The Role of the Topics for the Sentiment Analysis Task on a Mexican Tourist Collection

Juan David Jurado-Buch[1], Lázaro Bustio-Martínez[2] and
Miguel Ángel Álvarez-Carmona[3,*]

[1]*Servicio Nacional de Aprendizaje Centro Sur Colombiano (SENA), Nariño, Colombia.*

[2]*Universidad Iberoamericana, Ciudad de México (IBERO), 01219 CDMX, Mexico*

[3]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad de Transferencia Tecnológica Tepic (CICESE-UT3), 63173, Nayarit, Mexico.*

Abstract

This paper presents an approach to determine the polarity and attractive type from Mexican opinions to participate in the Rest-Mex 2022 evaluation forum. The purpose of the task is to determine satisfaction level (1-5) of tourists and determine the places that they visited. The proposed approach consists of extract a list of topics of the opinions. For this, we apply the LDA algorithm. With this approach, a weighted average of 0.80 was obtained, which is considerably higher than the baseline proposed by the organizers, which is 0.45.

Keywords

Sentiment analysis, Mexican Spanish, LDA, Rest-Mex.

1. Introduction

Due to the covid pandemic, various economic sectors were severely damaged. One of these sectors was Tourism [1][2][3][4]. As a result, in 2021, the Rest-Mex Evaluation forum emerged, whose purpose is to generate data on Mexican tourism and share it to motivate the international scientific community to work on this sector [5, 6].

In this second edition of Rest-Mex, the organizers proposed, among other tasks, Sentiment analysis [7], which consists in determining two labels given an opinion. First, it is needed to determine the polarity of the opinion, which is expressed as a number between 1 and 5, where 1 represents the lowest level of polarity and 5 is the highest level. The task for obtaining the second label consists in determining the type of place that tourists visited, which can be a Hotel, a Restaurant, or a Tourist Attraction.

One of the main tools that can help extract and represent the essential ideas of the opinions is Topic Extraction [8]. In this work we explore the hypothesis that finding the hidden topics in the collection of sentiment analysis will helps to represent and predict the polarity and the place for the Rest-Mex 2022 collection.

IberLEF 2022, September 2022, A Coruña, Spain

malvarez@cicese.edu.mx (M. Álvarez-Carmona)
0000-0003-4421-5575 (M. Álvarez-Carmona)

CEUR Workshop Proceedings (CEUR-WS.org)

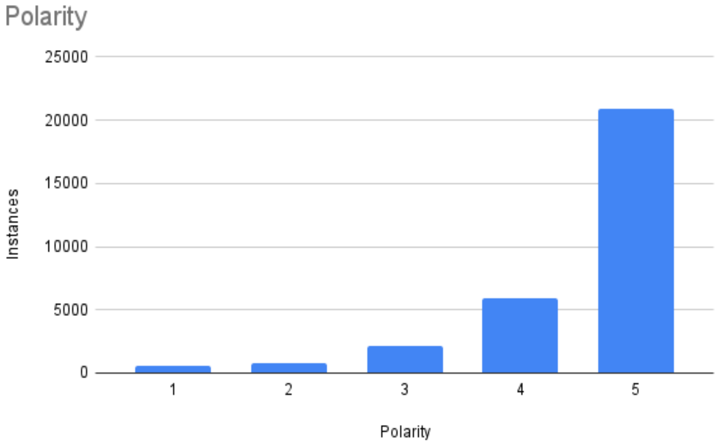

Figure 1: Distribution of the class for polarity

The rest of the document is organized as follows; In section 2, the methodology followed in this work is described. In section 3, the results and their analysis are presented. Finally, section 4 presents the conclusions of this work.

2. Methodology

This section presents the database that we use for demonstrating our hypothesis, and the proposed methodology for representing the touristic opinions. Also, the methodology is described.

2.1. Sentiment Analysis Collection

The training collection used for this task was built and shared by the Rest-Mex 2022 organizers [7] and consists of a total of 30,211 opinions. Each opinion contain a tittle, the opinion itself and the labels of polarity and attraction.

