# Sentiment Analysis in Spanish Reviews: Dataket Submission on REST-Mex 2022

Gabriel Missael Barco[1], Gil Estéfano Rodríguez Rivera[1] and
Delia Irazú Hernández Farías[2]

[1]Division de Ciencias e Ingenierias, Universidad de Guanajauto, Mexico
[2]Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico

**Abstract**

This paper presents our approach and results in the Rest-Mex 2022 shared task. In particular, we participated in the sentiment analysis track, classifying tourism reviews into two aspects: the polarity measured by the number of stars assigned in the review (from 1 to 5) and the type of attraction being reviewed (namely Hotel, Restaurant, or Attractive). Given that the dataset was imbalanced, we took advantage of corpora in the literature for data augmentation purposes. We used a pre-trained BERT model for the polarity sub-task, and we fine-tuned the model on the domain in hand. Our model excelled in classifying the low stars reviews (1 and 2) because of the data augmentation for balancing. We used a more traditional approach based on a bag-of-words with different machine learning algorithms for the attraction type sub-task. According to the official results, we obtained very competitive results while keeping the computational cost lower.

**Keywords**

Spanish Sentiment Analysis, Fine-tuning, Transfer Learning, BERT, Dataset augmentation

## 1. Introduction

Social media serve as a platform for sharing our ideas, experiences, and opinions. Exploiting such content could significantly benefit from a wide range of perspectives like evaluating the perception of some products to the user experience when using some service, among others. In Natural Language Processing (NLP), the area dedicated to studying the subjective information in a given piece of text is the Sentiment Analysis (SA) [1]. It is one of the most active fields in NLP, and many different approaches have been exploited, from rule-based to deep learning methods. However, most of the advances in this area have been made for the English language.

For what concerns other languages, several efforts have been made to promote the research in this area. Regarding Spanish (one of the most widely spoken languages around the world) [2], some efforts have been made considering traditional word-based [3] to deep learning-based approaches [4]. In [5, 6], literature reviews in this topic can be found. Besides, some shared tasks dedicated to Sentiment Analysis in Spanish have been organized such as the TASS[1] and the REST-MEX 2021[2] [7, 8, 9].

*IberLEF 2022, September 2022, A Coruña, Spain*


[1]http://tass.sepln.org/
[2]https://sites.google.com/cicese.edu.mx/rest-mex-2021

In this paper, we describe our participation in the second edition of REST-Mex, which includes a sub-task focused on sentiment analysis in the domain of Mexican tourist texts. The proposed approach exploits BERT in two different settings under both constrained and unconstrained frameworks. First, for the latter one, we enriched the official training data by using available resources in the literature. In the second proposal, we attempt to emulate how a person infers the polarity of a given piece of text; in this case, we rely only on the official training data. Both systems obtained competitive results, particularly the former one, which reached the best results in F-score terms of Macro F-score, Macro Precision, and the F-score in the classes corresponding to the most negative sentiment.

## 2. Dataket at Rest-Mex 2022

### 2.1. Task description

This year, the shared task *REST-Mex 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts* proposed three sub-tasks aiming to promote the research and development of intelligent systems dedicated to the tourism domain in Spanish written texts [10]. The first sub-task, *Recommendation Systems* aims to determine the degree of satisfaction that the tourist will have when visiting a given place. In the second one, *Epidemiological Semaphore Prediction* the goal is to predict the semaphore color for controlling the activities according to the severity of the Covid-19 pandemic by considering news content. And finally, the *Sentiment Analysis* which attempts to determine the polarity degree (ranging from 1 up to 5) of a given opinion regarding a Mexican tourist place. In addition, in this sub-task was also required to identify the type of opinion considering three different categories: hotel, restaurant, and attraction. The organizers provided data for tests and training and established no restrictions on using additional resources.

### 2.2. Our proposal

In the following paragraphs, we describe in more detail the proposed approaches for addressing the sentiment analysis and the classification of the opinions according to their type.

#### 2.2.1. Sub-task: Polarity Classification

We had two different approaches to solving this sub-task. Both rely on a pre-trained BERT-based model, taken from the Hugging Face models page [11], namely "*nlptown/bert-base-multilingual-uncased-sentiment*". This pre-trained model already has a classification head on top of the BERT architecture that does something very similar to what we are trying to achieve. It predicts the number of stars (from 1 to 5) of a product review. This model works in six languages: English, Dutch, German, French, Spanish, and Italian. It was fine-tuned on $629,000$ data points, from which $50,000$ were in Spanish.

