A Transfer Learning Model for Polarity in Touristic Reviews in Spanish from TripAdvisor

Daniel Mendoza[1], Jorge Ramos-Zavaleta[2] and Adrian Rodríguez[1]

[1]*Monterrey Institute of Technology and Higher Education (ITESM)*
Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, Mexico
[2]*Center for Research in Mathematics (CIMAT)*
De Jalisco s/n, Valenciana, 36023 Guanajuato, Mexico

Abstract

REST-MEX 2022 (Sentiment Analysis Track) is one of the IberLEF 2022 tasks, dedicated to provide good polarity classification for reviews of tourists that traveled around Mexico and leave a review for specific places in TripAdvisor from 2002 to 2021. Polarity analysis of reviews can help to develop intelligent systems for mexican tourism to improve the quality of service and the supply of new touristic places for foreign and domestic tourism. For this task, three different approaches were considered but given the results a traditional transfer learning by using a BERT model provided the best results.

Keywords

NLP, Transfer Learning, Opinion Polarity.

1. Introduction

In the last years there has been a tremendous hype about artificial intelligence and its potential to transform business . However, many organizations have struggled to see real benefits to their bottom lines due to AI initiatives [1].

Rapid progress in the advances of deep learning for tasks like image processing and speech recognition in the last years has been impressive and while deep learning certainly has shown valuable applications in business operations and business analytics still has not yet been widely adopted across organizations.

Neural language models seen as distributed word representations were first proposed as a solution for the curse of dimensionality by [2]. Some years later, [3] demonstrated the value of using deep learning for learning distributed representations of words from large unlabeled corpora, then transferred the learnt knowledge to multiple tasks learned simultaneously through further training on labeled datasets (transfer learning).

While deep learning and transfer learning have shown great results still some other methods based on more conceptual features like frequency of words or topics discovery is still used. For this task, a TF-IDF method was applied as a baseline to compare with transfer learning approach for this data.

IberLEF 2022, September 2022, A Coruña, Spain

2. Data

The data set consists of 30,212 rows. 70% of the original data set was used for released for training while the other 30% remaining is used as a test set. Each opinion. Each row of the dataset contains 4 columns:

- **Title:** The title that the tourist himself gave to his opinion. Data type: Text
- **Opinion:** The opinion issued by the tourist. Data type: Text
- **Polarity:** The label that represents the polarity of the opinion. Data type: [1, 2, 3, 4, 5]
- **Attraction:** The label of the type of place of which the opinion is being issued. Data type: [Hotel, Restaurant, Attractive]

The polarity goes from 1, which means the highest degree of dissatisfaction, to 5, which is the highest degree of satisfaction. It can be interpreted in the following way:

1. Very bad
2. Bad
3. Neutral
4. Good
5. Very good

One important thing to mention is that data is not well balanced [4, 5, 6]. Data is heavier in polarities defined as Good and Very good as for attractions Hotel holds more than half of reviews.

In figure 1 is plotted the distribution of the reviews for each type of attraction. For this data can be noticed that Hotel attraction carries the bigger weight in the distribution with a 54.83% of the reviews while Attractive and Restaurant only holds the 17.2% and 27.97% respectively.

Also by polarity an unbalance appears favouring the larger values (Good, Very good) that jointly with the unbalance of attraction's type creates a major issue for the correct training of the data. In figure 2 an small multiples plot shows this joint unbalance of the two categories.

3. Approaches

For this specific task we applied 3 different approaches. The first one, by using TF-IDF because we think that would work as a good baseline [7], and the other two approaches are based on transfer learning as we consider them as the best options to achieve the best results for the metrics stablished for the competition.

The 2 approaches that use the transfer learning approach are somewhat different, for one of the approaches we froze the layer of the BERT model trained and extracted the resulting embeddings, then we perform PCA to reduce the number of features and finally we provide these embeddings to an XGBoost model. The other approach, is a traditional transfer learning by using the BERT layers as input to train the new data in a sequential fashion for the reviews data.

