Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 499-511 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of Rest-Mex at IberLEF 2025: Researching Sentiment Evaluation in Text for Mexican Magical Towns

Resumen de la tarea Rest-Mex en IberLEF 2025: Investigando la Evaluación de Sentimientos en Textos sobre los Pueblos Mágicos de México

Miguel Á. Álvarez-Carmona,[1,2] **Ángel Díaz-Pacheco,**[4] **Ramón Aranda,**[1,2]
Ansel Y. Rodríguez-González,[1,3] **Lázaro Bustio-Martínez,**[5] **Vitali Herrera-Semenets**[6]
[1]Centro de Invetsigación en Matemáticas
[2]Secretaría de Ciencia, Humanidades, Tecnología e Innovación
[3]Centro de Investigación Científica y de Educación Superior de Ensenada
[4]Universidad de Guanajuato, [5]Universidad Iberoamericana
[6]Centro de Aplicaciones de Tecnologías de Avanzada
{miguel.alvarez, arac}@cimat.mx, angel.diaz@ugto.mx,
ansel@cicese.edu.mx, lazaro.bustio@ibero.mx, vherrera@cenatav.co.cu

Abstract: The REST-MEX 2025 shared task, hosted at IberLEF, represents the fourth edition of an ongoing effort to benchmark natural language processing systems in the domain of Mexican tourism. The task focuses on the automatic classification of user-generated reviews into three axes: *polarity* (from 1 to 5), *type of service* (hotel, restaurant, or attraction), and the identification of one of 40 predefined *Pueblos Mágicos* (Magical Towns). This year, the challenge attracted a record number of submissions and introduced greater complexity by expanding the dataset and encouraging research into advanced fine-tuning strategies, prompt engineering, and data-centric methods. A total of 32 teams submitted 70 valid runs, exploring a diverse range of methods including Transformer-based models, transfer learning from previous REST-MEX editions, class binarization schemes, and instance selection strategies based on contextual embeddings. This paper presents an overview of the task, the dataset characteristics, the evaluation protocol, and a comparative analysis of the results.
Keywords: Rest-Mex 2025, Sentiment Analysis, Mexican Tourist Text, Magical Towns Classification.

Resumen: La tarea del REST-MEX 2025, organizada en IberLEF, representa la cuarta edición de un esfuerzo continuo por establecer referencias comparativas de sistemas de procesamiento de lenguaje natural en el ámbito del turismo en México. La tarea se centra en la clasificación automática de reseñas generadas por usuarios en tres ejes: *polaridad* (de 1 a 5), *tipo de servicio* (hotel, restaurante o atractivo), y la identificación de uno de los 40 *Pueblos Mágicos* predefinidos. Este año, el reto atrajo un número récord de envíos e incrementó su complejidad al ampliar el conjunto de datos y promover la investigación en estrategias avanzadas de ajuste fino, ingeniería de prompts y métodos centrados en los datos. En total, 32 equipos realizaron 70 envíos válidos, explorando una amplia variedad de enfoques que incluyen modelos basados en Transformers, aprendizaje por transferencia a partir de ediciones anteriores de REST-MEX, esquemas de binarización de clases y estrategias de selección de instancias basadas en representaciones contextuales. Este artículo presenta una visión general de la tarea, las características del conjunto de datos, el protocolo de evaluación y un análisis comparativo de los resultados.
Palabras clave: Rest-Mex 2025, Análisis de sentimientos, Textos Turísticos Mexicanos, Clasificación de Pueblos Mágicos.

ISSN 1135-5948 DOI 10.26342/2025-75-37

1 Introduction

In recent years, the role of user-generated content in the tourism sector has grown exponentially (Arce-Cardenas et al., 2021; Diaz-Pacheco et al., 2024; Olmos-Martínez et al., 2024), with online reviews becoming a central source of information for travelers and a critical feedback channel for service providers (Álvarez-Carmona et al., 2022b). Within this context, the accurate interpretation of opinions expressed in natural language has gained importance as a tool for improving service quality, detecting trends, and informing policy in tourism development (Díaz-Pacheco et al., 2023). In the case of Mexico, a country with a rich and diverse touristic offering, the *Pueblos Mágicos* (Magical Towns) initiative stands as a strategic program aimed at promoting cultural, historical, and natural destinations (Guerrero-Rodriguez et al., 2023). Understanding how tourists perceive and evaluate their experiences in these towns offers valuable insights for both local governments and the tourism industry (Díaz-Pacheco et al., 2024).

To address this challenge from a computational perspective, the REST-MEX shared task was launched in 2021 (Álvarez-Carmona et al., 2021) under the IberLEF framework (González-Barba, Chiruzzo, and Jiménez-Zafra, 2025), with the goal of fostering research in sentiment analysis and thematic classification of Spanish-language tourism reviews. The 2021 edition focused on polarity prediction across five levels, setting a baseline for model development and evaluation in this domain. In 2022, the task expanded to include the classification of review types (e.g., hotels, restaurants, attractions) (Álvarez-Carmona et al., 2022). The 2023 edition further consolidated the task by refining the annotation process, increasing dataset size, and introducing the country classification (Álvarez-Carmona et al., 2023).

The 2025 edition marks the fourth iteration of REST-MEX, and introduces several innovations inspired by both technical trends and community feedback. First, the dataset has 208,051 training annotated reviews drawn from real-world platforms. Second, the 2025 edition emphasized advanced modeling techniques such as prompt engineering, transfer learning from prior competition models, binary classification per class, and centroids-based instance selection. These methodological directions reflect the growing maturity of the field and the shift toward data-centric and model-agnostic design. Additionally, this year introduces a finer-grained geographic classification task: instead of identifying the country of origin or general location, as in 2023, systems are now required to classify the specific "Pueblo Mágico" mentioned in the review. This refinement poses new challenges in distinguishing nuanced linguistic patterns tied to local identity, while offering greater value to stakeholders interested in place-specific sentiment trends (Castillo-Ortiz et al., 2024).

