Balancing of Tourist Opinions for Sentiment Analysis Task

Andrea Bethsabe García-Gutiérrez[1,*], Pablo Emilio López-Ávila[1], Pedro Adair Gallegos-Ávila [1], Ramón Aranda[2,3] and Miguel Ángel Álvarez-Carmona[1,3]

[1]Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, Nuevo León, Mexico, 66629
[2]Centro de Investigación en Matemáticas (CIMAT), Sede Mérida, Yucatán, Mexico, 97302
[3]Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), CDMX, Mexico, 03940

Abstract

This article presents a proposal for the treatment of an unbalanced tourist database with emphasis on minority classes for its classification, in this case, one based on BERT, called BETO. This methodology originally forms part of the thesis project of the authors, with the objective of balancing data with a tourist focus and being able to measure the impact that it has on the classification of texts.

Keywords

Unbalanced data, Oversampling, Subsampling, Tourism, Minority classes, BETO, Spanish

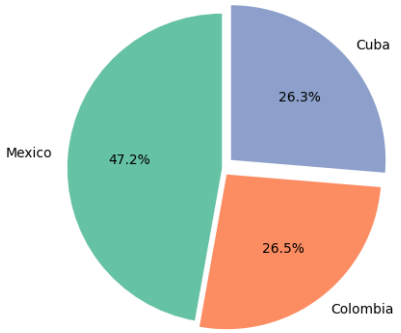
1. Introduction

The tourism sector contributes around 8% of the Gross Domestic Product (GDP) [1]. In addition, based on the Quarterly Indicators of Tourism Activity, tourism GDP in the fourth quarter of 2022 registered an increase of 1.3% compared to the third quarter of 2022, according to government estimates [2]. Interaction through social networks has increased in recent years and tourists are part of it. When a tourist stays in a hotel, visits a tourist attraction, or eats in a restaurant, they have the possibility of expressing a comment based on their experience, be it positive or negative [3, 4, 5, 6]. The analysis of these comments can help the owners or managers of the commented places to make decisions to improve the tourist experience [7].

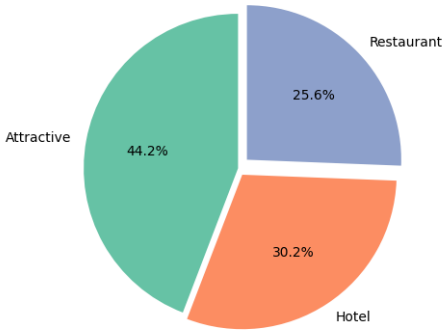
Since 2021, the Rest-Mex team has served as an evaluation forum that seeks to specialize in the analysis of texts from the tourism sector to solve different tasks in Mexican Spanish. In the 2023 edition, there are two tasks where the opinions of tourists, which are the object of analysis, obtained from the TripAdvisor site [8, 9, 10].

This article describes the participation of the Dataverse team in the sentiment analysis task. First, a **Dataset description** is shown, followed by **Proposed methodology** for class

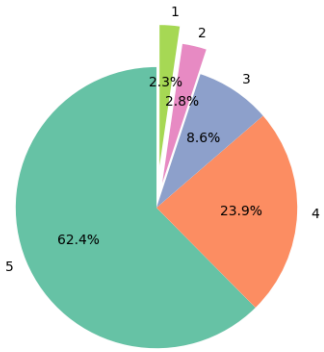
IberLEF 2023, September 2023, Jaén, Spain
*Corresponding author.
andrea.garcia@cimat.mx (A. B. García-Gutiérrez); pablo.lopez@cimat.mx (P. E. López-Ávila); pedro.gallegos@cimat.mx (P. A. G.); arac@cimat.mx (R. Aranda); miguel.alvarez@cimat.mx (M. : Álvarez-Carmona)
0009-0003-6264-3589 (A. B. García-Gutiérrez); 0009-0006-8214-5693 (P. E. López-Ávila); 0009-0004-9412-3980 (P. A. G.); 0000-0001-8269-3944 (R. Aranda); 0000-0003-4421-5575 (M. : Álvarez-Carmona)



(a) Distribution by “Country”



(b) Distribution by “Type”



(c) Distribution by “Polarity”

Figure 1: Distribution of the reviews according to the classes, according to the labels.

balance and their classification. Subsequently, the **Results** obtained are shown, ending with the **Conclusions and future work**.

