A Proposal and Comparison of Supervised and Unsupervised Classification Techniques for Sentiment Analysis in Tourism Data

Antonio Rico-Sulayes[1], Julian Monsalve-Pulido[2]

[1]*Universidad de las Américas Puebla México*
[2]*Universidad Pedagógica y Tecnológica de Colombia*

Abstract

In recent years, the growth of digital communication has influenced, sometimes significantly, the way many industries and businesses evaluate and attempt to improve their performance. The tourism sector is not an exception to this trend. Social networks and travel planners have become main sources of data for sentiment analysis and prediction in this industry. Within the context of an international competition targeting these tasks with Spanish data, Rest-Mex 2022, we have implemented and compared techniques from two very different approaches, supervised and unsupervised learning. In our comparison, the latter has been proven better, but more importantly, the exercise has helped us identify specific areas in which the latter, much less commonly used, can be improved.

Keywords

Sentiment Analysis, Senticnet, Tourism.

1. Introduction

In the area of sentiment analysis [1], the way of studying the opinions expressed by users has changed over time, using quantitative or qualitative indicators that measure the quality of a product or service. Currently, to catalog something, information is extracted from large volumes of data generated by people freely and spontaneously [2][3]. The identification or detection of emotions or sentiments is a complex process that results from the analysis of physical and psychological reactions that are expressed through different behaviors, which may create ambiguous variables in natural language [4]. For Rest-Mex 2022 challenge, two different sentiment classification models were explored, a supervised and an unsupervised classification model. Within the context of this competition, this article is organized as follows. In section 2, relevant related research studies are presented. The following section describes the data of the competition. Section 4 presents the methodological approaches applied to respond to the challenge. Finally, the articles closes with a presentation of results and some conclusions.

IberLEF 2022, September 2022, A Coruña, Spain

2. Related Works

The main sources for sentiment analysis in the tourism sector are social networks and travel planners. In [5], the authors conducted a review of hospitality and tourism studies, analyzing social networks in order to collect, examine, summarize and interpret data derived from these sources. In their survey, they notice that there has been a growth in social network analysis, along with an increased use of multiple analytical methods, including precision tests. These findings suggest there is a significant commitment, along with a scientific drive, to perform data analysis of comprehensive and established social networks. The study points out the need for an expansion of approaches to include common analytical methods, such as text and sentiment analysis. In their systematic survey study, Padmaja and Sudha [6] explain the importance of targeting large amounts of data in tourism development in order to predict the future of the field and develop the potential to forecast its outcomes with the help of large databases and a broad range of machine learning techniques. Their research study focuses on the integration of big data and sentiment analysis with the help of machine learning techniques.

The two comprehensive studies just cited make emphasis in the need to widen the range of approaches and techniques that are being used in the various classification tasks related to the tourism industry [7]. Therefore, the current research article attempts to do this by bringing together two diverse classification frameworks, supervised and unsupervised learning, to respond to the second of three tasks in Rest-Mex 2022 challenge [8]. This task was also part of Rest-Mex 2021 challenge, in which it was the second of two tasks only [9, 10].

3. Description of the Data for Analysis

The data provided by the organizing team in their call for participation, REST-MEX: Recommendation System for Text Mexican Tourism [8], consists of 30,212 rows. This represented the training set, which corresponded to 70% of the original data set. The other 30% was used as a test set. Each row contained 4 columns:

1. Title: The title that the tourist himself gave to his opinion. Data type: Text.
2. Opinion: The opinion issued by the tourist. Data type: Text.
3. Polarity: The label that represents the polarity of the opinion. Data type: [1, 2, 3, 4, 5].
4. Attraction: The label of the type of place for which the opinion was issued. Data type: [Hotel, Restaurant, Attractive].

The polarity goes from 1, which means the highest degree of dissatisfaction, to 5, which is the highest degree of satisfaction. It can be interpreted in the following way: [1. Very bad; 2. Bad; 3. Neutral; 4. Good; and 5. Very good] [11].

Below, Figure 1 shows a general summary of the distribution of the data collected and provided as training data for the challenge. A few facts seem relevant regarding this distribution. Firstly, there is an uneven overall distribution for attraction type with 55% reviews for hotels, 28% reviews for restaurants, and 17% of the comments about attraction sites. Secondly, as mentioned in the data description, each of the records are manually classified by polarity according to intensity from 1 to 5, with the following distribution: Polarity 5 with 69%; Polarity 4 with 19%,

Polarity 3 with 7% and Polarity 1 and 2 with a polarity of 2% each. It is evident that the data has a strong lean towards positive polarity since 89% of the comments are with a positive intensity. Thirdly, the distribution of polarity according to the type of attraction is displayed, where the highest number of positive polarity is evident in hotel comments compared to the general distribution in attraction sites and restaurants.

