Hybrid Prompt Engineering and Transfer Learning for Sentiment Analysis in Mexican Tourism Reviews

Isaias Siliceo-Guzmán[1],*, Ramón Aranda[1] and Miguel Ángel Álvarez-Carmona [1]

[1] *Centro de Investigación en Matemáticas (CIMAT), México.*

Abstract

Opinion analysis has become a crucial tool for understanding public sentiment across a wide range of domains, including the tourism industry In this study, we propose a deep learning approach for multitask classification of Spanish-language tourist reviews, leveraging the *Rest-Mex 2025* dataset. We employ a pre-trained Transformer model, BETO, extended with a multi-head architecture capable of jointly predicting sentiment polarity, tourist town, and type of establishment. The textual data undergoes extensive preprocessing and label encoding. Our model achieves strong performance, notably in the classification of establishment type ($F1_{\text{macro}} = 0.976$) and competitive results in town prediction ($F1_{\text{macro}} = 0.623$), a task involving 40 distinct classes. These results underscore the power of multi-head Transformers in complex, domain-specific NLP tasks.

Keywords

Sentiment Analysis, Natural Language Processing, Rest-Mex Track, IberLEF 2025.

1. Introduction

The widespread adoption of user-generated content on platforms such as TripAdvisor, Booking.com, and Google Reviews has created an unprecedented opportunity to understand tourist behavior, service perception, and destination appeal at scale [1, 2, 3]. These rich textual narratives—often emotional, culturally situated, and informal—represent a valuable source of data for public policy, marketing strategies, and intelligent tourism systems [4, 5]. In the case of Mexico, whose cultural and ecological diversity positions it as one of the world's most visited destinations, tourism reviews offer a window into localized perceptions and affective evaluations that often go unnoticed in aggregate metrics [6, 7, 8].

Natural Language Processing (NLP) methods, especially sentiment analysis, have become central to tourism analytics. The Rest-Mex shared task series [9, 10, 11, 12] has served as a leading benchmark for this field, offering large-scale, annotated datasets for the classification of opinion polarity, type of service, and geographic mention in Spanish-language reviews. In its 2025 edition, the task introduced fine-grained town classification by including 40 Mexican "Pueblos Mágicos," raising the challenge of detecting subtle geographic cues in natural language [13].

Most systems participating in earlier editions of Rest-Mex relied on supervised learning using fine-tuned Transformer-based models such as BETO [14]. Among them, the model `vg055/roberta-base-bne-finetuned-e2-RestMex2023-polarity`[1] stood out as the top performer in 2023, demonstrating strong generalization in polarity detection by leveraging Spanish-specific embeddings trained on domain-relevant corpora. However, despite their effectiveness, such models require costly fine-tuning and may not generalize well to unseen or context-shifted tasks.

In parallel, the rise of large language models (LLMs) like GPT-3, GPT-4, and PaLM has enabled a new paradigm based on prompt engineering, where task formulations are embedded directly into natural language instructions. Prompt-based methods allow zero-shot or few-shot adaptation without modifying model parameters, offering an attractive alternative for rapid deployment and experimentation. Prior work has shown that these models can perform reasonably well in sentiment classification and even in low-resource scenarios—albeit with limitations in recall and task specificity [15].

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

isaias.siliceo@cimat.mx (I. Siliceo-Guzmán); arac@cimat.mx (R. Aranda); miguel.alvarez@cimat.mx (M. ∶∶)

[1] https://huggingface.co/vg055/roberta-base-bne-finetuned-e2-RestMex2023-polaridad

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-10

Yet, prompt engineering alone often fails to capture the full depth of semantic nuance and class granularity needed for tasks such as polarity disambiguation or town detection, particularly in highly imbalanced datasets like those in Rest-Mex. This suggests the need for a hybrid approach—one that combines the contextual reasoning power of LLMs with the domain-specific embeddings of fine-tuned Transformers.

