The Sintax Surfers team at Rest-Mex 2025: Solving the sentiment analysis task using pre-trained transformers-based models

Marcelo Alejandro Huerta-Espinoza[1,*,†], Juan Carlos Villagomez-Garcia[1,†] and Ansel Yoan Rodríguez-González[1,†]

[1] *Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, C. Ney M. González 7, Ciudad del Conocimiento, 63173 Tepic, Nayarit, México*

Abstract

The widespread availability of online reviews and the pervasive role of social media have reshaped how travelers plan vacations. In this context, user-generated content has become a key resource, strongly influencing consumer behavior. In tourism, sentiment-driven decision-making is especially relevant, making sentiment analysis a valuable tool for extracting insights from reviews and social media posts. However, adapting these techniques to domain- and language-specific contexts remains challenging. To address this, Rest-Mex was developed as an evaluation framework for analyzing tourism-related texts in Mexican Spanish, a variety often underrepresented in NLP resources. This paper presents the Sintax Surfers team's approach to the Rest-Mex task, which involved training a decision tree-based model to predict polarity, location (town), and type of place from online reviews. Our model achieved accuracies of 68.49% for town, 71.40% for polarity, and 95.99% for type of place, demonstrating the effectiveness of lightweight, interpretable models for domain-specific sentiment analysis in under-resourced language varieties.

Keywords

Tourism, Sentiment analysis, Natural language processing, Deep learning, Transformers-based models

1. Introduction

The proliferation of online reviews and the ubiquitous presence of social media have profoundly transformed how travelers make vacation-related decisions. Today, consumers heavily rely on user-generated content when planning their trips. Tourists frequently prioritize peer reviews over traditional factors such as price or location [1, 2, 3]. Positive reviews can often mitigate perceived disadvantages, whereas negative evaluations may significantly reduce the appeal of otherwise attractive destinations [4, 5, 6]. The influence of this behavioral shift is so substantial that hotels, travel agencies, and even search engines have incorporated review systems to enhance user engagement [7]. Specialized platforms such as Booking, TripAdvisor, and Lonely Planet have become key sources of information within the tourism sector [8, 9].

This dynamic has led to an exponential increase in user-generated data, particularly in the form of travel reviews. These reviews now serve as a critical source of information for consumers and a powerful marketing tool for businesses [10]. Consequently, companies need to implement automated systems capable of processing the vast amount of data being collected [11]. This need is especially pressing in domain-specific applications such as tourism, where sentiment-driven decision-making plays a pivotal role [12]. In this context, sentiment analysis has emerged as a promising solution. Defined as the study of individuals' opinions, attitudes, and emotions toward a given entity, whether a product, service, tourist attraction, or destination, sentiment analysis has become an essential tool in

IberLEF 2025, September 2025, Spain

*Corresponding author.
†These authors contributed equally.
marcelo.huerta@cicese.edu.mx (M. A. Huerta-Espinoza); jvillagomez@cicese.edu.mx (J. C. Villagomez-Garcia); ansel@cicese.edu.mx (A. Y. Rodríguez-González)
0009-0002-5040-9640 (M. A. Huerta-Espinoza); 0009-0003-0494-2636 (J. C. Villagomez-Garcia); 0000-0001-9971-0237 (A. Y. Rodríguez-González)

the tourism industry. It enables the extraction of meaningful insights from user-generated content such as online reviews and social media posts [13].

Rest-Mex arises as a response to this challenge, representing the first evaluation forum since 2021 (with subsequent editions in 2022 and 2023) aimed explicitly at text analysis in the tourism domain, focusing on Mexican Spanish [14, 15, 16]. This paper presents the methodology adopted for the Sentiment Analysis Task for RestMex at IberLEF 2025 [17, 18], which consists of three subtasks. Given a tourist review and its corresponding title, the goal is to classify: (1) the location (specifically, the Pueblo Mágico) being referred to, (2) the sentiment polarity on a scale from 1 to 5, and (3) the type of tourist attraction discussed.

