What if We Use Fewer Data to Classify Tourist Opinions in Spanish?

Federico Sandoval[1]

[1]Algiedi Solutions, Cholula, Mexico, 72760

Abstract

This paper investigates the feasibility of using smaller amounts of data to accurately classify tourist opinions in Spanish. The classification of tourist reviews is important for businesses in the tourism industry, but data collection and processing can be time-consuming and costly. To test the effectiveness of smaller datasets, we conducted experiments using a machine learning approach to classify polarity, type, and country in a dataset of tourist reviews in Spanish. Our results show that it is possible to achieve good levels of accuracy with smaller datasets.

Keywords

Rest-Mex, Sentiment Analysis, Few opinions, data imbalance, Spanish opinions

1. Introduction

Tourism is a key sector for many countries, providing employment opportunities and contributing significantly to their economy [1, 2, 3]. With the advent of social media and online review platforms, tourists are increasingly sharing their experiences and opinions about destinations, attractions, and services [4, 5]. This wealth of information presents a valuable opportunity for tourism industry stakeholders to gain insights into customer preferences, improve service quality, and enhance the overall tourist experience [6, 7].

However, the large volume of user-generated content (UGC) presents a challenge for extracting useful insights from it [8]. Natural language processing (NLP) techniques have been widely used to classify and analyze UGC in various languages, including Spanish [9]. Traditionally, NLP models require large amounts of data to achieve high accuracy in classification tasks. However, collecting and annotating large amounts of data can be time-consuming and expensive [10, 11, 12, 13].

This raises the question: What if we use fewer data to classify tourist opinions in Spanish? Can we still achieve acceptable accuracy? This question is particularly relevant for small businesses or organizations with limited resources to collect and annotate large amounts of data [14, 9].

To address this question, we conducted an experiment to compare the performance of NLP models trained on different amounts of data in classifying tourist opinions in Spanish. We used a dataset of tourist reviews in Spanish from a popular online review platform and trained several models with varying amounts of data.

IberLEF 2023, September 2023, Jaén, Spain

fsandoval@algiedi.com.mx (F. Sandoval)

Our findings indicate that it is possible to achieve acceptable accuracy in classifying tourist opinions in Spanish even with fewer data. The results have implications for small businesses and organizations with limited resources, as they can still benefit from the insights gained from analyzing UGC, without having to invest in large-scale data collection and annotation efforts.

In this paper, we describe our experiment, present our findings, and discuss the implications of our results for tourism industry stakeholders. We also discuss the limitations of our study and suggest directions for future research. Overall, our paper contributes to the growing body of research on NLP techniques for analyzing UGC in the tourism industry, and highlights the potential for using smaller datasets to achieve meaningful insights.

2. Rest-Mex 2023 Corpus

The organizers of Rest-Mex 2023 [15] have curated a train collection comprising 251,702 opinions extracted from TripAdvisor. The dataset includes three classification labels:

1. Polarity
2. Type
3. Country

The polarity classification encompasses five classes, ranging from class 1 representing the most negative polarity to class 5 denoting the most positive polarity. Table 1 displays the distribution of these classes, revealing an evident class imbalance.

Table 1
Distribution of Polarity Classes

Class	Instances
1	5772
2	6952
3	21656
4	60227
5	157095

The classification of the type of place includes three classes: Attractive, Hotel, and Restaurant. Table 2 showcases the distribution of these classes. Although the imbalance is not as pronounced as observed in polarity, the table indicates some degree of imbalance.

Table 2
Distribution of Type Classes

Class	Instances
Attractive	111188
Hotel	76042
Restaurant	64472

The classification based on the country of origin of the visited place encompasses three classes: Mexico, Cuba, and Colombia. Table 3 illustrates the distribution of these classes.

Table 3
Distribution of Country Classes

Class	Instances
Colombia	66703
Cuba	66223
Mexico	118776

3. Simple instances selection

In this study, we investigate the impact of using reduced amounts of data on the classification of tourist opinions in Spanish. Specifically, we aim to determine whether randomly selecting 100, 500, 1000, and 5000 opinions from each polarity class can yield a balanced database while maintaining classification performance.

3.1. Classifier

For sentiment analysis, we utilize a BERT-based classifier that specifically utilizes the Beto-cased model. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a highly effective pre-trained language model that has shown exceptional performance across various natural language processing tasks.

Model: We employ the Beto-cased model, which is a variant of BERT trained specifically on Spanish text. This model captures detailed information and retains the capitalization of words, enabling better understanding of the context.

Max Length: To handle input sequences efficiently, we set a maximum sequence length of 32 tokens. Any input longer than this is either truncated or divided into smaller segments based on BERT's tokenization approach.

Optimizer: The Adam optimizer is used, as it is a popular choice for training deep neural networks. Adam combines adaptive learning rates with momentum, resulting in efficient optimization and convergence.

Learning Rate: We set the learning rate to 5×10^{-5}, a commonly used value for fine-tuning BERT models. This value strikes a balance between convergence speed and detailed optimization.