After the results for the 3 approaches the best results were obtained by traditional transfer learning approach, so we consider to train two different version under this approach to our submission for the contest.

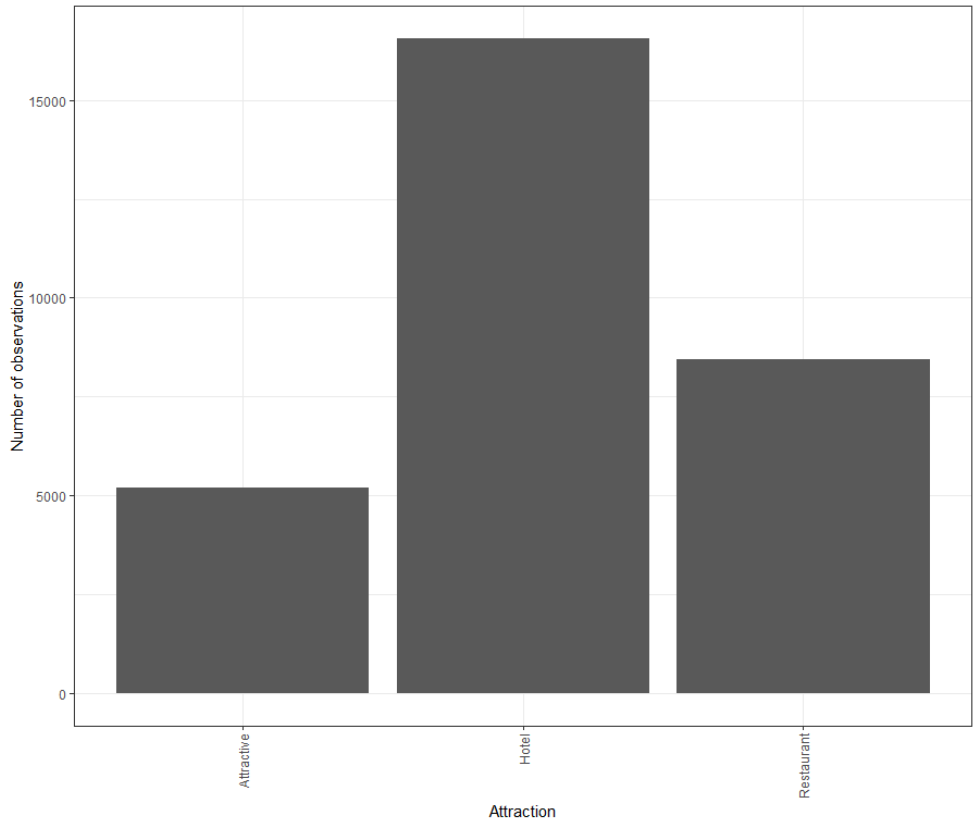

Figure 1: Distribution of reviews for type of attraction.

3.1. TF-IDF

The TF-IDF method has been widely used in the past for modeling text data, for example in [8] is used together with the KNN classifier to classify text data.

To perform TF-IDF for text classification and the search through the documents is necessary to implement a weight matrix. This matrix contains the values of relation between each unique words and documents. This matrix is the initial object for the algorithm, and allows to calculate the individual importance for each document that is in the search.

To allow the search between the documents, each document is represented as a vector in an n-dimensional vector space. Then the weight matrix is then modeled as an $N \times M$ matrix, where N denotes the number of each unique word in a sample of all the documents and M represents the number of documents that is going to be classified.

