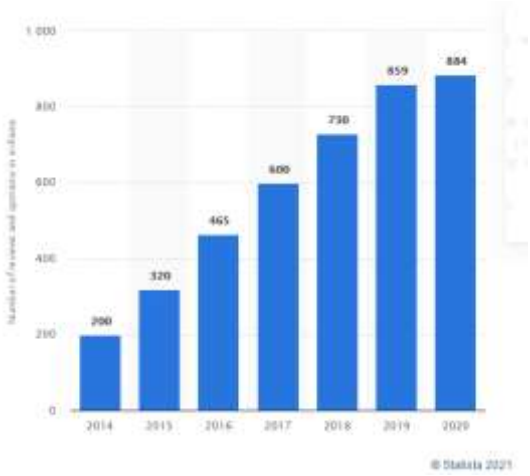

Fig. 1. Number of reviews, expressed in millions, on TripAdvisor.
Source: Statista.com

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.

The paper 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 Diaz-Garcia, Ruiz, & Martin-Bautsta (2018) 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 Quiroga, Ayala, Pinto, Tovar, & Beltrán (2016), 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 Amores Fernández (2016), 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 F_1 of 85% much higher than version 3.0.

The author in Mostafa (2020) 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 Ticona Nina (2019), 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 Chen, Xu, Zheng, Yu, & Luo (2020) extract the keywords from the comment to obtain lists of concepts and keywords through the Microsoft Knowledge Graph. In the research Shoeb & Ahmed (2017), 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 Moreno Ortiz, Pineda Castillo, & Hidalgo García (2010), 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.

An approach based on Aspect-Based Sentiment Analysis on reviews extracted from tourism websites, which has been proposed by Soumaya, Driss, Soufiane, Abderrahmane, & Mohamed (2023), uses nine supervised classification methods to classify online comments into three types of polarity: negative, positive, and neutral. Next, they apply an aspect-based sentiment analysis and used the nine classification methods getting good results.

In Shujun & Song (2022) presented a multifeature fusion approach for sentiment analysis of tourism reviews. The authors implemented three modules, one for text, other for images and the third that fusion both. The module fusion images and texts achieved better performance in tourism reviews.

The authors in Puh & Bagić Babac (2022) analyzed sentiment and ratings expressed in customer reviews on TripAdvisor using machine learning and deep learning methods. The algorithms implemented of machine learning were Naïve Bayes and support vector machines (SVMs). They got better results using deep learning methods mainly through bidirectional long short-term memory (BiLSTM). The aim the paper was to predict tourist satisfaction.

An improved version of the Naïve Bayes classifier is presented in Romano, Contu, Mola et al. (2023), which is called Threshold-based Naïve Bayes (Tb-NB) classifier, to predict the sentiment of the reviews about the Sardinian hotels on Booking.com. Tab-NB is compared with the classifiers: logistic regression, random forest, traditional Naïve Bayes, decision trees, linear discriminant analysis, and support vector machine (SVM), getting good experimental results. The authors selected several features that permitted getting a good performance of Tb-NB for the classification of sentiments in positive and negative opinions.

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. 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*), which, although it was not initially designed for the Spanish language, can use the corpus in Spanish as es cess_ esp (Martí, Taulé, Márquez, & Bertran, 2007), which has 500,000 words and 610 files (Navas-Loro & Rodríguez-Doncel, 2020). The data structure to be used will be trees, graphs and json mainly for the representation of information. For CNN classifier it will be use TensorFlow and Keras.

Reviews are sourced from TripAdvisor.com in Spanish language; Fig. 2 shows the form in which the comments are expressed, according to Valdivia, Luzón, & Herrera (2017) and Valdivia, Luzón, & Herrera (2017b), users sometimes 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.

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

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.