The remainder of this paper is organized as follows: first, we describe the dataset used for the task; next, we outline the architectures proposed for each subtask; and finally, we present the results obtained on the evaluation dataset.

2. Proposed Method

In this section, we describe the dataset used for the Sentiment Analysis Task and the architectures used to tackle this task.

2.1. Dataset Description

The dataset comprises a collection of 208,501 documents in Spanish, each containing a title and its corresponding review. Each document represents a user-generated review of a tourist attraction, hotel, or restaurant located in Mexico. The reviews were sourced from the TripAdvisor website, and the dataset was provided with pre-assigned labels.

As a first step in our analysis, we examined the word count distribution across the corpus. Table 1 presents summary statistics obtained after concatenating the title and review for each document. "Text length" refers to the total number of characters, including letters, digits, spaces, and punctuation marks, while "word count" denotes the number of words per document.

Table 1
Text length and word count in the train dataset

Metric	Text length	Word count
mean	383.350	65.411
std	293.434	51.245
min	54	2
25%	202	34
50%	286	48
75%	472	81
max	8299	1487

As part of our exploratory analysis, we investigated the distribution of word frequencies within the corpus. Figure 1 displays a word cloud representing the most frequent tokens. As is commonly observed in natural language datasets, the most frequent terms are stop words, such as “de,” “y,” “la,” “el,” “en,” “que,” “es,” “un,” “a,” and “muy,” among others.

Subsequently, we examined the distribution of instances across the different classes. Figure 2 illustrates the class distribution within the training dataset. Among the classification tasks, the Type class (Figure 2b) exhibits the most balanced distribution across its categories. In contrast, the Polarity (Figure 2a) and Town (Figure 2c) classes show a marked imbalance. Notably, the Town class includes a high number of distinct labels (a total of 40), which may introduce additional challenges during the model training process.

Figure 1: Word Cloud of the train corpus.

2.2. Preprocessing

To prepare the text for processing with transformer-based architectures, the following steps were applied to the reviews:

- Conversion to lowercase
- Typo correction
- Removal of whitespaces
- Removal of special characters

In some cases, the title and the review text were concatenated into a single input; such instances are explicitly indicated.

2.3. Sentiment Analysis Task

In this section, we describe the training procedure for each subtask. The main strategy involved using pre-trained transformer-based models for the feature extraction phase, followed by a dense neural network layer for the classification phase.

