Natural Language Processing Approaches for Spanish Tourist Review Analysis: Insights from the REST-MEX 2025 Shared Task

Alvaro Cabrera-Barrio[1], Yuri Melissa Garcia-Niño[1,*] and Victor Alfredo Apaza-Mamani[1]

[1]*Universidad Carlos III de Madrid, Av. de la Universidad, 30, 28911 Leganes, Madrid, España*

Abstract

In this paper, we describe our participation in the IberLEF 2025 shared task on analyzing TripAdvisor reviews of Mexican Magical Towns. The goal is to classify each review by sentiment polarity from 1 to 5, type of site (hotel, restaurant, or attraction), and the associated Magical Town from a predefined list. To address these subtasks, we explored several NLP techniques for Spanish language input, including Support Vector Machines (SVM), fine-tuned transformer models and prompting strategies with large language models. Our models were good but hindered by the difficult task of classifying between 40 diferent Magical Towns, as we will further develop in this document.

Keywords

Sentiment Analysis, Text classification, Natural Language Processing, Magical Towns,

1. Introduction

The rapid growth of travel-review platforms has made it possible to collect vast amounts of textual data reflecting tourists' perceptions and satisfaction[1, 2, 3, 4]. TripAdvisor reviews of Mexico's Magical Towns, towns designated for their cultural heritage and touristic value, contain insights that can guide local stakeholders in enhancing visitor experiences and shaping promotional activities. However, extracting these insights automatically is challenging, especially when dealing with Spanish-language nuances and tourism-specific terminology.

Recent initiatives such as REST-MEX have emphasized the importance of robust sentiment evaluation frameworks for tourism texts in Spanish [5, 6, 7]. Within IberLEF 2025, the REST-MEX shared task requires systems to assign each review a sentiment score from 1 (very negative) to 5 (very positive), determine whether it describes a hotel, restaurant, or attraction, and identify the correct Magical Town among 40 to 60 options [8, 9]. Addressing these subtasks involves dealing with ambiguous expressions of sentiment, overlapping mentions of site categories, and diverse ways users refer to locations.

We decided to approach this challenge as a learning experience, so we applied different NLP techniques to each sub-task independently. First, we used fine-tuning techniques for sentiment polarity and review type classification, leveraging their ability to capture more complex contextual patterns within the reviews. Next, we used more traditional techniques like SVM, just for educational purposes and to be able to compare their performance. In parallel, we employed prompting on large language models to assign reviews to the 40 Magical Towns and polarity, trusting these models' capacity to infer geographic and contextual relationships from a handful of examples (few-shot). By evaluating each technique in isolation, we could compare their strengths and weaknesses and then combine the best results from each approach to produce the final classifications.

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

alvcabre@pa.uc3m.es (A. Cabrera-Barrio); 100555181@alumnos.uc3m.es (Y. M. Garcia-Niño); 100449365@alumnos.uc3m.es (V. A. Apaza-Mamani)

2. Materials and methods

This section outlines the different techniques we employed during our participation in the shared task. Our approach focused on exploring a range of methodologies, from traditional machine learning to more recent advancements in language modeling. Specifically, we experimented with classical methods such as Support Vector Machines (SVM), fine-tuning of transformer-based models like RoBERTa (a BERT variant pre-trained for Spanish), and prompting with various large language models (LLMs). Each technique was applied with the goal of understanding its strengths and limitations in the context of sentiment analysis, site-type classification, and the assignment of Magical Towns.

For this contest, we were provided with a dataset consisting of 208,051 TripAdvisor reviews written in Spanish and covering 40 different tourist destinations categorized as Magical Towns. The dataset was sufficiently large and well-structured, allowing us to train and evaluate our models without the need for artificial data augmentation. This availability of real-world user-generated content enabled us to test various approaches under realistic and meaningful conditions.