As it can be seen, the Figures 1 and 2 show the instances distribution for polarity and attraction label. For both labels, the distribution of classes is very unbalanced. This makes the task more complicated.

2.2. Pre-processing

In order to extract the most important characteristics of the texts, a pre-processing phase was applied. The transformations that were carried out are:

1. Uppercase was converted to lowercase.
2. Stop-words were removed.
3. Punctuation marks were removed.
4. The digits were replaced by the letter 'd'.

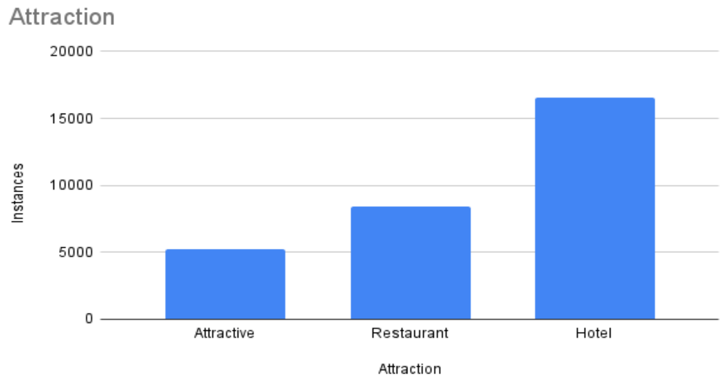

Figure 2: Distribution of the class for polarity

5. To avoid influencing the dates, the words for the months of the year were also removed.
6. Stemming was applied to the tokens in the texts.
7. Removed tokens that appear less than 50 times in the entire collection.

2.3. Topic Modeling on the Sentiment Analysis

For this task, the idea is to represent the opinions in order to classify the polarity and the attraction visited. For this work, we propose to extract the hidden topics to build a vector because, in the context of text modeling [9][10], the topic probabilities provide an explicit representation of a document [8], and this could be a good representation for the sentiment analysis task.

For this propose, we use the well-known Latent Dirichlet Allocation (LDA) method. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities [8].

The firs step for the representation proposed is to extract the LDA topics from the news. For this, we use the implementation of Gensim library in Python 3.0. We set the number of topics (k) in 100 [1].

It is proposed to use a simple deep neural network to decide both polarity and attractiveness. In particular, a neural network with 10 hidden layers ensures that the best relationship is found between the outputs of the participating systems and the real class of each instance. Table 1 summarizes the principal characteristics of the Deep Learning algorithm. This architecture was chosen due to works as in [1] have achieved good results with this proposal.

[1] Empirically selected

Table 1
Characteristics of the applied Deep Learning algorithm

Hidden Layers	10
Neurons per layer	1000
Activation function	Relu
Neurons of the final layer	5 (for polarity) or 3 (for attraction)
Final layer	Softmax
Loss function	Categorical Cross Entropy
Optimizer	Adam
Epochs	50

Table 2
Training data set results

Task	Acc	Macro-F	F 1	F 2	F 3	F 4	F 5	MAE	Measure
Polarity	64.52	0.33	0.16	0.16	0.22	0.31	0.79	0.47	
Task	**Acc**	**Macro-F**	**F Attractive**	**F Restaurant**	**F Hotel**	-	-	-	
Attraction	88.92	0.87	0.88	0.83	0.91				
Measure									0.77

3. Experiments and results

In this Section, the obtained results are presented. First, the results derived from the training corpus are displayed. Subsequently, the official data published by the organizers of Rest-Mex 2022 are presented.

3.1. Training data set results

10-fold cross-validation was performed for the training corpus to obtain the results. In addition, the final model was generated with all the training data to send the results and participate in the Rest-Mex 2022 forum.