## First approach

In this approach, we decided to fine-tune the model mentioned above with the provided data for training to improve the performance over the task in hand, i.e., to predict a sentiment over places instead of products (the domain with which it was developed).

We identified that the official training dataset is highly unbalanced, as shown in Figure 1b. Attempting to solve this, we decided to extract some reviews from the well-known Yelp Dataset [12], an extensive dataset containing reviews of some businesses with the same format as in the official training set for the shared task: a rating among 1 up to 5 for written opinions. We only selected reviews from restaurants and hotels (such as those included in the official training data); the subset polarity distribution is as shown in Figure 1a. An essential aspect to highlight is that the Yelp Dataset contains only text written in English; then, we used the "*googletrans*" Python library [13] to automatically translate the reviews to Spanish. Although the translation was not perfect, the irregularities in translation could provide variance in the final dataset, helping to avoid overfitting in the fine-tuning process. After merging both datasets, the final polarity distribution is as shown in Figure 1c, which is more balanced and less likely to have class bias. The entire augmented dataset has $75,713$ reviews. The Jupyter Notebook with all the data extraction and translation is publicly accessible in the GitHub repository [14].

For the fine-tuning process, we used the augmented dataset, and the Transformers Python packages available on HuggingFace [15]. The entire fine-tuning pipeline is available in the Jupyter Notebook "*polarity_bert.ipynb*" within the GitHub repository [14]. We filtered out a total of $4,532$ ( $15\%$) instances from the original training set with the objective to have a *testset* without data augmentation and used the $71,181$ remaining instances for experimental purposes. We split the augmented dataset in three subsets: train ($51,249$ instances), validation ($12,813$ instances), and test ($7,119$ instances).

We trained the model for six epochs, with a batch size of $8$, and used a linear decay for the learning rate ranging from $1e-5$ to $0$. We used the default tokenizer provided with BERT. Finally, we used the Adam optimizer and a sparse categorical cross-entropy loss. The accuracy of the validation set remains almost constant within all six epochs, as observed in Figure 2. Also, we observed an increase in the loss function in the validation set. That is not necessarily convenient; however, the MAE presents progress in the training and validation sets.

After training, we evaluated the model before and after fine-tuning, both in the augmented and original test data partitions, as shown in Figure 3. We observed an improvement in both accuracy and MAE in both test sets, and the confusion matrix is also better in the model after fine-tuning. This last detail is more clearly observed in the augmented test set.

## Second approach

Attempting to emulate the process behind a person reading a review and trying to guess the score associated with it. We have the intuition that first is to read the title and then the opinion: each can act as an individual indicator of the score. Because of the text length, the essential information should be expected to come from the opinion. However, there are cases when there's enough ambiguity in the opinion, such that there's a lot of uncertainty regarding the score guessing process; in such situations, the title of the review becomes a relevant information

