A Single Model Based on Beto to Classify Spanish Tourist Opinions Through the Random Instances Selection

Juan David Jurado-Buch[1], Ever Sebastian Minayo-Díaz[1], Jhony Alexander Tello[1], Kaily Estefanía Chaucanes[1], Laura Valentina Salazar[1], Mauricio Daniel Oquendo-Coral[1] and Miguel Ángel Álvarez-Carmona[2,3,*]

[1]*Servicio Nacional de Aprendizaje Centro Sur Colombiano (SENA), Nariño, Colombia*
[2]*Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, Mexico*
[3]*Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), CDMX, Mexico*

Abstract

In this study, we developed a single Beto classification model capable of predicting 45 different classes that describes the polarity, type, and country of tourist opinions in Spanish. In addition, we proposed a novel function to balance the imbalanced database of tourist opinions, allowing us to achieve better results with reduced data. Specifically, we show that using only 27 % of the total training data leads to better results compared to using the entire dataset, with even competitive results obtained using just 2 % of the data. Notably, our proposed method achieved a top 8 ranking at the Rest-Mex 2023 forum. Overall, our results highlight the effectiveness of our proposed function in improving the performance of machine learning models trained on imbalanced datasets, especially in the context of tourist opinions.

Keywords

Rest-Mex, Sentiment Analysis, Beto, Type prediction, Country prediction, Mexican tourism

1. Introduction

Opinion classification is a major problem in the field of text mining [1, 2, 3]. In particular, in the field of tourism [4, 5], the classification of opinions can provide valuable information for decision-making in the tourism industry and improve the experience of tourists [6]. However, this problem is complex, especially when you have multiple classes and want to predict different aspects of sentiment.

As a result, the Rest-Mex initiative emerged [7]. Rest-Mex is an international evaluation forum specialized in Natural Language Processing applied to the tourism sector.

For the 2023 edition [8], the organizers have proposed an extension to the sentiment analysis task that has been developed since 2021 [9]. On this occasion, the task is to predict, given an opinion of a tourist place, the:

1. Polarity of the opinion: an integer value between 1 and 5.
2. Type of tourist place: it can be an attraction, a hotel, or a restaurant.

IberLEF 2023, September 2023, Jaén, Spain
miguel.alvarez@cimat.mx (M. Á. Álvarez-Carmona)

3. Country of the tourist place: it can be Mexico, Cuba, or Colombia.

This year's train data collection for the polarity task has more than 250,000 opinions and the test one with more than 100,000. Typically, 3 different models would be built with the data, one for each trait to be predicted (polarity, type, and country). However, with a collection in the order of hundreds of thousands of data, the training and test phase for 3 different models could be very slow [10, 11, 12].

In this work, we focus on the construction of a textual classifier of tourism opinions of 45 classes, where the polarity, the type of place, and the country of origin of the tourist place are predicted. One of the most important challenges when building a textual opinion classifier is the imbalance in the distribution of classes [13]. In many cases, some classes have too few instances, which can negatively affect the classifier's ability to correctly identify those classes in a single model.

Another important problem is class imbalance. In this way, it is necessary to apply an instance selection method to the data [14].

To address this problem, we propose a random instance selection method to balance data. The approach consists of randomly selecting an equal number of instances of each class in the training set. The idea is that, by balancing the number of instances of each class, the classifier has more opportunities to learn the characteristics of the less represented classes and, therefore, improves its generalizability [15, 16].

To assess the effectiveness of our method, we conducted experiments using a Rest-Mex dataset. The results obtained show that our random instance selection method improves the evaluation of the classifier for all classes, especially those with a small number of instances.

In summary, our work presents a solution to the instance selection challenge in the context of multi-class tourism opinion classification. The results obtained suggest that our method can be useful in different opinion analysis applications in the tourism industry.

The rest of the paper is organized as follows: Section 2 describes the Rest-Mex corpus for the 2023 edition. Section 3 shows the methodology proposed in this work. 4 shows the results obtained with this proposal. Finally, 5 lists the conclusion of this work.

2. Dataset

The train collection built by the organizers of Rest-Mex 2023 consists of 251,702 opinions from TripAdvisor.

