An Embeddings Based Recommendation System for Mexican Tourism. Submission to the REST-MEX Shared Task at IberLEF 2021

Jean Arreola, Lizeth Garcia, Jorge Ramos-Zavaleta, and Adrian Rodríguez

Center for Research in Mathematics, Monterrey, México
jean.arreola@cimat.mx
lizeth.garcia@cimat.mx
jorge.ramos@cimat.mx
adrian.rodriguez@cimat.mx

Abstract. REST-MEX 2021 (Recommendation System for Text Mexican Tourism) is one of the IberLEF 2021 tasks, dedicated to generate recommendation systems for tourist sites based on an user's profile's affinity compared to each place description. Considering the importance of tourism in the economy, it is vitally important to generate Spanish resources that allow the generation of systems that help to develop intelligent systems in tourism. Considering the above, we proposed a system based on distributed representations of texts, using the BERT approach. We did not use any handcrafted features or external datasets as prior information.

Keywords: BERT · Embedding · Recommendation systems · Mexican tourism

1 Introduction

Tourism is an important economic sector in Mexico, OECD says the sector directly accounts for 8.5 percent of GDP [5]. Tourism also represents one of the main activities in Nayarit. According to the Secretariat of Tourism, Riviera Nayarit was the 5th most visited beach destination in Mexico in 2019, thanks to Nayarit offers many beaches along a 200-mile stretch of the Pacific Coast [6].

The importance of tourism lies in the fact that it contributes significantly to Nayarit employment, foreign direct investment, and economic growth.

With the new advent of technologies, an increasing number of tourists search for information on the Internet to help themselves make their travel decisions

IberLEF 2021, September 2021, Málaga, Spain.

[7]. However, information on the internet has grown exponentially, and tourists are usually overwhelmed by the large quantity of travel information that can be found. The implementation of Recommender systems can help tourists in managing large amounts of available information and ease their travel decisions.

Natural Language Processing (NLP) is an artificial intelligence area that can help restore tourism by developing systems that consider the user and destination information to recommend the places where the user will have better tourist experiences. The advent of Transformer-based pre-trained language models has greatly improved the accessibility of the average user to high-performing models. [2]

The ease of use of pre-trained NLP models justifies their use in diverse applications. This is explained because now practitioners from disciplines outside the computer science realm could fine-tune their own models for their own domain specific downstream tasks. [2]

2 Data

The dataset used in this work was provided by the contest organizers. This collection was extracted for tourists who traveled to the most representative places of Nayarit, Mexico, and who shared their satisfaction on TripAdvisor between 2010 and 2020. Each class of satisfaction is an integer between [1, 5], where 1 represents the most negative satisfaction and 5 the most positive.

The dataset is divided in two:

User information: Gender (Male or female) and place of origin. It was taken into account the state of origin if the tourist is Mexican, if not, they will have the 'foreign' label. The description that the same user put on TripAdvisor (Some users do not share a description) and opinions that he has put of other places (not necessarily Mexican) on TripAdvisor.

Place information: A brief text description of the place and a series of representative characteristics of the place as a type of tourism that can be done there (adventure, beach, relaxation, etc.) and other features like the type of tourist groups that visit these places (family, couple, friends),and some other characteristics to be considered by the potential tourists.

This corpus consists of $2,263$ instances with $2,033$ tourists and 18 famous Mexican places. Approximately, 70% of the corpus is used for training purposes, while the remaining 30% is used for test. The detailed statistics of the satisfaction rate in the training dataset are shown in Table 1. In this Table, it can be see we face an imbalanced classification problem, where 83 percent of the

observations have a 4 or 5 rating.

Label	Tourists	Percentage
1	45	2.84
2	53	3.35
3	267	10.56
4	457	28.89
5	860	54.36

Table 1. Description of dataset for training

3 Modelling Approach

For the contest, we tried several approaches, from a model based on creating features of the ratings of the places to the incorporation of NLP for the representation of the comments and descriptions of the tourist sites. In the end, a model with NLP was the most promising. We generated two variants of the same model, where the most relevant change was in the way that the vector word representations were obtained.

3.1 Word2Vec

In [9] Mikolov introduced the Skip-gram model, an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data. Unlike most of the previously used neural network architectures for learning word vectors, training of the Skip-gram model does not involve dense matrix multiplications, which makes training extremely eficcient.

The word embeddings computed by Mikolov's method are very interesting because the output vectors explicitly encode many linguistic regularities and patterns that are useful for machine learning methods that are not meant originally to deal with text data.

In [4] the Word2Vec model is extended from the original Skip-gram model, and is stated how varying some hyperparemeters as subsampling and negative sampling, can help achieve better word vector representations, also is shown how applying linear combinations of word representations can also produce new meaningful vectors which could be helpful to represent sentences or documents, although recently exists extensions specfic for those purposes.

3.2 BERT

BERT stands for Bidirectional Encoder Representations from Transformers, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. These pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks [3].