2. Dataset description

There is a total of 251,702 tourist reviews obtained from TripAdvisor tagged as follows:

- Polarity: It is an integer between [1,5] where 1 is the most negative polarity and 5 is the most positive.
- Country: Represents the name of the country that was visited and is one of the following: Mexico, Cuba, or Colombia.
- Type: Represents the type of destination being reviewed and can be: Hotel, Restaurant, or Attractive.

The distribution of the data by class and by label type is depicted in Figure 1. As can be seen in Fig. 1b the Type reviews are not very unbalanced, for country reviews, Fig. 1a, there is a slight imbalance that can be dealt with. On the other hand, for polarity, Fig. 1c, there is a very

notorious imbalance, as seen in Fig. more than half of the reviews are from class 5, that is, very positive, in contrast to the data from class 1, very negative. and 2 negative, which present less than 5% of the reviews, and these represent the minority classes [11].

For the preprocessing part of the data, it was all lowercase; special characters, multiple spaces, numbers, stop words and words of length equal to or less than 3 were removed.

3. Proposed methodology

It is important to deal with minority classes, to address class imbalance to improve the performance of our model by avoiding possible bias, especially when there are highly imbalanced classes, because if the data set is biased towards one class, the model trained with the same data will be biased towards the same class[12].

There are several methods that can be used to address class imbalance. They are subsampling and oversampling. On the one hand, there is subsampling, which consists of reducing the data obtained by randomly taking reviews from the majority classes. On the other hand, there is oversampling, which is the increase in data, generally on the representation of minority classes where new instances can be generated or duplicate existing data of those classes randomly. In this last case, it has been found that it is not very efficient since new information is not introduced, it does not address the fundamental problem of the lack of information and real variability in this type of classes. However, generating new instances can alleviate this problem, but overfitting should be avoided[13].

In the case of subsampling, it is proposed to randomly obtain a certain number of reviews than the total of each one that coincides with the number of reviews of the "neutral" class, in this case reduce class 4 and 5. In the case of oversampling It is proposed to obtain the representative words of all classes with mutual information, which helps us to measure the dependence or association between two random variables. In this context, to measure the association between words within classes[14]. It can be defined as follows:

$$I(A, B) = log_2 \frac{P(A, B)}{(P(A)P(B))} \tag{1}$$

where $P(A, B)$ is the joint probability that word A and class B appear together. $P(A)$ and $P(B)$ are the marginal probabilities that they will appear in reviews. Mutual information measures the deviation of the joint probability of A and B from what would be expected if they were independent. A higher value of mutual information indicates a stronger association between the word and the class.

Once you have these words, ordered from the most representative to the least, synthetic opinions are generated, which consists of replacing one of the representative words that will generate a new review as a synonym from a Web dictionary or a FastText embedding.

The synonyms are obtained from the Word Reference virtual dictionary, and the embeddings from the Spanish version of FastText trained by Common Crawl, which is an organization that has been doing web scraping since 2008 and makes its data public, and with the free encyclopedia Wikipedia [15].

Having already the possible words that can be chosen for the generation of new instances, it should be considered that by having representative words of each class, if only those values

were taken, there could be a risk of bias and being over adjusted both in the classes. In general, as well as in the context of the data that is available, this could be observed using techniques that help us to lower the dimension of the entire data set and see the behavior of the new data. To mitigate the above, it is proposed to include the hyperparameters temperature and probability.

The generation function requires the following hyperparameters:

- The database of representative words
- The class in which to generate new data
- The maximum number of words that can be in the new data
- What other classes are taken into account for the generation
- Temperature
- Probability
- Synonym source type

First, the number of words that the text will have is randomly chosen, there is a minimum of 4 and a maximum of the total number of words that can be included, given as a hyperparameter.

The probability refers to whether or not one of the representative words is considered as a synonym for the generation of the new text. The procedure consists of first giving an integer and then choosing a random number between 1 and that given number. Having that value, it is verified if it matches the number given at the beginning. For example, if you give the number 2, there is a 50% probability that the given word will be considered, on the other hand, if the number 1 is given, there is a 100% probability of including it in the new data.

