Combining Fine-Tuned BERT and Classical MLP for Mexican Tourism NLP Tasks: Participation of CIMAT-CC in REST-MEX 2025

Gustavo-Hernández-Angeles[1,*], Diego-Paniagua-Molina[1], César-Aguirre-Calzadilla[1] and Uziel-Isaí-Luján-López[1]

[1]*Mathematics Research Center (CIMAT-Centro de Investigación en Matemáticas), Graduate Program in Statistical Computing, Nuevo León, Mexico*

Abstract

This article details our participation in REST-MEX 2025, an evaluation task focused on sentiment analysis of tourism reviews about Pueblos Mágicos in Mexico. We addressed three subtasks: sentiment polarity classification, destination type identification, and Pueblo Mágico recognition. For the sentiment polarity and Pueblo Mágico recognition tasks, we employed fine-tuned BERT-based Transformer models. For the destination type identification, we explored an approach based on Word2Vec embeddings classified by a Multilayer Perceptron (MLP). We present a data analysis, a detailed methodology for each task—including implementation and fine-tuning/training procedures—the results obtained on the official dataset, and the conclusions drawn from our participation.

Keywords

Sentiment analysis, Natural Language Processing, BERT, Multilayer Perceptron.

1. Introduction

Natural Language Processing (NLP) has become a cornerstone of modern information processing, enabling the extraction of meaning and structure from large volumes of textual data [1]. Among its many applications, sentiment analysis—a task focused on identifying and categorizing opinions and emotions expressed in text—has gained prominence due to its relevance across diverse sectors, including marketing, politics, healthcare, and tourism [2, 3, 4]. Despite this relevance, most NLP tools have historically been developed for the English language, leaving a significant gap in resources and methodologies for Spanish, a language spoken by hundreds of millions across various dialectal and cultural contexts [5, 6].

Tourism, in particular, stands to benefit from robust sentiment analysis frameworks. In Mexico, tourism accounted for 8.6 % of the national GDP in 2023, according to government statistics [7, 8]. Understanding public opinion is essential not only for enhancing tourist experiences but also for shaping regional development strategies. Initiatives such as the "Pueblos Mágicos" program have successfully promoted lesser-known destinations, yet they lack sophisticated sentiment-analysis tools tailored to Spanish-language reviews [9].

To address this gap, the REST-MEX 2025 shared task—"Researching Sentiment Evaluation in Text for Mexican Magical Towns"—was launched as part of IberLEF@SEPLN 2025 [10] and organized by the Centro de Investigación en Matemáticas (CIMAT) [11]. This task focuses on classifying the sentiment polarity of Spanish-language reviews on a 1-to-5 scale, identifying the type of destination (hotel, restaurant, or attraction), and associating the text with one of 40 designated Pueblos Mágicos. The dataset provided includes over 200,000 annotated reviews divided into training and testing subsets.

IberLEF 2025: International Iberian Languages Evaluation Forum, September 2025, Spain.

*Corresponding author.

gustavo.hernandez@cimat.mx (Gustavo-Hernández-Angeles); diego.paniagua@cimat.mx (Diego-Paniagua-Molina); cesar.aguirre@cimat.mx (César-Aguirre-Calzadilla); uziel.lujan@cimat.mx (Uziel-Isaí-Luján-López)

0009-0001-3160-5294 (Gustavo-Hernández-Angeles); 0009-0006-6564-2794 (Diego-Paniagua-Molina); 0009-0005-8418-7355 (César-Aguirre-Calzadilla); 0009-0003-7360-5377 (Uziel-Isaí-Luján-López)

In this paper, we present our system developed for REST-MEX 2025. Our architecture leverages pre-trained Transformer models—specifically BERT—for the Polarity and Town classification tasks, and a custom-designed Artificial Neural Network for the Type classification task [12]. We describe the technical details of our approach, report on the experiments conducted, and analyze the performance achieved on the official competition dataset.

2. Related Work

Sentiment analysis in Spanish has gained increasing attention in recent years, particularly within the tourism domain. Several corpora and shared tasks have been developed to support this line of research, including TASS [13] and REST-MEX [14, 15, 16]. Studies by Guerrero-Rodríguez et al. [17] and Álvarez-Carmona et al. [15] have shown that ensemble methods combining multiple classifiers enhance polarity-detection performance on user-generated content from platforms such as TripAdvisor.

The introduction of Transformer-based architectures, especially BERT, has significantly improved performance over classical models in opinion-mining tasks. Spanish-adapted variants such as BETO [5] have demonstrated robust results when fine-tuned for polarity and topic classification. In REST-MEX 2021, the top-ranking system employed BETO fine-tuned to predict star ratings from Mexican tourism reviews [14]. Subsequent editions in 2022 and 2023 confirmed this trend, with the majority of participating systems relying on BERT variants, thereby consolidating the dominance of Transformer-based models in this task domain [15, 16].