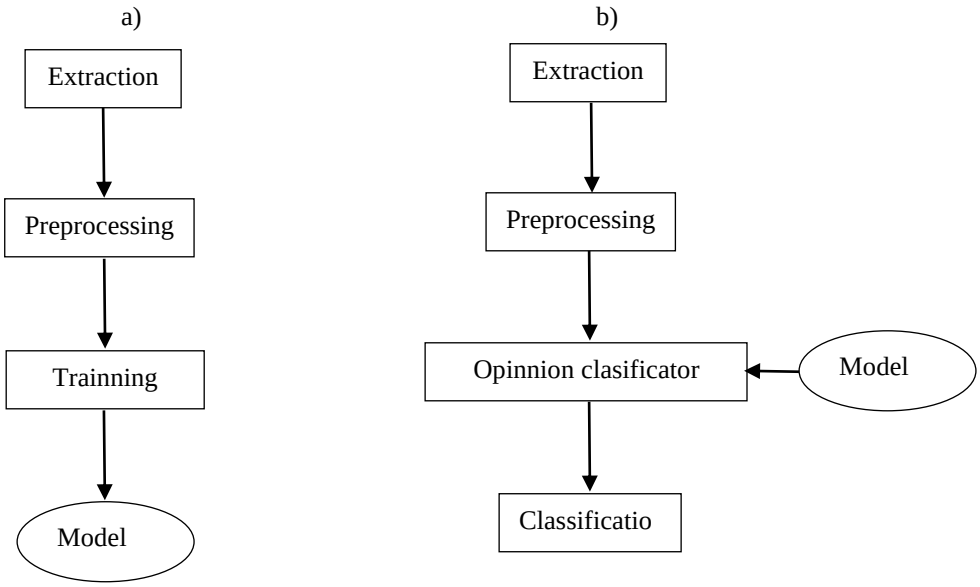

Fig. 3. Methodology for the classification of opinions a) Training b) Classification with the generated model.

3.2.1 Extraction

The data will read from csv file provided by Rest-Mex 2022 challenge (REST-MEX-Sentiment Analysis Task, n.d.), for this research the opinion and the polarity will be take into consideration.

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.).

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 Figure 4 shows the whole process using the aforementioned library with *Snowball Stemmer* .

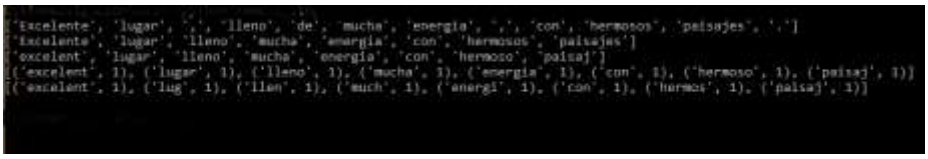

Fig. 4. Pre-processing using the Python NLTK library.

3.2.3 Training

The dataset was split randomly in cross validation 75% of the data for training and the rest 25% for test. The training data will be used for SVM (Suthaharan, 2016), Naïve Bayes (Murphy, 2006), Decision Tree (Myles, Feudale, Liu, Woody, & Brown, 2004) and CNN to generate a model, which will be loaded into the classification.

A CNN uses a sequential order of layers, with 5 layers, first Conv1D with 64 filters and a filter size of 5, then a MaxPooling1D layer with a pool size of 4, the next layer is a Flatten layer, the fourth layer is a Dense with 64 units and *ReLu* activation function, finally another Dense layer with 5 units and *softmax* as activation function.

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 Flores Limaylla & Peña Álvarez (2020). 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 and CNN 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:

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

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

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

$$P = \frac{TP}{TP + FP} \tag{2}$$

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

$$R = \frac{TP}{TP + FN} \tag{3}$$

- $F_{1:}$ is the harmonic mean that combines the precision and accuracy values as shown in equation 4.

$$F_1 = \frac{2 * P * R}{P + R} \tag{4}$$

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 CNN, SVM, *Naïve Bayes* and Decision Trees algorithms.

4.1 Dataset

The data used was obtained from the TripAdvisor.com page between 2002 to 2021 from tourist opinions to the most representative places, hotels, and restaurants, each opinion has polarity from 1 to 5 as we described before. This dataset was provided by REST-MEX 2022 for Sentiment Analysis Task.

Table 1 shows the total opinions classified by classes from 1 (very negative) to 5 (very positive), out of a total of 30210 opinions retrieved from the aforementioned challenge.

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

	1	**2**	**3**	**4**	**5**
Opinions	547	730	2121	5877	20935

4.2 Results

With the CNN, SVM and *Naïve Bayes algorithms* presented in section 3 and adding decision trees, the training and classification are carried out. Table 2 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. For the baseline (Carmona-Sanchez, 2021) proposed a model that obtained the 10th place in Rest-Mex 2021.

Table 2. Results of the experiment using the 4 algorithms mentioned.

Algorithms	**Accuracy**	**Precision**	**Recall**	**F$_1$ _**
CNN	98.83	0.98	0.98	0.98
Naïve Bayes	70.03	0.84	0.70	0.76
SVM	71.82	0.81	0.72	0.75
Decision tree	62.54	0.64	0.63	0.63
Baseline	45.71	-	-	0.17

5 Conclusions

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). Although SVM (71.82%) and *Naïve Bayes*

(70.03) are the most used in the consulted research the best performance was using a CNN with an accuracy of 98.83%, all the algorithms are above the baseline created. We can see that the decision tree algorithm does not have a good performance (62.54%) 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 but the use of Convolutional Neural Network increase the prediction of the polarity, so it is proposed as future work to enrich the dataset and use RESNET which is a CNN in residual neural network this algorithm achieve significant improvements [4]. In addition to implementing other artificial intelligence techniques such as Extreme Gradient Boosting (XGBoost), Random Forest and others with great performance reported in the literature.

References

Amores Fernández, M. A. (2016). Detection of the polarity of opinions based on new lexical resources. Central University "Marta Abreu" of Las Villas.

Carmona-Sanchez, G. (2021). Naive Features for Sentiment Analysis on Mexican Touristic Opinions Texts.