2.1. Dataset Analysis

First, we made an analysis of the provided data set. In this section, we are going to describe the general characteristics of the dataset used in the REST-MEX 2025 sentiment evaluation task for Mexican Magical Towns, and provide visual insights from it.

The distribution of words per review, as we can see in Fig. 1, shows that the majority of reviews are relatively short, with a predominant range between 20 and 200 words; longer reviews are rare. For site-type distribution (see Fig. 2), the dataset comprises three categories: Attraction, Hotel, and Restaurant. Restaurant reviews are the most abundant, this might indicate a higher frequency in reviewing dining experiences. All three categories represent a significant volume, so that the data distribution for this subtask is quite well distributed, influencing good results.

Finally, the data set exhibits a clear imbalance in sentiment polarity (see Fig. 3). The very positive class reviews has a dominance over all other classes, this is due to that visitors usually share experiences when they are highly satisfied. This imbalanced distribution will be taken into account for balancing through different techniques (such as oversampling, under-sampling) or specialized loss functions to improve classification performance for less represented classes.

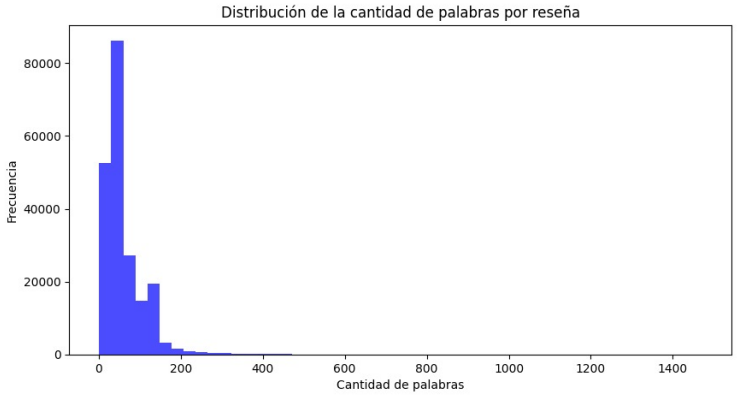

Figure 1: Distribution of the number of words per review. Number of words vs. Frequency.

2.2. SVM

We explored the performance of classical machine learning methods by employing Support Vector Machines (SVM) for both the sentiment polarity and type classification subtasks. For the implementation of the SVM models, we followed a structured approach described as follows:

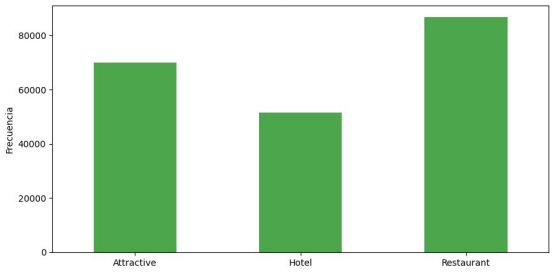

Figure 2: Distribution of site-type in reviews. Site-type vs. Frequency.

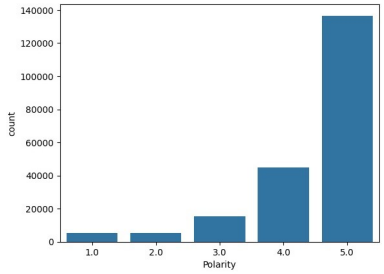

Figure 3: Polarity distribution in the reviews. Polarity vs. Frequency.

2.2.1. Data Preprocessing

The text data was preprocessed using standard techniques to clean and normalize the reviews, including:

- Conversion to lowercase.
- Removal of punctuation, numerical characters, null, and duplicate values.
- Stopword removal, stemming, and lemmatization.
- Tokenization.

2.2.2. Feature Extraction

We utilized the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method to convert the processed text data into numerical features suitable for SVM classification. TF-IDF captures in a efficient way the importance of words within the reviews, aiding the SVM model in differentiating between categories and sentiment levels.

