Pumas PCIC team at REST-MEX 2025:Classification strategies for sentiment analysis

León Felipe Dueñas-González[1,†], Cristian Enrique Olvera-Morales[1,†] and Helena Gomez-Adorno[2]

[1]*Posgrado en Ciencias e Ingeniería de la Computación (PCIC), Universidad Nacional Autónoma de México, Ciudad Universitaria, Coyoacán, 04510, México*

[2]*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, Ciudad de México, México.*

Abstract

This paper presents the participation of the Pumas PCIC team in the REST-MEX 2025 shared task, focusing on sentiment analysis and magical town detection using Spanish-language tourist reviews. The study explores traditional models (SVM and XGBoost) and transformer-based architectures (RoBERTa-bne and BETO) to address three classification tasks: sentiment polarity, place type, and town name. Emphasis was placed on mitigating class imbalance through data augmentation techniques such as back-translation and paraphrasing. The final transformer models demonstrated significant improvements in macro F1-score, especially in minority sentiment classes, validating the benefit of phased fine-tuning and contextual enrichment with metadata. The findings highlight the potential of multilingual pre-trained models and domain-specific adaptations in tackling real-world tourism review classification challenges.

Keywords

Class Imbalance, Domain Adaptation, Tourism Analytics, NLP

1. Introduction

For Mexico, tourism is a significant economic activity, benefiting both individuals and corporations, as well as the government. Mexico offers natural and cultural wealth to all its visitors[1, 2, 3, 4, 5]. The "Pueblos Mágicos" program was launched in 2001 during the administration of Vicente Fox. The first town to be named a "Pueblo Mágico" was Huasca de Ocampo in the state of Hidalgo[6]. To be considered a Magical Town, it must possess symbolic attributes, legends, history, significant events, everyday life—in short, magic that emanates from each of its sociocultural manifestations, which today represent a great opportunity for tourism[7].

In this shared task, unlike to other editions [8, 9, 10], we had access to a little over 297,000 reviews from *Tripadvisor*, which were delivered in two stages: in the first stage, we received 70% of the data (just over 208,000), and in the second stage, the remaining 30% (just over 89,000)[11, 12].

Regarding the test dataset, we had the following information:

- `Title`: The title that the tourist assigned to their opinion. Data type: `Text`.
- `Review`: The opinion issued by the tourist. Data type: `Text`.
- `Polarity`: The label representing the sentiment polarity of the opinion. Data type: `[1, 2, 3, 4, 5]`.
- `Town`: The 'Pueblo Mágico' where the review is focused. Data type: `Text`.
- `Region`: The state in Mexico where the town is located. Data type: `Text`.
- `Type`: The type of place the review refers to. Data type: `[Hotel, Restaurant, Attractive]`.

IberLEF 2025, September 2025, Zaragoza, Spain

[†] These authors contributed equally.

lduenas@ciencias.unam.mx (L. F. Dueñas-González); colvera@ciencias.unam.mx (C. E. Olvera-Morales)

https://github.com/Leocdmx (L. F. Dueñas-González); https://github.com/CristianOlvera (C. E. Olvera-Morales)

0009-0008-2194-1770 (L. F. Dueñas-González); 0009-0005-6278-2129 (C. E. Olvera-Morales); 0000-0001-7116-9338 (H. Gomez-Adorno)

2. Summary of Findings from Reviewed Articles

The first reviewed article corresponds to the work by Morales-Murillo et al. (2023), presented at Rest-Mex 2023 [13, 10]. This study addresses polarity classification, place type classification, and country of origin identification in Mexican tourist reviews using Transformer-based models and domain adaptation techniques.

The authors employed the **RoBERTa-base-bne** model, previously pretrained on large Spanish corpora. Three training strategies were implemented:

1. Direct fine-tuning on the data without robust preprocessing.
2. Domain adaptation followed by fine-tuning, which achieved the best results.
3. Strategy 2 complemented with *data augmentation* techniques.