This weight matrix can be characterized as a relational matrix of word-document, and where each matrix element a_{ij} represents the weight value of word i in the document j.[9]

To calculate the weights for this matrix is necessary to consider 2 elements: TF - term frequency of term i in document j and IDF - inverse document frequency of term i. Then the weights of the matrix can be calculated as

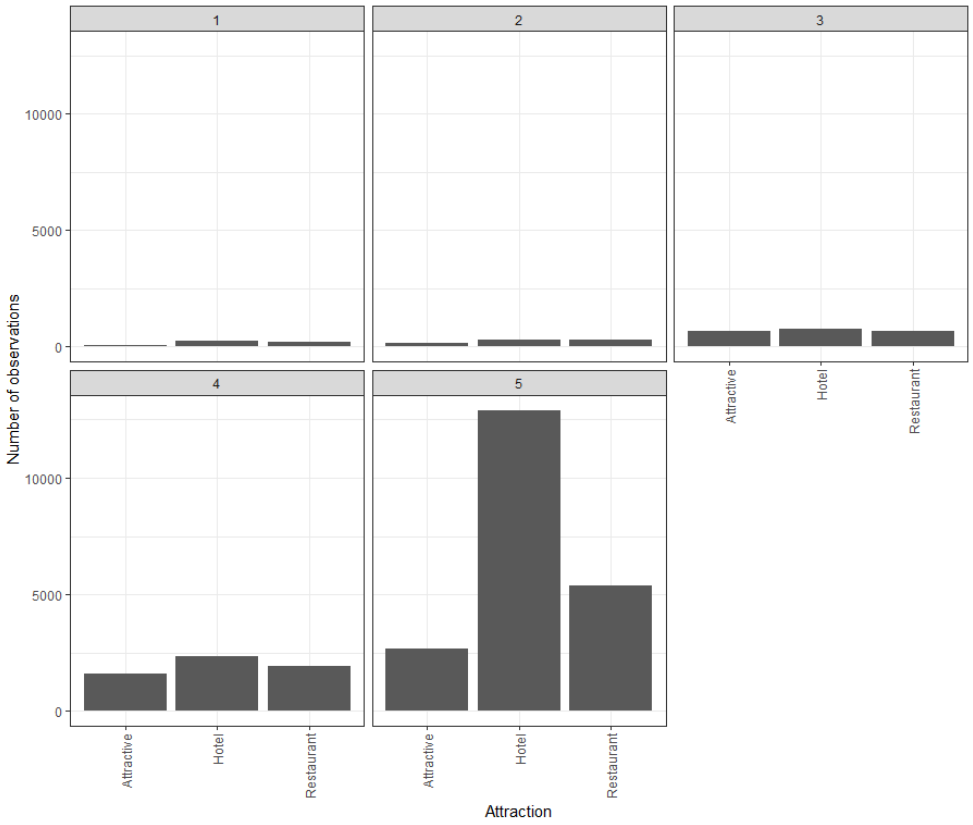

Figure 2: Distribution of reviews for type of attraction and polarity.

$$a_{ij} = tf_{ij} idf_i = tf_{ij} \times log_2 \left(\frac{N}{df_i} \right)$$

where N is the number of documents in the collection, tf_{ij} is the term frequency of term i in document j and df_i is the document frequency of term i in the collection.

3.2. BERT Embeddings + PCA + XGBoost

This is a similar approach to the one we used in the previous rest-mex 2021 for the recommendation task [10], but in this case, we were forced to apply a dimensional reduction technique given that we have a large number of observations (large considering the time and memory to train a Boosting learning method as XGBoost) and a large number of features (embedding dimensionality is 768).

This is a summarized version of the process for training in this approach:

- Load pretrained BERT model
- For every opinion/sentence we get an embedding representation

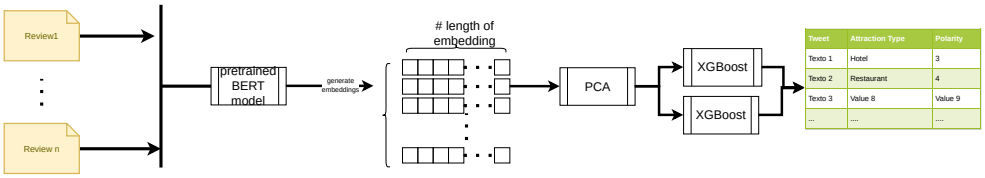

Figure 3: BERT Embeddings +PCA+XGboost schema.