2.2.3. Model Training

Two separate SVM classifiers were developed: one for Type Classification (Categories: Attractive, Hotel, Restaurant), and another for Sentiment Polarity Classification (Ratings from 1 to 5). The models were trained using scikit-learn's SVC implementation with linear kernels, which provided optimal performance given the size and characteristics of the dataset.

2.2.4. Results

The dataset for the type classification task was relatively balanced, enabling the SVM model to achieve robust performance. Evaluation metrics from the test set (313 samples) are shown in Table 1.

The confusion matrix for this task is:

$$\begin{bmatrix} 60 & 10 & 7 \\ 2 & 125 & 4 \\ 2 & 4 & 99 \end{bmatrix}$$

Table 1
SVM Results for Type Classification

Category	Precision	Recall	F1-Score	Support
Attractive	0.94	0.78	0.85	77
Hotel	0.90	0.95	0.93	131
Restaurant	0.90	0.94	0.92	105
Accuracy			**0.91**	313
Macro Avg.	0.91	0.89	0.90	313
Weighted Avg.	0.91	0.91	0.91	313

The model performed best at identifying hotel-related reviews, followed closely by restaurants, and exhibited minor confusion distinguishing attractions.

For the polarity classification task, we employed similar preprocessing and TF-IDF features. However, this task proved more challenging due to the subjectivity and subtlety of sentiment expressions. Overall performance was moderate, as shown below:

- **Accuracy**: 45%
- **Macro Average F1-Score**: 0.45

Class 1 (Very Negative): Precision = 0.53, Recall = 0.57, F1-Score = 0.55
Higher performance suggests that strongly negative sentiments, expressed with clear cues, were easier to detect.

Class 2 (Negative): Precision = 0.36, Recall = 0.35, F1-Score = 0.35

Class 3 (Neutral): Precision = 0.36, Recall = 0.35, F1-Score = 0.36

Class 4 (Positive): Precision = 0.40, Recall = 0.39, F1-Score = 0.40

Class 5 (Very Positive): Precision = 0.59, Recall = 0.59, F1-Score = 0.59
Highest performance among all classes, likely due to explicit positive expressions.

Misclassifications were most common between adjacent classes—especially between Classes 3 (Neutral) and 4 (Positive), and between Classes 1 (Very Negative) and 2 (Negative). These patterns emphasize the limitations of traditional SVM classifiers, which rely on surface-level lexical features and struggle with subtler contextual distinctions.

2.2.5. Discussion

The SVM approach demonstrated strong baseline performance in structured tasks like type classification, aided by clear lexical cues and a balanced dataset. In contrast, the sentiment polarity task revealed the limitations of classical methods when dealing with more nuanced language.

While SVM was successful in recognizing strongly negative and strongly positive sentiments, it struggled with intermediate categories. Therefore, in the following sections, we explore transformer-based fine-tuning and prompting methods, which are better suited to capture deeper contextual information and improve performance on sentiment classification.

2.3. Fine-tuning

For the sentiment polarity and site-type classification subtasks, we employed a fine-tuning approach using the `dccuchile/bert-base-spanish-wwm-cased` transformer model, a BERT variant pretrained specifically on large Spanish corpora [10], [11]. This choice was motivated by its proven effectiveness in capturing syntactic and semantic nuances in Spanish, which is critical for tourism-related user reviews.

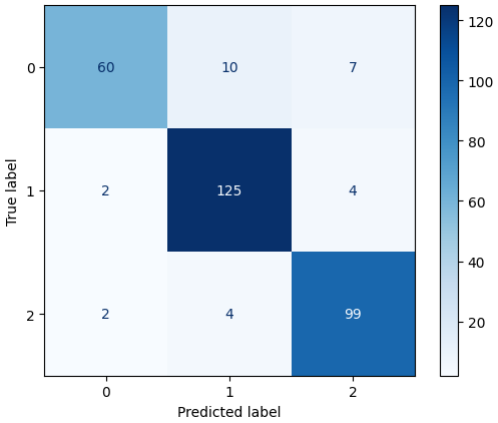

Figure 4: Site-type confusion matrix.