Table 2 shows the results obtained for the training set. The first row shows the results for the polarity detection. In this row, we can see that the MAE result is 0.47. Also, the proposal achieves 0.33 F-measure and 64.52 Accuracy. The best-class label for the different classes is for the polarity 5, as this is the majority class. On the other hand, the worst classes are 1 and 2 with 0.16. Also, the reason is that they are the minority classes.

For the attraction identification, the results are higher. F-measure achieves 0.87 and 88.92 of accuracy. The hotel class is the highest value with 0.91.

For this edition, the authors propose a measure to evaluate the sentiment analysis task. This measure is defined as shown in the equation 1.

$$measure = \frac{\frac{1}{1+MAE_p} + F_A}{2} \tag{1}$$

With the equation 1, the final measure is 0.77.

Table 3
Important topics for sentiment analysis

Polarity	Attraction
present	com
tiemp	cen
habit	plat
vent	postr
dia	menu
histori	hotel
visit	servici
muse	alberc
guanajuat	personal
conoc	buen
com	vist
cen	jardin
plat	plaz
postr	entrar
menu	cercan

Table 4
Test data set results

Task	Acc	Macro-F	F 1	F 2	F 3	F 4	F 5	MAE	Measure
Polarity	65.28	0.31	0.18	0.06	0.20	0.29	0.80	0.47	
Task	**Acc**	**Macro-F**	**F Attractive**	**F Restaurant**	**F Hotel**	-	-	-	
Attraction	93.11	0.92	0.94	0.88	0.95				
Measure									0.80

Table 3 shows the three most important topics according to the measure of information gain, both for polarity and attraction. For polarity, the most important topic refers to sales, time, and gifts. The second most important topic addresses words concerning cultural tourism such as museums, history, and knowing. The last topic talks about food.

On the other hand, for attraction, it can be seen that the first topic talks about food, possibly alluding to restaurants, and the second topic covers hotel services. Finally, the third topic talks about the description of places. The topics seem to represent the different attractions in the database.

3.2. Test data set results

Table 4 shows the results obtained for the test corpus for the proposal of this work. These results are also shown on the official Rest-Mex site [2]. In these results, it is possible to observe that the results are higher than the training results, except for class 2 for polarity. This indicates that the proposal does not tend to overfit.

Finally, Table 5 shows the result obtained by our proposal compared to the rest of the best

[2] https://sites.google.com/cicese.edu.mx/rest-mex-2022/results

Table 5
Final Rest-Mex Rank

Team	Measure
UMU-Team-Run-1	0.89
UC3M-Run1	0.89
CIMAT MTY-GTO-Run1	0.88
SENA-Team	0.80
Majority Class (Baseline)	0.45

three participants of the 2022 edition of Rest-Mex. This result is interesting, considering that, although a deep neural network was used for the final classification, it has highly comparable results with more complex architectures such as Transformers.

4. Conclusions

This work was proposed within the framework of Rest-Mex 2022 for the sentiment analysis track.

The proposal consists of representing the opinions of Mexican tourists through hidden topics to determine the polarity and type of attraction for each instance.

The results show that this idea has comparable results with more complex systems despite its simplicity. One of the advantages of this approach is that there is no overfitting, and even the results for the test partition are higher than those computed for the train partition.

References

[1] 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.

[2] M. Á. Alvarez-Carmona, R. Aranda, et al., Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022). doi:https://doi.org/10.1590/SciELOPreprints.3834.

[3] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: A study case in mexico, in: Mexican International Conference on Artificial Intelligence, Springer, 2021, pp. 184–195. doi:https://doi.org/10.1007/978-3-030-89820-5_15.

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

[5] 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.

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

[7] 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).

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[9] M. A. Alvarez-Carmona, A. P. López-Monroy, M. Montes-y Gómez, L. Villasenor-Pineda, H. Jair-Escalante, Inaoe's participation at pan'15: Author profiling task, Working Notes Papers of the CLEF (2015) 103.

[10] M. E. Aragón, M. A. A. Carmona, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets., in: IberLEF@ SEPLN, 2019, pp. 478–494.