For the Polarity classification, there are 5 classes, where class 1 represents the worst polarity and 5 is the best polarity. In Figure 1 it is possible to observe the distribution of these classes. This figure shows a clear imbalance. To measure this imbalance we use the same method used in [17]. Following this same method, an imbalance value of 56,896.36 is obtained.

To determine the Type of place there are 3 classes: Attractive, Hotel, and Restaurant. Figure 2 shows the distribution of this trait. In this case, there is not an imbalance as marked as for polarity, however, it is possible to appreciate that there is no balancing. Its imbalance value is 19,864.79.

Finally, to classify the Country of origin of the place that the tourist visited, there are 3 classes in the collection: Mexico, Cuba, and Colombia. In Figure 3 its distribution is presented. The

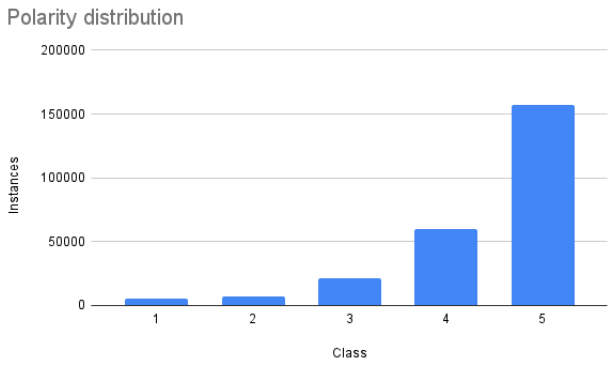

Figure 1: Polarity distribution

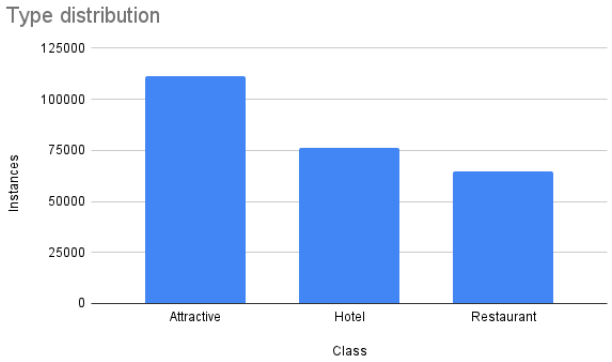

Figure 2: Type distribution

degree of imbalance of the Country trait is 24,661.36.

In this way, it can be seen that the trait with the least imbalance is that of Type, and that of polarity is significantly greater than the other two.

3. Proposed Methodology

The proposed methodology consists of 3 key steps. First, combine the classes of the 3 traits (Polarity, Type, and Country) to generate a single model. Second, generate a function that returns the number of instances to choose for trying to balance the database. Finally, classify the data.

Each of the 3 steps is described below.

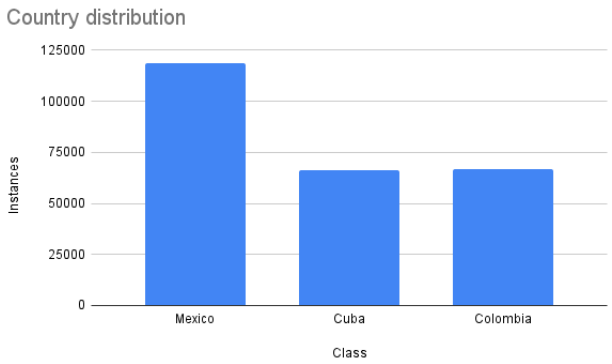

Figure 3: Country distribution

3.1. Combining classes

As mentioned in the 1 section, the idea of this work is to generate a single model to classify the 3 traits involved in the Rest-Mex sentiment analysis task. The reason is to propose a simpler system that has acceptable results than 3 different models.

For this, it is proposed to generate all the possible combinations for the 3 different traits of the collection. This is combining the 5 Polarity classes, the 3 Type classes, and the 3 Country classes. Given these numbers, there are 45 different combinations.

Table 1 shows the 45 possible classes ordered by the number of instances each combination has.

Table 1
45 different classes for a single trait in the Rest-Mex training corpus.