- Run PCA over embeddings to reduce dimensionality (For computation purpose)
- Train an XGBoost taking the first 100 PCA vectors as predictor variables and taking polarity as y

An schematic diagram of the steps involved could be visualized in Figure 3

Given that for this approach the embeddings are not objectivized to sentiment analysis, the results were better than TFIDF, but even applying hyperparameter optimization to XGBoost, we could not get a significative improvement in the metrics.

3.3. BERT Transfer Learning

For this approach we load a model and objectivize to classification using transfer learning, in this way we retrain the model on the given dataset for a certain number of epochs, as a result we got a new model (with their respective new weights) fine-tunned for polarity and attraction type prediction for the TripAdvisor spanish reviews.

- Load an empty BERT model using BERT architecture as a template from huggingfaces
- Preprocess data (tokenize,padding,truncation) in order to get the data in the way the model expect.
- CrossValidation Attraction Type Model (Epochs , learning rate)
- CrossValidation Polarity Model (Epochs , learning rate)
- Train both Models (using the best hyperparemeters obtained in CV) one used to predict the attraction type and another model to predict polarity
  - Save model weights for best metric
  - Save model weights for best/min loss
- Generate the results dataset mixing predictions from both models

An schematic diagram of this process is presented in figure 4.

4. Results and metrics

In the validation set we predict using the 2 trained models (Best Metric and Best/Min Loss) submitted. In table 1 the classification matrix is shown for the Polarity part in a validation set for the Best Metric trained model while in 3 is shown the classification matrix for the Best

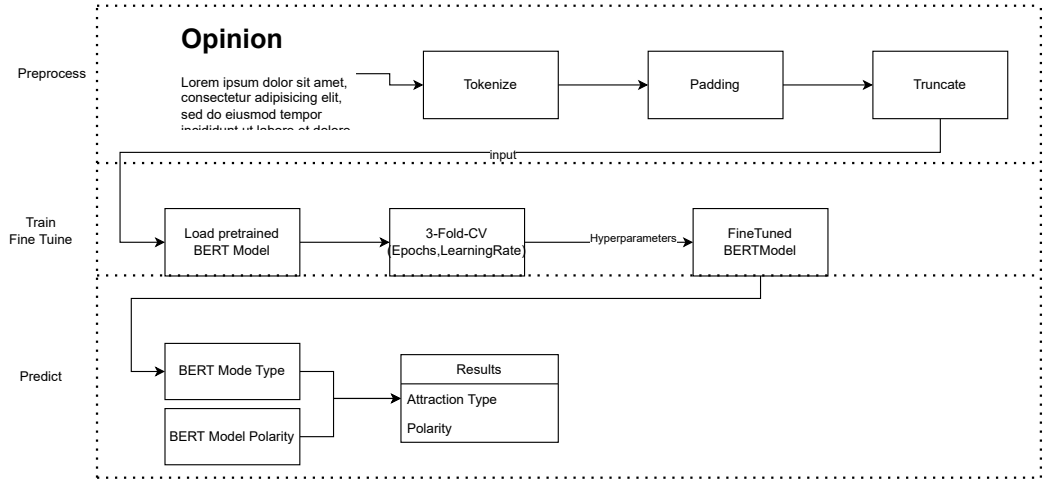

Figure 4: BERT Transfer Learning

PredictedPolarity	1	2	3	4	5
Polarity					
1	64	10	40	2	3
2	25	30	75	16	3
3	15	13	211	172	27
4	3	1	56	477	627
5	0	0	9	253	3911

Table 1
Confusion Matrix Polarity Best Metric Model (Validation Set)

Min/loss model. The classification for both models is similar but the Best Metric Model shows a better performance classifying the polarities 4 and 5.

