Recommendation System Rest-Mex 2022 for Mexican Tourism Using Natural Language Processing

Santiago Veigas Ramírez[1], Daniel Martínez Davies[1] and Isabel Segura Bedmar[1]

[1]Universidad Carlos III de Madrid, Spain

Abstract

This paper explores the use of Natural Language Processing for the Recommendation System for Mexican Tourism 2022 (Rest-Mex 2022). The aim is to accurately predict the rating a user would give a new place they have never visited before, in order to give better recommendations. Various natural language processing techniques were tried for achieving the best results. In particular, the BERT model produced the best results for processing reviews and generating a prediction score for given users.

Keywords

Natural Language Processing, Tourism, BERT.

1. Introduction

Tourism is an important sector of the Mexican economy, representing approximately 8.5 % of the country's GDP [1]. Additionally, Mexico was the sixth most visited country in 2017. Clearly, tourism is an important activity in the Mexico. As a result, appropriate tools for guaranteeing customer satisfaction when reserving a trip can boost this already strong sector.

Reviews serve as important feedback for gauging the experience held by tourists in a certain location or place. Consequently, these reviews can help identify trends in tourist satisfaction as well as finding locations that attract a targeted demographic. Inspections can be done manually, but these are often tedious and can quickly become an insurmountable task as the number of reviews grows. This is especially relevant when talking about an online tourism platform that stores thousands of reviews.

Natural Language Processing can be used to solve this task. By using artificial intelligence to find patterns in language, user intent and experience can be measured in a review [2], in order to map the users preferences with certain potential locations. This allows for accurately recommending new destinations to customers for their next trip, and increase their satisfaction with both the booking platform and the chosen destination.

2. Rest-Mex Data

The data was provided by the Rest-Mex organizers. It was collected from reviews on the tourism platform TripAdvisor. The users rated their experience of the trip on a scale from 1 to 5.

IberLEF 2022, September 2022, A Coruña, Spain

The given data was divided into three different sets: user reviews, place de- scription and recommendation ratings. Each set had its own attributes:

- **User reviews**: Review, Rating, Place, Average user rating. Each user review was given as an individual file.
- **Place Description**: Place, Description, Keywords.
- **Recommendation ratings**: User, Gender, Place, Location, Date, Type, Rating.

2.1. Data distribution

The original data provided by Rest-Mex was split into 70 % training data and 30 % test data [3]. After merging, but before pre-processing, a total of 1582 entries were identified.

A simple analysis regarding data distribution and correlation was performed. No correlation was found between the classification labels and other parameters such as 'gender', 'place', or 'type'.

Gender. The distribution of comments by gender appeared to be relatively balanced, with the exception of comments with 'Unspecified' (N/I) gender. Fe- male comments represented 53.41 % of the data, male comments 46.46 %, with 'Unspecified' attribute encompassing 0.13

Place. The distribution of comments by place was heavily skewed towards a single tag. 'Islas Marietas' appeared in 53.03 % of the total user comments, followed by 'Playa Los Muertos' and 'The Jazz Foundation' appearing in 9.35 % and 8.97 % of the entries, respectively.

Type. The distribution of comments by type of trip appeared to be skewed towards three different types. The two most prominent ones being 'Family' and 'Couple', which were represented in 37.35 % and 34.26 % of the entries, followed by 'Friends' with 21.26 %. The remaining types of trips appeared with less frequency in the comments: 'Alone' in 5.05 % of the comments, and 'Business' in the remaining 2.08

Label. The data distribution of the label to be classified was not balanced, having a heavy bias towards 5 and 4 star ratings. The top three star ratings accounted for over 93 % of the training data. Label '5 stars' was represented in 54.36 % of the data, '4 stars' in 28.88 %, '3 stars' in 10.55 %, '2 stars' in 3.3 %, and lastly '1 star' in 2.91 %. This over representation means that predicting edge cases is significantly more challenging, as the problem becomes a classification task with skewed class distributions. Thus any trained model will have less examples to work with for the minority classes.

3. Modeling approaches

3.1. Data pre-processing

Review-less users. A first filtering pass was done, removing all of the user reviews that contained the entry 'desconocido', as it appeared to be a byproduct of a SQL query that returned empty data.

User commentary merging. The training and test data provided by Rest- Mex contained different subsets of the total user reviews. Both sets of user reviews were merged, resulting in 26,442 total reviews.