The REST-MEX shared tasks have evolved across editions. In 2021, tasks included site recommendation and sentiment-polarity classification; the leading system fine-tuned BETO to predict polarity on a five-point scale [14]. In 2022, a new task—classification of destination type (hotel, restaurant, attraction)—was introduced. The winning system integrated handcrafted linguistic features (UMUTextStats) with BERT, achieving a Macro-F1 score of 0.89 [18]. Another competitive approach involved preprocessing techniques such as translation and data cleaning prior to BERT training [15]. REST-MEX 2023 further expanded the task set by incorporating country identification; results indicated that predicting destination type and country (F1 ≈ 0.99 and ≈ 0.93, respectively) was less complex than predicting sentiment polarity [16].

Overall, the REST-MEX track at IberLEF has significantly contributed to advancing sentiment analysis in Spanish, particularly in tourism-related contexts. These shared tasks have reinforced the effectiveness and widespread adoption of Transformer-based architectures for multilingual and domain-specific NLP applications.

3. Background

In this section we outline the theoretical and contextual foundations that support our system: (i) the Transformer architecture and its BERT adaptation; (ii) Spanish-trained BERT variants, with emphasis on BETO; (iii) the fundamentals of the Multilayer Perceptron (MLP); and (iv) the relevance of the *Pueblos Mágicos* program to Mexican tourism.

3.1. Transformer and BERT

The Transformer introduced a self-attention mechanism that replaces traditional recurrence and enables the modeling of long-range dependencies with full training parallelism [19]. [1]

BERT (*Bidirectional Encoder Representations from Transformers*) extends this architecture through self-supervised pre-training based on *Masked Language Modeling* (MLM) and *Next Sentence Prediction* (NSP), producing bidirectional contextual representations [12]. In numerous text-classification challenges—including TASS and REST-MEX—BERT systematically outperforms SVM, LSTM, or CNN

[1] The draft for this section was generated with assistance from the AI tool Gemini and subsequently reviewed, edited, and validated by the authors.

approaches when a moderately sized annotated corpus is available and task-specific fine-tuning is applied [14, 16].

BERT for Spanish

Models such as BETO [5], MarIA [6], and DistilBETO narrow the performance gap with English by training on large Spanish-language corpora. These variants retain the base BERT architecture (`bert-base`, ≈110 M parameters) but employ Spanish *WordPiece* vocabularies and weights adapted to the language. Comparative studies in sentiment polarity (TASS, MEX-A3T, REST-MEX) report absolute macro-F1 improvements of 3–8 points over BERT, as documented by [17, 15, 20], among others.

3.2. Multilayer Perceptron (MLP)

The MLP is a feed-forward neural network composed of one or more dense layers with non-linear activation functions (e.g. ReLU). Its universal approximation power makes it a lightweight alternative—compared with complex Transformers—when the input is already a fixed vector representation (TF–IDF, averaged embeddings, etc.) and the output involves a large number of classes [21]. For the *Type* identification task (3 labels), an MLP provides:

- **Efficiency**: fast training and inference even without a high-end GPU;
- **Simple regularization**: techniques such as Dropout or L2 mitigate overfitting under class imbalance;
- **Flexibility**: ability to combine lexical features with available metadata.

3.3. Sentiment Analysis in Tourism and *Pueblos Mágicos*

Sentiment analysis has become a strategic tool for understanding tourist perceptions and guiding decision-making in emerging destinations. In 2023, tourism accounted for 8.6 % of Mexico's GDP, and the *Pueblos Mágicos*[2] program has been instrumental in diversifying the offer beyond beach resorts. As of January 2025, the official register lists **177** towns after the addition of 45 new destinations [9].

Spanish tourist reviews often exhibit colloquial language, regional dialect blends, and code-switching (Spanish–English), increasing preprocessing complexity. In REST-MEX 2025 these reviews are labelled with: *Polarity* (1–5 stars), *Type* (hotel, restaurant, attraction), and *Town* (40 *Pueblos Mágicos* in the corpus). Capturing the underlying semantics and emotional nuances motivates the use of contextual-attention models (BERT) for *Polarity/Town*, whereas the 3-class granularity of *Type* encourages a lighter, robust architecture (MLP) based on high-dimensional sparse vectors.

3.4. Synthesis

In summary, the combination of (i) deep BERT representations, (ii) Spanish-adapted variants such as BETO, and (iii) an efficient MLP for high-cardinality multi-class classification forms the backbone of our system for the three REST-MEX 2025 subtasks. The tourism context of the *Pueblos Mágicos* not only lends practical relevance to the research but also introduces linguistic and class-imbalance challenges that justify the proposed hybrid strategy.

4. Data Analysis

This section provides a quantitative and qualitative overview of the dataset used in REST-MEX 2025, including its structure, class distributions, and textual characteristics, along with a summary of the preprocessing steps.

[2]Federal initiative launched in 2001 by the Secretaría de Turismo (SECTUR) to promote towns with high historical and cultural value.

4.1. Corpus Description

The official REST-MEX 2025 dataset consists of a total of **297,218** Spanish-language tourism reviews, divided into a training set of **208,052** examples and a test set of **89,166**. Each instance includes a short `Title`, a longer `Review`, and three classification labels: *Polarity* (1–5), *Type* (Hotel, Restaurant, Attraction), and *Town* (one of 40 designated *Pueblos Mágicos*).