This paper provides an overview of the 2025 edition, describing the dataset, task formulation, evaluation metrics, and an in-depth analysis of the systems submitted. We also examine the methodological trends observed across top-performing teams, providing insights for future research in sentiment analysis and tourism informatics.

2 Evaluation framework

This section outlines the construction of the corpus, highlighting particular properties, challenges, and novelties. It also presents the evaluation measures used for the task.

2.1 Rest-Mex 2025 corpus

The classification task challenges participating systems to predict three aspects of each tourist review: its sentiment polarity, the type of tourist service mentioned, and the specific Mexican "Pueblo Mágico" to which it refers. The dataset comprises reviews written by tourists who visited prominent destinations in Mexico and shared their opinions on TripAdvisor between 2002 and 2024. Each opinion is labeled with a polarity score on a five-point ordinal scale: {1: Very bad, 2: Bad, 3: Neutral, 4: Good, 5: Very good}. In addition to polarity, systems must classify the type of attraction discussed in the review, which can be one of three categories: *Attractive*, *Hotel*, or *Restaurant*. Lastly, unlike previous editions that focused on broader geographic classification (e.g., by country), the 2025 task requires identifying the exact "Pueblo Mágico" referenced in the review, chosen from a predefined list of the 40 most popular towns in the country.

The REST-MEX 2025 corpus consists of **297,217 opinions** shared by tourists who visited representative destinations in Mex-

Task	Class	Train	Test
Polarity	1 (Very bad)	5,441	2,281
	2 (Bad)	5,496	2,269
	3 (Neutral)	15,519	6,712
	4 (Good)	45,034	18,135
	5 (Very good)	136,561	59,769
	Total	208,051	89,166
Type	Attractive	69,921	30,110
	Hotel	51,410	21,725
	Restaurant	86,720	37,331
	Total	208,051	89,166
Towns (top-3)	Tulum	45,345	19,434
	Isla Mujeres	29,826	12,783
	San Cristóbal	13,060	5,597
	··· (40 towns)	208,051	89,166

Table 1: Distribution of instances in the REST-MEX 2025 training and test datasets across the three classification tasks.

ico. Following the established evaluation protocol, we used a 70/30 split to construct the training and test partitions. Specifically, **208,051 labeled instances** were allocated for training, while **89,166 labeled instances** were used for testing.

Table 1 shows the class distribution for the three subtasks of sentiment analysis: polarity prediction, service type classification, and geographical identification of the visited location (at the level of Mexican Magical Towns).

As observed, there is a notable **class imbalance** across all three classification axes, particularly in the polarity task where most reviews are positive (label 5). This imbalance poses a significant challenge for supervised learning models and calls for robust techniques to mitigate skewed learning (Álvarez-Carmona and Aranda, 2022).

Formally, the task can be defined as:

> Given a tourist opinion about a place in Mexico, the goal is to simultaneously determine:
>
> 1. The sentiment polarity, on a scale from 1 (very bad) to 5 (very good),
> 2. The type of place being reviewed: `Attractive`, `Hotel`, or `Restaurant`,
> 3. The specific **Magical Town** among a list of 40 representative destinations in Mexico.

This edition introduces a finer-grained level of geographic classification compared to the 2023 edition, moving from country-level to town-level resolution. This change increases the semantic complexity of the task and provides a more meaningful benchmark for geolocalized sentiment modeling in tourism.

2.2 Performance Measures

Previous editions of REST-MEX have faced challenges due to the significant class imbalance in the dataset, especially along the polarity axis, where most reviews are positive. This imbalance tends to obscure performance on minority (negative) classes. For this reason, in REST-MEX 2025 we propose a weighting scheme designed to explicitly favor correct predictions of underrepresented sentiments.

For polarity classification, the evaluation measure defined in Equation 1 assigns higher importance to classes with fewer examples. Formally:

$$R_P(k) = \frac{\sum_{i=1}^{|C|} \left(\left(1 - \frac{T_{C_i}}{T_C}\right) \cdot F_i(k) \right)}{\sum_{i=1}^{|C|} \left(1 - \frac{T_{C_i}}{T_C}\right)} \quad (1)$$

Here:

- k denotes a participating system.
- $C = \{1, 2, 3, 4, 5\}$ represents the set of polarity classes.
- T_C is the total number of training instances.
- T_{C_i} is the count of instances belonging to class i.
- $F_i(k)$ is the F1 score that system k achieved for class i.

This metric ensures that correctly predicting the most negative opinions (class 1) contributes more strongly to the final score than predicting frequent positive opinions (e.g., class 5).

For the **Type** prediction task, which involves three categories (Attractive, Hotel, and Restaurant), we apply the classical macro-averaged F1 measure as defined in Equation 2:

$$R_T(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3} \quad (2)$$

where:

- $F_A(k)$ is the F1 score for the Attractive class,
- $F_H(k)$ is the F1 score for Hotel,
- $F_R(k)$ is the F1 score for Restaurant.

Finally, the new subtask of this edition—the fine-grained classification of the visited town among **40 Mexican Magical Towns**—is evaluated using a similar macro-F1 scheme. Equation 3 shows the formulation:

$$R_{MT}(k) = \frac{\sum_{i=1}^{|MT|} F_{MT_i}(k)}{|MT|} \tag{3}$$

where MT denotes the set of 40 town classes, and $F_{MT_i}(k)$ is the F1 score obtained by system k for town i.

To compute the final overall measure, we define a weighted average of the three subtasks. Since polarity is considered the critical aspect for tourism sentiment analysis, it receives double the weight of type prediction and a slightly lower weight than town identification, reflecting the added complexity of distinguishing fine-grained locations. This aggregation is formalized in Equation 4:

$$Sent(k) = \frac{2 \cdot R_P(k) + R_T(k) + 3 \cdot R_{MT}(k)}{6} \tag{4}$$

This design aims to reward systems that perform robustly across all axes while emphasizing the importance of capturing nuanced sentiment and precise geographical context.

2.3 Measuring the Easiness of the Corpus

As an additional contribution of this work, we propose an approach to quantify the *easiness* of a corpus. This measure provides a descriptive and qualitative understanding of how distinguishable the classes are, based on their lexical overlap.