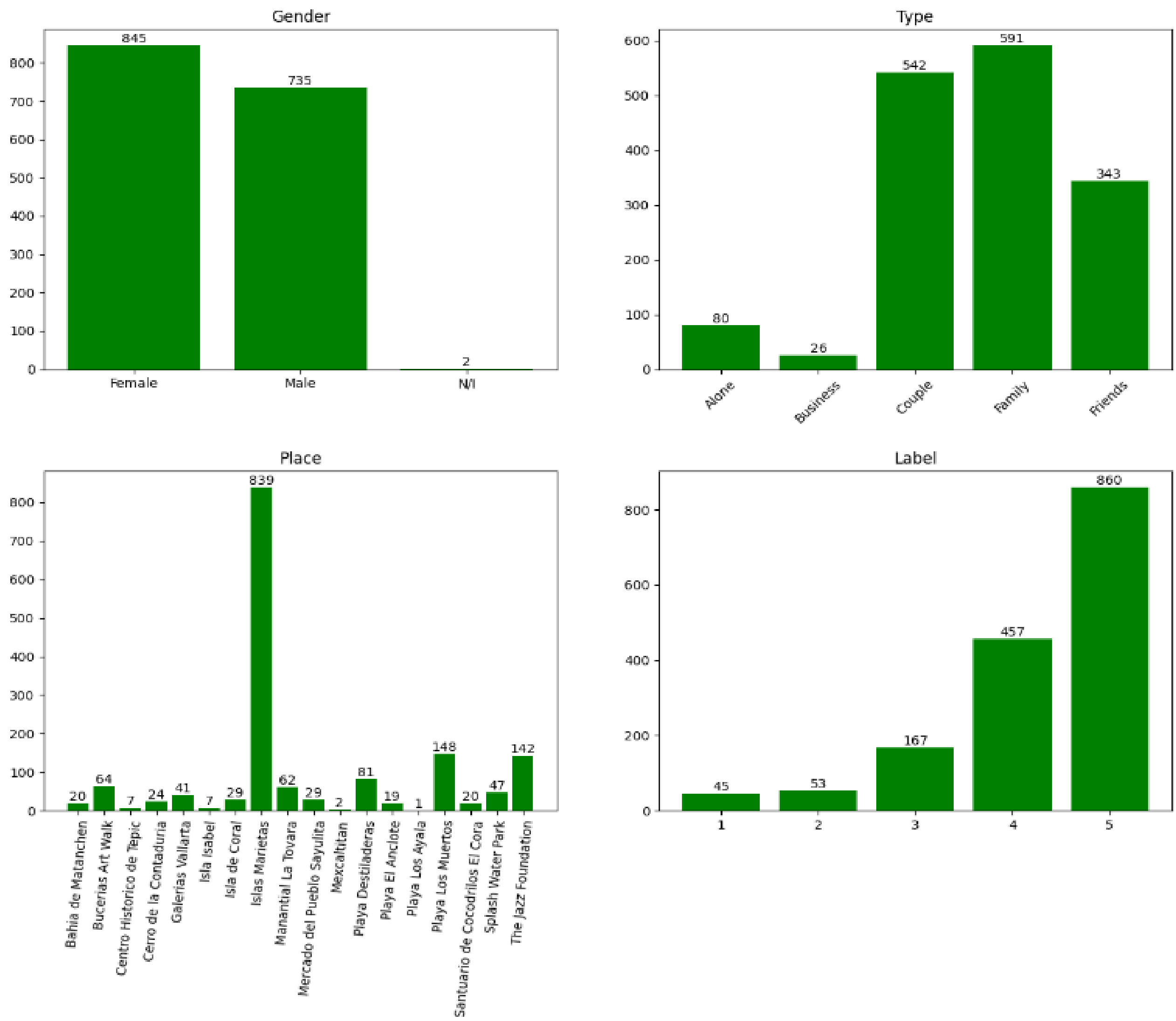

Figure 1: Data distribution.

User deduplication. The provided data did not contain information that would directly allow for the identification of the users from Trip Advisor. As such, the assumption that the data had undergone an anonymization process was made. Duplicate user comments were identified and associated with a singular user; there were many instances were supposedly different users had the same reviews.

These comments were filtered out from the training set, as this meant that reviews would have certain users over-represented, further skewing the data. A total of 603 duplicated users entries were found, from which 234 were unique; the remaining were removed to avoid the over representation of these users.

In certain instances, inconsistencies were identified, such as a single user having different genders after the deduplication process. However, we discarded the possibility of it being an error in the deduplication process by manually verifying the data.

Language filtering. Identifying the language was considered a relevant option after reviewing some of the approaches tried in previous editions of Rest-Mex [4]. Naturally, introducing reviews in a foreign language would give trouble to any model used. This meant that these reviews either had to be removed or translated into Spanish. Both approaches were tried, though the best results were obtained by simply ignoring the reviews that were not in Spanish.

The Google Translate API [5] was used both for identifying the language of the review, and when applicable, for translating the reviews into Spanish. It's worth noting that Google Translate automatically performed grammar correction before the translation step. Around 61 % of the user reviews were written in Spanish, 31 % in English, and 8 % in another language. Non-English reviews were not translated, and instead discarded.