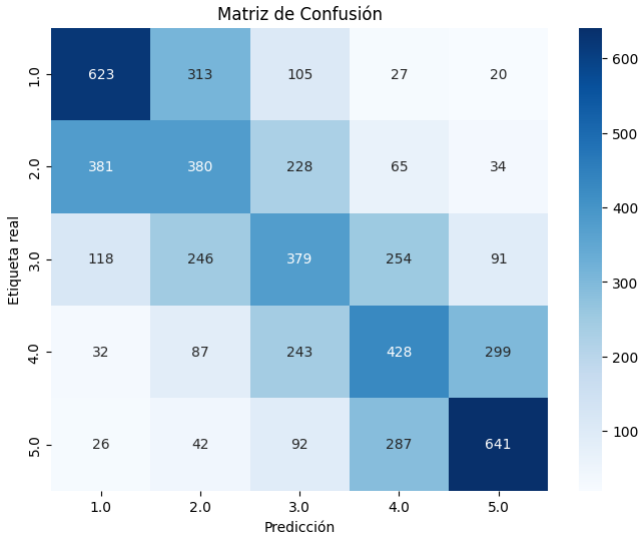

Figure 5: Polarity confusion matrix.

The fine-tuning process involved adding a classification head on top of the pre-trained encoder to predict one of five sentiment classes or one of three site types. The model was implemented using Hugging Face's `transformers` library and trained using PyTorch. To leverage available hardware efficiently, we dynamically assigned training to GPU if available:

```
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
```

We loaded the pre-trained model and set it to output five classes for polarity or three for site type, depending on the subtask, we run the model with 154227 instances, the maximum value we were allowed to run in Colab. Importantly, we froze the first six layers of the BERT encoder to reduce training time and overfitting risk, as the early layers primarily capture general linguistic information that does not need to be adjusted for our specific classification tasks:

```
for name, param in model.bert.encoder.layer[:6].named_parameters():
param.requires_grad = False
```

We used the AdamW optimizer with a learning rate of 1×10^{-5}, a common setting for transformer fine-tuning, along with CrossEntropyLoss as the objective function. This combination provides stable convergence in multi-class classification tasks:

```
optimizer = AdamW(model.parameters(), lr=1e-5)
criterion = nn.CrossEntropyLoss()
```

Training was conducted over a maximum of 10 epochs with early stopping based on validation loss to prevent overfitting. After each epoch, we evaluated model performance on a held-out validation set, computing both loss and accuracy. The best model (i.e., the one with the lowest validation loss) was saved for future use:

```
model.save_pretrained(best_model_dir)
tokenizer.save_pretrained(best_model_dir)
```

In addition to saving the model, we plotted loss and accuracy curves across epochs to visually inspect convergence and learning behavior. These visualizations confirmed that our model improved steadily during the first few epochs to around epoch 2, after,see figure 9 .

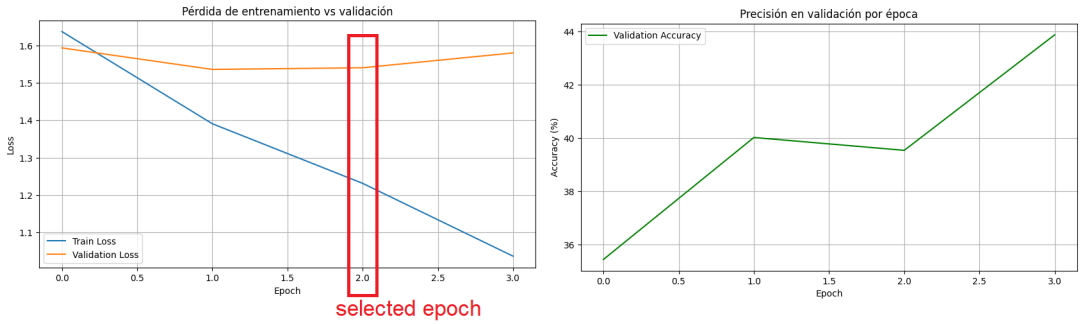

Figure 6: Training loss vs. validation loss applying fine tuning to "Polarity"