4.2. Label Distributions

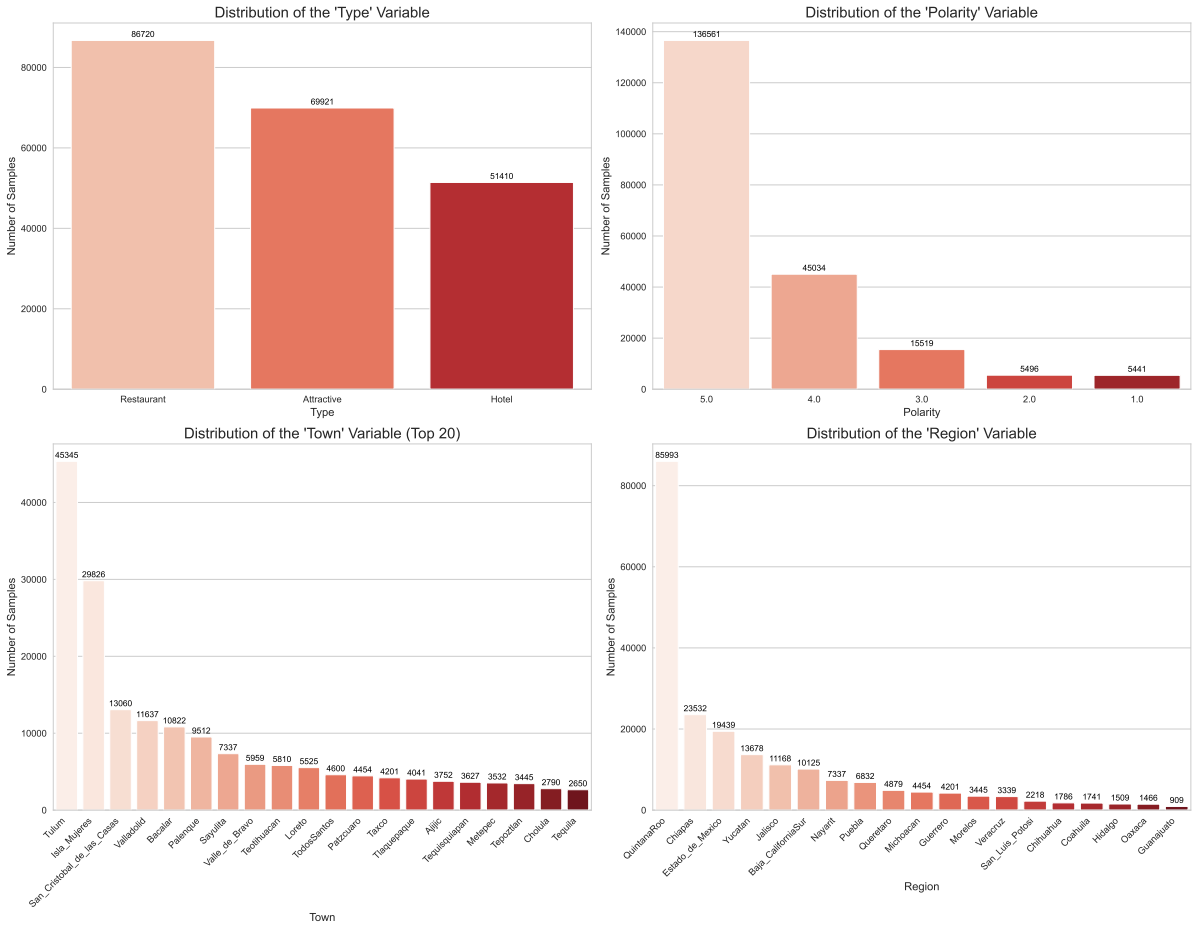

Figure 1: Label distributions for the REST-MEX 2025 dataset. Top: Type and Polarity. Bottom: Town and Region. The Python code used to generate this figure was developed with assistance from ChatGPT and GitHub Copilot.

Sentiment Polarity. The class distribution is strongly skewed towards positive opinions. As shown in Figure 1, more than 65% of the reviews are labeled with 5 stars, while 1- and 2-star reviews collectively account for less than 10% of the total. The mean polarity is 4.45, with a median of 5.

Destination Type. The `Type` label exhibits moderate imbalance: Restaurants dominate (86,720 reviews), followed by Attractions (69,921) and Hotels (51,410). This may reflect user interest bias or platform review frequency.

Town. A long-tailed distribution is observed for the `Town` variable. The most reviewed town is Tulum (45,345), followed by Isla Mujeres (29,826) and San Cristóbal de las Casas (13,060). On the opposite end, Tapalpa and Real de Catorce have fewer than 800 reviews each. This variability suggests challenges for classification models in minority towns.

Region. Although not used as a target label, the `Region` field is included as metadata. Reviews are unevenly distributed across states, with Quintana Roo (85,993) and Chiapas (23,532) representing the largest shares.

4.3. Textual Characteristics

The reviews range widely in length. Based on tokenization statistics:

- **Mean length:** 63.4 tokens
- **Median:** 45 tokens
- **95th percentile:** ∼180 tokens

Most texts are short and highly subjective, often containing informal language, emojis, emphasis with repeated characters, and occasional code-switching to English. These traits introduce noise and motivate careful preprocessing and robust tokenization strategies.