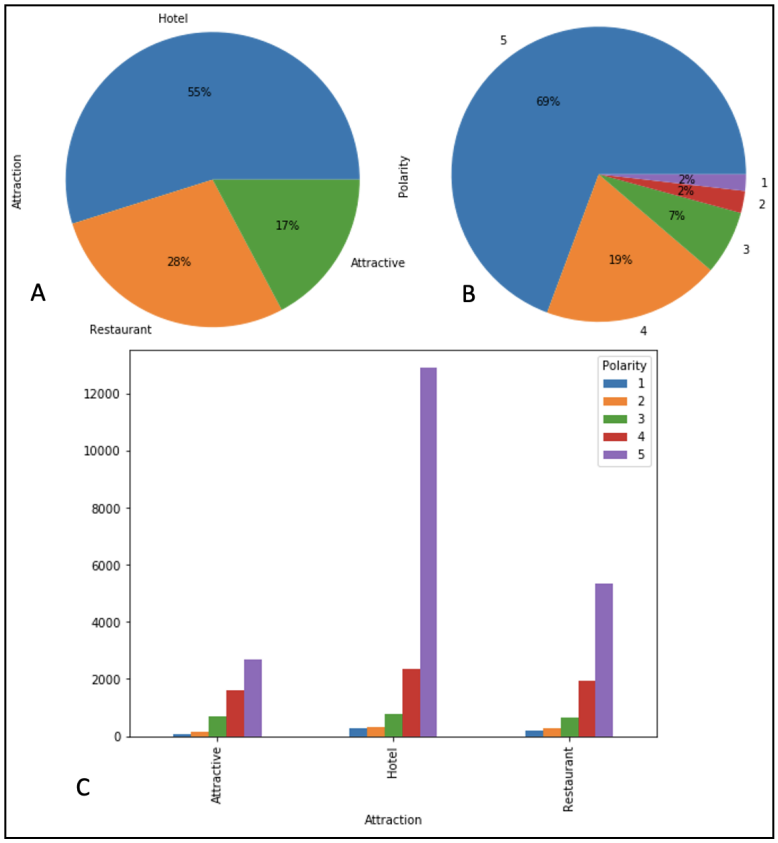

Figure 1: General data distribution

4. Methodology

For the participation in the Rest-Mex 2022 challenge, the strategy of applying two classification perspectives, supervised and unsupervised, had as its main goal to compare the classification results of these two approaches and analyze the most appropriate perspective for the identification of intensity polarities from 1 to 5 in the Spanish language in the context of tourists' opinions.

4.1. Supervised Classification

As it has been mentioned before, the current research article has explored and submitted both supervised and unsupervised solutions to the sentiment analysis task in the Rest-Mex 2022 challenge. In this subsection, we briefly present how the supervised solution submitted was produced.

For the prediction of polarity and type of attraction, we employed a Naïve Bayes Multinomial algorithm for both tasks. In order to choose this classifying algorithm, we performed various tests with different classifiers, which included the following methods: Naïve Bayes, Naïve Bayes Multinomial, SVMs, Multilayer Perceptron, Random Forests, and C4.5. We tested this particular set of algorithms, because we have attested their efficiency in very diverse forensic classifications tasks, such as authorship attribution [12], speaker identification [13] and bot detection [14].

As to the preprocessing of the training and testing sets, all titles and opinions were lowercased and punctuation was eliminated. For features, we used unigrams, bigrams and trigrams. We obtained these different length n-grams only from titles to predict polarity classes, and only from opinions to predict the type of attraction. Using a mutual information ranker on the testing data, we attempted to determine the optimal number of classifying features for a database of this size. By this means, we decided to use 2,000 features derived from titles and 5,000 features derived from opinions. As our supervised solution uses a Naïve Bayes Multinomial method, we will describe briefly the logic of this classifier.

When trying to determine the polarity or type of attraction (a class) for a particular title or opinion (an event), Naïve Bayes classifiers use information on what features (e.g., words or n-grams) appear or not in a post by some tourist or client (i.e., feature Boolean presence). Therefore, these classifiers function by predicting the likeliness each class has of producing a certain text [15]. In the context of predicting the number of stars a tourist attraction might get, we can imagine the following illustrative example. After looking at many clients' opinions with some known class (1, 2, 3, 4 or 5 *estrellas*, 'stars'), the algorithm may determine that if the word *excelente*, 'excellent', occurs in a new opinion, the probability of it being in the *5 estrellas*, '5-star', class will be higher and the chances of it being in the *1 estrella*, '1-star', class will be lower.