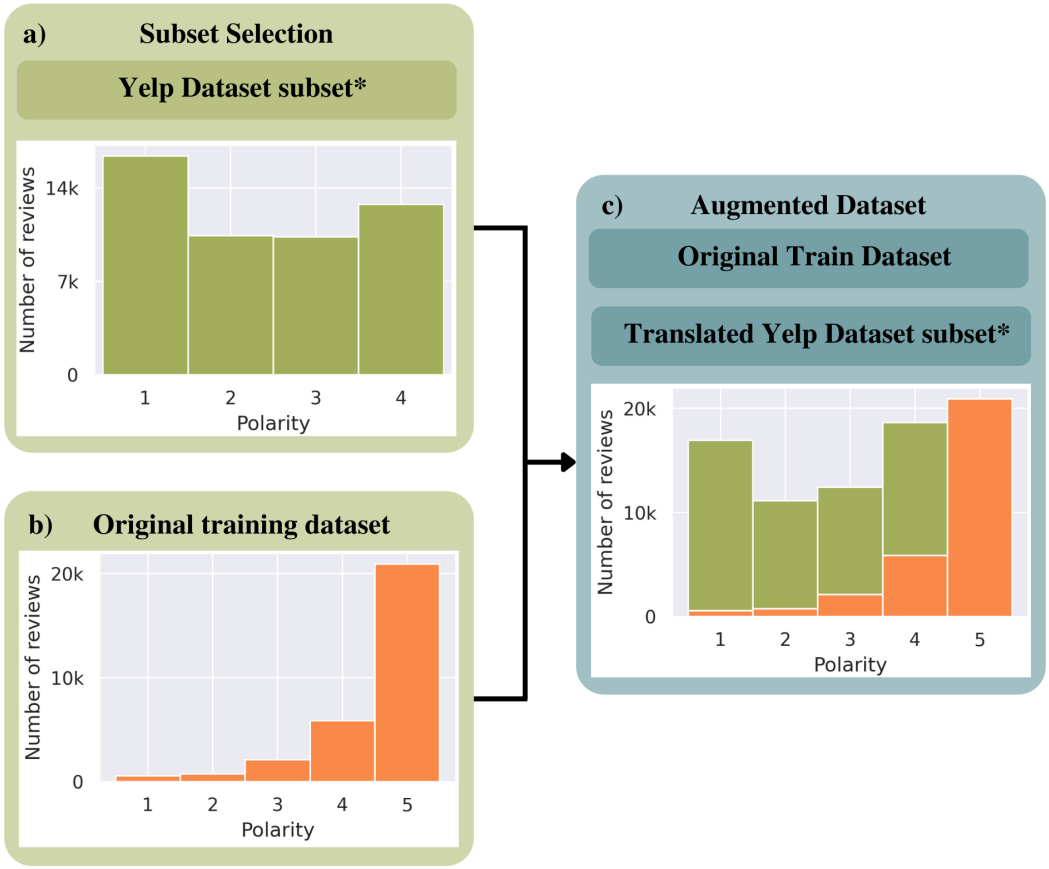


**Figure 1:** Original dataset augmentation with Yelp Dataset. Distribution of the polarity in each subset and final set shown.