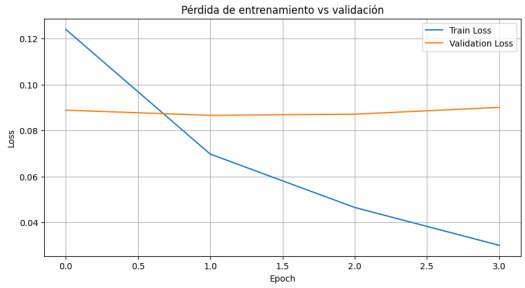

Figure 7: Training loss vs. validation loss applying fine tuning to "Type"

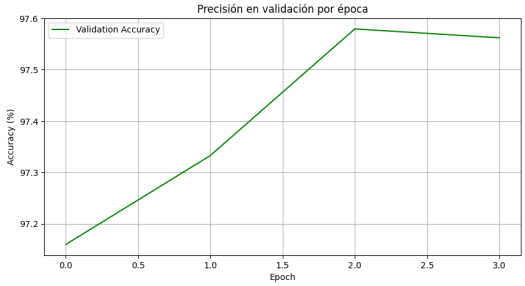

Figure 8: Training loss vs. validation loss applying fine tuning to "Type"

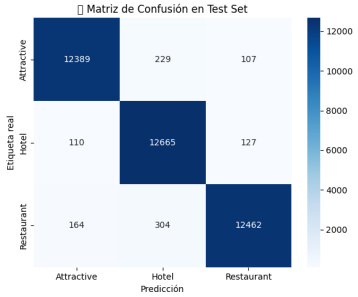

Figure 9: Confusion transformers matrix for "Type"

After training, the model was evaluated on a separate test set using standard metrics: accuracy, precision, recall, and F1-score. We also generated detailed classification reports for each class, allowing us to analyze performance per label and detect any potential imbalance-related issues.

The fine-tuning strategy proved to be especially effective for polarity classification, yielding significantly higher accuracy than the SVM baseline. For site-type classification, the model also performed well, achieving macro-F1 scores close to 0.96 in the official evaluation (see Table 4). However, as we will discuss in later sections, this approach was not scalable for the 40-class Magical Town classification subtask due to computational limitations.

Table 2
Comparison of results: Polarity vs Types using fine tuning

Métrica	Polarity	Types
Accuracy	0.6678	0.9730
Macro avg (F1)	0.6033	0.9730
Weighted avg F1	0.6601	0.9732

In summary, fine-tuning a Spanish BERT variant offered a strong balance of performance and practicality for sentiment and type classification subtasks. The combination of frozen layers, early stopping, and appropriate hyperparameter choices helped us train an effective model without extensive resources.

2.4. Prompting

Regarding prompting, we decided early on to apply it to the Magical Town classification subtask. This decision was motivated by the high complexity of the task, as it involves assigning each review to one of many possible classes. Although we had access to a large training dataset and it was theoretically possible to solve this subtask properly by fine-tuning a transformer model such as BERT in Spanish, we concluded that training a custom fine-tuned model for this multi-class problem was not feasible due to the computational constraints of our environment.

Instead, we opted for large language models and focused our efforts on prompting strategies. Specifically, we used Google's Gemini model through its API and conducted extensive testing with several versions, both with and without reasoning capabilities. Our experiments aimed to find the right balance between response quality, execution speed, and cost. For the final submission, we selected the `2.5-flash` model, which was still in preview at the time.