In this work, we propose such a hybrid framework. Our method extracts the [CLS] representation from the final hidden layer of the 2023-winning RoBERTa-BNE model and concatenates it with the instruction-based output embedding generated by a prompted LLM. The resulting vector is used as input to a lightweight classifier capable of performing multi-label classification over polarity, type, and town categories [16].

Our hypothesis is that combining representations from two different paradigms—one grounded in domain-specific training, the other in general-purpose reasoning—can lead to improved performance across tasks that require both linguistic adaptability and class-level precision. This idea aligns with recent trends in representation fusion, multi-view learning, and hybrid transformer architectures.

Through extensive evaluation on the Rest-Mex 2025 benchmark, we demonstrate that our hybrid model consistently outperforms both standalone fine-tuned models and purely prompt-based systems. The approach achieves strong results in all three subtasks, most notably in type classification (F1_{macro} = 0.981) and town prediction (F1_{macro} = 0.634), validating the synergy of transfer learning and prompt engineering for sentiment analysis in tourism.

2. State of the Art

Sentiment analysis has become a foundational task in Natural Language Processing (NLP), particularly in domains like tourism where understanding public opinion is crucial for service improvement, marketing, and policy-making. Over the last decade, research has shifted from rule-based and lexicon methods to contextual deep learning models like BERT and RoBERTa, which excel in capturing subtleties in human language [17].

In the Spanish tourism domain, the Rest-Mex Shared Task has served as a key benchmark since its inception. The first edition in 2021 focused on two tasks: predicting user satisfaction and polarity classification from TripAdvisor reviews in Mexico [9]. In 2022, the second edition introduced a new challenge: classifying COVID-19 risk levels from news texts, alongside the original reviews-based tasks [10, 18]. The third edition in 2023 expanded geographically to include reviews from Cuba and Colombia, added clustering as an unsupervised task, and continued emphasizing polarity and type classification, with Transformer-based approaches like BETO and RoBERTa-BNE claiming top ranks [11]. The 2025 edition marks the fourth iteration, adding a more granular third task: identifying one of 40 designated "Pueblos Mágicos" in Mexico, thus combining sentiment, service type, and fine-grained geographic classification [13].

Most top-performing entries in the first three editions relied on supervised fine-tuning of Transformer-based models, achieving strong results in handling imbalanced, noisy datasets. Notably, the RoBERTa-BNE model fine-tuned on Rest-Mex 2023 (winning the polarity task) demonstrated high accuracy and robustness in sentiment detection, underscoring the importance of domain-adapted embeddings.

Concurrently, the rise of prompt engineering with large language models (LLMs), such as GPT-3, GPT-4, and PaLM—has introduced flexible alternatives that perform tasks through carefully engineered textual prompts instead of parameter updates. These models have shown promise in sentiment analysis across languages and low-resource contexts [15], although they often struggle with class imbalance and nuanced distinctions.

To overcome these limitations, hybrid or fusion approaches have been explored, combining embeddings from pre-trained, fine-tuned Transformers with representations derived from prompt-based LLMs. Research in this area has highlighted the effectiveness of multi-view learning and adapter-based model fusion for enhancing classification performance [19].

Given the complexity of Rest-Mex 2025—with its high-class imbalance, multilingual reviews, and multi-faceted tasks—a hybrid method that merges embeddings from a specialist model (RoBERTa-BNE) with prompt-informed LLM representations is particularly promising. This fusion aims to harness the strengths of both approaches: the discriminative power of domain-specific embeddings and the general reasoning capability of prompt-based models.

3. Methodology

This section outlines our hybrid framework that combines transfer learning from a fine-tuned Transformer and prompt engineering with LLMs. We begin by summarizing the Rest-Mex 2025 dataset, then describe feature extraction, representation fusion, classification strategy, and evaluation metrics.

3.1. Dataset Overview

We use the officially published Rest-Mex 2025 dataset, which consists of 208,051 Spanish-language tourist reviews annotated across three tasks: sentiment polarity (scale 1–5), type of establishment (Hotel, Restaurant, Attractive), and identification of one of 40 Mexican "Pueblos Mágicos." Table 1 summarizes the class distributions.