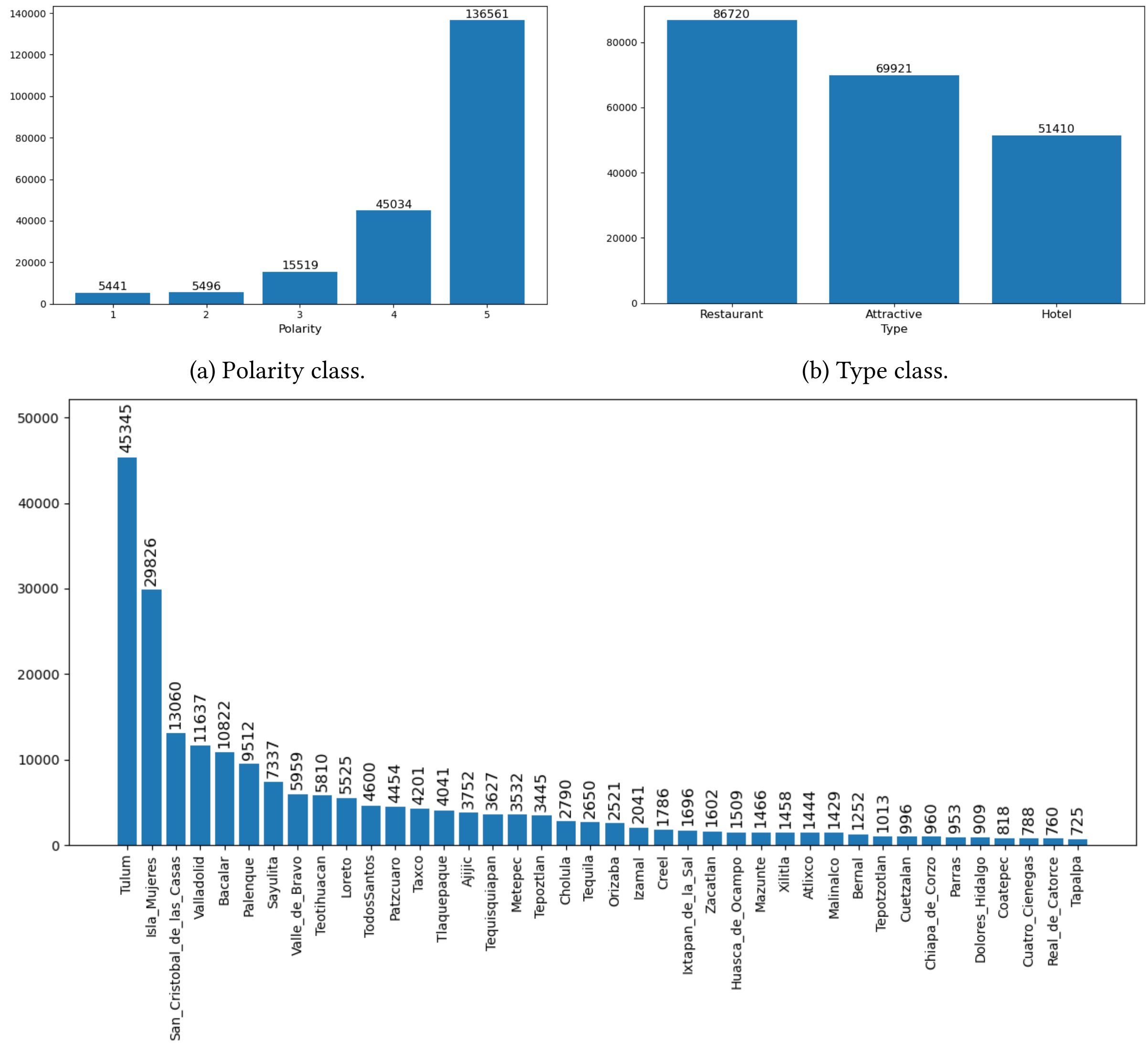
2.3.1. Polarity Substask

In the Polarity subtask, we employed the RoBERTa-base-bne model [19], a variant of the RoBERTa architecture [20]. RoBERTa-base-bne is a transformer-based masked language model specifically designed for the Spanish language.

To prepare the input text, the RoBERTa model requires tokenizing documents using its own tokenizer, so it was use. The architecture of the proposed neural network is based on the RoBERTa base model, whose output embeddings are concatenated with a fully connected neural layer consisting of 512 units. This is subsequently followed by an additional dense layer of 256 units, which serves to further refine the learned representations before passing them to the final classification head.

To improve the model's generalization capabilities and reduce the risk of overfitting, a dropout rate of 0.2 was applied to the fully connected layers. The training process was conducted using the AdamW optimization algorithm, with a learning rate set to 0.00002. The model was trained for a total of 8 epochs, a setting chosen to balance convergence and computational efficiency.

This architecture aims to leverage the powerful contextual representations produced by RoBERTa, while the additional dense layers facilitate task-specific adaptation and improved discriminative performance.



(a) Polarity class.

(b) Type class.

(c) Town class.