Formally, the probability of seeing the word *excelente*, as a classifying feature f, when exposed to a document from some class c, $P(f|c)$, is computed in a Naïve Bayes Multinomial classifier as the number of documents (if Boolean presence is used) with feature f among the ones that belong to class c, $f^{in}c$, divided by the number of instances of all features (F) in all the documents (D) in the training data. Namely,

$$P(f \mid c) = \frac{f^{in}c}{\sum_{f \epsilon F} \sum_{d \epsilon D} f^{in}d} \tag{1}$$

4.2. Unsupervised Classification

The unsupervised classification was carried out through a strategy of using a linguistic resource, with the aim of calculating the polarity of an opinion in an intensity spectrum according to the scale of the data to be analyzed. Senticnet 5.0 [16], developed at Nanyang Technological

University Singapore, aims to combine symbolic and sub-symbolic artificial intelligence to automatically discover primitive, conceptual concepts from the text and link them to common sense concepts and named entities in a new representation.

The total polarity (TP) of an opinion is given by:

$$TP = \sum_{i=1}^{N} \frac{Pt_i + Pc_i}{N} \tag{2}$$

Where Pt is the polarity of the title, Pc is the polarity of the comment and N the number of variables identified in the comment.

The total value of the polarity is assigned a value from 1 to 5 depending on the intensity, as shown in Figure 2.

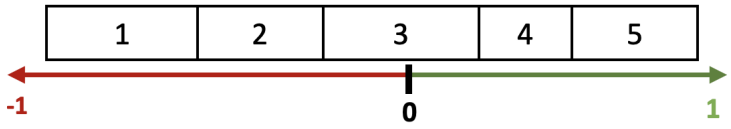

Figure 2: Polarity value

The value is given in the ranges from -1 to 1 according to classification, a scale of 1 = value $(\geq -1 \wedge < -0.5)$; 2 = value $(\geq -0.5 \wedge < -0.1)$; 3 = value $(\geq -0.1 \wedge < 0.1$; 4 = value $(\geq 0.1 \wedge < 0.5)$; and 5 = value $(\geq 0.5 \wedge \leq 1)$.

5. Analysis Results

For the validation of the classification algorithms, the challenge proposed two general metrics. In the first one, polarity is defined as a natural number in a range (1, 5), which is evaluated with MAE. In the prediction of type of attraction, there are 3 classes (attractive, hotel and restaurant), for which the Macro measure F is applied (3). For the second general metric, the final measure of the task, the average of the inverse of MAE and the Macro F1 are combined as shown in (4).

$$MacroF1_k = \frac{F1_A(k) + F1_H(k) + F1_R(k)}{3} \tag{3}$$

$$Sentiment_{Res_k} = \frac{\frac{1}{1+MAE_k} + MacroF1_k}{2} \tag{4}$$

Using the final classification results in the challenge, the Accuracy was 67.61% for the supervised classification and 58.64% for the unsupervised classification. It was evident that the supervised classification obtained a better Accuracy due to the training carried out on the model with the same data to be classified.

In a comparison of the two classification perspectives, Figure 3 shows that the precision of classes 1, 2 and 3 were not adequately classified through the unsupervised process. On the other hand, it is also evident that classes 4 and 5 have a similar outcome in the two types

of classification. This comparison clearly points out that the polarity intensity values in the unsupervised process should be improved by refining the classification in the lower end of the range from 1 to 5, and this problem was partly derived from the scarcity of training data in that end of the spectrum.

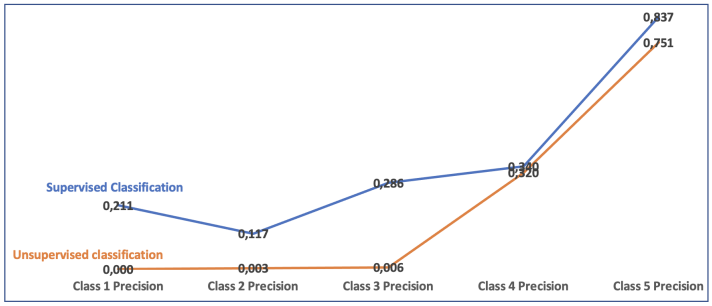

Figure 3: Class precision comparison

In the prediction classification of the three classes of the type of tourist comments (Attractive, Hotel, and Restaurant), better results were obtained in the supervised classification process, with an Accuracy of 96.5% against 47.43% of the unsupervised classification. It is evident that the linguistic resource that was used for the unsupervised classification is not adapted correctly for classifications other than feelings.