The measure extends the idea introduced in (Álvarez-Carmona et al., 2018; Álvarez-Carmona et al., 2024), where the difficulty of a paraphrase detection corpus (with only two classes) was assessed by comparing the vocabulary shared between paraphrase pairs and non-paraphrase pairs. The core intuition is that if texts belonging to different classes share many words, the classification problem becomes inherently more challenging.

In the context of REST-MEX, where each classification axis involves multiple classes, we adapt this principle as follows: we define a corpus as *easier* if texts in a given class have a larger proportion of words that do not appear in texts from any other class. Conversely, a corpus is *harder* if most words are shared across classes, making discrimination based on lexical cues more difficult.

Formally, let each class be represented by the set of all unique words appearing in its texts. For a class x, its *exclusive vocabulary* is defined as the set difference between the words in x and the union of words in all other classes. This is expressed in Equation 5:

$$easiness_{Class}(x, C) = C_x \setminus \bigcup_{i=1, i \neq x}^{|C|} C_i \tag{5}$$

Here:

- C denotes the set of all classes in the corpus.
- C_x is the set of words belonging to class x.

To quantify the easiness of the entire corpus, we aggregate the number of exclusive words for all classes and normalize this sum by the total number of unique words in the collection. This yields the global easiness score, as shown in Equation 6:

$$easiness(C) = \frac{\sum_{i=1}^{|C|} |easiness_{Class}(i, C)|}{\left|\bigcup_{i=1}^{|C|} C_i\right|} \tag{6}$$

In this formulation:

- A value of 1 indicates the maximum easiness: all words are exclusive to their respective classes, and there is no lexical overlap.
- A value of 0 indicates the minimum easiness: all classes share the same vocabulary, making discrimination solely by lexical means impossible.

This metric provides an intuitive, corpus-level descriptor of how separable the categories are when relying on vocabulary, and

Trait	Class	$easiness_{Class}$
Polarity	1	4989
	2	4107
	3	9041
	4	22986
	5	74062
	$easiness(Polarity)$	**0.38**
Type	Attractive	61516
	Hotel	29155
	Restaurant	30549
	$easiness(Type)$	**0.47**
Town	Tulum	29232
	Isla Mujjeres	15207
	San Cristobal	5291
	$easiness(Town)$	**0.16**

Table 2: Easiness for sentiment analysis.

can be useful for anticipating classification challenges or interpreting model performance.

2.3.1 Easiness of the sentiment analysis corpus

Applying the proposed method described in Equation 6 yielded the results presented in Table 2.

Overall, the **easiness score for the Polarity trait is 0.38**, indicating a moderate level of lexical separability between sentiment classes. Within Polarity, Class 5 (Very Good) exhibits the highest number of exclusive words, making it the easiest class to distinguish. This is consistent with the intuition that positive reviews often employ distinct, highly repetitive vocabulary (e.g., "excelente," "increíble," "maravilloso") that is less likely to appear in negative or neutral reviews. Conversely, Class 1 (Very Bad) and Class 2 (Bad) have the smallest counts of exclusive words, confirming that negative sentiments tend to share a larger proportion of their vocabulary with other classes—especially Class 3 (Neutral) and Class 4 (Good). This lexical overlap explains why Polarity is the most challenging trait for classification.

For the Type trait, the easiness score is **0.47**, reflecting a relatively higher level of discriminability. Among Type classes, the *Attractive* category has the largest exclusive vocabulary, suggesting that users describing tourist attractions tend to use more distinctive terms (e.g., references to sites, landmarks, or activities). In contrast, the *Hotel* and *Restaurant* categories have smaller exclusive sets, likely due to shared vocabulary when describing service, food, and accommodations. This pattern supports the observation that classifying Type is generally easier than Polarity but still subject to moderate lexical overlap.

Finally, the Town trait—reflecting the identification of the specific **Pueblo Mágico**—achieved an easiness score of only **0.16**, markedly lower than the other traits. Even the town with the highest exclusive vocabulary, Tulum, has fewer unique words than the most distinctive Polarity or Type classes. This suggests that many reviews of different towns share similar general terms related to tourism (e.g., "playa," "hotel," "comida," "gente"). The low easiness score underscores the considerable challenge of fine-grained geographic classification in this domain.

In summary, the analysis reveals clear patterns:

- **Polarity** is the most complex axis, primarily due to negative reviews lacking distinctive lexical markers.
- **Type** benefits from clearer separation, especially for Attractions.
- **Town** classification is the most difficult overall, with high lexical similarity across towns.

These insights help contextualize the evaluation results and provide a rationale for why certain subtasks yielded higher or lower performance across participating systems.

3 Overview of the Submitted Approaches

This section presents the results obtained by the participants for the three sub-tasks.

3.1 Sentiment Analysis Overview

For this edition, 32 teams submitted 70 systems to the sentiment analysis task.

Table 3 presents the overall ranking, displaying only the best run of each team.[1]

The UDENAR team (Jurado-Buch, 2025) achieved the top overall performance in

[1]To see the complete results of all runs, visit the official leaderboard: `https://sites.google.com/cimat.mx/rest-mex-2025/results`

REST-MEX 2025 with an innovative strategy that reformulated the classification tasks into a series of binary subtasks. Instead of training a single multiclass model per task, their approach constructed individual binary classifiers for each class across the three axes: polarity, type, and town. This transformation resulted in more than fifty binary classification models. For each classifier, they balanced the dataset by combining all positive instances with an equal number of negative examples selected either randomly or using centroid-based sampling over [CLS] embeddings derived from a RoBERTa model previously fine-tuned in REST-MEX 2023. The predictions from all binary classifiers were then concatenated and passed into a ten-layer multilayer perceptron, which produced the final multiclass label.

This method proved particularly effective at addressing class imbalance, especially for rare classes, as semantically similar negative samples improved the model's discriminative capability. Experiments demonstrated that centroid-based sampling outperformed random selection across all tasks, yielding the highest macro F1-scores of the competition.