Figure 2: Class distribution in train dataset

2.3.2. Type Subtask

For the Type subtask, we also employed the RoBERTa-base-bne model [19], as mentioned in the previous section the use of the RoBERTa-base-bne model requires the RoBERTa tokenizer [20]. The overall architecture consists of a sequence of processing steps: input text is first tokenized using the RoBERTa tokenizer, followed by feature extraction through the RoBERTa-base-bne model. The output embeddings are then passed to a fully connected layer for classification. The classification layer comprises five neurons with a softmax activation function to predict class probabilities.

he training configuration included a batch size of 16, a dropout rate of 0.2 applied to the output of the RoBERTa model, and the AdamW optimizer [21] with a learning rate of 0.0001. The model was trained for a total of 9 epochs.

2.3.3. Town Subtask

For the Town subtask, three different strategies were explored: Baseline Strategy: As a baseline, the BETO [22] model was used with minimal modifications. A classification head was added to the BETO model, with the number of output neurons corresponding to the number of target classes. The model was fine-tuned using a batch size of 8, for 3 epochs, with the AdamW optimizer. No further changes were made to the dataset.

Undersampling Strategy: In this approach, the dataset was undersampled so that each class contained approximately the same number of samples, equal to the average number of samples across all classes (or the nearest feasible number). The BETO model was then trained using the same hyperparameters as in the baseline strategy.

Combined Input Strategy: In this strategy, the Title and Review fields were concatenated into a single text input. This combined text was then used for both training and inference with the BETO model.

Only the best performance was reported in the Results section

2.4. Sentiment Analysis Evaluation

The organizers of the Rest-Mex challenge proposed a custom evaluation metric to assess model performance. This metric, defined in Equation 4, combines the results of different subtasks through a weighted average of F1 scores. The components of the metric are defined as follows:

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

Where:

- k is a forum participant system.
- $C = \{1, 2, 3, 4, 5\}$.
- $F_i(k)$ is the F-measure value for the class i obtained by the system k.

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

Where:

- $F_a(k)$ represents the F measure obtained by the system k for the Attractive class.
- $F_H(k)$ represents the F measure obtained by the system k for the Hotel class.
- $F_R(k)$ represents the F measure obtained by the system k for the Restaurant class.

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

Where:

- MTL represents the list with all Magical Towns.
- $F_{MTL_i}(k)$ represents the F measure obtained by the system k for the Town class.

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

Where:

- $Res_p(k)$ represents the result given by Equation 1.
- $Res_t(k)$ represents the result given by Equation 2.
- $Res_M T(k)$ represents the result given by Equation 3.

In addition to the Rest-Mex evaluation metric, we also report Accuracy, Precision, and F1 score to assess our models' performance. These metrics provide a broader understanding of model behavior and help evaluate both the overall effectiveness and the reliability of the predictions.

3. Results

The test dataset consists of a total of 89,166 pairs of titles and their corresponding reviews. Both the title and the review text were joined to create the input required by the pre-trained models. As explained in previous sections, the Rest-Mex organizers proposed an equation to evaluate model performance. According to the reported results [17], the Track Score is: $Sentiment(k) = 64.123$.

Table 2 presents the results reported by the Rest-Mex organizers. The model that achieved the lowest F1 score was the one developed for the Town subtask. This outcome may be attributed to the high number of distinct classes in the corpus, as well as the pronounced class imbalance between the most and least frequent labels.

Table 2
Models performance metrics

SubTask	Accuracy	Precision	F1 Score
Town	68.492	59.571	57.119
Polarity	71.398	59.484	58.762
Type	95.998	96.086	95.855

A similar issue appears in the Polarity subtask, where the F1 score is also low. Although the number of classes in Polarity is significantly smaller compared to Town, the substantial imbalance among its classes suggests that class distribution plays a more critical role in model performance than the number of classes alone.

This may also explain why the accuracy scores for both subtasks are relatively similar, despite the differences in F1 score. It is likely that, in the Polarity task, the model disproportionately favored majority classes, leading to inflated accuracy at the expense of minority class performance due to the more severe imbalance.