Steps: The step size, represented as epsilon (ϵ), is set to 1×10^{-8}. This parameter controls the level of noise added to the learning rate update, ensuring stability during the training process.

Epochs: The classifier is trained for 4 epochs, where each epoch represents a complete pass through the entire training dataset. This choice balances the model's learning capacity with computational resources.

By utilizing BERT-based models with these specific configurations, our aim is to leverage the contextual representation capabilities of BERT for accurate sentiment analysis on Spanish text. The selected settings provide a strong foundation for training and optimizing the classifier.

4. Results

Table 4 shows the results obtained from the training corpus. For this, we use a 70/30 partition.

4.1. Train results

Table 4
Train results

Instances	F(Polarity)	F(Type)	F(Country)
100	0.49	0.94	0.52
500	0.50	0.93	0.74
1000	0.47	0.91	0.76
5000	0.53	0.95	0.83

From the results, we can observe the following trends:

- **Polarity Classification:** As the number of instances increases from 100 to 5000, the F-measure for polarity classification improves, with the highest F-measure of 0.53 achieved with 5000 instances. This suggests that increasing the training data helps in capturing the nuances of sentiment analysis more accurately.
- **Type Classification:** The F-measure for type classification remains consistently high across different numbers of instances, ranging from 0.91 to 0.95. This indicates that the classifier performs well in accurately categorizing opinions into attractive, hotel, and restaurant types, regardless of the training data size.
- **Country Classification:** Similar to the type classification, the F-measure for country classification remains consistently high, ranging from 0.52 to 0.83. This suggests that the classifier effectively recognizes the country of origin of the visited places, irrespective of the number of training instances.

Based on these results, it is evident that increasing the number of instances generally leads to improved performance in polarity classification. However, for type and country classification, the classifier achieves high F-measures even with smaller training datasets.

These findings highlight the effectiveness of the BERT-based classifier and demonstrate its robustness across different classification tasks. By utilizing a reduced number of instances, we can achieve competitive classification performance, potentially reducing the computational resources required for training without significant loss in accuracy.

Further analysis and evaluation on larger datasets and real-world scenarios would be valuable to validate the generalization capabilities of the classifier.

4.2. Rest-Mex 2023 official results

Table 5 presents the results obtained from the Rest-Mex 2023 forum. The analysis reveals that utilizing 5000 instances for each class yields the most favorable outcomes. This amounts to a total of 25,000 instances, approximately 10% of the entire dataset. These findings demonstrate the effectiveness of this approach, yielding compelling and competitive results.

Table 5
Test results

Instances	Final Rank	F(Polarity)	F(Type)	F(Country)
100	0,52	0,37	0,92	0,49
500	0,60	0,41	0,93	0,74
1000	0,60	0,41	0,91	0,74
5000	0,66	0,49	0,95	0,81
BaseLine-Beto-No-Fine-Tuning	0,38	0,24	0,83	0,34
BaseLine-majority	0,14	0,15	0,20	0,21
BaseLine-minority	0,11	0,00	0,20	0,21

Moreover, the achieved results surpass the baselines proposed by the organizers. Out of a total of 17 participants, our approach secured the 13th position, showcasing its superior performance.

5. Conclusions

The findings of this study demonstrate that it is feasible to achieve reasonable results in sentiment analysis for tourist opinions in Spanish by utilizing a small percentage of the original data.

One prominent characteristic observed across various tourism collections is the inherent data imbalance, with negative polarities representing the minority classes. Therefore, it becomes imperative to address this data imbalance issue.

For the Rest-Mex 2023 edition, the organizers compiled a database consisting of over 250,000 opinions, with more than 50% exclusive to class 5. This presents a challenge for oversampling techniques. Consequently, in this research, we explored the application of sub-sampling methods.

Specifically, we employed a random selection of data based on polarity class, considering sample sizes of 100, 500, 1000, and 5000 (the highest available value).

The most favorable results were obtained when utilizing 5000 instances per class, totaling 25 instances. This subset, constituting only 10% of the total data, enabled the development of a competitive model, surpassing the baselines and yielding acceptable results.

Such an approach proves to be an ideal solution in scenarios where there are constraints on execution time and available memory. By utilizing a smaller representative sample, the computational resources required are reduced without compromising the performance significantly.

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] 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 sys-

tematic review and future research directions, Journal of King Saud University-Computer and Information Sciences (2022).
[3] 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.
[4] 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.
[5] 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).
[6] 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.
[7] 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).
[8] 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.
[9] M. Á. Álvarez-Carmona, E. Villatoro-Tello, L. Villaseñor-Pineda, M. Montes-y Gómez, Classifying the social media author profile through a multimodal representation, in: Intelligent Technologies: Concepts, Applications, and Future Directions, Springer, 2022, pp. 57–81.
[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] 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.
[12] 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.
[13] 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 environments, Information Sciences 603 (2022) 42–59.
[14] M. A. Alvarez-Carmona, R. Aranda, A. Diaz-Pacheco, J. de Jesús Ceballos-Mejıa, Generador automático de resumenes cientıficos en investigación turıstica, Research in Computing

Science (2022).
[15] 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).