When implementing the prompting strategy, we chose not to perform any data preprocessing. Since large language models (LLMs) are designed to handle raw, unstructured text effectively, applying traditional preprocessing techniques such as lowercasing, stemming, or stopword removal is neither necessary nor particularly beneficial in this context.

To enhance the model's performance in classification, we employed a few-shot prompting strategy, where the model is guided by multiple input-output examples. This approach has been widely recog-

nized as effective for enabling language models to perform tasks without additional training, especially in settings where labeled data is scarce or fine-tuning is not practical [12].

In addition, we explored both fine-tuning and prompting methods for the sentiment polarity subtask, aiming to improve the initial results. For fine-tuning, we used a BERT based transformer model, freezing the initial layers and training it for a 5-class classification task. This setup allowed us to directly compare the two paradigms: prompting and fine-tuning. During implementation and evaluation, prompting initially yielded better results; however, as we will show later in this paper, this advantage didn't hold on the test dataset.

3. Results

Two different pipeline configurations were used at the time of submitting the results. In the first configuration, we applied fine-tuning for both polarity classification and review type prediction, while prompting was used for the Magical Town classification. In the second configuration, we replaced the fine-tuned model for polarity classification with a prompting-based approach, keeping the other components unchanged.

Table 3
Comparison of system configurations for submitted runs

Run	Polarity	Type	Magical Town
AVYus_1	BERT fine-tuning	BERT fine-tuning	Prompting
AVYus_2	Prompting	BERT fine-tuning	Prompting

The evaluation of the system follows the guidelines established in the Sentiment Analysis Track provided by the contest organizers. The complete dataset for the task was divided by the organizers into two subsets: one for training our models and the other for evaluating the results. The evaluation process relies on standard metrics such as precision, recall, and F1-score, which are applied across multiple subtasks, including polarity classification, type prediction (Attractive, Hotel, Restaurant), and the identification of Magical Towns [13].

The official evaluation procedure defined by the organizers includes three main subtasks: polarity classification, type prediction, and Magical Town identification. Each subtask is assessed using an F-measure-based metric, and the final score is computed as a weighted average of the three.

The polarity classification score is obtained by averaging the F-measure across all sentiment classes:

$$Res_P(k) = \frac{\sum_{i=1}^{|C|} F_i(k)}{|C|} \tag{1}$$

For the type prediction subtask, which involves three categories (Attractive, Hotel, and Restaurant), the macro-averaged F-measure is used:

$$Res_T(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3} \tag{2}$$

The Magical Town identification subtask evaluates the system's performance across all known towns in a predefined list (MTL), averaging their F-measures:

$$Res_{MT}(k) = \frac{\sum_{i=1}^{len(MTL)} F_{MTL_i}(k)}{len(MTL)} \tag{3}$$

Finally, the overall evaluation score combines the three subtasks, assigning higher importance to polarity and Magical Town classification by applying weights of 2 and 3, respectively:

$$Sentiment(k) = \frac{2 \cdot Res_P(k) + Res_T(k) + 3 \cdot Res_{MT}(k)}{6} \quad (4)$$

Table 4 shows the rankings of our two runs in the contest, which placed 33rd and 34th, respectively—well above the baseline and around the middle of the overall leaderboard.

Table 4
Results of all teams runs

Place	Run	Track Score	Macro F1 (Polarity)	Macro F1 (Type)	Macro F1 (Town)	Accuracy (Polarity)
1st	UDENAR_1	0.7254	0.6445	0.9877	0.6917	78.53
2nd	Axolotux_E_T3	0.7226	0.6395	0.9876	0.6896	78.20
3rd	Axolotux_E3	0.7197	0.6389	0.9874	0.6848	78.32
10th	LyS_6_3	0.6706	0.5823	0.9801	0.6226	73.74
33rd	AVYus_1	0.5847	0.5875	0.9578	0.4584	68.73
34th	AVYus_2	0.5785	0.5372	0.9596	0.4284	69.79
–	Baseline	0.0901	0.1584	0.1967	0.0089	65.54
68th	PMOTE-UC-2	0.0525	0.0787	0.1306	0.0089	3.94
69th	PMOTE-UC-3	0.0425	0.0487	0.1306	0.0089	5.99