Subsequently, it is decided if it is with a dictionary or FastText, in any of both, a list of similar words based on is obtained and a word is chosen according to a certain temperature, which will be the random number between 0 and the minimum between a given number and the total number of synonyms that are taken into account. For example, if the number 1 is given, the minimum between the two is taken, and the random one between 0 and 1 is more likely that a word is similar to the representative one, in this case it can be the first or second value of the list of synonyms, that is, the most similar ones, on the other hand, if a large value is given, it gives more possibility of taking words that are further away from the representative word that is found. On the other hand, in order not only to have words from one class, a limited random number of words that can be included are taken from the representative words of the respective classes chosen to enter the model. To finish this part, all the values generated are taken to form the text strings where the length of each one varies and the type of words that are also present.

The classification model used is BERT (Bidirectional Encoder Representations from Transformers) is a language model based on neural networks that has revolutionized natural language processing (NLP). BERT is a pre-trained model that uses the Transformer architecture, which is an attention-based neural network. Unlike previous language models that were trained in a unidirectional fashion (left to right or right to left), BERT is bidirectional, which means that it can capture the context of words in both directions[16, 17].

BERT training is done in a task known as “pretraining language modeling”. During this stage, BERT is trained to predict hidden words in masked sentences and to predict the relationship between pairs of sentences. This massive training allows you to learn contextualized representations of words, capturing information from the context that surrounds them. in the Spanish version that is called BETO.

Training details:

- 3 epochs
- Adam Optimizer using a learning rate of $5e^{-5}$
- Batch size of 32 elements
- Unfreezed BETO weights

4. Results

In this edition of Rest-Mex, the Dataverse team evaluated models with different metrics, however, minority classes are given more weight. The team placed seventh in the competition with a Sentiment Track Score of 0.7173586609. The results of the metrics obtained are the following:

Table 1
Table with macro results and MAE of Polarity.

	Sentiment Track Score	Macro F1 (Polarity)	Macro F1 (Type)	Macro F1 (Country)	MAE (Polarity)
BETO_balanced_classes	0.717358661	0.520697485	0.976026081	0.915223349	0.420236782
Median	0.698477823	0.504605613	0.973658078	0.896611023	0.364054402
Average	0.623386446	0.455474842	0.898391624	0.759450539	0.551673786

Table 2
Table with the F1-score results of the minority classes.

	F1(1)	F1(2)	F1(Colombia)	F1(Cuba)
BETO_balanced_classes	0.598301222	0.296312364	0.897817548	0.916741626
Median	0.561103783	0.293753558	0.87726678	0.897431923
Average	0.478650654	0.277144544	0.706332701	0.742940687

The result of the classification of the test reviews that was sent, the dictionary option was taken into account, with temperature 100 and probability 1, they were balanced with respect to polarity and countries. From the results of the metrics obtained, as can be seen in the Table1, it can be seen that in most of the metrics, with the exception of the MAE in the median, both the average and median values of all the participants were exceeded. Something similar happens in the F1 scores of the minority classes as seen in the Table 2.

5. Conclusions and future work

It can be concluded that the oversampling and subsampling techniques are very useful together with the BERT model, it produced good results, but particularly not the best ones. However, oversampling shows that it can be an effective strategy to improve the performance of models in unbalanced data sets. Nevertheless, it is important to note that oversampling must be applied carefully, as excessive generation of synthetic examples can lead to overfitting and degradation of model performance. Tests were needed with the combinations of temperature and probability, which could have better results in this competition.

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", which originated this work.

References

[1] 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.

[2] S. de Turismo, Resultados de la actividad turística marzo 2023 (2023).

[3] 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).

[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] 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).

[6] 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.

[7] 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).

[8] 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.

[9] 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) 289–299.

[10] M. Á. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-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).

[11] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñiz-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).
[12] G. E. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD explorations newsletter 6 (2004) 20–29.
[13] Y. Ma, H. He, Imbalanced learning: foundations, algorithms, and applications (2013).
[14] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, Computational linguistics 16 (1990) 22–29.
[15] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, arXiv preprint arXiv:1802.06893 (2018).
[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[17] N. Sabharwal, A. Agrawal, N. Sabharwal, A. Agrawal, Bert algorithms explained, Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing (2021) 65–95.