The dataset was recoded for polarity (from 1–5 to 0–4), and three independent models were constructed: one for polarity, one for place type, and one for country classification. In the third strategy, **oversampling** was applied to balance minority classes in polarity, which improved overall performance.

Aspect-Based Sentiment Analysis Using CNN

The second article, published by Pérez et al. (2022) [14], proposes a traditional approach based on deep convolutional neural networks (CNN) for Aspect-Based Sentiment Analysis (ABSA), using the dataset from the SemEval-2016 Task 5 challenge.

The preprocessing stage involved the removal of mentions, URLs, emoticons, and special characters, followed by lemmatization using spaCy. Vector representations were generated with `fastText` and `word2vec` embeddings.

Two architectures were designed for aspect detection:

- AB1-CNN (sequential model),
- AB2-CNN (multi-channel model).

And three for sentiment classification:

- P1-CNN and P3-CNN (sequential models),
- P2-CNN (multi-channel model).

Key results include an F1-score of 0.6540 for aspect detection using AB1-CNN, and an accuracy of 0.7969 for sentiment classification with P2-CNN. In real-world field tests, combining models achieved F1-scores up to 0.9333. However, a drop in precision was observed due to polarity class imbalance, highlighting a significant challenge for future research.

Transformer-Based Approach for ABSA

The third article, by Gómez et al. (2024) [15], proposes a more recent model based on Transformer architectures, specifically **BETO**, a BERT model pretrained on large Spanish corpora. The objective was to enhance ABSA performance through a simple architecture with minimal preprocessing and an additional linear layer for both aspect detection and sentiment classification.

Two context selection strategies were evaluated:

- **SDR** (syntactic dependency relations),
- **WW** (a 5-word window around the aspect), the latter proving more effective.

In addition to BETO, other models such as RoBERTa, RoBERTuito, and GPT-2 were compared. Results show that BETO outperformed others in aspect detection, while RoBERTa combined with WW performed best in sentiment classification. Despite the simplicity of the architecture, the results reached state-of-the-art performance in the evaluated tasks.

2.1. Overall Conclusion

The three reviewed studies provide complementary approaches to the problem of sentiment analysis in Spanish-language texts. The work by Morales-Murillo et al. (2023) demonstrates the effectiveness of domain adaptation with Transformer models in a real-world tourism context. Meanwhile, the study by Pérez et al. (2022) shows that traditional CNNs remain useful, especially in scenarios with limited computational resources. Finally, Gómez et al. (2024) present a substantial improvement by leveraging the power of Transformer models and modern context selection strategies, setting a new methodological standard in the field.

3. Distribution of Variables and Classes

In order to better understand the structure of the provided dataset, a detailed exploratory analysis was conducted. This analysis included the graphical representation of each relevant variable, which allowed us to observe their respective distributions. Additionally, the distribution of the classes and the composition of the different present groups were examined, with the aim of identifying possible imbalances, relevant patterns, or particular characteristics that could influence the subsequent modeling process.

Figure 1 shows the distribution of records according to the type of place referenced in each opinion within the dataset. It can be observed that the *Restaurant* category concentrates the highest number of opinions, with a total of 86,720 records, followed by *Attractive* with 69,921 records, and finally *Hotel*, with 51,410 records. This distribution suggests that the gastronomic experience represents a central component in the reviews given by tourists, surpassing even opinions about accommodations and tourist attractions.