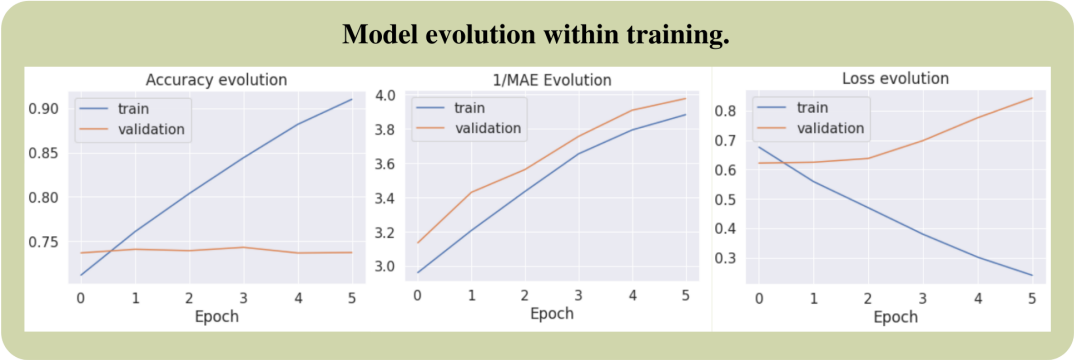


**Figure 2:** Accuracy, MAE, and loss evolution of train and validation sets during the training phase.

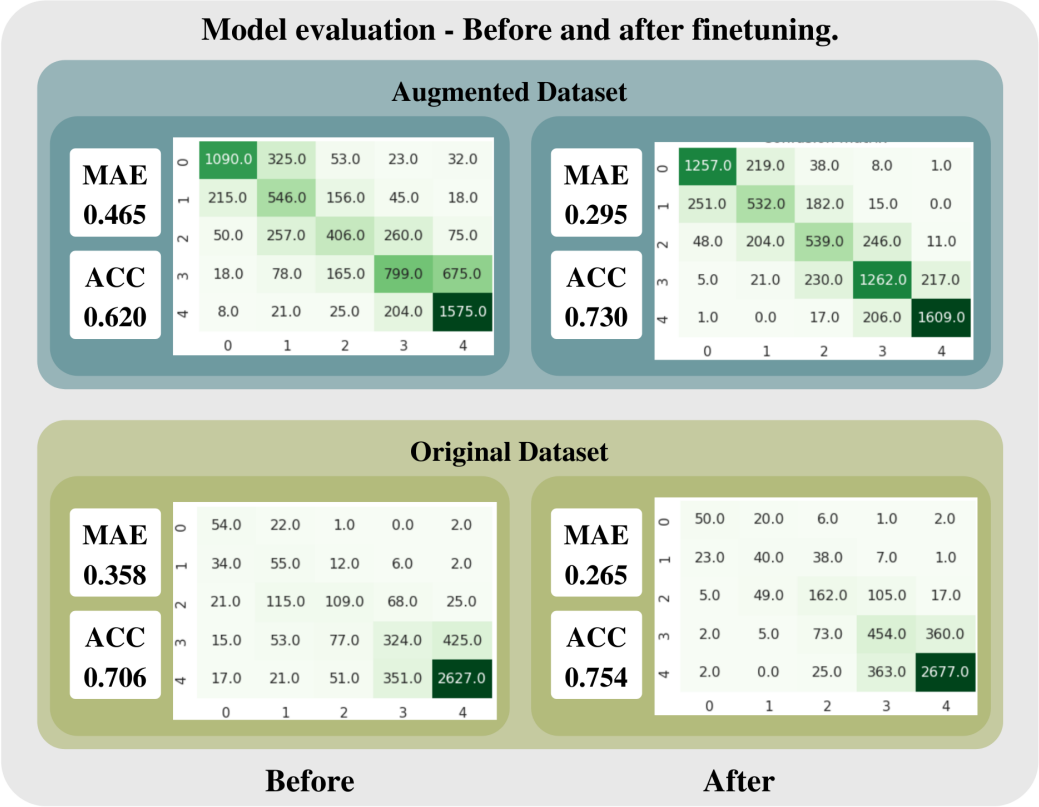


**Figure 3:** Model evaluation before and after fine-tuning, evaluated in augmented and original data partitions.

source. The link between our intuition and the proposed architecture can be tracked as follows:

1. BERT is analogous to the person reading both the title and the opinion and then guessing the corresponding label. The guess and its uncertainty could be considered as an analogy to the output probabilities on each part of the review.
2. The extra features are the proportions of words on the opinion associated with specific emotions. For doing so, we exploited EmoLex [16]. These features aim to consider the emotional factor to try to improve the uncertainty of the BERT-generated features. We considered that we could improve this process by not counting only total matches between words on the text and the lexicon mentioned above but also taking into account grammar mistakes, elongated words, or applying other pre-processing steps.
3. The Multilayer perceptron puts together the patterns of the uncertainties on the BERT pre-processing and the proportion of emotions extracted from the opinion. The perceptron's output is the final score the model assigns to the review.

For experimental purposes, we separated only the official dataset consisting of $30,212$ data points on $6,042$ points (20%) for the test and $24,170$ points (80%) training subsets. The training