4.4. Preprocessing Summary

The original `Title` and `Review` fields were merged into a single column (`read_text`). We then applied a standard cleaning routine:

1. **Mojibake correction** using a custom decoding function.
2. **Lower-casing** and Unicode NFD normalization to remove diacritics.
3. **Noise filtering:** removal of non-alphabetic characters using regex.
4. **Stop-word removal** using the NLTK Spanish stop-word list.
5. Resulting tokens were re-joined and stored in a new column `clean_text`.

The cleaned datasets were saved as `train_clean.csv` and `test_clean.csv`, and used for all downstream experiments.

4.5. Implications

The strong polarity imbalance led us to explore weighted loss functions. For the `Type` task, the long-tailed label distribution and varying text lengths informed the choice of a lightweight architecture with regularization mechanisms to avoid overfitting on rare classes.

5. Methodology

This section outlines the methodological approaches implemented to address the three subtasks of the REST-MEX 2025 challenge: sentiment polarity classification (`Polarity`), prediction of the type of destination mentioned (`Type`), and identification of the specific *Pueblo Mágico* (`Town`).

For the Polarity and Town tasks, we adopted advanced Transformer-based models, specifically BETO—a Spanish-pretrained variant of BERT—capable of capturing the linguistic intricacies of the domain more effectively.

In contrast, the Type classification task posed lower complexity and involved a balanced label set. Therefore, we opted for a classical machine learning approach using a Multilayer Perceptron (MLP) as the primary model.

5.1. Sentiment Polarity Classification

The objective of this task is to assign a sentiment label ranging from 1 (very negative) to 5 (very positive) to each tourist review. We implemented a pre-trained Transformer model and fine-tuned it on the task-specific training data provided in the competition.

5.1.1. BERT

The core model employed for this task was BERT, specifically the 'dccuchile/bert-base-spanish-wwm-cased' variant. This model is pre-trained on a large Spanish-language corpus using Whole Word Masking as its training strategy. Its selection was motivated by the model's established performance across a range of Spanish-language NLP tasks.

The architecture retains the standard BERT-base configuration: 12 encoder layers, 768 hidden units per layer, 12 attention heads per self-attention mechanism, and approximately 110 million pre-trained parameters.

Leveraging BERT's self-attention layers allows the model to capture complex contextual relationships between words, which is essential for accurately identifying sentiment in tourism reviews. A linear classification layer was added on top of the [CLS] token representation from the final hidden layer to perform sequence-level classification. This layer was trained to map the contextualized representation to one of five polarity classes.

5.1.2. Implementation and Fine-tuning

The official REST-MEX 2025 training set served as the starting point. Initial preprocessing addressed encoding issues in the 'Title' and 'Review' columns, which were concatenated into a unified input field named 'Texto_Leido'. No aggressive preprocessing such as stopword or accent removal was applied, in order to preserve as much contextual information as possible.

The polarity labels, originally ranging from 1 to 5, were shifted to a 0–4 scale to meet the requirements of PyTorch's `CrossEntropyLoss`. The dataset was partitioned into training and validation sets in an 80/20 split.

Tokenization was handled using the `AutoTokenizer` utility, with sequences truncated to a maximum length of 128 tokens. Shorter sequences were padded to ensure input uniformity.

Model training was conducted using Hugging Face's Transformers library and its `Trainer` class. The main training parameters were configured as shown in Table 1:

Table 1
Training configuration for the sentiment polarity model

Parameter	Value
output_dir	Path for model outputs and checkpoints
evaluation_strategy	epoch
num_train_epochs	3
per_device_train_batch_size	16
per_device_eval_batch_size	32
load_best_model_at_end	True
metrics	precision, recall, and F1-score

5.2. Magical Town Identification

The goal of this task was to identify the specific *Pueblo Mágico* referenced in each tourist review, selecting from the set of localities included in the REST-MEX 2025 database. Given the nature of this multi-class classification task with a high number of categories, a pre-trained Transformer model was again selected.

5.2.1. BERT

The selected model for the *Pueblo Mágico* identification task was, once again, BETO. We used the same variant as in the Polarity task: 'dccuchile/bert-base-spanish-wwm-cased'. The underlying architecture remained unchanged from the previously described configuration.

We hypothesized that BETO's capacity to model context and semantics would be advantageous in this setting, particularly in cases where location names are implied rather than explicitly mentioned. The model's ability to extract relevant contextual cues from descriptions and characteristics was considered essential for this classification task.

For the final classification layer, a linear head was added on top of the [CLS] token representation from the final hidden layer. This layer was trained to map each review's contextual embedding to one of the defined *Town* labels.

5.2.2. Implementation and Fine-tuning

The dataset provided by REST-MEX 2025 was used as the input source. Data preprocessing followed the same steps as in the Polarity task, with one key difference: encoding the target variable `Town`. A dictionary was created to map each unique *Pueblo Mágico* name to a numerical identifier (0-39), producing a `label` column used as the model target. An 80/20 split was applied for training and validation sets, respectively.

Tokenization was carried out using the `AutoTokenizer` from the 'dccuchile/bert-base-spanish-wwm-cased' model, with a maximum sequence length of 128 tokens. Sequences exceeding this limit were truncated, and shorter sequences were padded as needed.