6. Conclusions

The application of two types of classification perpectives for sentiment analysis in Spanish in a tourism context, with data provided in the challenge, helped us identify the advantages and disadvantages that can be found in the polarity classification process, especially with values of intensity. Some linguistic resources available in the literature can be used, but it is necessary to adapt them to the classification context, using hybrid techniques through the use of supervised models.

In the task of type of attraction detection, the supervised model is clearly more effective with an Accuracy of 96.5% compared to the unsupervised model, which obtained 47.43%. In general, the supervised models can be better in this type of classification tasks, but the computational expense and demands for the training process remains a challenge.

A simplistic interpretation would lead us to conclude that supervised methods are better at these types of tasks and there is no need to expand the work on unsupervised learning techniques in this context, but the situation is much more complex than that. While we were testing our supervised models with the training data, we noticed that if we used the entire data set provided in the 2022 challenge, we obtained certain performance, but the accuracy figures dramatically deteriorated if we changed the data distribution by, for example, making the data set more balanced. In that sense, we knew that if the testing data had a very different distrubutional pattern among its classes, our results for the supervised model would fall significantly. At the end, the testing data turned out to be very similar in its distribution to the training data and

our results for the supervised model were very close to what we had predicted. However, this means that supervised models, although very flexible and adaptable to different tasks, tend to get overfit easily. Unsupervised models should be more resistant to this overfitting due to their independence from training data. This is just one example, among various unsupervised learning advantages, which still make it worth developing.

References

[1] E. Cambria, A. Hussain, Sentic computing, marketing 59 (2012) 557–577.

[2] Y. Susanto, A. G. Livingstone, B. C. Ng, E. Cambria, The hourglass model revisited, IEEE Intelligent Systems 35 (2020) 96–102. doi:https://doi.org/10.1109/MIS.2020.2992799.

[3] C. Salazar, J. Aguilar, J. Monsalve-Pulido, E. Montoya, Análisis de sentimientos/polaridad en diferentes tipos de documentos, Revista Ibérica de Sistemas e Tecnologias de Informação E41 (2021) 344–357.

[4] M. Minsky, The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind, Simon and Schuster, 2007.

[5] F. Mirzaalian, E. Halpenny, Social media analytics in hospitality and tourism, Journal of Hospitality and Tourism Technology (2019). doi:https://doi.org/10.1108/JHTT-08-2018-0078.

[6] N. Padmaja, T. Sudha, A systematic review of application of big data analytics in tourism sector, Journal of Computational and Theoretical Nanoscience 16 (2019) 1832–1838. doi:https://doi.org/10.1166/jctn.2019.8107.

[7] M. Á. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Science (2022). doi:https://doi.org/10.1177/01655515221100952.

[8] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[9] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[10] M. A. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodrıguez, A. Y. Rodrıguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, Computación y Sistemas 26 (2022). doi:https://doi.org/10.13053/CyS-26-2-4055.

[11] R. Guerrero-Rodriguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanaju-

ato, mexico, Current Issues in Tourism (2021) 1–16. doi:https://doi.org/10.1080/13683500.2021.2007227.

[12] A. Rico-Sulayes, Reducing vector space dimensionality in automatic classification for authorship attribution, Revista Científica de Ingeniería Electrónica, Automática y Comunicaciones 38 (2017) 26–35.

[13] C. A. Cervantes Méndez, A. Rico-Sulayes, Sing a high-level feature annotated corpus for speaker recognition: A mixed approach with text classification techniques, Entorno UDLAP. Revista de conocimiento e Innovación 13 (2021) 50–61.

[14] M. J. D. Torres, A. R. Sulayes, Detection of bot accounts in a twitter corpus: Author profiling of social media users as human vs. nonhuman, Lengua y Habla 25 (2021) 76–86.

[15] A. Rico-Sulayes, Authorship Attribution on Crime-Related Social Media: Research on the darknet in forensic linguistics, Aracne, 2018.

[16] E. Cambria, S. Poria, D. Hazarika, K. Kwok, Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings, in: Proceedings of the AAAI conference on artificial intelligence, 2018.