BERT and Data Augmentation for Sentiment Analysis in TripAdvisor Reviews

Enrique Santibáñez-Cortés[1,*], Azael Carrillo-Cabrera[1,*], Yair Antonio Castillo-Castillo[1,*], Daniela Alejandra Moctezuma-Ochoa[2] and Victor Hugo Muñíz-Sánchez[1]

[1]*Research Center in Mathematics (CIMAT) campus Monterrey, Mexico*

[2]*Research Center in Geospatial Information Sciences (CentroGEO), Mexico*

Abstract

Sentiment analysis is one of the most studied topics over years in the Natural Language Processing community. Usually, to solve this task, the solutions must classify the input text into a set of categories, ranging from negative (or very negative) to positive (or very positive). In this paper, we proposed a system for the sentiment analysis challenge in the Rest-Mex track at IberLEF 2022. The objective is to classify tourist opinions about places, restaurants, and hotels in Mexico through the data acquired by the TripAdvisor platform. Our solution explore deep learning (DL) architectures according to the subtasks related with this challenge. Those architectures includes convolutional neural networks (CNN), contextualized vector representations of the text based on BERT models and a particular data augmentation scheme.

Keywords

Sentiment Analysis, NLP, BERT, data augmentation, Mexican tourism

1. Introduction

As we know, Twitter is a very popular micro-blogging platform used in almost all countries, it has transformed the way people share information, opinions, and interests, on social media. Also, Twitter's users can get information about people or organizations of interest from them. Regarding this, Sentiment Analysis is one of the most studied problems in the Natural Language Processing research community, where a lot of efforts have been done through the years to improve the sentiment analysis classification problem. One of the most employed data to test the sentiment analysis task is Twitter which has emerged as the main data employed in the related literature. Spite there are very methods in the literature, still, the overall performance in short texts remains low, with over 70% accuracy as was reported in [1]. Some reason for this level of performance is that sentiment analysis is a very challenging task due to linguistic phenomena such as negation, irony, sarcasm, the subjectivity of the opinions, etc. also these could be maximized in informal short texts.

Formally, sentiment analysis is the task to associate a polarity to a given text, these polarities

IberLEF 2022, September 2022, A Coruña, Spain

enrique.santibanez@cimat.mx (E. Santibáñez-Cortés); azael.carrillo@cimat.mx (A. Carrillo-Cabrera); yair.castillo@cimat.mx (Y. A. Castillo-Castillo)

usually are labeled as *positive, negative*, and *neutral*. This polarity could be assigned considering the whole sentence or any aspect of it. As it happens in almost everything related to language, most of the huge advances have been done in English, nevertheless, growing efforts are also done in Spanish and other languages. Usually, the methods proposed to deal with this task are based on lexical, syntactic, and semantic features. The lexical and syntactic ones could be expressed through lexicons, emoticons, and exclamation marks, among others. With the Deep Learning bang currently most of the proposed words in the literature used this approach to deal with sentiment analysis. The authors in [2] propose a word embeddings method with an unsupervised learning algorithm using a large Twitter corpus. The contextual semantic relationships and co-occurrence features between words were computed. For building the sentiment feature set, these prior embeddings are combined with n-grams and lexicon data to feed a convolutional neural network.

In [3] is proposed another sentiment analysis method based on sentiment diffusion patterns which tries for analyzing how information diffusion is affected by sentiments in social networks when sentiment polarities change or differ from a tweet to its retweets. A transformer-based word representation is proposed in [4]. This word representation is used as input to a BiLSTM with an attention mechanism network to classify the sentence sentiment. This word representation is called DICE from Deep Intelligent Contextual Embedding. DICE is used to address language ambiguity, capture deep relationships between words, and tackle polysemy, semantics, and syntax issues.

In this paper, we tackle the sentiment analysis of Tripadvisor data. To do this our proposed methodology considers two approaches, the first one employs the BERT Transformer architecture trained in Spanish. The second one uses a set of data augmentation techniques trying to improve the unbalanced dataset.

The manuscript is organized as follows, Section 3 describes the proposed methodology for sentiment analysis problem. Section 4 is shown the obtained results and finally, Section 5 draws our main findings and conclusions.