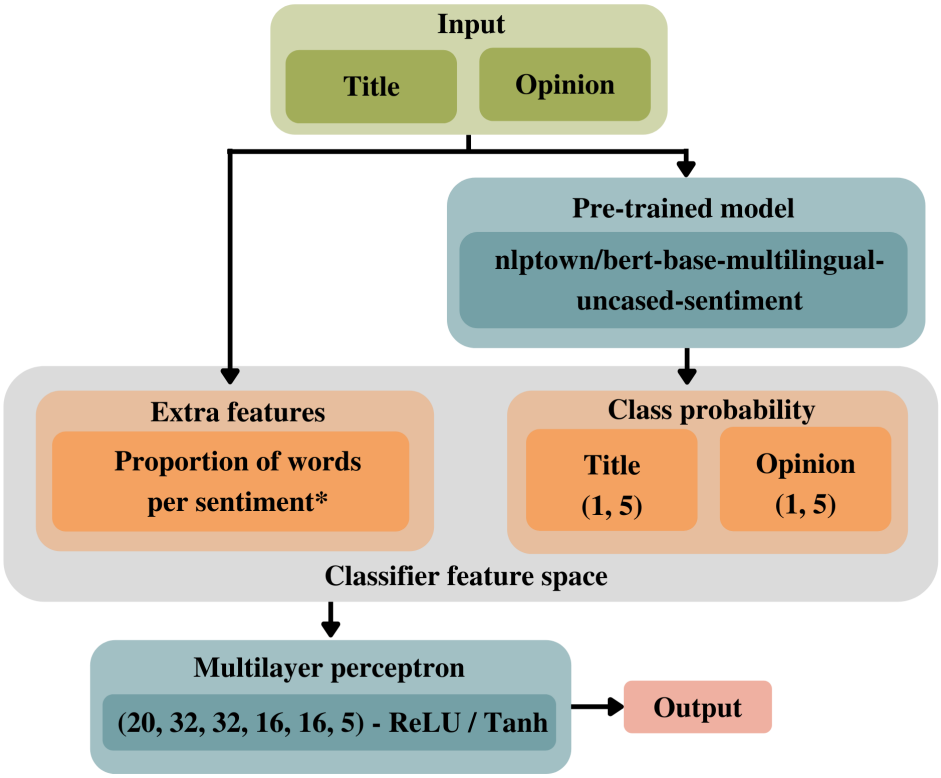


**Figure 4:** Second approach architecture for polarity classification.

consisted of a maximum of 128 epochs with a validation split of 30%. However, the process was stopped early on the 28-th epoch due to the presence of overfitting. The overfitting couldn't be improved even with the inclusion of two 10% dropout layers. The evaluation results are found in Figure 5.

We also used the "*nlptown/bert-base-multilingual-uncased-sentiment*" for our second approach, but without finetuning it to the new data.

We merged both the Title and the Opinion in a five-dimensional vector with the probabilities of belonging to one category. Each five-dimensional probability vector is the output of the aforementioned BERT-based model: we used the text of interest (be it the Opinion or the Title part of the review) on the model as an input, and the vector of probabilities comes from its output. The entire architecture is shown in Figure 4.

The idea behind the second approach is to propose a less computationally expensive method with a MAE similar to the one of the first approach.

#### 2.2.1. Sub-task: Type of Opinion.

We evaluated a BERT-based approach similar to the previous subtask, using a BERT model with a pre-trained head for zero-shot classification [17], and a more traditional model based on

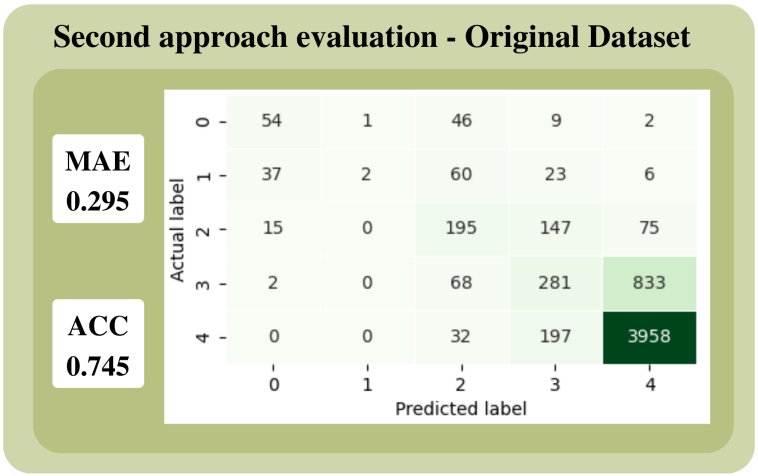


**Figure 5:** Model evaluation after the training on the second model.

different configurations of a bag-of-words together with supervised machine learning classifiers. Being the latter the one obtaining the highest performance, we decided to use it because it has a good performance and is less computationally expensive than the BERT model. However, we acknowledge that performing fine-tuning in the BERT model for classification with the three classes of this sub-task could have resulted in a similar or slightly better performance, to the detriment of the computation and training cost.

A bag-of-words model with different settings was used together with standard classifiers. We assessed the performance of the most frequent 500, 1000, 5000, and 10000 uni-grams, bi-grams, and tri-grams with a term-frequency and binary-weighted schema. As classifiers, we evaluated the Scikit-learn [18] implementation of Random Forest (RF), k-Nearest Neighbors (5NN), and Decision Tree (DT) with default parameters and a Support Vector Machine in which we applied GridSearch for parameter optimization. Besides, we use a Convolutional Neural Network (CNN) with two convolution filters and a dense layer with RELU as activation function and ADAM as optimizer. Finally, we included an ensemble for all the classifiers mentioned above with majority voting.

The final configuration is based on the 500 most frequent uni-grams with a term-frequency weighted scheme together with a SVM with the following parameters: kernel=RBF, C=10, and gamma=0.01. For participating in the shared task, we decided to use the same settings in combination with the two approaches described above for the first sub-task.