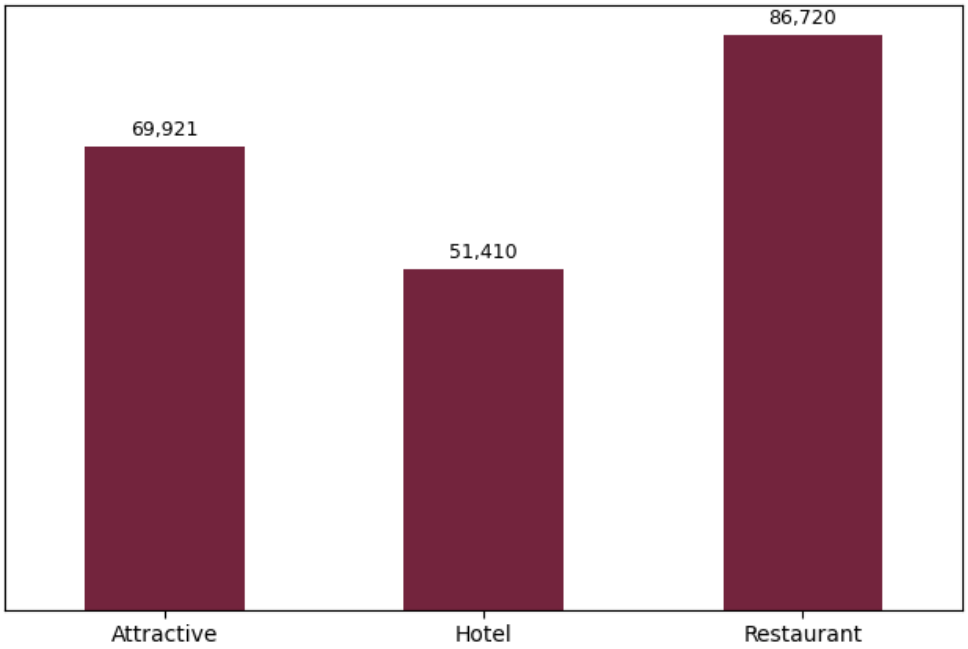

Figure 1: Distribution of records according to type.

Figure 2 illustrates the distribution of records according to the sentiment polarity expressed in the reviews. A strong bias towards highly positive evaluations can be observed, with polarity 5 being the most frequent category, totaling 136,561 records. This is followed in decreasing order by polarity 4 (45,034 records), polarity 3 (15,519 records), while negative evaluations — polarities 1 and 2 — represent a minimal percentage of the total, with 5,441 and 5,496 records, respectively.

This behavior demonstrates a clear asymmetry in user perception, suggesting a general tendency to

provide favorable opinions. From a modeling perspective, this highly imbalanced distribution could affect the performance of classification algorithms, especially in detecting negative reviews. Therefore, it will be necessary to consider class balancing or adjustment techniques in later stages of the analysis.

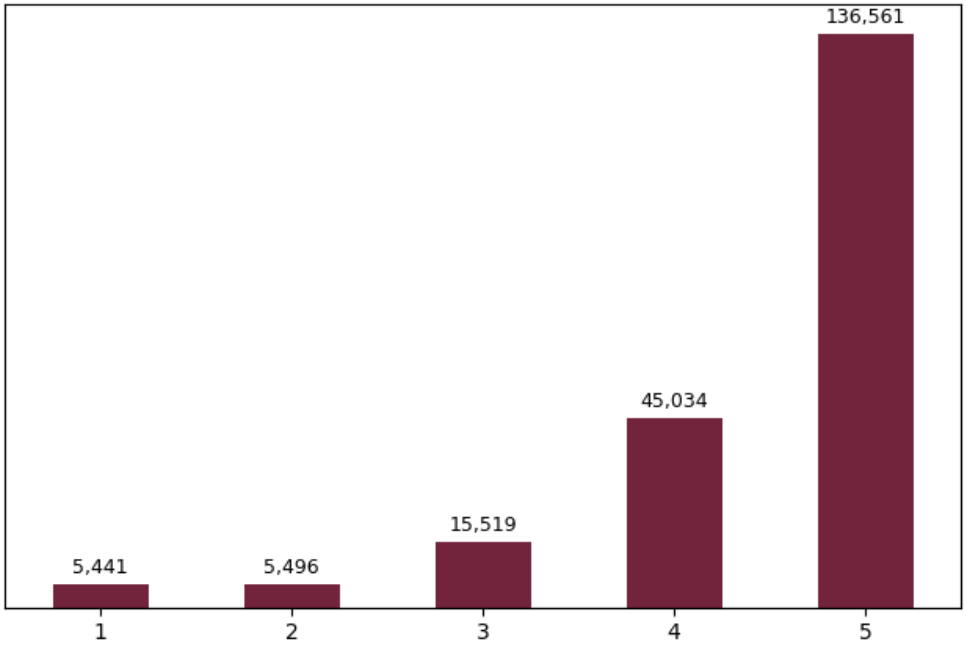

Figure 2: Distribution of records according to sentiment polarity.