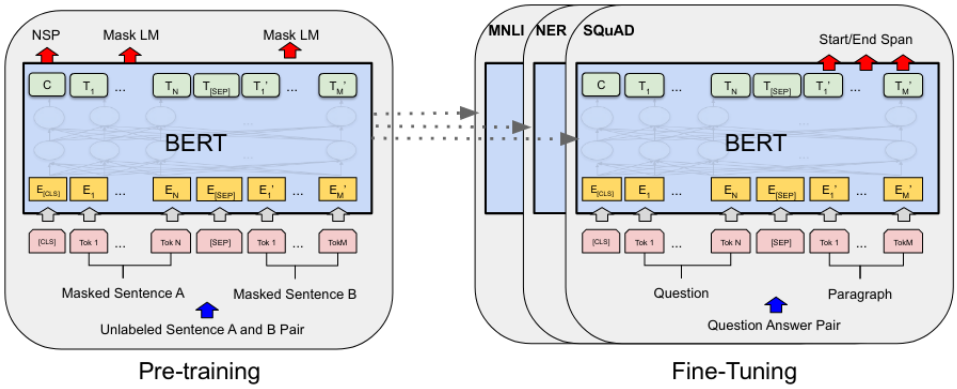

Fig. 1. Fine Tuning BERT for different NLP Tasks

In [2] a fine tuned BERT Model with Spanish data is presented and compared against an mBERT (multilingual BERT Model), showing better results in some NLP in spanish Tasks (Natural Language Inference (XNLI), Paraphrasing (PAWS-X) , Named Entity Recognition (NER), Part-of-Speech Tagging (POS), Document Classification (MLDoc). These results are shown in Table 2.

Model	XNLI	PAWS-X	NER	POS	MLDoc
Best mBERT	78.50	89.00	87.38	97.10	95.70
es-BERT uncased	80.15	**89.55**	82.67	98.44	**96.12**
es-BERT cased	**82.01**	89.05	**88.43**	**98.97**	95.60

Table 2. es-BERT VS mBERT. Comparison table found in [2]

3.3 Data Preprocessing

When dealing with user-generated content, it is common to find quality problems in the texts. In these cases, the dataset contains some issues as poor spelling and the presence of multiple languages in comments. We handled these errors

differently in each of the two proposed approaches.

In the first approach, the followings steps were applied to each review of the corpus:

1. We identified the language of the review.
2. An entity extractor was implemented. Places and organizations are likely to be detected as misspelled, so they need to be identified and added to the dictionary.
3. We removed symbols and some numbers like coordinates.
4. A spell checker was applied to clean the review. A different engine is used depending on the language.
5. Finally, reviews in another language were translated to Spanish.

The preprocessing flow can be seen in figure 2.

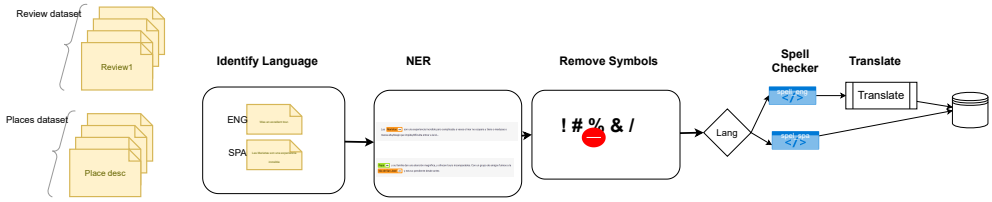

Fig. 2. Preprocessing of the data.

3.4 First Approach

Once the preprocessing criteria were applied, and with the information clean and completely in Spanish, we integrated the information for each user taking advantage of the user ratings.

A Doc2Vec model was trained using both the information of the reiews and the places description and the model was applied to each review. A centroid embedding was generated for each user by taking the mean of his reviews embeddings.

The data presented the cold-start problem so there are users with no reviews, even when there are several attempts to solve this like in [8]. For this work, a global embedding centroid was generated by taking the mean of all the reviews embeddings and imputed for them. Figure 3 shows the centroids' generation process.