Training was performed using the Hugging Face Transformers library and the `Trainer` class. To address class imbalance resulting from uneven mention frequencies among towns, a class weighting strategy was applied. We computed class-specific weights using `compute_class_weight` with the 'balanced' option from `scikit-learn`. These weights were integrated into the loss function (`CrossEntropyLoss`) via a custom `WeightedTrainer` wrapper.

The main training parameters are summarized in Table 2.

Table 2
Training configuration for BERT model on Magical Town identification task.

Parameter	Value
output_dir	./resultados_beto_town
evaluation_strategy	epoch
save_strategy	epoch
num_train_epochs	3
per_device_train_batch_size	16
per_device_eval_batch_size	32
warmup_steps	500
weight_decay	0.01
load_best_model_at_end	True
metric_for_best_model	accuracy

5.3. Destination Type Classification

For the Type task, which aims to identify whether a review refers to a "Hotel", "Restaurant", or "Attraction", we explored an alternative approach to Transformer-based models. Specifically, we implemented a combination of Word2Vec vector representations and a Multilayer Perceptron (MLP) classifier.

5.3.1. Multilayer Perceptron (MLP)

The input representation for the MLP consisted of vectors generated via Word2Vec embeddings with a dimensionality of 100. The neural network was structured with three hidden layers comprising 256, 128, and 64 neurons respectively ('hidden_layer_sizes=(256, 128, 64)'). The ReLU (Rectified Linear Unit) activation function was used for all hidden layers due to its computational efficiency and its

capacity to mitigate the vanishing gradient problem, enabling the model to learn complex, non-linear representations of the input data.

The MLP's output layer was designed for multi-class classification, predicting one of the three destination types. A Softmax activation was applied at the output to produce class probabilities, with the model trained using a log-loss (cross-entropy) function in combination with the Adam optimizer.

This architecture was selected as an efficient and effective solution for this specific task, especially when paired with dense input representations generated via Word2Vec. The MLP was trained to learn the underlying patterns in the document vectors and to accurately classify the type of destination.

5.3.2. Implementation and Training

The implementation and training process consisted of two main stages: generating vector representations using Word2Vec and training the MLP classifier.

The MLP was trained using the `scikit-learn` library, while Word2Vec embeddings were generated with Gensim. Data manipulation was handled using `pandas` and `NumPy`.

As noted earlier, the Adam optimizer was employed. The `MLPClassifier` in `scikit-learn` was used, which implicitly applies log-loss (cross-entropy) as the default loss function. The batch size was automatically set as `min(200, n_samples)`. The number of iterations (epochs) was capped at 300. L2 regularization was also applied with an `alpha` value of 1×10^{-4} to reduce overfitting.

The trained MLP was subsequently used to classify the test set review vectors, generating the predictions for the Type classification task.

Word2Vec Representation: The Word2Vec model was trained using the parameters listed in Table 3.

Table 3
Training parameters of the Word2Vec model.

Parameter	Value
vector_size	100
window	3
min_count	5
sg	1 (Skip-gram)
epochs	10

6. Results

This section presents the experimental results obtained. We report performance across the three subtasks: sentiment polarity classification, destination type identification, and recognition of specific *Pueblos Mágicos*. For each task, we provide detailed evaluation metrics—including precision, recall, and F1-score—on both the training and official test sets whenever applicable. The goal is to assess the generalization capabilities of our models under real-world conditions and to identify strengths and limitations in each approach.

6.1. Sentiment Polarity Results

Tables 4 and 5 show the classification performance of our model for the sentiment polarity task using 5 ordinal labels (from 1 to 5), on the training set and on the official test set, respectively.

On the training set, the model achieved a macro F1-score of 0.6953 and a weighted F1 of 0.7780, with particularly strong performance on the dominant class (label 5, F1 = 0.86), followed by class 1 (F1 = 0.79). In contrast, classes 2 and 4 had relatively lower F1-scores, likely due to class imbalance and label ambiguity in moderate sentiment expressions.

On the test set, we observed a moderate drop in overall performance. The macro F1-score decreased to 0.6057 and accuracy fell to 73.75%. Nonetheless, the classifier preserved consistent patterns, maintaining high precision and recall for class 5 (F1 = 0.85), and competitive results for class 1. However, performance on classes 2 and 4 remained weak, with class 2 notably achieving an F1-score of only 0.39, confirming the difficulty of distinguishing mid-range sentiment levels in real-world reviews.

These results highlight the effectiveness of the model in capturing polar sentiment extremes, while suggesting the need for specialized handling of ambiguous or neutral cases. Future improvements could explore ordinal classification loss functions or contrastive learning approaches tailored to sentiment gradation.

Table 4
Classification metrics for the `Polarity` task using the training set. The model distinguishes five sentiment levels (1 to 5) and exhibits strongest performance on class 5, which dominates the dataset.