Class	Instances	Class	Instances	Class	Instances
5-Attractive-Mexico	29353	4-Restaurant-Cuba	4979	2-Restaurant-Mexico	959
5-Attractive-Colombia	29175	3-Hotel-Cuba	3904	2-Restaurant-Cuba	775
5-Hotel-Mexico	24787	3-Attractive-Mexico	3333	2-Attractive-Mexico	706
5-Restaurant-Mexico	23378	3-Hotel-Mexico	3297	4-Restaurant-Colombia	630
5-Attractive-Cuba	12929	4-Hotel-Colombia	3218	1-Restaurant-Cuba	598
4-Attractive-Mexico	10979	3-Restaurant-Mexico	2654	2-Hotel-Colombia	497
5-Restaurant-Cuba	10750	3-Attractive-Cuba	2602	1-Attractive-Mexico	484
4-Attractive-Colombia	10542	3-Restaurant-Cuba	2304	2-Attractive-Cuba	430
5-Hotel-Colombia	9428	3-Attractive-Colombia	2148	1-Hotel-Colombia	337
5-Hotel-Cuba	8891	2-Hotel-Cuba	1682	2-Attractive-Colombia	333
5-Restaurant-Colombia	8404	2-Hotel-Mexico	1455	1-Restaurant-Colombia	315
4-Attractive-Cuba	7832	1-Hotel-Cuba	1444	1-Attractive-Cuba	195
4-Hotel-Mexico	7654	1-Hotel-Mexico	1292	3-Restaurant-Colombia	166
4-Restaurant-Mexico	7485	3-Hotel-Colombia	1248	1-Attractive-Colombia	147
4-Hotel-Cuba	6908	1-Restaurant-Mexico	960	2-Restaurant-Colombia	115

In this way, it has a different distribution of a single trait with 45 different classes. However,

although it may be easier to classify instances with a model of 45 classes than 3 models of 5, 3, and 3 classes respectively, the class imbalance problem has not yet been resolved. If we measure the class imbalance with the new collection of Table 1 we obtain a degree of 7,532.97. This degree is lower than that of the 3 individual traits, but some balancing is still necessary.

3.2. Instances selection

To make a balance of the data, it is proposed to make a random selection of the data.

The idea is to take the number of instances of a class j as parameter k as a reference and from this value, for all classes with a number of instances greater than k select randomly k instances.

To select the reference value, the equation 1 is proposed.

$$k(dataset, i) = instances\left(dataset, \frac{len(d(dataset))}{2^i}\right) \tag{1}$$

Where $d(x)$ is a function that takes a database x and returns an ordered list of the classes within x. $len(x)$ is a function that returns the length of a class list of x. $instances(x, y)$ is a function that returns the number of instances of the $y-$th class in an ordered list of classes in a data set x.

The parameter i is used to regulate the degree of balancing, if $i = 0$ then the class with the largest number of instances will be taken as a reference value, which would make the database maintain the same instances, if otherwise if i is large enough to take the minority class, all other classes would randomly select as many instances as the class with the fewest instances.

In the case of the distribution of Table 1, if $i = 6$ is taken, 115 instances of each class would be randomly selected. This would generate a new fully balanced database with 5175 instances.

It is proposed to experiment with all possible values of i, that is, with $i \in [0, 6]$ and with $i \in \mathbb{Z}$.

3.3. Classifier

To classify the data, it is proposed to use a transformer based on Bert but specialized for Spanish such as Beto.

Table 2 shows the features used in the implementation used for the classifier

Table 2
Parameters of the Beto classifier

Model	Beto-cased
Max length	64
Classes	45
Optimizer	Adam
Learning rate	$5e-5$
Steps	$1e-08$
Epochs	4

4. Results

To obtain the results within the training database, it is proposed to make a 70/30 partition for training and testing respectively. This partition was made respecting the distribution of the original partition for each value of *i*.

The idea is to be able to observe how the different values of *i* within the ordering affect the classification performance.

Table 3 shows the results for each value of *i*. It can be seen as the used percentage of the original database, with $i = 1$ drops to 27% of the original data. This value reaches a little more than 2 % for higher values of *i*.

It can be seen how the imbalance value decreases as the value of *i* increases. However, this is also reflected when the Accuracy of the models is calculated. If the entire database is used, values of more than 63 for Accuracy and 0.38 for F-measure are achieved. This difference is very likely due to the degree of imbalance in the data.