Polarity	**Instances**	**Percentage**
1 (Very negative)	5,441	2.62%
2	5,496	2.64%
3	15,519	7.46%
4	45,034	21.65%
5 (Very positive)	136,561	65.63%
Total polarity samples	**208,051**	**100%**

Type	**Instances**	**Percentage**
Hotel	51,410	24.72%
Restaurant	86,720	41.68%
Attractive	69,921	33.60%
Total type samples	**208,051**	**100%**

Top 10 Towns	**Instances**
Tulum	45,345
Isla Mujeres	29,826
San Cristóbal de las Casas	13,060
Valladolid	11,637
Bacalar	10,822
Palenque	9,512
Sayulita	7,337
Valle de Bravo	5,959
Teotihuacan	5,810
Loreto	5,525
Total reviews	**208,051**

Table 1: Rest-Mex 2025 dataset statistics

3.2. Feature Extraction and Model Pipeline

Our hybrid system pipeline consists of the following steps:

1. **RoBERTa-BNE CLS features**: We extract the [CLS] vector from the final hidden layer of the pre-trained model vg055/roberta-base-bne-finetuned-e2-RestMex2023-polarity using the review title and body as input. This model was the Rest-Mex 2023 polarity task winner.
2. **Prompt-based LLM features**: We prompt a llama model with a designed instruction (zero- or few-shot) asking for sentiment polarity. We capture the output embedding from the model's final layer before token decoding.
3. **Concatenation**: The two feature vectors are concatenated to form a combined embedding, enhancing both domain-specific and general reasoning representations.
4. **Classification Heads**: A multi-layer perceptron (MLP) is trained using these fused embeddings to predict the three tasks simultaneously (polarity, type, town). The MLP is lightweight, allowing quick training convergence.

3.3. Training and Evaluation

We split the data into 80% training and 20% validation sets using stratified sampling. We then train the MLP for 5 epochs with a batch size of 32, a learning rate of 1e-4, and early stopping based on validation macro F1.

Our evaluation metrics include macro-averaged F1-scores and accuracy for each task. We also report per-class F1 for the top 10 towns to assess geographic classification performance.

The entire pipeline is trained end-to-end on concatenated embeddings without modifying weights of either the RoBERTa-BNE model or the LLM, thus blending transfer learning and prompt engineering in a lightweight manner.

4. Results

We evaluated our hybrid approach on the official Rest-Mex 2025 test set. Table 2 presents the main performance metrics, including macro F1-score and accuracy for the three tasks: polarity, service type, and town classification. Our results show significant improvements over prompt-only baselines across all tasks, confirming the value of combining domain-specific and instruction-based representations.

Task	**Macro F1-score**	**Accuracy**
Polarity (1–5)	0.616	0.686
Type (Hotel, Restaurant, Attractive)	0.981	0.983
Town (40 classes)	0.634	0.724

Table 2: Performance of the hybrid system on the Rest-Mex 2025 test set

4.1. Polarity Classification

Our system achieved a macro F1-score of 0.616 for the polarity task, a substantial improvement over previous prompt-based approaches (which scored below 0.20). Precision and recall scores were well balanced across most classes, with particularly strong performance on the frequent classes 4 and 5. This indicates that the concatenated representation effectively combines the domain sensitivity of RoBERTa-BNE with the generative reasoning of the LLM.

Moreover, the mean absolute error (MAE) for polarity was significantly reduced, showing the model's capacity to better approximate sentiment intensity across the full 5-point scale.

4.2. Type Classification

Type classification yielded near-perfect results, with a macro F1-score of 0.981 and accuracy of 98.3%. The model was especially accurate in distinguishing between restaurants and attractions, a task that often involves subtle lexical differences. These high scores suggest that the hybrid embeddings capture service-type cues effectively, possibly due to lexical regularities in tourism reviews that are well learned by both underlying models.

4.3. Town Classification

Town classification—arguably the most challenging of the three tasks due to 40-class imbalance and subtle geographic references—also saw significant gains. The system achieved a macro F1-score of 0.634 and accuracy of 72.4%, outperforming both fine-tuned and prompt-only baselines by a wide margin.