Class	Precision	Recall	F1-score	Support
1	0.7195	0.8654	0.7857	1070
2	0.6489	0.5452	0.5926	1139
3	0.7182	0.5850	0.6448	3072
4	0.5257	0.6743	0.5908	9073
5	0.8975	0.8306	0.8628	27257
Accuracy			0.7715	41611
Macro avg	0.7020	0.7001	0.6953	41611
Weighted avg	0.7918	0.7715	0.7780	41611

Table 5
Evaluation metrics on the test set for the `Polarity` classification task. Despite class imbalance, the model performs well on extreme sentiment categories (1 and 5), with reduced performance in middle classes.

Class	Precision	Recall	F1-score
1	0.7438	0.6222	0.6776
2	0.3520	0.4290	0.3867
3	0.4871	0.6127	0.5427
4	0.6244	0.4844	0.5455
5	0.8188	0.8831	0.8497
Macro avg	0.6052	0.6063	0.6057
Accuracy			0.7375

6.2. Destination Type Results

Tables 6 and 7 present the classification performance for the `Type` prediction task, using only the training set and the complete dataset (train + test), respectively. As described in Section *Methodology*, this task was tackled with a Multilayer Perceptron (MLP) classifier fed with Word2Vec vector representations.

On the training set, the MLP achieved an overall accuracy of 96% and a macro F1-score of 0.95. The performance across all three destination types—*Attractive*, *Hotel*, and *Restaurant*—was consistently strong, with F1-scores ranging from 0.94 to 0.97. This indicates that the model effectively captured the semantic patterns associated with each category from the vectorized representations.

When evaluated on the complete dataset, which includes the test set, the model maintained high generalization performance. It achieved a macro F1-score of 0.9437 and an accuracy of 94.6%. Notably, the classes *Attractive* and *Restaurant* obtained the highest F1-scores (0.9565 and 0.9500, respectively), confirming the robustness of the model even on unseen data.

These results demonstrate the effectiveness of the Word2Vec + MLP pipeline in distinguishing between types of tourism destinations. The high precision, recall, and F1-scores in both training and test

scenarios highlight the suitability of dense word embeddings combined with deep learning architectures for fine-grained text classification tasks in the tourism domain.

Table 6
Classification performance for the Type prediction task using the training set. The model was trained with Word2Vec embeddings and an MLP classifier, achieving high accuracy and balanced performance across the three classes.

Class	Precision	Recall	F1-score	Support
Attractive	0.97	0.96	0.97	14000
Hotel	0.95	0.92	0.94	10305
Restaurant	0.95	0.97	0.96	17306
Accuracy			0.96	41611
Macro avg	0.96	0.95	0.95	41611
Weighted avg	0.96	0.96	0.96	41611

Table 7
Classification results for the Type prediction task using the official train/test split. The MLP model maintains robust generalization performance across all destination types.

Class	Precision	Recall	F1-score
Attractive	0.9509	0.9623	0.9565
Hotel	0.9218	0.9275	0.9246
Restaurant	0.9562	0.9438	0.9500
Macro avg	0.9430	0.9445	0.9437
Accuracy			0.9460

6.3. Magical Town Identification Results

Tables 8 and 9 present the classification results for the Town prediction task on the training set and the official train/test split, respectively. This task posed a substantial challenge due to the high number of classes (up to 40 different *Pueblos Mágicos*) and severe class imbalance.

In the training set, the model achieved a macro F1-score of 0.70 and a weighted F1 of 0.74, with notable performance on well-represented towns such as *Tulum* (F1 = 0.84), *Teotihuacan* (F1 = 0.84), and *San Cristóbal de las Casas* (F1 = 0.72). Some towns with fewer training examples (e.g., *Tapalpa*, *Dolores Hidalgo*) showed moderate to low F1-scores, indicating data scarcity as a limiting factor.

When evaluated on the test set (Table 9), the model obtained a macro F1-score of 0.5958. Although performance dropped relative to the training set, the classifier still performed competitively for several towns. For instance, *Tulum*, *Teotihuacan*, and *Chiapa de Corzo* maintained strong F1-scores above 0.75, showcasing the model's ability to generalize well in high-frequency cases. However, low recall values in underrepresented classes highlight the difficulty in managing class imbalance in real-world deployment.

Overall, these results suggest that while the model generalizes reasonably well for dominant towns, future work should consider strategies such as data augmentation or cost-sensitive learning to improve performance in minority classes.

Table 8
Classification performance for the Town prediction task on the training set. The table reports precision, recall, and F1-score for each Pueblo Mágico, highlighting the model's ability to distinguish between numerous classes in a highly imbalanced dataset.