2. Dataset

The corpus given by the IberEval organizers is conformed by 28,632 opinions labeled with the level of pleasure (polarity) and assessed place (attraction). The problem's objective is to determine the polarity with values between 1 to 5, where 1 reflects the most negative value and 5 the most positive, and intermediate values are reflected as less or more positive or negative. Furthermore, determining the kind of place (hotel, restaurant, and entertainment). Analyzing the dataset, we observed that both data are unbalanced, that is for polarity and for attraction some of their labels are more frequent than others [5, 6, 7, 8](see Figure 1).

Also, we observed that opinions contained emojis symbols, (🍃, 🌹), English words (it is important because the opinion is in Spanish), abbreviations (*km, cm, a/c, etc., Sr.*), symbols (%, #, +, $), grammatical errors (*xq, x, ivamos, muuuy, filolononoooonn*), and some acquisition errors (the tag *'Leer menos'*, white spaces). Considering this, we did pre-processing and data cleaning steps trying to improve the data to get better performance of our classification models.

To illustrate better this dataset, the following is an example of the type of opinions from the

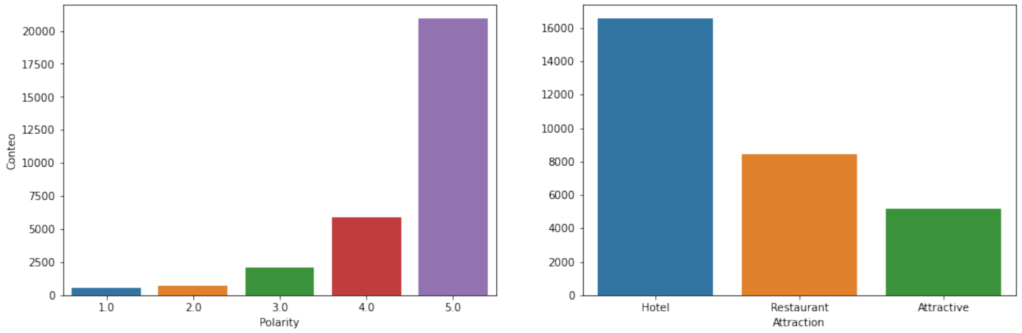

Figure 1: Classes frequencies for both, polarity and attraction problems.

dataset:

Polarity 5, hotel: *Excelentes vacaciones en compañía de amigos y familia, excelentes juegos x parte del equipo de JoySquad y en la playa,* 100% *recomendado para pasar un tiempo increíble con la familia o con tus amigos, sin dudar de la mejor atención! Atención con Kimberly excelente* 😄

3. Proposed Methodology

The proposed methodology considers two approaches, the first one employs the BERT Transformer architecture trained in Spanish. The second one uses a set of data augmentation techniques trying to improve the unbalanced dataset, these techniques are *BO-TextAutoAugment*.

3.1. Data pre-processing

In Section 2 some examples of the opinions in the dataset as well were remarked on the need for pre-processing the data by a specific task. Considering this, the steps for the pre-processing are:

- Lower case transformation
- punctuation cleaning, weird symbols removal
- Repeated words
- Grammatical errors
- Emojis replacement
- Spanish translation

Furthermore, we did punctuation, weird symbols, and repeated word lists which were observed in a sample of the dataset for each of the two tasks. These lists are shown in Table 3.1. To handle the emojis symbols, we count the frequency of emojis in all the opinions, and only the 80% most frequent were considered. This set of most used emojis was replaced by their meaning specified with words, the 20% less frequent was replaced by the special character ([*emoji_k*]).

Table 1
Examples of modifications done in the dataset in the pre-processing step.

Original	Modified	Original	Modified
1°	primero	a/c	aire acondicionada
x	por	ok	esta bien
q	que	gdl	guadalajara
otel	hotel	sr.	señor
aprox	aproximadamente	am.	de la mañana
$	pesos	ene	enero

Also, there were 200 opinions and some words written in the English language, for instance, *check-in, check-out, valet parking, wifi*, all of which were translated into Spanish, to put all the texts into the same language.

3.2. Metrics and Evaluation

To measure the models' performance we split the provided dataset into two sets, 80% for training and 20% for the test. The metrics used were established by the competition, which is for polarity the mean absolute error (MAE), and for attraction, the F1-Macro score. For the overall evaluation were considered the mean of the inverse of MAE and the F1-Macro scores (see Equation 1).

$$Sentiment_{\text{Res}} = \frac{1}{2}\left(\frac{1}{1 + MAE_{\text{polarity}}} + F1 - M_{\text{attraction}}\right) \tag{1}$$

Where $Sentiment_{\text{Res}}$ is the overall sentiment score, the MAE_{polarity} is the Mean Absolute Error per polarity class, and $F1 - M_{\text{attraction}}$ is the F1-Macro score for the attraction class.