### 2.2.2. Manual annotation of data.

Aiming to evaluate the proposed method during the development phase, we decided to generate a subset of data (denoted as *ownTestPartition*) from the official test partition provided by the task's organizers. We randomly selected 200 hundred samples that the authors manually annotated according to both *polarity* and *type of opinion*. Table 1 shows the distribution obtained in the *ownTestPartition*. As we can notice, it is very similar to the one in the official training set.

| Polarity | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Count | 5 | 8 | 23 | 54 | 110 |

| Type of attraction | Hotel | Restaurant | Attractive |
|---|---|---|---|
| Count | 110 | 52 | 38 |

**Table 1**
Class distribution on the samples selected in the *ownTestPartition* obtained from manual annotation

## 2.3. Obtained results

**Our results during development.**
We evaluated the performance of four different models in the polarity sub-task using as metrics the MAE (Eq. 2.3 shows the used formula). The model are: the base NLPTown model, the first approach model finetuned in the original train set, the one finetuned in the augmented train set, and the model of the second approach. As it is shown in Table 2, the best model is obtained from the first approach, corresponding with fine-tuning the base NLPTown model with the augmented train set with Yelp data.

| **Model** | **Accuracy** | **MAE** |
|---|---|---|
| Base NLPTown | 0.670 | 0.400 |
| First Approach (Original dataset) | 0.670 | 0.370 |
| First Approach (Augmented dataset) | 0.700 | 0.325 |
| Second Approach | 0.650 | 0.425 |

**Table 2**
Obtained results during the development phase in the manually annotated subset of samples from the official test partition for the first sub-task. MAE and accuracy of the base model and the three different trained models.

Table 3 shows the obtained results in Macro F-score for the type of opinion tasks in both the official training data and in the *ownTestPartition*. As can be noticed, the SVM achieves the best results in both datasets. We observed a drop in the performance in most classifiers when comparing the obtained results in the official training data and the manual annotation one. This drop in performance could be because of the underlying error associated with manual classification: that is, the manual classification made by the authors is not necessarily correct. We obtained the best classification rates for the *Hotel* class, while the lower ones are on the *Attractive*. This difference could be related to the unbalanced characteristic of the dataset. It is essential to mention that, in this subtask, we did not apply data augmentation like in the polarity classification task.

**Official Results**

The official results of the track of sentiment analysis are available at [19]. We show a summary of our results of the polarity sub-task in Table 3. For the polarity sub-task, the ranking metric is the mean absolute error, defined as:

| | | RF | 5NN | DT | SVM | CNN | ENS |
|---|---|---|---|---|---|---|---|
| Hotel | Training | 0.96 | 0.7 | 0.92 | 0.97 | 0.96 | 0.97 |
| | *ownTestPartition* | 0.93 | 0.81 | 0.90 | 0.93 | 0.92 | 0.93 |
| Restaurant | Training | 0.91 | 0.65 | 0.83 | 0.94 | 0.92 | 0.93 |
| | *ownTestPartition* | 0.85 | 0.64 | 0.72 | 0.84 | 0.79 | 0.84 |
| Attractive | Training | 0.96 | 0.62 | 0.93 | 0.96 | 0.95 | 0.96 |
| | *ownTestPartition* | 0.91 | 0.76 | 0.82 | 0.93 | 0.88 | 0.91 |

**Table 3**
Obtained results during the development phase in both official training data and in the manually annotated subset of samples from the official test partition.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t| \tag{1}$$

For the second sub-task, the metric is the Macro F-score, defined as:

$$MacroF1_k = \frac{F1_A(k) + F1_H(k) + F1_R(k)}{3} \tag{2}$$

Finally, the ranking for the whole task is a combination of the MAE and the Macro F-score, given by:

$$Sentiment_{Res_k} = \frac{1}{2}\left(\frac{1}{1 + MAE_k} + MacroF1_k\right) \tag{3}$$

We ranked in the ninth-place with our best run, compounded by our first approach for the polarity sub-task and our unique approach for the second, with a final ranking score of 0.87537. The first place on the track got a 0.89238, an absolute difference of only 0.01701 compared with our results. It is essential to highlight that our approach achieves the best scores within the polarity sub-task in the Macro F-measure, Macro Precision, Class 1 F-measure, and Class 2 F-measure, with the corresponding values shown in Table 4.