Town	Precision	Recall	F1-score	Support
Sayulita	0.59	0.76	0.67	1467
Tulum	0.91	0.78	0.84	9069
Isla Mujeres	0.89	0.72	0.80	5965
Patzcuaro	0.70	0.67	0.69	891
Palenque	0.73	0.75	0.74	1903
Valle de Bravo	0.76	0.63	0.69	1192
Ixtapan de la Sal	0.49	0.75	0.59	339
Creel	0.65	0.81	0.72	357
Taxco	0.78	0.77	0.78	840
Valladolid	0.69	0.67	0.68	2328
Izamal	0.61	0.79	0.69	408
San Cristobal de las Casas	0.75	0.68	0.72	2612
Atlixco	0.67	0.69	0.68	289
Tequisquiapan	0.60	0.61	0.61	725
Ajijic	0.62	0.78	0.69	750
Teotihuacan	0.82	0.86	0.84	1162
Tequila	0.63	0.73	0.68	530
Bacalar	0.70	0.72	0.71	2165
TodosSantos	0.67	0.72	0.69	920
Parras	0.50	0.77	0.61	191
Coatepec	0.77	0.77	0.77	164
Huasca de Ocampo	0.74	0.86	0.80	302
Tepoztlan	0.61	0.74	0.67	689
Cholula	0.68	0.79	0.73	558
Cuatro Cienegas	0.72	0.82	0.77	158
Metepec	0.57	0.70	0.63	706
Loreto	0.71	0.74	0.72	1105
Orizaba	0.61	0.77	0.68	504
Tlaquepaque	0.55	0.72	0.62	808
Cuetzalan	0.56	0.71	0.63	199
Bernal	0.55	0.73	0.63	250
Xilitla	0.79	0.85	0.82	292
Malinalco	0.64	0.65	0.65	286
Real de Catorce	0.60	0.78	0.68	152
Chiapa de Corzo	0.71	0.85	0.78	192
Mazunte	0.54	0.79	0.64	293
Tepotzotlan	0.54	0.78	0.64	203
Zacatlan	0.69	0.71	0.70	320
Dolores Hidalgo	0.55	0.71	0.62	182
Tapalpa	0.51	0.76	0.61	145
accuracy			0.74	41611
macro avg	0.66	0.75	0.70	41611
weighted avg	0.76	0.74	0.74	41611

Table 9
Classification performance for the Town prediction task on the test set. The results reflect the model's generalization capacity across 40 different Pueblos Mágicos, based on unseen data.

Town	Precision	Recall	F1-score
Tulum	0.7820	0.9053	0.8391
Isla Mujeres	0.7209	0.8819	0.7933
San Cristóbal de las Casas	0.6207	0.7221	0.6676
Valladolid	0.6611	0.6829	0.6718
Bacalar	0.6785	0.6560	0.6671
Palenque	0.6926	0.7095	0.7009
Sayulita	0.7014	0.6126	0.6540
Valle de Bravo	0.5431	0.5816	0.5617
Teotihuacan	0.8141	0.7646	0.7886
Loreto	0.7291	0.6799	0.7037
TodosSantos	0.6800	0.5864	0.6297
Patzcuaro	0.6468	0.6306	0.6386
Taxco	0.6752	0.6941	0.6845
Tlaquepaque	0.5964	0.5061	0.5476
Ajijic	0.6636	0.5463	0.5993
Tequisquiapan	0.4929	0.4265	0.4573
Metepec	0.6290	0.4601	0.5314
Tepoztlán	0.6378	0.5170	0.5711
Cholula	0.6611	0.5441	0.5969
Tequila	0.6326	0.5452	0.5856
Orizaba	0.6098	0.4955	0.5467
Izamal	0.6857	0.5703	0.6227
Creel	0.6235	0.5048	0.5579
Ixtapan de la Sal	0.5983	0.3488	0.4407
Zacatlán	0.5662	0.5867	0.5763
Huasca de Ocampo	0.6749	0.6030	0.6370
Mazunte	0.6672	0.4302	0.5231
Xilitla	0.7104	0.6597	0.6841
Atlixco	0.6003	0.5315	0.5638
Malinalco	0.5588	0.4554	0.5018
Bernal	0.5624	0.4048	0.4708
Tepotzotlán	0.5253	0.4057	0.4578
Cuetzalan	0.5105	0.4449	0.4755
Chiapa de Corzo	0.7372	0.6461	0.6886
Parras	0.5917	0.4130	0.4864
Dolores Hidalgo	0.5064	0.4529	0.4782
Coatepec	0.5714	0.5747	0.5731
Cuatro Ciénegas	0.5651	0.5107	0.5365
Real de Catorce	0.5754	0.4783	0.5223
Tapalpa	0.5516	0.3977	0.4622
Macro avg (Town)	0.6313	0.5642	0.5958

7. Conclusions and Future Work

This paper has presented the approach of the CIMAT CC team to the REST-MEX 2025 shared task, where we addressed three distinct NLP challenges within the domain of Mexican tourism: sentiment polarity classification, destination type identification, and *Pueblo Mágico* recognition. Our participation demonstrated the value of combining state-of-the-art Transformer architectures with well-established classical models, selecting the appropriate methodology according to the complexity and structure of each subtask.

For sentiment polarity classification, the fine-tuned BERT model achieved a macro F1-score of 0.6057 and an accuracy of 73.75% on the test set. These results confirmed the model's strength in capturing semantic subtleties, particularly at the extremes of the sentiment scale. However, performance remained lower for mid-scale values, where expressions of sentiment tend to be more ambiguous. This observation suggests the need for models that explicitly account for ordinal relationships between classes.

In contrast, the destination type classification task was tackled using a Multilayer Perceptron (MLP) fed with Word2Vec embeddings. This classical approach yielded a macro F1-score of 0.9437 and an accuracy of 94.6%, indicating that dense vector representations coupled with a well-configured neural classifier remain highly effective in tasks with balanced data and clearly separable class boundaries.