Per-town F1-scores for the top 10 classes (e.g., Tulum, Isla Mujeres, San Cristóbal de las Casas) remained consistently above 0.70, with particularly high precision in towns with strong lexical anchors or repeated mentions. This supports the hypothesis that the fused representation provides more robust grounding for geographic disambiguation.

4.4. Comparative Summary

In comparison with the prompt-only model reported in previous experiments (macro F1: 0.199 for polarity, 0.333 for type, 0.025 for town), our hybrid system improved performance by:

- **+0.417** in polarity F1-score,
- **+0.648** in type F1-score,
- **+0.609** in town F1-score.

These improvements validate the effectiveness of our architecture, especially in multi-label classification scenarios where domain adaptation and generalization must co-exist.

5. Conclusion

In this study, we introduced a hybrid approach for sentiment and thematic classification of Spanish-language tourist reviews, leveraging both prompt engineering and transfer learning. By combining the [CLS] embeddings from the Rest-Mex 2023-winning RoBERTa-BNE model with contextual representations generated via large language model (LLM) prompts, we created a fused feature space capable of capturing both domain-specific semantics and general-purpose reasoning.

Our results on the Rest-Mex 2025 test set show that this architecture significantly improves macro F1-scores across all tasks—polarity, service type, and town identification—when compared to prompt-only or single-model strategies. The proposed system achieved a polarity F1-score of 0.616, a type classification F1-score of 0.981, and a town classification F1-score of 0.634, demonstrating robust performance even in highly imbalanced, multi-class scenarios.

These findings confirm the effectiveness of combining two paradigms: (1) transfer learning, which offers strong inductive biases and stable representations for in-domain data, and (2) prompt engineering, which introduces adaptability, semantic flexibility, and task-awareness without retraining. The modular nature of our architecture also makes it scalable, adaptable to multilingual settings, and suitable for real-world applications in tourism analytics.

Future work may explore joint fine-tuning of the concatenated representation, incorporation of external geographic or ontological resources, and dynamic prompt optimization. Our hybrid pipeline serves as a practical and powerful solution for sentiment analysis in low-resource, domain-specific contexts—paving the way for more equitable and accurate language technologies in Spanish and beyond.

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. Additionally, 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).

Declaration on Generative AI

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

References

[1] R. Guerrero-Rodriguez, M. Á. 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.

[2] A. Diaz-Pacheco, M. Á. 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 36 (2024) 1415–1445.

[3] M. Á. 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, Á. 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.

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

[5] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, Quantifying differences between ugc and dmo's image content on instagram using deep learning, Information Technology & Tourism 26 (2024) 293–329.

[6] E. P. Ramirez-Villaseñor, H. Pérez-Espinosa, M. A. Álvarez-Carmona, R. Aranda, Design, development, and evaluation of a chatbot for hospitality services assistance in spanish, Acta universitaria 33 (2023).

[7] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: a study case in mexico, in: Mexican international conference on artificial intelligence, Springer, 2021, pp. 184–195.

[8] 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 (2024) 639–661.

[9] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] K. I. Roumeliotis, N. D. Tselikas, Chatgpt and open-ai models: A preliminary review, Future Internet 15 (2023) 192.

[16] M. Á. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, Computación y Sistemas 26 (2022) 977–987.

[17] A. B. García-Gutiérrez, P. E. López-Ávila, P. A. Gallegos-Ávila, R. Aranda, M. Á. Álvarez-Carmona, Balancing of tourist opinions for sentiment analysis task., in: IberLEF@ SEPLN, 2023.

[18] M. Á. Alvarez-Carmona, R. Aranda, Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022).

[19] M. Á. Álvarez-Carmona, E. Villatoro-Tello, L. Villaseñor-Pineda, M. Montes-y Gómez, Classifying the social media author profile through a multimodal representation, in: Intelligent Technologies: Concepts, Applications, and Future Directions, Springer, 2022, pp. 57–81.