3.3. Generated models

Taking advantage of the number of elements of each class, we established as a default classification the most frequent class; that means, for polarity classification by default the predicted result was *5* and for the attraction was *Hotel*. It is important to state that this procedure was established by observing the training data, for the test only was assigned the most frequent label observed in training, which means, this information is not computed with the test data.

Our best solution for attraction classification was based on a Convolutional Neural Network (CNN) as shown in Figure 2, with Adam optimizer, the cross-entropy loss function, size batch of 16, and run in 50 epochs.

With this model, we reached an F1-Macro score of 0.98 which we considered a very outstanding result. For this, we spent more time on the polarity classification solution.

To tackle the polarity classification two approaches were designed. For the first one, we used the pre-trained models currently available, specifically the **RoBERTuito** was used which is a pre-trained language model for social media text in Spanish [9]. We fine-tuned this model to specialize it on the review data. Using this approach we achieved an MAE of 0.24 in the

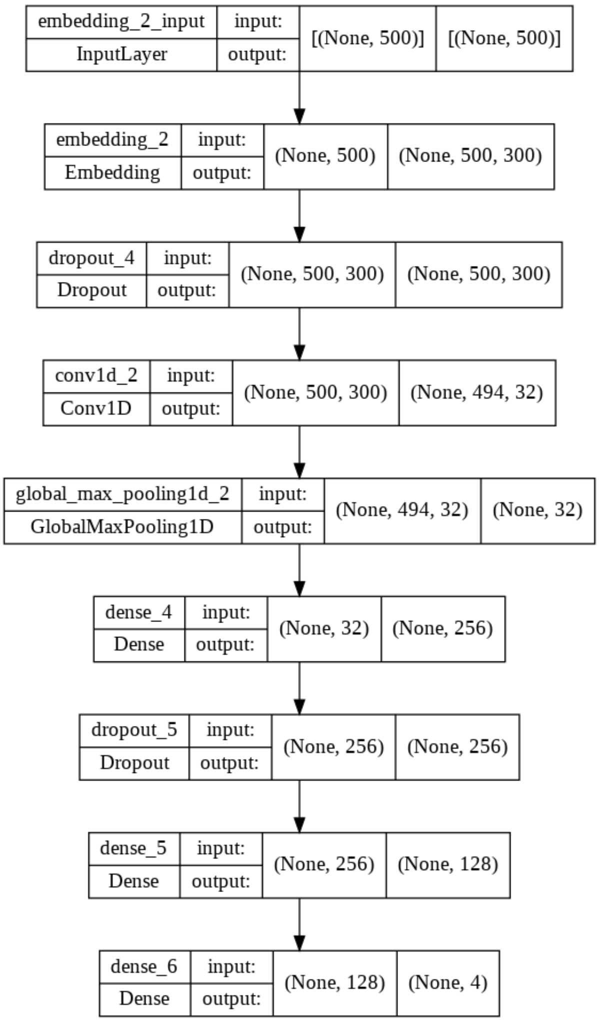
embedding_2_input | input: | output: | InputLayer | [(None, 500)] | [(None, 500)]
embedding_2 | input: | output: | Embedding | (None, 500) | (None, 500, 300)
dropout_4 | input: | output: | Dropout | (None, 500, 300) | (None, 500, 300)
conv1d_2 | input: | output: | Conv1D | (None, 500, 300) | (None, 494, 32)
global_max_pooling1d_2 | input: | output: | GlobalMaxPooling1D | (None, 494, 32) | (None, 32)
dense_4 | input: | output: | Dense | (None, 32) | (None, 256)
dropout_5 | input: | output: | Dropout | (None, 256) | (None, 256)
dense_5 | input: | output: | Dense | (None, 256) | (None, 128)
dense_6 | input: | output: | Dense | (None, 128) | (None, 4)

Figure 2: CNN architecture from the Attraction classification problem.

validation set and 0.26 on the test set. We submitted this approach because we consider that it is an outstanding model and is in the current state-of-the-art approaches.