Figure 3 shows the distribution of the number of records per *Pueblo Mágico* in the analyzed dataset. A notable concentration of reviews is observed in a small group of tourist towns, led by Tulum with 45,345 records, followed by Isla Mujeres with 29,826, and San Cristóbal de las Casas with 13,060. These three locations alone represent a significant proportion of the total number of observations, suggesting high visibility and tourist activity in these destinations.

In contrast, a large number of *Pueblos Mágicos* have considerably fewer reviews, with several — such as Tapalpa, Real de Catorce, Cuatro Ciénegas, among others — below 1,000 records. This uneven geographical distribution can significantly impact the model's ability to generalize predictions to underrepresented destinations. Therefore, it is advisable to consider normalization strategies, stratified sampling, or frequency-based weighting in later stages of the analysis.

This observation also highlights the relevance of certain destinations within the national tourism ecosystem and reinforces the need to assess potential regional biases when building predictive or classification models.

4. Metodology

Before we started preprocessing the train data, we ran a standard model of SVM and XGBoost for each category, so we could compare ahead in the processes if it did improve the score. It was decided to compare with SVM as it is more robust to noise and nonlinear data.It supports multiclass tasks directly During the development we incorporated two techniques that we will be using in the following tests that we considered important. The first one consists in combine the columns "Tittle" and "Review", by merging this two will be adding more context to the data, enriching the representation of the text and enhancing the ability of the classification model to capture the meaning and relationship between keywords. The second is the use of the method Term Frequency Inverse Document Frequency (TF-IDF) which captures the frequency of each word and penalizes the most common words in all the corpus. This will aid the model's ability to differentiate classes and find relevant information or patterns. Three variants for each model where tested, the first variants for SVM and XGBoost has a baseline of

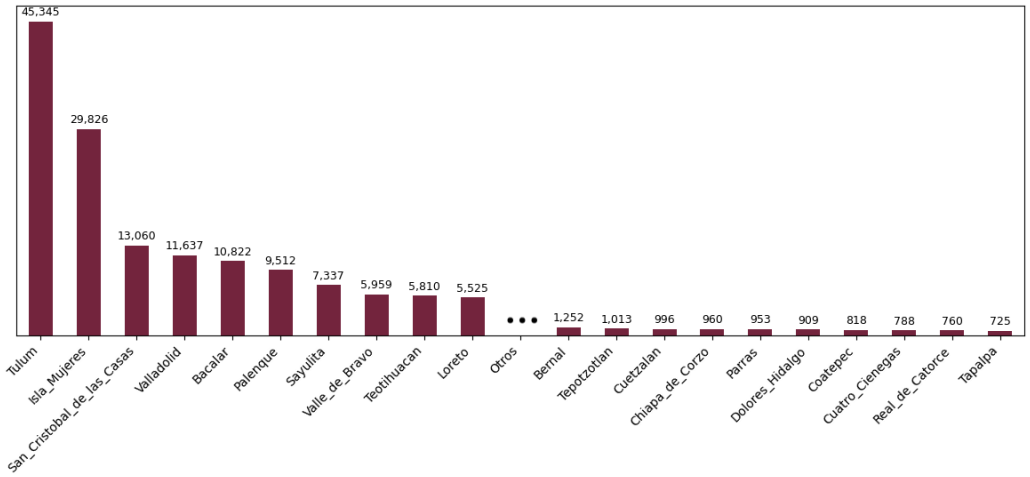

Figure 3: Distribution of the number of records per *Pueblo Mágico*.