The Axolotux team (Minutti-Martinez, Escalante-Ramirez, and Olveres, 2025) developed an approach integrated multiple pretrained models, specifically RoBERTa and LLaMA version 3.2, into an ensemble framework designed to exploit model diversity. Key strategies included experimenting with input formatting alternatives, such as dual-sentence encoding versus simple concatenation, extensive hyperparameter tuning, and the application of weighted voting schemes to aggregate predictions.

This carefully constructed ensemble demonstrated that combining complementary models yields performance improvements compared to single-model approaches or domain adaptation alone.

The Pandas Rojos team (Siliceo-Guzmán, Aranda, and Álvarez-Carmona, 2025) proposed a hybrid framework that combines prompt engineering and transfer learning to enhance sentiment and thematic classification of Spanish-language tourism reviews. Their method extracted the [CLS] embedding from the vg055/roberta-base-bne-finetuned-e2-RestMex2023-polarity model, which was the winner of the polarity task in REST-MEX 2023 (Morales-Murillo et al., 2023). This embedding was concatenated with contextual embeddings generated by a prompted llama model, yielding richer representations that integrate domain-specific knowledge with general-purpose information without requiring additional fine-tuning of the base models.

The LyS team (Imran, Rasheed, and Gómez-Rodríguez, 2025) presented a hybrid framework that addressed class imbalance through oversampling and back-translation from structurally similar and dissimilar languages. They fine-tuned two transformer-based models, roberta-base-bne and twitter-xlm-roberta-base, on the augmented datasets to classify sentiment polarity, destination type, and town identity.

The FisBio team (García-Espinosa, Flores-Luna, and Moreno-Sánchez, 2025) developed a multi-task system based on the multilingual BERT model to jointly predict sentiment, destination type, and Magical Town classification. Their architecture employed a shared BERT backbone with three classification heads trained using equal loss weights.

The Corpus Christi team (Hernández-Angeles et al., 2025) combined fine-tuned BERT models with a Word2Vec-embedded multilayer perceptron to tackle sentiment polarity, destination type, and town classification.

The MictalMapper team (Monsivais Borjón and Álvarez-Carmona, 2025) proposed MultiHeadBETO, a multitask Transformer model with a shared BETO encoder and three output heads for sentiment, type, and town classification. Their system achieved a competitive performance for town prediction, and moderate performance for sentiment polarity. The study demonstrated that shared representations improve generalization, while suggesting future enhancements such as geographic embeddings and task-specific attention.

A wide variety of approaches were proposed by the teams obtaining competitive results in REST-MEX 2025, illustrating the diversity of techniques currently explored for sentiment analysis and thematic classification of Spanish-language tourist reviews. Most systems shared the goal of simultaneously addressing the three core subtasks: predicting sentiment polarity on a five-point scale, classifying the type of establishment

(hotel, restaurant, or attraction), and identifying the corresponding "Pueblo Mágico."

Several teams applied multitask learning frameworks using transformer-based models such as BETO, RoBERTa, or multilingual BERT (Almanza-Gonzalez et al., 2025; Phu and Van, 2025; Llanes Guilarte et al., 2025; Toapanta-Bernabé et al., 2025; Romero-Canton and Aranda-Romero, 2025; Gallardo-Hernández, Aranda, and Diaz-Pacheco, 2025; Torres-Santana and Balan-Euan, 2025; Balan-Euan and Torres-Santana, 2025; Carlos-Martínez and Pool-Cen, 2025). These systems often incorporated strategies like class weighting, SMOTE oversampling, or curriculum learning to address the strong class imbalance inherent in the corpus. For example, VerbaNexAI (Almanza-Gonzalez et al., 2025) reported strong performance in site type classification (macro F1 = 0.9043), while Hammer Squad (Llanes Guilarte et al., 2025) demonstrated that multilingual BERT-based models can be adapted effectively with progressive unfreezing and learning rate scheduling.

Other teams explored hybrid pipelines integrating transformer embeddings with classical classifiers. The lephuquy team (Phu and Van, 2025) combined BERT and XLM-RoBERTa embeddings with XGBoost classifiers and instruction-tuned LLaMA models for geocultural reasoning, though their town classification underperformed due to reliance on region metadata. Similarly, the Pumas PCIC team (Dueñas-González, Olvera-Morales, and Gomez-Adorno, 2025) compared SVMs, XGBoost, and pre-trained transformers, concluding that contextual features and data augmentation improved sentiment prediction accuracy.

In contrast to transformer-centric methods, several teams emphasized lighter or more interpretable alternatives. The Plenitas team (Martínez-López et al., 2025) used a MultiOutput Random Forest classifier enhanced with metadata, achieving 68.48% accuracy in polarity prediction. The Pepe el Gordo team (Agudelo Fuentes and Rojas Miranda, 2025) applied Random Forests over balanced Bag-of-Words representations, showing that undersampling and oversampling can yield solid baselines in imbalanced domains.

Prompt engineering emerged as an attractive lightweight alternative in low-resource settings. The Algiedi team (Sandoval, 2025) evaluated zero-shot and few-shot prompting with LLMs like GPT-3.5 and GPT-4, demonstrating moderate success in broad categories but limited effectiveness in fine-grained sentiment or town identification. Similarly, AVYus (Cabrera-Barrio, García-Niño, and Apaza-Mamani, 2025) assessed prompting strategies alongside SVMs and BERT fine-tuning, arguing that hybrid methods can leverage complementary strengths.

Other systems proposed novel architectures or data augmentation pipelines. PMOTE-UC-CUJAE (Leguen-de Varona et al., 2025) applied probabilistic oversampling with Ledoit-Wolf covariance shrinkage to augment RoBERTa embeddings, training MLP classifiers without expensive fine-tuning. The FrogCode team (Torres Torres and Serrano Cárdenas, 2025) employed a hierarchical attention network with bidirectional GRUs, outperforming baselines even for underrepresented classes. The NLP-GTO system (Hernández-Baca et al., 2025) combined MiniBERT embeddings with a genetic algorithm and Self-Organizing Map classifiers, achieving high performance in attraction type prediction.