In table 2 and table 4 the results for the classification of attraction for the Best Metric and Best Min/Loss models are shown respectively. Both models have very similar results and no clear difference can be distinguished in the classification of each type of attraction.

PredictedAttraction	Attractive	Hotel	Restaurant
Attraction			
Attractive	1003	3	3
Hotel	1	3342	26
Restaurant	3	32	1630

Table 2
Confusion Matrix Attraction Best Metric Model (Validation Set)

Finally for both models the sentiment was calculated in the validation set by giving the next results

- $Sentiment_{Res_k}$ (Best/Min Loss) 0.8959

PredictedPolarity2	1	2	3	4	5
Polarity					
1	70	28	14	4	3
2	32	56	44	14	3
3	17	49	233	122	17
4	3	5	82	616	458
5	0	1	14	416	3742

Table 3
Confusion Matrix Polarity Submission 2 (Validation Set)

PredictedAttraction2	Attractive	Hotel	Restaurant
Attraction			
Attractive	1008	0	1
Hotel	2	3331	36
Restaurant	6	23	1636

Table 4
Confusion Matrix Attraction Submission 2 (Validation Set)

- $Sentiment_{Res_k}$ (Best Metric) 0.8983

Where $Sentiment_{Res_k} = \frac{\frac{1}{1+MAE_k}+MacroF1_k}{2}$ and $MacroF1_k = \frac{F1_A(k)+F1_H(k)+F1_R(k)}{3}$

5. Conclussions and Competition Results

For the competition even when both transfer learning approaches achieved a good result in the training data the best results where obtained by the direct transfer learning approach. The fourth and fifth place in the final rank was achieved by considering small variations of this approach. The difference between the top 5 rank results is minimal, maybe because a similar approach was used by the other participants probably by using a different base model or using a different combination of epochs/batch size or other hyperparameters.

The final results for both models trained and a baseline using the majority class are shown in table 5. It can be seen that both approaches are similar in their results and no clear winner can be chosen between the two except for the macro F-Measure where the direct transfer learning approach under the Best Metric training got a better result.

Approach	MAE	Accuracy	Macro F-Measure Polarity	Accuracy Attraction	Macro F-Measure Attraction
TF Best Metric	0.267	75.85	0.542	98.88	0.989
TF Best Min Loss	0.269	76.29	0.495	98.624	0.986
Baseline (Majority Class)	0.476	70.03	0.165	54.877	0.236

Table 5
Final results for both models and a baseline considering a majority class approach to compare.

There are other approaches and preprocessing that we could not take into account because of lack of time; those approaches could possibly helped to add some power for the final score,

but we consider to use them in future versions for the contest.

One of these approaches that we could not consider for this occasion is to avoid the limitation for the number of input tokens. Other possible improvements that can be made are the use of paddings as if we were considering a CNN model and also the selection of the first and final tokens in order for the input to obtain a better source of information for the model.

References

[1] T. Fountaine, B. McCarthy, T. Saleh, Building the ai-powered organization, Harvard Business Review 97 (2019) 62–73.

[2] Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, Advances in Neural Information Processing Systems 13 (2000).

[3] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, 2008, pp. 160–167.

[4] 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:http://hdl.handle.net/10045/117505.

[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] 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.

[7] 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:10.1080/13683500.2021.2007227.

[8] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, volume 242, Citeseer, 2003, pp. 29–48.

[9] T. Bruno, M. Sasa, D. Donko, Knn with tf-idf based framework for text categorization, volume 69, 2013. doi:10.1016/j.proeng.2014.03.129.

[10] J. Arreola, L. Garcia, J. Ramos-Zavaleta, A. Rodrıguez, An embeddings based recommendation system for mexican tourism. submission to the rest-mex shared task at iberlef 2021 (2021).