max_features = 5000 and ngrams range on 1 and 2. The second we adjusted the previous parameters to max_features = 10000 and ngrams on 1,3. The last one, we add the class_weight to balanced on SVM and "scale_pos_weight" so we could penalize the minority classes. The results are shown as follows:

Model	Accuracy	Macro F1
SVM STANDARD	0.73	0.50
SVM VARIANTE 1	0.73	0.51
SVM VARIANTE 2	0.70	0.50
XGBOOST STANDARD	0.72	0.46
XGBOOST VARIANTE 1	0.72	0.47
XGBOOST VARIANTE 2	0.64	0.48

Table 1: Comparison of classification metrics (Polarity)

Model	Accuracy	Macro F1
SVM STANDARD	0.61	0.44
SVM VARIANTE 1	0.64	0.52
SVM VARIANTE 2	0.61	0.47
XGBOOST STANDARD	0.60	0.43
XGBOOST VARIANTE 1	0.62	0.52
XGBOOST VARIANTE 2	0.58	0.44

Table 2: Comparison of classification metrics (Town)

Model	Accuracy	Macro F1
SVM STANDARD	0.96	0.95
SVM VARIANTE 1	0.96	0.96
SVM VARIANTE 2	0.96	0.96
XGBOOST STANDARD	0.95	0.94
XGBOOST VARIANTE 1	0.95	0.94
XGBOOST VARIANTE 2	0.94	0.94

Table 3: Comparison of classification metrics (Type)

Preprocessing was a fundamental stage to assure quality and consistency of data. While examine the data, we detect some texts were written in other languages, so in order to detect this other data not necessarily fundamental for Spanish analysis, we run a language detector with "Lang detect" and the result threw that 0.18024% of the data was in other language, most of it were in English and in

polarity 5. Since the majority class for polarity is 5, it was decided not to use this information and to remove it from the corpus. Then, the text was standardized removing URL, emails and possessive symbols in English, symbols like ®, ~, ©. and # were also eliminated since doesn't apport context. For the hashtags we leave the word that followed it. Some reviews had words that seemed to be part of a button that displayed more information, this text "…Más" wasn't supposed to be relevant so it was removed too with some variants presented in most reviews. Character normalization to UTF 8 was also applied, emoticon substitution for words, use of some slang words were substituted for more reliable words. For tokenization and lemmatization, spaCy library was used, due to its high efficiency and accuracy in Spanish, in addition to its ease of integration into processing pipelines. It allows a fast and sufficiently accurate analysis for classification tasks such as those addressed in this work. Based on Polarity, additional feature extraction for the SVM model was incorporated to complement the TF-IDF vector-based representation. Emotional and structural attributes, such as the presence of exclamations and words with negative connotation, were included to capture the subjective intensity of opinions and improve polarity prediction. Excessive use of exclamation marks were more present in polarity one and five. Character length did not represent a characteristic of any polarity and words such as "bad, bad, lousy" were relevant on polarity one and two. Then based on the results of the first runs we train the model with all the previously mentioned preprocessing on the data and made some other experiments adding class weights resulting on 0.49% on macro F1 only for polarity. Next, trying t to solve the imbalance in classes , it was added the "RandomOverSampler" with a seed of 42 which did not improve the previous score with 3% less. For the last adjustment on both models, for Polarity the input embedding was trained using a representation that combined the lemmatized text vectorized with TF-IDF, the variable Region encoded with One Hot Encoding and the emotional and structural features (negative words, text length and exclamation marks). This combination allowed us to capture both semantic features and subjective intensity and geographical context. Type and Town on the other hand were only trained with region and the TF-IDF vectors, this configuration proved to be effective for classifying the categories of place type and magic town without adding the additional subjective intensity characteristics, which were less relevant for these tasks. Additionally for Town we detected that 20.03% of the reviews contained the value of "Town" so we develop an additional function to classify it if this was the case. The results for the final 2 models are as follows:

Model	Task	Accuracy	MF1-score
SVM	Polarity	0.70	0.50
XGBoost	Polarity	0.64	0.48
SVM	Town	0.87	0.83
XGBoost	Town	0.58	0.44
SVM	Type	0.95	0.95
XGBoost	Type	0.94	0.94

Table 4: Comparación de desempeño entre SVM y XGBoost en clasificación de Polarity, Town y Type

The previous models showed good results for some classes, but they presented some limitations capturing the context semantics and long-range relationships within text. In comparison, Transformer models leverage attentional mechanisms that allow modeling complex and contextual dependencies more effectively. This is especially relevant in a domain with informal language and colloquial expressions such as Spanish tourist reviews, where the meaning of a sentence may depend on its surrounding context. This facilitates knowledge transfer by being pre-trained on large corpora, which improves generalization to new examples. Therefore, we chose to integrate Transformers models to take advantage of these in the classification task, obtaining notable improvements in precision and recall with respect to traditional methods. In order to solve the problem of imbalanced data in polarity classification, augmented data techniques were implemented for minority classes such as one, two and three. The dataset was split into 90% train and 10% validation. We applied to train data the back translation technique using Helsinki-NLP/opus-mt to automatically translate to English and backwards to Spanish,

also T5 was used to paraphrasing. This generated 30,431 new examples for training which were cleaned and preprocessed after the previous process generated some noise. The pre trained models used in this work were RoBERTa-bne and BETO both in Spanish, which help us to capture linguistic particularities. To reinforce the problem address for polarity, CrossEntropyLoss function was used to compensate the inter-class imbalance. To calculate the weights, the formula $w_c = \frac{N}{n_c}$ where N is the total number of examples in the corpus and n_c is the number of examples for class c. Tokenizing the combination of title and review, the model input was also enriched by adding additional contextual information from the region and type columns. We applied two runs trained using the Hugging Face Trainer class with a carefully selected configuration. A learning rate of 2e-5 and a batch size of 16, applying a weight_decay of 0.01 to regularize the model. To facilitate the selection of the best model, the parameter load_best_model_at_end=True was set and f1_macro was used as the main metric were used for both training and validation with the only difference that the first run was trained with the train set plus the augmented data during 6 epoch and the second for four epochs only with train data and a fine tuning for 3 epoch with only the augmented data.

5. Automatic Classification System for Tourist Reviews

An automatic classification system was developed to identify both the type of site reviewed and the *Pueblo Mágico* associated with each opinion. The reviews could refer to hotels, restaurants, or attractions, and the dataset required an initial preprocessing phase. This phase included the cleaning of special characters, normalization of feminine ordinals, and mapping of the target variables to facilitate their use in machine learning models.

For the task of site type classification, the base model used was **RoBERTa-large-bne**, a large-scale version of the RoBERTa model adapted to Spanish. This model was pre-trained on a 570 GB corpus of clean, deduplicated text collected by the Spanish National Library between 2009 and 2019. A two-stage domain adaptation process was applied. In the first stage, the model was adjusted using the training dataset specific to the task, and in the second stage, an external corpus was used to further align the model with the tourism domain. Once the domain adaptation was completed, fine-tuning was performed for a three-class classification task, using five training epochs, a learning rate of 2e-5, and a batch size of 64 samples per step. The model achieved an average F1 score of **0.982** on the validation set and **0.976** on the test set.

For the identification of the associated *Pueblo Mágico*, a two-stage strategy was implemented. The first stage involved detecting direct mentions of the town's name within the text, including name variants and counting occurrences to determine the most likely match. This stage achieved an F1 score of **0.9753**. For cases without a direct mention, a second strategy was applied, based on feature extraction using ID-TF over the text, title, and region, followed by classification with a multiclass SVM model capable of distinguishing between 40 different *Pueblos Mágicos*.