As can be observed, the first run achieved marginally better results than the second, which is notably surprising as it contrasts with the validation outcomes obtained during the model implementation and prompting phases, where prompting demonstrated a slight advantage of 3% over our fine-tuned model.

4. Conclusion

In this work, we explored various natural language processing techniques for classifying Spanish-language tourist reviews, applied to the REST-MEX 2025 challenge. The evaluated methods included traditional approaches such as SVM, fine-tuning of Spanish-specialized transformer models (DistilBERT), and prompting techniques (Gemini) with large language models.

The results showed that traditional methods like SVM perform well on tasks with balanced classes and clear lexical features, such as classifying the type of site (hotel, restaurant, attraction). However, they perform poorly when classifying sentiment polarity.

Fine-tuning BERT-based models for Spanish significantly improved performance, especially in sentiment classification, thanks to their ability to understand complex contexts and linguistic nuances. This approach also proved effective for site type classification, outperforming SVM in precision and F1 metrics. However, due to computational limitations, this method was not scalable for the 40-category classification of Magical Towns.

Finally, prompting with large language models proved to be a promising alternative for classifying complex multi-class tasks, such as identifying the specific magical town, leveraging their few-shot inference capabilities.

Overall, combining the strengths of each technique can be an effective strategy to tackle complex problems in the analysis of Spanish-language tourist reviews, maximizing both precision and efficiency according to the specific subtask.

Declaration on Generative AI

We declare that the present manuscript has been written entirely by the authors and that no generative artificial intelligence tools were used in its preparation, drafting, or editing.

References

[1] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-Gonzalez, D. Fajardo-Delgado, M. G. Sánchez, H. Pérez-Espinosa, J. Martínez-Miranda, R. Guerrero-Rodríguez, L. Bustio-Martínez, Ángel Díaz-Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University - Computer and Information Sciences 34 (2022) 10125–10144. URL: https://www.sciencedirect.com/science/article/pii/S1319157822003615. doi:https://doi.org/10.1016/j.jksuci.2022.10.010.

[2] M. Á. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-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.

[3] A. Diaz-Pacheco, M. A. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, Journal of Experimental & Theoretical Artificial Intelligence 0 (2022) 1–31. doi:10.1080/0952813X.2022.2153276.

[4] R. Guerrero-Rodriguez, M. A. 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 26 (2023) 289–304. URL: https://doi.org/10.1080/13683500.2021.2007227. doi:10.1080/13683500.2021.2007227. arXiv:https://doi.org/10.1080/13683500.2021.2007227.

[5] 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.

[6] 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).

[7] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

[8] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of rest-mex at iberlef 2025: Researching sentiment evaluation in text for mexican magical towns, volume 75, 2025.

[9] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.

[11] J. K. Tripathy, S. C. Sethuraman, M. V. Cruz, A. Namburu, P. Mangalraj, R. N. Kumar, S. S. Ilango, V. Vijayakumar, Comprehensive analysis of embeddings and pre-training in nlp, Computers and Electrical Engineering 93 (2021) 107231. doi:10.1016/j.compeleceng.2021.107231.

[12] S. Ekin, Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices, TechRxiv (2023). URL: https://doi.org/10.36227/techrxiv.22683919.v1. doi:10.36227/techrxiv.22683919.v1.

[13] REST-MEX 2025, Rest-mex 2025: Data and evaluation, 2025. URL: https://sites.google.com/cimat.

mx/rest-mex-2025/data-and-evaluation, accessed: 2025-06-02.