For the *Pueblo Mágico* identification task, a BERT-based model was again employed. Despite the challenge posed by a high number of classes and strong class imbalance, the model attained a macro F1-score of 0.5958. Performance was highest for towns with greater representation in the dataset (e.g., "Tulum", "Teotihuacan"), but declined for those with few training instances. This underscores the difficulty of high-cardinality classification under data scarcity.

Overall, our results suggest that while BERT models offer robust solutions for semantically rich tasks, classical models such as MLPs can outperform on more structured, low-ambiguity tasks. The hybrid strategy employed here provides a flexible and effective framework for addressing heterogeneous NLP problems in domain-specific contexts.

Future work will focus on addressing class imbalance in the Town classification task, potentially through advanced data augmentation or loss re-weighting techniques such as focal loss. In the polarity classification task, incorporating ordinal regression or contrastive learning may help improve accuracy on intermediate sentiment levels. Finally, multi-task learning approaches that simultaneously optimize for all three tasks may offer benefits through shared representation learning, enhancing generalization across related subtasks in the tourism review domain.

Acknowledgments

We thank the organizers of REST-MEX 2025 and IberLEF 2025 for coordinating the task and providing the dataset. We also highlight that undertaking this challenge afforded us a pleasant experience and valuable learning.

Declaration on Generative AI

During the preparation of this work, we made limited use of generative AI tools—namely ChatGPT (OpenAI) and Gemini (Google)—for general text drafting, language refinement, and code assistance. All AI-generated material was subsequently reviewed, edited, and validated by us, and we take full responsibility for the final content.

References

[1] D. Jurafsky, J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2023. Draft version, available online: https://web.stanford.edu/~jurafsky/slp3/.

[2] B. Liu, Sentiment Analysis and Opinion Mining, volume 5 of *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool, 2012.
[3] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5 (2014) 1093–1113. doi:10.1016/j.asej.2014.04.011.
[4] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-Gonzalez, D. Fajardo-Delgado, M. G. Sánchez, H. Pérez-Espinosa, J. Martínez-Miranda, R. Guerrero-Rodríguez, L. Bustio-Martínez, Ángel Díaz-Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University - Computer and Information Sciences 34 (2022) 10125–10144. URL: https://www.sciencedirect.com/science/article/pii/S1319157822003615. doi:https://doi.org/10.1016/j.jksuci.2022.10.010.
[5] J. Cañete, G. Chaperón, R. Fuentes, J. Pérez, C. Bizarreta, Spanish pre-trained bert model and evaluation data, in: Proceedings of the PML4DC Workshop at ICLR 2020, 2020.
[6] A. Gutiérrez-Fandiño, M. G. Pino, R. Pérez, et al., Maria: Spanish language models, corpora and benchmark, Procesamiento del Lenguaje Natural 68 (2022) 39–60.
[7] A. Diaz-Pacheco, M. A. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, Journal of Experimental & Theoretical Artificial Intelligence 0 (2022) 1–31. doi:10.1080/0952813X.2022.2153276.
[8] S. de Turismo de México, Informe de resultados del sector turismo 2023, https://www.gob.mx/sectur/documentos/informe-turismo-2023, 2024. Accessed: 3 Jun 2025.
[9] S. de Turismo de México, Programa *Pueblos Mágicos*: Listado oficial de localidades 2025, https://www.gob.mx/sectur/acciones-y-programas/pueblos-magicos, 2025. Accedido: 3 jun 2025.
[10] 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.
[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. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186.
[13] E. Martínez-Cámara, J. Villena-Román, F. García-Sánchez, et al., Overview of tass 2020: Sentiment analysis and emotion detection in spanish, in: Proceedings of the SEPLN 2020 Workshop on TASS, 2020, pp. 13–27.
[14] M.-A. Álvarez Carmona, et al., Rest-mex 2021: Resources and evaluations for tourism sentiment analysis in spanish, in: Proceedings of IberLEF 2021, 2021, pp. 638–651.
[15] M.-A. Álvarez Carmona, V. Guzmán-Flores, L. González-Gurrola, G. Bel-Enguix, Overview of the rest-mex 2022 shared task on sentiment and tourism text classification, in: Proceedings of IberLEF 2022, 2022, pp. 684–703.
[16] V. Guzmán-Flores, M.-A. Álvarez Carmona, et al., Rest-mex 2023: Adding country identification to tourism sentiment tasks, in: Proceedings of IberLEF 2023, 2023, pp. 712–728.
[17] G. Guerrero-Rodríguez, K. Hernández-Figueroa, A. Sandoval-Sánchez, Combining ensembles for fine-grained sentiment analysis of tripadvisor reviews, in: Proceedings of the TASS Workshop at SEPLN 2021, 2021, pp. 123–132.
[18] L. García-Villalba, E. Martínez-Cámara, Umutextstats at rest-mex 2022: Combining linguistic statistics with bert for tourism sentiment tasks, in: Proceedings of IberLEF 2022, 2022, pp. 704–714.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 5998–6008.
[20] 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. URL: http://dx.doi.org/10.1016/j.jksuci.2023.101746. doi:10.1016/j.jksuci.2023.101746.

[21] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. https://www.deeplearningbook.org.