Merging user reviews. The reviews of each user were merged as a single review per user and into a single row. Subsequently, all the rows were combined into a single file. This allowed for a single vector of strings, containing all of the reviews of a user, to be used as input.

The processed user reviews, the place descriptions and the recommendation ratings were all joined together. This meant that the data could now be trained as a natural language processing problem. However, many of the input parameters were not considered relevant. Namely: the user gender, the date, the type and the location were removed. The remaining parameters were merged together into a single column as a vector of strings.

3.2. Bert

Bidirectional Encoder Representations from Transformers (BERT) is a natural language processing method used for classifying a concrete output using a vector of strings as input [6]. BERT uses a masked language approach where, by hiding a word or token in a sentence, the BERT model looks to predict the unknown word.

Layers can be added to BERT models depending on the task to solve. An encoder is used for processing the input. For classification, a decoding layer can be added. For the task at hand, a pre-trained BERT model was used with Spanish as the input language.

3.3. SVM

Support vector machines are a form of supervised learning for tasks involving classification [7]. By creating hyper-planes, support vector machines are able to divide data into classes.

For natural language processing, the input text can be encoded as a vector. Taking this vector as input, a SVM can create a hyper plane for each class in order to correctly classify the given text [8].

A support vector machine was tried for correctly classifying the given reviews. However, since the results were not as good as the BERT model, it was not used for the submission of the Rest-Mex competition.

Table 1
Results Rest-Mex 2022

Team	MAE	Accuracy	Macro F-measure	Macro Recall	Macro Precision
UC3M1	0.717	48.899	0.222	0.2165	0.231
Baseline	0.7416	53.3040	0.139	0.106	0.2
UC3M2	0.746	52.7166	0.138	0.106	0.198

3.4. Final approach

The best results were obtained using a pre-trained BERT model, particularly Google's BERT, which is available through Hugging-Face [9]. Not all parameters were modified, in order to avoid forcing the destruction of the cached model. This would have inadvertently implied training from scratch.

In order to achieve betters predictions for this model, the following hyper- parameters were modified:

- **Number of hidden layers**: 1
- **Activation function**: RELU
- **Evaluation steps**: 1300
- **Epochs**: 3
- **Training batch size**: 8
- **Evaluation batch size**: 1
- **Optimizer**: AdamW [10]

Due to the small amount of data entries, a singular hidden layer was used. Our attempts to increase this value resulted in models that classified all of the entries as the majority class with high confidence (always 5 stars). On the other hand, the size of the hidden layer remained unchanged to prevent the deletion of the cache that contained the weights of the pre-trained model.

The resulting model had trouble predicting classes that were not a 5. Since the BERT model generates a confidence value for each class, it is possible to predict a 3 or a 4 even when the model places a 5 as the best classification. This is done by comparing the confidence value for each class.

Since the initial distribution for each class is known, one of the approaches chosen was to calculate a confidence interval so that the predicted output was a five in 54 % of the cases, a four 30 % of the cases and a three 11 % of the cases. Nevertheless, this approach excluded one and two star ratings. However, since these outcomes only accounted for 5 % of the data distribution, this approach could still produce strong results. This approach was submitted as UC3M1.

4. Evaluation and Results

The two models with the best results for the test data were submitted. The results obtained in the Rest-Mex competition can be seen below in Table 1.

The achieved results were not particularly impressive, even though different approaches were tried. The mean average error produced was barely under the baseline of the most common class. Nevertheless, the results achieved by the other participants were not significantly better. For better results, an increase in the size of the training data would be ideal, especially for a BERT model. In addition, it is hard to tell whether the discrepancies in the training data affected the obtained results.

Other natural language approaches could be tried in future attempts, such as a convolutional neural network applied for text classification. A more thorough pre-processing of the training data could also be tried to eliminate bad examples for classification, or to artificially remove any possible biases from the data.

5. Conclusions

Although the BERT model produced the best results, these were not particularly good. However, it is hard to tell what the cause is for the lackluster results.

Nevertheless, the REST-MEX 2022 competition allowed for the exploration of different natural language processing techniques with real world data with a feel for the data pre-processing needed in order to properly train a model.

References

[1] 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.

[2] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodrıguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Sciences (2022). doi:https://doi.org/10.1177/01655515221100952.

[3] 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).

[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:https://doi.org/10.26342/2021-67-14.

[5] M. O. Prates, P. H. Avelar, L. C. Lamb, Assessing gender bias in machine translation: a case study with google translate, Neural Computing and Applications 32 (2020) 6363–6381.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[8] Y. Li, K. Bontcheva, H. Cunningham, Adapting svm for natural language learning: A case study involving information extraction, Natural Language Engineering 15 (2009) 241–271.

[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[10] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).