Efforts to incorporate linguistic preprocessing and domain adaptation were also notable. SINAI-UGPLN (Toapanta-Bernabé et al., 2025) normalized dialectal and noisy Spanish text to improve model generalization, while DSVS (Vázquez-Santana et al., 2025) implemented curriculum learning and automatic loss weighting. The LKE Alliance (Reyes Peralta et al., 2025) combined TF-IDF features with frozen BETO encoders in a modular pipeline, and ELTSA-CUJAE (Rivero-Tapia, Simon-Cuevas, and Maestre Peña, 2025) proposed ensemble fusion via the Zimmerman–Zysno operator. Finally, UNISON-LCC (Robles-Robles et al., 2025) built a parallel NLP pipeline incorporating named entity recognition for hierarchical town classification.

Additional contributions further illustrated the spectrum of methods explored in the shared task. The Abit team (Tolulope Olalekan et al., 2025) combined logistic regression and fine-tuned Spanish BERT to establish baseline multimodal sentiment analysis pipelines, emphasizing the persistent challenge of class imbalance, especially in town prediction. Sintax Surfers

(Huerta-Espinoza, Villagomez-Garcia, and Rodríguez González, 2025) demonstrated that pre-trained transformers like RoBERTa-base-bne and BETO can yield high accuracy while highlighting the difficulties of scaling to many labels in town identification. The BRIAN team (Lezama-Sánchez, Tovar Vidal, and Reyes-Ortiz, 2025) explored an LSTM-based architecture, showcasing the potential of sequential models for Spanish text processing despite resource constraints and limited training epochs. Finally, the ReviewWizards team (Vazquez and Arreola, 2025) integrated Doc2Vec embeddings with XGBoost classifiers and heuristic town labeling strategies, offering an innovative combination of dense semantic representations and knowledge-guided rules, though their system showed mixed performance across metrics.

In addition to evaluating macro F1-scores and accuracy metrics, we conducted a statistical analysis to assess whether the performance differences among the top seven ranked systems were significant. The analysis was performed in Python using the Nemenyi post-hoc test, which compares multiple classifiers across tasks while controlling for Type I error in pairwise comparisons. Results indicated that although there were observable variations in average ranks and individual metric scores, none of the pairwise differences between the first seven submissions reached statistical significance at the conventional confidence level ($\alpha = 0.05$). This finding suggests that, despite architectural diversity and methodological innovations, the leading systems achieved comparable overall performance on the REST-MEX 2025 dataset.

Across these contributions, several commonalities emerged: most teams relied on transformer-based architectures to encode contextual representations, applied balancing techniques to mitigate label skew, and experimented with multitask learning or ensemble strategies to improve robustness. Despite differing levels of computational resources and architectural complexity, all systems contributed valuable insights into the challenges of multilingual, multidimensional sentiment analysis in the tourism domain.

The baseline is the Majority baseline, which simply predicts the majority class for all instances.

Table 4 presents the top-performing results for each class across the three traits. Notably, the UDENAR team achieved the highest performance in all classes, except for classes 1 and 3 of polarity and Atlixco, Chiapa de Corzo, and Real de 14 for Magical Town, where the proposal from the Axolotux team outperformed.

4 Perfect Ensemble and System Selection

To explore the collective potential of all submitted systems, we constructed a Perfect Ensemble (Álvarez-Carmona et al., 2022a; Álvarez-Carmona et al., 2018), where an instance is considered correctly classified if any system predicts it correctly, yielding the theoretical upper bound of performance across sentiment, type, and town classification. However, combining over sixty systems raises the question of whether a smaller subset can achieve comparable results.

To address this, we propose a set-theoretic selection method where each system is represented by the instances it uniquely solves, quantified by Equation 7. Systems contributing fewer exclusive correct predictions are iteratively removed.

$$Contributes(k) = \left| Correct_k \setminus \bigcup_{j \neq k} Correct_j \right| \tag{7}$$

Table 3 shows that the Sub Perfect Ensemble of just 16 systems achieves identical results to the full ensemble. This demonstrates the effectiveness of the method in identifying a compact yet highly complementary subset of models, detailed in Table 5.

5 Interesting opinions

Some of the interesting opinions in the collection are those that were correctly classified by all 70 participating systems. These are the easiest instances to classify. The same applies to instances that none of the systems were able to classify correctly, which could indicate the most challenging instances to classify or instances whose labels are incorrect for some reason.

Table 6 shows the number of interesting opinions per class. It can be observed that for polarity, no class was correctly classified by all systems. However, there are instances