It should be noted that, when *i* takes the values 1 and 2, the F-measure rises to 0.41, although the Accuracy is better for $i = 1$. Even with $i = 3$ the result of F-measure is very close to that obtained with $i = 0$. This is an interesting result since with a small percentage of data it is possible to arrive at similar results. In addition, they are competitive results taking into account that there are 45 classes.

Finally, when $i > 4$ the difference in the results is larger.

Table 3
Class imbalances for different *i* values

i	Instances	Percentage	Class Imbalance	Accuracy	F-measure
0	251702	100.00	7,532.97	63.87	0.38
1	69908	27.77	858.26	49.62	0.41
2	25037	9.94	149.99	43.17	0.41
3	14258	5.66	51.12	39.34	0.37
4	7400	2.93	7.96	33.71	0.32
5	6583	2.61	4.71	33.98	0.34
6	5175	2.05	0.00	31.64	0.28

4.1. Test partition results

For this edition, the organizers of Rest-Mex propose some evaluation metrics that give greater weight to correctly classify the negative classes of polarity.

To assess the effectiveness of the polarity classifier, the organizers propose the equation 2. This metric gives the additive inverse of importance according to the percentage of instances of a class in the test collection.

$$Res_P(k) = \frac{\sum_{i=1}^{|C|}((1 - \frac{T_{C_i}}{T_C}) * F_i(k))}{\sum_{i=1}^{|C|} 1 - \frac{T_{C_i}}{T_C}} \tag{2}$$

To evaluate Type and Country traits, they propose the equations 3 and 4. These metrics are macro F-measures of each trait.

$$Res_A(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3} \tag{3}$$

$$Res_C(k) = \frac{F_{Mex}(k) + F_{Cub}(k) + F_{Col}(k)}{3} \tag{4}$$

Finally, to obtain a unique value per participant, they propose a combination of the results as indicated by the equation 5. It is important to mention that in the same way, greater weight is given to the result of polarity than to the other two traits.

$$Sentiment(k) = \frac{2 * Res_P(k) + Res_A(k) + Res_C(k)}{4} \tag{5}$$

This way of evaluating the results seems ideal for the method proposed in this work, since the main objective is to find the best possible result, balancing the corpus to a certain degree.

Figure 4 shows a summary of the results obtained in the test partition of the forum.

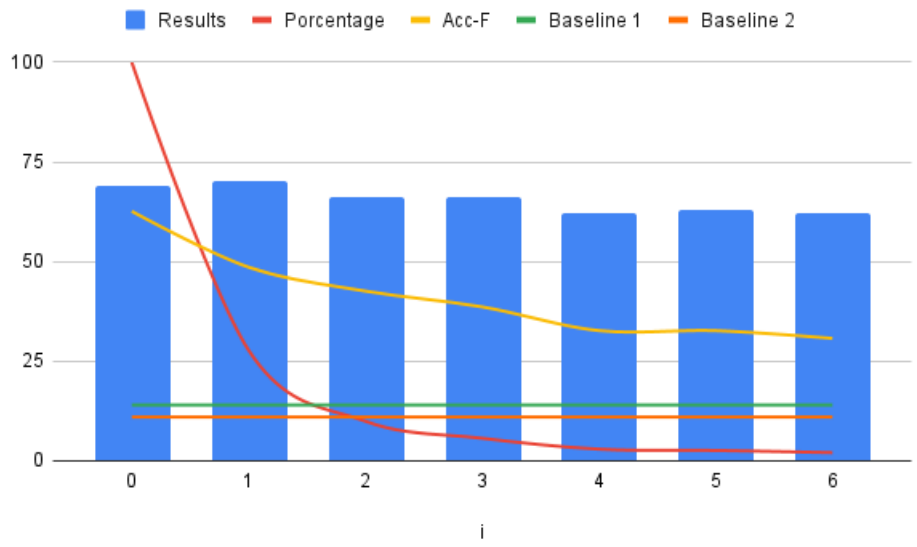

Figure 4: Results for different i values

The bars represent the value obtained with the equation 5 for each value of i.

As you can see, the best result obtained is 0.70, which is achieved when $i = 1$. That is when we worked with 27% of the training data. It is also important to note that when $i > 1$ the results remain competitive at the value obtained when $i = 0$, which is 0.69.