In the Type subtask, the best performance was achieved across all metrics. This suggests that class imbalance may significantly affect the other tasks. Additionally, the relatively small number of unique labels in this subtask could further contribute to the high performance observed.

4. Conclusion

Sentiment analysis is a valuable tool for tourists, as it enables quick access to client opinions about services, locations, products, and more. In recent years, natural language processing techniques—particularly attention-based models—have emerged as effective solutions for tasks of this nature.

In this paper, we presented our proposed method and the results obtained in the Rest-Mex Sentiment Analysis Task. The results reveal significant challenges in the Town and Polarity subtasks, where the large number of labels and pronounced class imbalance negatively affected both the training and classification phases. In contrast, the Type subtask yielded the highest performance across key evaluation metrics, including accuracy, precision, and F1 score.

As future work, we propose exploring advanced fine-tuning strategies and rebalancing techniques, such as EDA, back-translation, or the generation of synthetic instances, to enhance performance, particularly in the Town and Polarity subtasks. Additionally, further investigation could focus on the use of data augmentation methods, including EDA, Random Oversampling (ROS), and synthetic data generation approaches. Other rebalancing strategies, such as oversampling and undersampling, as well as the adoption of alternative architectures for model training, could also be explored.

Acknowledgments

The authors gratefully acknowledge the financial support provided to the first author by SECIHTI

which supported his graduate studies during the development of this work.

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] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico, International Journal of Tourism Cities 10 (2023) 639–661. URL: http://dx.doi.org/10.1108/IJTC-09-2022-0223. doi:10.1108/ijtc-09-2022-0223.

[2] 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, International Journal of Interactive Multimedia and Artificial Intelligence In Press (2023) 1–14. URL: http://dx.doi.org/10.9781/ijimai.2023.10.003. doi:10.9781/ijimai.2023.10.003.

[3] 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.

[4] L. A. Book, S. Tanford, Y.-S. Chen, Understanding the Impact of Negative and Positive Traveler Reviews: Social Influence and Price Anchoring Effects, Journal of Travel Research 55 (2016) 993–1007. URL: https://journals.sagepub.com/doi/10.1177/0047287515606810. doi:10.1177/0047287515606810.

[5] P. Sema, Does social media affect consumer decision-making, MBA Student Scholarship 24 (2013).

[6] S. Tanford, E. L. Kim, Risk versus Reward: When Will Travelers Go the Distance?, Journal of Travel Research 58 (2019) 745–759. URL: https://journals.sagepub.com/doi/10.1177/0047287518773910. doi:10.1177/0047287518773910.

[7] V. Katsoni, D. Laloumis, The influence of online reviews on customers and travel agencies, Zeszyty Naukowe Małopolskiej Wyższej Szkoły Ekonomicznej w Tarnowie (2013) 123–130. Publisher: Małopolska Wyższa Szkoła Ekonomiczna w Tarnowie.

[8] R. S. Almazán, R. O. González, Plataformas online en la 'nueva normalidad': el caso tripadvisor/online platforms and 'new normality': the tripadvisor case, Revista Rosa dos Ventos-Turismo e Hospitalidade 13 (2021).

[9] 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.

[10] A. Sharma, S. Dey, Mining marketing intelligence from online reviews using sentiment analysis, International Journal of Intercultural Information Management 5 (2015) 57. URL: http://www.inderscience.com/link.php?id=72544. doi:10.1504/IJIIM.2015.072544.

[11] G. I. Flores Limaylla, E. P. Peña Alvarez, Aprendizaje automático para la optimización de procesos de marketing digital en el sector turístico (2020).

[12] Q. Ye, R. Law, B. Gu, The impact of online user reviews on hotel room sales, International Journal of Hospitality Management 28 (2009) 180–182.

[13] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams engineering journal 5 (2014) 1093–1113.

[14] 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. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism (2021).
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[21] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. URL: https://arxiv.org/abs/1412.6980. arXiv:1412.6980.
[22] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC at ICLR 2020, 2020.