Rank	Institute	Country	Team	Sent	Res_P	Res_T	Res_{MT}
-	-	-	*Perfect Assembly*	0.96	0.99	1.00	0.93
-	-	-	*Sub Perfect Assembly*	0.89	0.92	1.00	0.84
1st	UDENAR	Col	UDENAR	0.73	0.64	0.99	0.69
2nd	INFOTEC/UNAM	Mex	Axolotux	0.72	0.64	0.99	0.69
3rd	CIMAT	Mex	Pandas Rojos	0.69	0.62	0.98	0.63
HM	U. Coruña/University Islamabad	Esp/Pak	LyS	0.68	0.60	0.98	0.63
HM	UNAM	Mex	FisBio	0.67	0.58	0.97	0.62
HM	CIMAT	Mex	CorpusChristi	0.65	0.60	0.94	0.59
HM	CIMAT	Mex	MixtlanMapper	0.65	0.55	0.96	0.61
HM	CENATAV/Ibero/CIMAT	Cub/Mex	Hammer Squat	0.65	0.58	0.97	0.58
HM	CICESE-UAT	Mex	Sintax Surfers	0.64	0.59	0.96	0.57
HM	BUAP/CIMAT	Mex	LKE Alliance	0.62	0.62	0.97	0.50
HM	UNISON	Mex	UNISON	0.62	0.48	0.94	0.60
HM	Ibero	Mex	ZIZEK	0.61	0.47	0.97	0.59
HM	UIT/Vietnam National University	Vnm	lephuquy	0.61	0.63	0.98	0.47
HM	CIMAT	Mex	FrogCode	0.60	0.54	0.94	0.53
HM	Universidad Carlos III	Esp	AVYus	0.58	0.59	0.96	0.46
HM	CUJAE	Cub	ELTSA-CUJAE	0.57	0.54	0.90	0.47
HM	UNAM	Mex	Pumas PCIC	0.56	0.58	0.98	0.40
HM	UC/Plénitas/Universidad de Camaguey	Esp/Cub/Mex	UC-UCO-Plenitas	0.52	0.32	0.92	0.51
HM	CICESE-UAT	Mex	HuertaEspinoza	0.51	0.59	0.20	0.57
HM	Universidad Tecnológica de Bolívar	Col	VerbaNexIA	0.45	0.18	0.90	0.48
HM	ITM	Mex	MouseSquad	0.38	0.48	0.75	0.18
HM	BUAP/UAM	Mex	BRIAN	0.34	0.18	0.88	0.26
HM	IPN	Mex	DSVS	0.32	0.42	0.93	0.04
HM	IPN/CIMAT	Mex	ReviewWizards	0.28	0.19	0.36	0.31
HM	UG	Mex	NLP-GTO	0.28	0.27	0.86	0.09
HM	Centro Geo	Mex	GeoLab	0.17	0.23	0.36	0.06
HM	UC/Plénitas/Universidad de Camaguey	Cub/Esp	PMOTE-UC-CUJAE	0.16	0.41	0.13	0.01
HM	CIC/F.U. Oye-Ekiti/ Ladoke Akintola	Mex/Nga	Abit	0.13	0.20	0.33	0.02
HM	Algiedi Solutions	Mex	Algiedi	0.13	0.20	0.33	0.03
HM	SEP Mérida	Mex	The last	0.13	0.20	0.33	0.03
HM	Ibero/TecNM	Mex	Pepe el gordo	0.13	0.20	0.33	0.02
HM	SINAI/Universidad de Guayaquil	Esp/Ecu	UGPLN	0.13	0.19	0.33	0.02
BL	-	-	Mayority	0.09	0.16	0.20	0.01

Table 3: Performance for the REST-MEX 2025 Sentiment Analysis task (all runs as reported).

Class	Best F result	Team
1	0,6842	Axolotux-E3
2	0,4722	UDENAR
3	0,6044	Axolotux-E-T3
4	0,5545	UDENAR
5	0,8883	UDENAR
Attractive	0,9882	UDENAR
Hotel	0,9844	UDENAR
Restaurant	0,9904	UDENAR
Atlixco	0,6695	Axolotux-E-T3
Chiapa de Corzo	0,8078	Axolotux-E-T3
Real de 14	0,6678	Axolotux-E-T3
The rest	-	UDENAR

Table 4: Performance for the Sentiment Analysis task per class.

Team	*P*	*T*	*MT*
UDENAR	✓	✓	✓
Axolotux	✓	✓	✓
Pandas Rojos	✓	✓	✓
VerbaNexIA	✓		
MouseSquad	✓		
Hammer Squat	✓		✓
PMOTE-UC-CUJAE	✓		
lephuquy		✓	
UC-UCO-Plenitas		✓	
FisBio	✓	✓	
MixtlanMapper		✓	✓
ReviewWizards			✓
Algiedi			✓
The last			✓
GeoLab			✓
BRIAN			✓

Table 5: The best 16 systems selected for the assembly.

from all polarity classes that were incorrectly classified by all systems. For example:

Polarity 1: *Lo único interesante es sacar la típica foto con las ruinas delante del mar, pero es difícil ya que los turistas salen en la foto continuamente.*

In the case of the type attribute, it is notable that the class *Attractive* has 1581 instances that were correctly classified by all systems. This could be because this class was inherently the easiest within this attribute.

Class	Correct by All	Wrong by All
1	0	51
2	0	10
3	0	19
4	0	5
5	0	2
Attractive	1581	0
Hotel	0	0
Restaurant	0	0
Towns	0	5058

Table 6: Interesting opinions.

For the town classification task, the large number of instances not correctly predicted by any system likely reflects the difficulty of this attribute, especially when the text contains no explicit clues about the geographic location. For example:

> **Magical Town Valladolid:** *Es una lástima que siendo un lugar tan bonito tenga personal tan poco capacitado y con 0 actitud de servicio*

6 Conclusions

This paper described the design and results of the Rest-Mex shared task collocated with IberLef 2025. For the task, 32 teams participated. Mainly, the members of these teams come from institutes in countries such as Mexico, Spain, Cuba, Colombia, Vietnam, Nigeria and Pakistan. 70 systems were received to be evaluated to solve the task proposed in the Rest-Mex 2025.

The sentiment analysis task aimed to identify the polarity, type, and country of an opinion. The team that achieved the best performance was (Jurado-Buch, 2025). This team represents the University of Nariño in Colombia. They proposed a method based on decomposing each multiclass problem into a set of independent binary classifiers, one per class. Each binary classifier was trained on a balanced dataset combining all positive examples of the target class and a matching number of negative examples selected either randomly or through centroid-based sampling of embeddings from a fine-tuned RoBERTa model. The outputs of all classifiers were then concatenated and passed to a multilayer perceptron to produce the final prediction. This approach obtained the highest macro F1-scores reported in the competition.

Also, the results indicate that distinguishing between opinions of hotels, restaurants, and attractions is a task that can have very high results, close to 100 %. The country trait is a little more difficult.

A method was proposed to measure the easiness of a corpus, which appears to effectively reflect the performance of the approaches aiming to solve different tasks.

Finally, it is shown that there is significant complementarity between the participating systems. A system was also proposed to select among the participating systems in such a way that the same results can be achieved by combining only 16 systems, compared to the original 70 systems of the sentiment analysis task.