6. Results

As a result, we obtained an improvement in F1 macro and in classes 2 and 3, which validates the phased approach as an alternative strategy in the face of imbalance [Table 5]

7. Discussion

Transformers demonstrated significant advantages in modeling complex contextual relationships and outperformed traditional models in accuracy and recall, consolidating their choice as the final solution in this challenge.

Classs	Precision	Recall	F1-Score
1	0.7302	0.5802	0.6466
2	0.4408	0.5346	0.4832
3	0.5489	0.5692	0.5589
4	0.4561	0.6198	0.5255
5	0.8896	0.7802	0.8313
Accuracy	0.7176 (total: 20769)		
Macro avg	0.6131	0.6168	0.6091
Weighted avg	0.7538	0.7176	0.7303

Table 5: Performance by class in post-corrected classification

8. Conclusions

Throughout the development of solutions for this contest, we developed an SVM classifier due to its simplicity and easy interpretation. However, the model did not achieve the expected results, especially for the class imbalance, linguistic variability, and the final result for town prediction. Despite hyperparameter tuning and preprocessing techniques, performance remained below expectations for metrics such as F1 score and recall for minority classes. For “town” classification task, we decided to use an SVM classifier due to the large number of classes and the constraints of time and computational resources. Although it managed to generate acceptable predictions on the tests, its final results did not exceed performance expectations on the official contest test set. Although our hybrid approach on Transformers and SVM generated a solution, the results obtained with SVM on the "town" task suggest that a transformer model would have represented a more robust and coherent option with the rest of our architecture, possibly achieving more consistent and competitive performance.

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] 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.

[2] R. Guerrero-Rodríguez, M. A. Álvarez-Carmona, R. Aranda, et al., Big data analytics of online news to explore destination image using a comprehensive deep-learning approach: a case from mexico, Information Technology & Tourism 26 (2024) 147–182. URL: https://doi.org/10.1007/s40558-023-00278-5. doi:10.1007/s40558-023-00278-5.

[3] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism, Journal of King Saud University-Computer and Information Sciences 35 (2023) 101746.

[4] A. Diaz-Pacheco, M. A. Álvarez-Carmona, A. Y. Rodríguez-González, H. Carlos, R. Aranda, Measuring the difference between pictures from controlled and uncontrolled sources to promote a destination. a deep learning approach (2023).

[5] C. M. A. Molinar, R. G. Sánchez, I. M. Carrillo, Influencia de la inseguridad en la competitividad turística: el caso de méxico, RITUR-Revista Iberoamericana de Turismo 15 (2025) 34–49.

[6] A. M. F. Poncela, Una revisión del programa pueblos mágicos, CULTUR: Revista de Cultura e Turismo 10 (2016) 3–34.
[7] M. L. D. Torres, J. C. S. Salinas, J. M. Arias, H. V. Lapaz, J. de Dios Romero Palop, Big data y turismo en méxico: Pueblos mágicos, 2016. URL: https://datatur.sectur.gob.mx/Documentos%20compartidos/2016_5.pdf.
[8] 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.
[9] 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) 289–299.
[10] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, V. Muñiz-Sánchez, A. P. López-Monroy, F. Sánchez-Vega, L. Bustio-Martínez, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023) 425–436.
[11] 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.
[12] 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.
[13] V. G. Morales-Murillo, H. Gómez-Adorno, D. Pinto, I. A. Cortés-Miranda, P. Delice, Lke-iimas team at rest-mex 2023: Sentiment analysis on mexican tourism reviews using transformer-based domain adaptation (2023).
[14] J. Pérez, M. Sánchez, L. Torres, Deep learning approach for aspect-based sentiment analysis of restaurants reviews in spanish, Journal of Computational Linguistics Research 12 (2022) 45–58.
[15] C. Gómez, L. Ramírez, S. Ortega, Enhancing spanish aspect-based sentiment analysis through deep learning approach, Natural Language Processing Advances 15 (2024) 22–39.