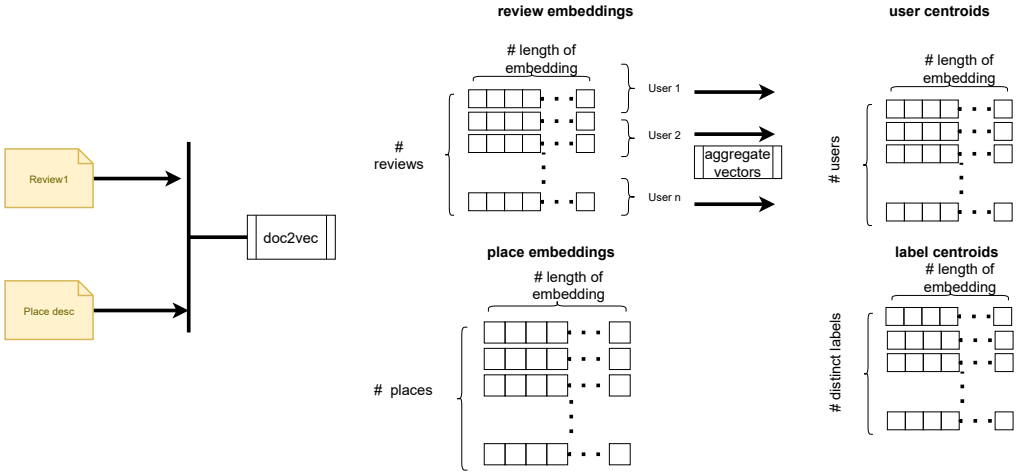

Fig. 3. Reviews's centroids and places's embeddings.

The Doc2Vec model was also applied to the Place information column of the dataset. The obtained embeddings were matched with the reviews' centroid embeddings through similarity metrics, and these embeddings were assigned to the design matrix. Finally, for the other user variables, a hot encoding was applied to be incorporated in the design matrix to be modeled through a Neural Network with one hidden layer and ordinal encoding to deal with the unbalanced problem of the data.

3.5 Second Approach

The second approach took fewer steps because we did not include the users' reviews. This because the resulting vectors are pretty large, and incorporate the reviews data into the data matrix as a new column did not provide any difference. Nevertheless, this approach showed better results in the training and test data than the doc2vec approach with reviews.

In this case, we used the processed data for places' descriptions considered in the first approach. Also, we redecode the variable from the traveler's origin by using a dummy transfomation indicating if the traveler was local(Mexican) or foreign.

For the categorical variables of users (Gender, place of origin, and type of travel), we enrich the categories descriptions to help the BERT model to capture the users' features into the vectors. We wrapped each categorical variable with extra text in a way that highlights the category; for example, for the dummy variable of the place of origin, if the traveler is local, the text generated is *I'm*

a Mexican traveler and the text generated is *I'm a foreign knowing Mexico* in other case.

Once we wrapped each variable, we concatenate these new variables with the place's description, and the BERT model was applied to this larger text.

Then, we applied a simple neural network with one hidden layer with ordinal encoding, and an XGBoost model. Both models can deal with the unbalanced problem for this specific problem.

4 Results and discussion

We presented two modeling approaches for the REST-MEX competition. The results for both systems were outstanding, by achieving the first and second place of the task.

The results of both systems and the baseline of the contest are presented in the table 3.

System	**MAE**	**RMSE**	**Accuracy**	**F-Measure**	**Recall**	**Precision**
BERT+XgBoost	0.317	0.755	77.287	0.505	0.529	0.498
Doc2Vec+NN	0.329	0.765	76.220	0.473	0.494	0.462
Baseline model	0.733	1.238	53.811	0.140	0.108	0.2

Table 3. Contest overall results for both systems

Even though the BERT model presents the best result, it can still be improved by considering the user's reviews. We could not consider the reviews in the BERT model because the model output is 768 entries long, and when the idea of the centroids generated was integrated the BERT approach did not cause any significant improvement.

In future work, we would like to consider a different approach to integrate the reviews in the BERT model to deal with the cold start problem and find a way to reduce the size of BERT vectors to decrease the training and prediction time.

References

1. Álvarez-Carmona, Miguel Á and Aranda, Ramón and Arce-Cárdenas, Samuel and Fajardo-Delgado, Daniel and Guerrero-Rodríguez, Rafael and López-Monroy, A. Pastor and Martínez-Miranda Juan and Pérez-Espinoza, Humberto and Rodríguez-González, Ansel. (2021). Overview of Rest-Mex at IberLEF 2021: Recommendation

System for Text Mexican Tourism. In Procesamiento del Lenguaje Natural (Vol. 67)
2. Cañete, José and Chaperon, Gabriel and Fuentes, Rodrigo and Ho, Jou-Hui and Kang, Hojin and Pérez, Jorge. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. PML4DC at ICLR 2020.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
4. Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.
5. OECD (2017), Tourism Policy Review of Mexico, OECD Studies on Tourism, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264266575-en
6. SECTUR (2019), Resultados de la Actividad Turística Enero 2019, Secretaría de Turismo. http://www.datatur.sectur.gob.mx/SitePages/versionesRAT.aspx
7. S. Praveenkumar. (2014). Internet Marketing in Tourism. In Indian Journal of Applied Research (Vol. 4, issue 11).
8. Suryana, N., Basari, H., and Bin, A. S. (2018). An understanding and approach solution for cold start problem associated with recommender system: A literature review. Journal of Theoretical & Applied Information Technology, 96(9).
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.
10. Wang, Y., Chan, S. C. F., and Ngai, G. (2012). Applicability of demographic recommender system to tourist attractions: a case study on trip advisor. In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (Vol. 3, pp. 97-101). IEEE.