Acknowledgements

The authors gratefully acknowledge the support provided by the Mexican Academy of Tourism Research (AMIT) for the project *"Balancing Tourism Text Data with Artificial Intelligence for Sentiment Analysis: A Specialized Language Model Approach"* funded through the *Research Projects 2024* call.

This work was also supported by the project *"Text Generation for Data Balancing in Sentiment Classification: Application to Tourism Data"* under the *CICIMPI 2024* call of the Centro de Investigación en Matemáticas (CIMAT).

Our special thanks go to all of Rest-Mex's participants, the organizers, and their institutions.

References

Agudelo Fuentes, L. and R. S. Rojas Miranda. 2025. Balanced bag-of-words classification of spanish tourism reviews using random forests. In *IberLEF@SEPLN*.

Almanza-Gonzalez, D., J. Serrano Castañeda, J. C. Martinez-Santos, and E. Puertas. 2025. Application of multitask bert models for the classification of sentiments, types of places, and magical towns in spanish tourist reviews. In *IberLEF@ SEPLN*.

Álvarez-Carmona, M. Á., R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, and A. Y. Rodríguez-González. 2021. Overview of rest-mex at iberlef 2021: recommendation system for

text mexican tourism. *Procesamiento del Lenguaje Natural*.

Álvarez-Carmona, M. Á., R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, and A. P. López-Monroy. 2022a. A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one. *Computación y Sistemas*, 26(2).

Álvarez-Carmona, M. A., R. Aranda, A. Rodriguez-Gonzalez, D. Fajardo-Delgado, M. G. A. Sanchez, H. Perez-Espinosa, J. Martinez-Miranda, R. Guerrero-Rodriguez, L. Bustio-Martinez, and A. D. Pacheco. 2022b. Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University-Computer and Information Sciences*.

Álvarez-Carmona, M. A., R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, and H. Carlos. 2024. Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news. *Journal of Information Science*.

Álvarez-Carmona, M. Á., Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, and L. Bustio-Martínez. 2022. Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts. *Procesamiento del Lenguaje Natural*, 69:289–299.

Álvarez-Carmona, M. A., M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, and L. Villaseñor-Pineda. 2018. Semantically-informed distance and similarity measures for paraphrase plagiarism identification. *Journal of Intelligent & Fuzzy Systems*, 34(5):2983–2990.

Álvarez-Carmona, M. Á. and R. Aranda. 2022. Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias.

Álvarez-Carmona, M. Á., Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, V. Muñiz-Sánchez, A. P. López-Monroy, F. Sánchez-Vega, and L. Bustio-Martínez. 2023. Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts. *Procesamiento del Lenguaje Natural*, 71:425–436.

Arce-Cardenas, S., D. Fajardo-Delgado, M. Á. Álvarez-Carmona, and J. P. Ramírez-Silva. 2021. A tourist recommendation system: a study case in mexico. In *Mexican International Conference on Artificial Intelligence*, pages 184–195. Springer.

Balan-Euan, M. E. and E. A. Torres-Santana. 2025. Balanced text classification for tourism data in mexico using nlp and custom sampling techniques. In *IberLEF@ SEPLN*.

Cabrera-Barrio, Á., Y. M. García-Niño, and V. A. Apaza-Mamani. 2025. Natural language processing approaches for spanish tourist review analysis: Insights from the rest-mex 2025 shared task. In *IberLEF@ SEPLN*.

Carlos-Martínez, H. and J. Pool-Cen. 2025. Holistic classification of tourism reviews: A structured prediction approach with energy-based models. In *IberLEF@ SEPLN*.

Castillo-Ortiz, I., M. Á. Álvarez-Carmona, R. Aranda, and Á. Díaz-Pacheco. 2024. Evaluating culinary skill transfer: A deep learning approach to comparing student and chef dishes using image analysis. *International Journal of Gastronomy and Food Science*, 38:101070.

Diaz-Pacheco, A., M. Á. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, and R. Aranda. 2024. Artificial intelligence methods to support the research of destination image in tourism. a systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(7):1415–1445.

Díaz-Pacheco, Á., R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, and R. Aranda. 2023. 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(9):101746.

Díaz-Pacheco, Á., R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-

González, and R. Aranda. 2024. Quantifying differences between ugc and dmo's image content on instagram using deep learning. *Information Technology & Tourism*, 26(2):293–329.

Dueñas-González, L. F., C. E. Olvera-Morales, and H. Gomez-Adorno. 2025. Pumas pcic team at rest-mex 2025: Classification strategies for sentiment analysis. In *IberLEF@SEPLN*.

Gallardo-Hernández, A. Z., R. Aranda, and A. Diaz-Pacheco. 2025. A multi-task beto-based framework with synthetic data augmentation for sentiment and contextual classification of spanish tourist reviews. In *IberLEF@ SEPLN*.

García-Espinosa, D. A., L. E. Flores-Luna, and A. Moreno-Sánchez. 2025. Multitask bert architecture for sentiment analysis and classification of mexican tourism reviews in rest-mex 2025. In *IberLEF@SEPLN*.

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. 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*.

Guerrero-Rodriguez, R., M. Á. Álvarez-Carmona, R. Aranda, and A. P. López-Monroy. 2023. Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico. *Current issues in tourism*, 26(2):289–304.

Hernández-Angeles, G., D. Paniagua-Molina, C. Aguirre-Calzadilla, and U. I. Luján-López. 2025. Combining fine-tuned bert and classical mlp for mexican tourism nlp tasks: Participation of cimat-cc in rest-mex 2025. In *IberLEF@ SEPLN*.

Hernández-Baca, A., M. A. Rojas-Andrade, J. N. Figueroa-Ramírez, and J. R. Prieto-Valdivia. 2025. Rest-mex 2025: Sentiment analysis and magical towns detection task. In *IberLEF@SEPLN*.

Huerta-Espinoza, M. A., J. C. Villagomez-Garcia, and A. Y. Rodríguez González. 2025. The sintax surfers team at rest-mex 2025: Solving the sentiment analysis task using pre-trained transformers-based models. In *IberLEF@ SEPLN*.