The worst result obtained is when $i = 6$ with 0.62, which is a competitive result considering that it only uses 2 % of the training data.

Within the figure, it can also see a red curve, which represents the percentage used for training. The yellow curve represents the difference between the Accuracy and the F-Measure, which tends to decrease as i is higher. The green line and the orange line represent the baselines proposed by the organizers.

Table 4 shows the results obtained for each trait. In this table, it can be seen that the best result for accuracy is obtained with $i = 0$ for the three traits, however, when $i = 1$ a better result is obtained for F-measure for Polarity, in addition to obtaining the same result for Type. Only the Country trait, obtains better results from F-measure when $i = 0$.

Table 4
Test partition results for different i values

i	**Final**	Acc_{Pol}	F_{Pol}	Acc_{Type}	F_{Type}	Acc_{C}	F_{C}
0	0.69	**70**	0.48	**97**	**0.97**	**90**	**0.90**
1	**0.70**	61	**0.50**	**97**	**0.97**	88	0.88
2	0.66	51	0.44	96	0.96	84	0.83
3	0.66	57	0.46	96	0.95	82	0.82
4	0.62	43	0.39	96	0.95	80	0.80
5	0.62	53	0.42	95	0.95	78	0.78
6	0.62	46	0.40	95	0.95	80	0.79

5. Conclusions

This work presents a single model for sentiment analysis. The data is labeled to classify the Polarity, Type, and Country from a tourist opinion. The main idea is to generate the 45 combinations to build a model from these classes.

Our main proposal is a function that selects a number of opinions to balance the data by randomly selecting instances from the database.

Experiments give evidence that using 27 % of the total training data generates equal or better results than using 100 %.

Also, it is concluded that it is possible to use only 2 % of the data and reach competitive results for Polarity, Type, and Country.

With this method, it was possible to obtain place 8 out of a total of 17 systems participating in Rest-Mex 2023.

Acknowledgments

The authors thank the Mexican Academy of Tourism Research (AMIT) for their support of the project "Creation of a labeled database related to tourist destinations for training artificial intelligence models for classifying relevant topics" through the call "I Research Projects 2022"

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] A. Diaz-Pacheco, M. Á. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, Journal of Experimental & Theoretical Artificial Intelligence (2022) 1–31.

[3] M. A. Alvarez-Carmona, R. Aranda, A. Rodriguez-Gonzalez, D. Fajardo-Delgado, M. G. A. Sanchez, H. Perez-Espinosa, J. Martinez-Miranda, R. Guerrero-Rodriguez, L. Bustio-Martinez, A. D. Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University-Computer and Information Sciences (2022).

[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 26 (2023) 289–304.

[5] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico, International Journal of Tourism Cities (2023).

[6] G. Carmona-Sanchez, A. Carmona, M. A. Alvarez-Carmona, Naive features for sentiment analysis on mexican touristic opinions texts., in: IberLEF@ SEPLN, 2021, pp. 118–126.

[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] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

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

[10] 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 103 (2015).

[11] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: a study case in mexico, in: Advances in Soft Computing: 20th Mexican International Conference on Artificial Intelligence, MICAI 2021, Mexico City, Mexico, October 25–30, 2021, Proceedings, Part II 20, Springer, 2021, pp. 184–195.

[12] L. Bustio-Martínez, M. A. Álvarez-Carmona, V. Herrera-Semenets, C. Feregrino-Uribe, R. Cumplido, A lightweight data representation for phishing urls detection in iot environ-

ments, Information Sciences 603 (2022) 42–59.
[13] M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain, volume 6, 2018.
[14] 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.
[15] M. A. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, L. Villaseñor-Pineda, Semantically-informed distance and similarity measures for paraphrase plagiarism identification, Journal of Intelligent & Fuzzy Systems 34 (2018) 2983–2990.
[16] M. E. Villa-Pérez, M. A. Alvarez-Carmona, O. Loyola-Gonzalez, M. A. Medina-Pérez, J. C. Velazco-Rossell, K.-K. R. Choo, Semi-supervised anomaly detection algorithms: A comparative summary and future research directions, Knowledge-Based Systems 218 (2021) 106878.
[17] 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.