Since we must deal with unbalanced data, the second approach tries to handle this issue through data augmentation (DA) techniques. In the NLP area, there are different approaches to do data augmentation, these could be classified according to their data generation as; rules-based, interpolation, and model-based [10].

Usually, more than one DA technique is used in a sequential way, and these set of techniques is called *DA policies*.

A policy $\mathscr{P}$ of fixed size N is defined as a set of DA operations:

$$\mathscr{P}_N = \{\mathscr{O}_1, \cdots, \mathscr{O}_N, \}, \tag{2}$$

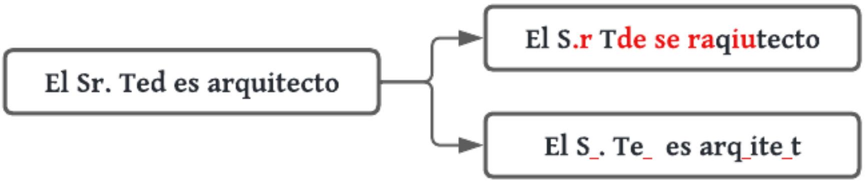

Figure 3: An example of a rule-based DA technique is noise injection where are swap and eliminate characters from the texts in a random way.

where the operation $\mathcal{O}$ is an individual component responsible to apply a DA method. In particular, in this work, we used the *BO-TextAutoAugment* technique to find the best DA approach [11].
The considered model is the CNN architecture shown in Figure 2. With this second approach, we achieved an MAE of 0.26 in training and 0.27 in the test set. It is important to say that the worst results were reached without considering the DA scheme.

4. Results

Our first submission was called CIMAT2020_Beto, it achieves 8th place, and the second one called CIMAT_*BO-TextAutoAugment* reaches the 12th place with an overall performance of 0.869 and 0.882, respectively. In comparison with the best solutions, we can see the really small gaps between metrics values with approximately 2% of difference. The best results and also our results are shown in Table 4. In this table, it can be seen how close the first scores are between them.

Table 2
Best scores and our scores in sentiment analysis competition on REST-MEX 2022.

Solution	**MAE**	**F1-Macro**	$Sentiment_{Res}$
	(*polarity*)	(*attraction*)	
1st. UMU-Team	0.258	0.990	0.892
2nd. UC3M	0.260	0.988	0.890
3rd. CIMAT MTY-GTO	0.264	0.988	0.889
8th. CIMAT2020_Beto	0.267	0.977	0.882
12th. CIMAT_*BO-TextAutoAugment*	0.315	0.977	0.869
More frequent class assignation	0.476	0.145	0.386

5. Conclusions

This work presented the solutions for sentiment analysis, for both opinions and types of attraction. We propose a methodology of pre-processing and data labeling, a pre-trained deep learning network, and also data augmentation scheme. We reached 8th and 12th places in each sub-task, respectively. The most difficult and time-consuming task we did was reviewing the dataset to clean and pre-process it.

References

[1] D. Zimbra, A. Abbasi, D. Zeng, H. Chen, The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation, ACM Transactions on Management Information Systems (TMIS) 9 (2018) 1–29. doi:https://doi.org/10.1145/3185045.

[2] Z. Jianqiang, G. Xiaolin, Z. Xuejun, Deep convolution neural networks for twitter sentiment analysis, IEEE Access 6 (2018) 23253–23260. doi:https://doi.org/10.1109/ACCESS.2017.2776930.

[3] L. Wang, J. Niu, S. Yu, Sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis, IEEE Transactions on Knowledge and Data Engineering 32 (2019) 2026–2039. doi:https://doi.org/10.1109/TKDE.2019.2913641.

[4] U. Naseem, I. Razzak, K. Musial, M. Imran, Transformer based deep intelligent contextual embedding for twitter sentiment analysis, Future Generation Computer Systems 113 (2020) 58–69. doi:https://doi.org/10.1016/j.future.2020.06.050.

[5] 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.

[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 (2021) 1–16. doi:https://doi.org/10.1080/13683500.2021.2007227.

[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. 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] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, 2021. arXiv:2111.09453.

[10] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. H. Hovy, A survey of data augmentation approaches for NLP, CoRR abs/2105.03075 (2021). URL: https://arxiv.org/abs/2105.03075. arXiv:2105.03075.

[11] E. S. Cortés, *BO − TextAutoAugment*: Aumento de Datos automático en NLP usando Optimización Bayesiana, Master's thesis, Mathematics Research Center, CIMAT, 2022.