Imran, M., T. Rasheed, and C. Gómez-Rodríguez. 2025. Enhancing transformer-based sentiment analysis for the rest-mex 2025 challenge: A hybrid strategy with oversampling, back translation, and transformers. In *IberLEF@ SEPLN*.

Jurado-Buch, J. D. 2025. Binary fine-tuning with centroid-based sampling and transformer fusion: The top-scoring system at rest-mex 2025. In *IberLEF@SEPLN*.

Leguen-de Varona, I., J. Madera, A. Simon-Cuevas, L. Lastre Figueroa, and Y. Martínez-López. 2025. Pmote-uc-cujae at rest-mex 2025: Evaluation of probabilistic data augmentation models for mining mexican tourist reviews. In *IberLEF@ SEPLN*.

Lezama-Sánchez, A. L., M. Tovar Vidal, and J. A. Reyes-Ortiz. 2025. Place classification and sentiment analysis in reviews of mexican magical towns using lstm networks. In *IberLEF@ SEPLN*.

Llanes Guilarte, D. S., V. Herrera-Semenets, L. Bustio-Martínez, and M. Á. Álvarez-Carmona. 2025. Bert-based models for joint sentiment, type, and location classification of spanish tourist reviews. In *IberLEF@ SEPLN*.

Martínez-López, Y., I. d. l. M. Leguén de Varona, M. Bethencourt, D. Rodríguez Fernández, J. Madera, A. Y. Rodríguez González, C. de Castro Lozano, J. M. Ramírez Uceda, and J. C. Arévalo Fernández. 2025. Uc-ucoplenitas team - exploring in the rest-mex 2025: Researching sentiment evaluation in text for mexican magical towns. In *IberLEF@ SEPLN*.

Minutti-Martinez, C., B. Escalante-Ramirez, and J. Olveres. 2025. Multitask analysis of spanish travel reviews: Sentiment, destination, and topic classification with roberta and llama ensembles. In *IberLEF@SEPLN*.

Monsivais Borjón, J. J. and M. Á. Álvarez-Carmona. 2025. Multitask classification

of mexican tourist reviews using a multi-head transformer model based on beto. In *IberLEF@SEPLN*.

Morales-Murillo, V. G., H. Gómez-Adorno, D. Pinto, I. A. Cortés-Miranda, and P. Delice. 2023. Lke-iimas team at rest-mex 2023: Sentiment analysis on mexican tourism reviews using transformer-based domain adaptation.

Olmos-Martínez, E., M. Á. Álvarez-Carmona, R. Aranda, and A. Díaz-Pacheco. 2024. 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(2):639–661.

Phu, Q. L. and T. D. Van. 2025. Lpq team at rest-mex 2025: Bert and llm approaches in tourism review classification. In *IberLEF@ SEPLN*.

Reyes Peralta, A. E., E. A. Padilla Luis, V. Muñiz Sánchez, and D. Pinto. 2025. Tourist reviews analysis: An integral approach with traditional models and fine-tuned llms. In *IberLEF@ SEPLN*.

Rivero-Tapia, M. A., A. Simon-Cuevas, and R. Maestre Peña. 2025. Eltsa-cujae at rest-mex 2025: A novel ensemble learning approach with transformers for mining mexican tourist reviews. In *IberLEF@ SEPLN*.

Robles-Robles, G. S., J. J. Ramirez-Ramirez, A. D. Durazo-Bartolini, G. Balderrama-Dominguez, L. H. Hernandez-Gutierrez, V. H. Ramirez-Rios, J. A. Nava-Banda, G. Gutierrez-Navarro, J. D. Garcia-Ruiz, M. A. Castro-Lerma, J. A. Flores-Briones, M. I. Melendez-Rivera, C. A. Flores-Alvarez, J. A. Morales-Nuñez, and M. Toledo-Acosta. 2025. A parallel nlp pipeline with ner-enhanced hierarchical classification: Sentiment analysis for mexican magical towns. In *IberLEF@SEPLN*.

Romero-Canton, A. and J. R. Aranda-Romero. 2025. A test of mutual information features in multi-task classification spanish tourist reviews. In *IberLEF@ SEPLN*.

Sandoval, F. 2025. Prompt engineering for sentiment analysis in tourism: The case of mexican pueblos mágicos. In *IberLEF@SEPLN*.

Siliceo-Guzmán, I., R. Aranda, and M. Á. Álvarez-Carmona. 2025. Hybrid prompt engineering and transfer learning for sentiment analysis in mexican tourism reviews. In *IberLEF@SEPLN*.

Toapanta-Bernabé, M. d. C., M. Á. García-Cumbreras, L. A. Ureña-López, K. G. Bajaña-Bastidas, and S. L. Urgiles-Manzano. 2025. Sinai-ugpln at rest-mex iberlef 2025: Multilevel analysis of dialectal and noisy spanish text for sentiment classification. In *IberLEF@ SEPLN*.

Tolulope Olalekan, A., A. Tewodros, O. Olumide Ebenezer, A. Olaronke Oluwayemisi, A. Oluwatobi Joseph, O. Temitope Dasola, and G. Sidorov. 2025. Multimodal sentiment analysis in spanish tourist reviews: A data quality-aware approach. In *IberLEF@ SEPLN*.

Torres-Santana, E. A. and M. E. Balan-Euan. 2025. Multi-dimensional classification system using pre-trained transformer models for multilingual text analysis. In *IberLEF@ SEPLN*.

Torres Torres, E. F. and E. D. Serrano Cárdenas. 2025. Hierarchical attention networks for multilabel sentiment analysis in spanish reviews of mexican magic towns. In *IberLEF@SEPLN*.

Vazquez, L. and J. Arreola. 2025. Multi-task text classification of tourist reviews using doc2vec. In *IberLEF@ SEPLN*.

Vázquez-Santana, D., S. Damián-Sandoval, C. Yáñez-Márquez, and E. Felipe-Riverón. 2025. Dsvs at rest-mex 2025: A multitask approach for sentiment, place type, and town classification in spanish reviews. In *IberLEF@ SEPLN*.