Chen, W., Xu, Z., Zheng, X., Yu, Q., & Luo, Y. (2020). Research on Sentiment Classification of Online Travel Review Text. *Applied Sciences*, 10, 5275. https://doi.org/10.3390/app10155275

Diaz-Garcia, J., Ruiz, D., & Martin-Bautsta, M. (2018). Unsupervised Opinion Mining on Twitter. In XVIII Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2018). Granada, Spain.

Filieri, R., Acikgoz, Y., Ndou, V., & Dwivedi, Y. (2020). 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.

Flores Limaylla, G. I., & Peña Álvarez, E. P. (2020). Automatic learning for the optimization of digital marketing processes in the tourism sector.

Gandhi, U. D., Kumar, P.M., Babu, G.C. & Karthick, G. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). Wireless Personal Communications. https://doi.org/10.1007/s11277-021-08580-3.

Ghanem, J. (2017). Conceptualizing “the Tourist”: A critical review of UNWTO definition.

Han, H., Bai, X., Liuet, J. (2018). Attention-based ResNet for Chinese Text Sentiment Classification. In Proceedings of the 2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018) (pp. 495–499). Atlantis Press. https://doi.org/10.2991/csece-18.2018.108

Hariguna, T., Sukmana, H. T., & Kim, J. I. (2020). Survey Opinion using Sentiment Analysis. *Journal of Applied Data Sciences*, 1, 35-40.

INEGI. (n.d.). Sistema de Cuentas Nacionales de México. Turismo: Producto Interno Bruto turístico, Base 2013. Retrieved from https://www.inegi.org.mx/temas/turismosat/#Informacion_general

Kinstler, L. (2018). How TripAdvisor changed travel. The Guardian. Retrieved from https://www.theguardian.com/travel/2018/aug/17/how-tripadvisor-changed-travel

Martí, M., Taulé, M., Márquez, L., & Bertran, M. (2007). Cess-ece: annotated corpus of Spanish and Catalan. Roman Arena: A New Nordic, *Journal of Romance Studies*, 1.

Moreno Ortiz, A., Pineda Castillo, F., & Hidalgo García, R. (2010). Hotel User Rating Analysis with Sentitext: a domain-independent sentiment analysis system.

Mostafa, L. (2020). Machine learning-based sentiment analysis for analyzing the travelers' reviews on Egyptian hotels. In Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 405-413). Springer.

Murphy, K. P. (2006). Naive Bayes Classifiers. University of British Columbia, 18.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18, 275-285.

Navas-Loro, M., & Rodríguez-Doncel, V. (2020). Spanish corpora for sentiment analysis: a survey. *Language Resources and Evaluation*, 54, 303-340. https://doi.org/10.1007/s10579-019-09458-1

Puh, K., & Bagić Babac, M. (2022). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights*. https://doi.org/10.1108/JHTI-02-2022-0078

Quiroga, M. A. R., Ayala, D. V., Pinto, D., Tovar, M., & Beltrán, B. (2016). Aspect-based sentiment analysis: A model to identify the polarity of user reviews. *Research in Computing Science*, 115, 171-180.

REST-MEX-Sentiment Analysis Task. (n.d.). Retrieved April 8, 2023, from https://sites.google.com/cicese.edu.mx/rest-mex-2022/tracks/sentiment-analysis-task

Romano, M., Contu, G., Mola, F. (2023). Threshold-based Naïve Bayes classifier. Advances in Data Analysis and Classification. https://doi.org/10.1007/s11634-023-00536-8

Shoeb, M., & Ahmed, J. (2017). Sentiment analysis and classification of tweets using data mining. *International Research Journal of Engineering and Technology (IRJET)*, 4.

Shujun, W., & Song, S. (2022). Sentiment Classification of Tourism Reviews Based on Visual and Textual Multifeature Fusion. *Wireless Communications and Mobile Computing*, 2022, 1-10. https://doi.org/10.1155/2022/9940817

Soumaya, O., Driss, M., Soufiane, A., Abderrahmane, D., & Mohamed, A. (2023). Customer Sentiment Analysis in Hotel Reviews Through Natural Language Processing Techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(1). http://dx.doi.org/10.14569/IJACSA.2023.0140162

Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 207-235). Springer.

Ticona Nina, R. (2019). Opinion mining based on supervised learning in the evaluation of tourist destinations in the Puno region.

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis in TripAdvisor. *IEEE Intelligent Systems*, 32, 72-77.

Valdivia, A., Luzón, M. V., & Herrera, F. (2017b). Sentiment analysis on TripAdvisor: Are there inconsistencies in user reviews? In International Conference on Hybrid Artificial Intelligence Systems (pp. 15-25). Springer.

Vazquez, K. L., Tovar, M., Vilariño, D., & Beltrán, B. (2016). An algorithm to detect the polarity of opinions in the domains of laptops and restaurants. *Research in Computing Science*, 128, 91-98.