## 3. Conclusions

This paper described our proposal for participating in the sentiment analysis sub-task in the REST-Mex 2022 edition. This sub-task has two different objectives: determining the polarity degree and identifying the opinion type of a given touristic review. In particular, we used a multilingual BERT model and fine-tuned version that only uses Spanish data. The model we used was trained on a similar task [11]. The results show that even though the model was already domain-specific for reviews, further fine-tuning in the tourism review domain leads to better performance. That may be because of the slight change of content in the reviews from product to place, the shift from multilingual inputs to monolingual Spanish inputs, and the usage of new training data.

| Metric | Run 1 | Run 2 |
|---|---|---|
| Ranking | 9 | 13 |
| Final Rank | 0.8753 | 0.8662 |
| MAE | 0.269 | 0.300 |
| Accuracy | 75.769 | 74.609 |
| Macro F-score | **0.567** | 0.432 |
| Macro Recall | 0.575 | 0.512 |
| Macro Precision | **0.561** | 0.406 |
| Class 1 F-score | **0.616** | 0.475 |
| Class 2 F-score | **0.376** | 0.092 |
| Class 3 F-score | 0.484 | 0.368 |
| Class 4 F-score | 0.485 | 0.350 |
| Class 5 F-score | 0.875 | 0.878 |

**Table 4**
Official obtained results from both runs for the sentiment analysis track. The numbers in bold correspond with those metrics in which our model got the best performance among all participants.

| Accuracy | Macro F-score | F-score | | |
|---|---|---|---|---|
| | | *Hotel* | *Restaurant* | *Attractive* |
| 96.55 | 0.963 | 0.974 | 0.944 | 0.970 |

**Table 5**
Official obtained results in the attraction type sub-task.

We proposed two different data settings for determining the sentiment: constrained and unconstrained. For the latter one, we take advantage of the Yelp Open Dataset, a well-known dataset for sentiment analysis tasks. An automatic translation to Spanish was carried out over a subset of samples from this dataset. The obtained results, particularly in the less represented classes in the official train set, show that having a more balanced dataset is essential for the class representation in the predictions. Also, having more data available for training may result in better performance when dealing with big models, such as BERT. For what concerns the type of attraction sub-task, we compared the performance of state-of-the-art models against classic NLP ones (such as a bag-of-words). The obtained results were very similar; for participating in the shared task, we selected the less computational expensive one, i.e., the bag-of-words.

Our unconstrained system obtained very competitive results in the polarity classification task, particularly in the underrepresented classes. For future work, we are interested in evaluating other data augmentation techniques and exploiting other information resources to improve the performance of the second proposal for the sentiment analysis task.

## References

[1] F. A. Pozzi, E. Fersini, E. Messina, B. Liu, Sentiment Analysis in Social Networks, 1st ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2016.

[2] M. A. 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.
[3] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, E. A. Villaseñor, A Case Study of Spanish Text Transformations for Twitter Sentiment Analysis, Expert Systems with Applications 81 (2017) 457–471.
[4] N. C. Dang, M. N. Moreno-García, F. De la Prieta, Sentiment Analysis Based on Deep Learning: A Comparative Study, Electronics 9 (2020).
[5] C. Henriquez Miranda, J. Guzman, A Review of Sentiment Analysis in Spanish, TECCIENCIA 12 (2016) 35–48.
[6] M. Navas-Loro, V. Rodríguez-Doncel, Spanish Corpora for Sentiment Analysis: A Survey, Language Resources and Evaluation 54 (2020).
[7] 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.
[8] 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.
[9] 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.
[10] 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).
[11] NLPTown, bert-base-multilingual-uncased-sentiment, First commit on February 14, 2020. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.
[12] Yelp, Yelp open dataset, ???? URL: https://www.yelp.com/dataset.
[13] H. SuHun, Googletrans, ???? URL: https://py-googletrans.readthedocs.io/en/latest/.
[14] G. M. Barco, G. E. Rodríguez Rivera, D. I. Hernández Farías, Dataket at rest-mex 2022, ???? URL: https://github.com/Dataket/Rest-Mex-2022-DCI-UG.
[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2019. URL: https://arxiv.org/abs/1910.03771. doi:10.48550/ARXIV.1910.03771.
[16] S. Mohammad, Nrc word-emotion association lexicon, ???? URL: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.
[17] MoritzLaurer, mdeberta-v3-base-mnli-xnli, First commit on December 5, 2021. URL: https:

//huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli.
[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[19] REST-MX, Rest-mx 2022 results, May 17, 2022. URL: https://sites.google.com/cicese.edu.mx/rest-mex-2022/results?authuser=0.