LPQ Team at Rest-Mex 2025: BERT and LLM Approaches in Tourism Review Classification

Le Phu Quy[1,2,*], Dang Van Thin[1,2]

[1] *University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

This study addresses the Rest-Mex 2025 challenge by developing a multi-task framework for Spanish tourism review analysis, focusing on sentiment polarity (1–5 scale), destination type classification (hotel/restaurant/attraction), and Magical Town identification. We explore transformer-based models (BETO, XLM-RoBERTa), hybrid architectures (BERT embeddings with XGBoost), domain adaptation, ensemble strategies, and LLaMA-3 fine-tuned specifically for Magical Town recognition. The approach provides a scalable pipeline for enhancing destination analytics through advanced NLP techniques.

Keywords

Hope classification, Spanish language, English language, sentiment analysis, fine-tuning BERT

1. Introduction

Mexico's Pueblos Mágicos (Magical Towns) are vibrant destinations where history, culture, and economic development intersect, attracting millions of travelers each year. These towns are celebrated for their rich artisan traditions, historic landmarks, and breathtaking natural scenery, making them essential to Mexico's tourism industry. As digital platforms like TripAdvisor and social media grow in influence, travelers share their experiences more widely than ever, creating a wealth of reviews that reveal diverse sentiments, destination preferences, and regional identities. However, analyzing these texts poses challenges, as they are written in various Spanish dialects and often include irony or local slang[1, 2, 3, 4].

Unlike past editions [5, 6, 7], the Rest-Mex 2025 [8, 9] shared task addresses these complexities by introducing three key objectives. First, sentiment analysis helps determine the emotional tone of a review, using a rating scale from 1 to 5. Second, destination classification identifies whether a review refers to a hotel, restaurant, or tourist attraction. Lastly, geolocation detection pinpoints which Pueblo Mágico the review describes. These tasks extend beyond academic interest—they play a crucial role in shaping sustainable tourism strategies, improving infrastructure, and preserving the unique cultural heritage of Mexico's 40 designated Pueblos Mágicos. By extracting meaningful insights from online reviews, researchers and policymakers can enhance visitor experiences while ensuring these towns retain their distinctive charm for future generations.

In this work, we present an evaluation of modeling strategies for tourism analytics, benchmarking several approaches. First, we optimize baseline BERT models—including BETO (Spanish BERT), Roberto (RoBERTa Spanish), and XLM-RoBERTa—to serve as our fundamental framework. Additionally, we explore embedding-XGBoost hybrids, where sentence embeddings derived from these BERT variants are fed into XGBoost classifiers fine-tuned with focal loss to better capture minority classes. To further enhance destination classification, we introduce a domain-adapted BETO model, trained on 15GB of Mexican tourism texts to effectively capture region-specific expressions. We also employ BERT ensemble models by fine-tuning BETO, Roberto, and XLM-RoBERTa independently and aggregating their outputs

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

23521316@gm.uit.edu.vn (L. P. Quy); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0000-0001-8340-1405 (D. V. Thin)

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

via both soft and hard voting, supplemented by metadata features such as regional keywords. Finally, we fine-tune LLaMA-3 using LoRA, further enriching our approach.

2. Related Work

Tourism review analytics has evolved significantly over the years, moving from early rule-based[10] and traditional machine learning methods to sophisticated deep learning approaches. Early work in this field focused on manual feature engineering for sentiment extraction and destination classification, but these methods struggled with the variability and cultural nuances present in user-generated content. The advent of transformer-based models, especially Spanish-specific variants like BETO[11] and Roberto has greatly enhanced our ability to capture complex expressions and regional idioms in tourism reviews.

More recent studies have leveraged hybrid architectures that combine the semantic strength of transformer models with the robustness of classical classifiers. In particular, embedding-XGBoost hybrids—where sentence embeddings from BETO, Roberto, and XLM-RoBERTa are fed into XGBoost classifiers fine-tuned[12] with focal loss have been successful in addressing challenges related to class imbalance and minority class emphasis[13]. Additionally, domain adaptation[14] via pre-training on extensive tourism-specific corpora has proven effective for capturing regional expressions, enhancing the performance of models in destination classification tasks.

Ensemble methods[15] and large language models (LLMs) have further pushed the boundaries in tourism review analysis. Independent fine-tuning of various BERT variants followed by ensemble aggregation has demonstrated notable improvements in tasks like Magical Town identification. Moreover, fine-tuning LLaMA-3 using parameter-efficient methods like LoRA on culturally annotated datasets has enabled more precise disambiguation of similar regional references. Our work builds on these independent modeling strategies, offering a modular approach that isolates and leverages the unique strengths of each method to achieve state-of-the-art performance in tourism NLP.

3. Methodology

3.1. Transformer-Based Classification Models

The Transformer architecture has become a cornerstone in Natural Language Processing due to its effective use of self-attention mechanisms. These mechanisms allow the model to capture the contextual relationships between tokens, while positional encoding preserves the sequential order. Multi-head attention further enables the parallel extraction of distinct patterns and representations from the input text. Such principles form the theoretical basis of our classification approach.

Implemented Model Variants: In our experiments, we employ four Transformer-based models:

- **BETO**: A Spanish language model adapted from BERT, which utilizes dynamic gradient accumulation to address instability in gradient updates during training.
- **Roberto**: A variant similar in foundation to BETO, optimized for our specific domain requirements.
- **XLM-RoBERTa**: A robust multilingual model fine-tuned using conventional truncation settings to handle regional vocabulary and context.
- **Domain-Adapted BETO**: This model undergoes an additional phase of continual pre-training on 18 million tokens from the hospitality domain, enhancing its understanding of tourism-related texts.

3.2. Hybrid Embedding-XGBoost Framework

In our approach, the process is divided into two main phases: extracting information from text and then using that information to make predictions. First, a transformer model transforms the raw text into a vector, which serves as a semantic summary that efficiently captures the context and meaning of the original content. This vector eliminates the need for continuous, heavy computation in the subsequent stages. For sentiment analysis, the extracted embedding is passed to an XGBoost model designed to predict ratings while naturally respecting their ordinal relationship. This means that the model is set up so that a higher rating is always treated as more positive than a lower one, ensuring that the predictions follow the expected order and are consistent with the natural ranking of sentiments.

For town or geographical classification, a similar XGBoost model is employed, but it is fine-tuned to focus on features that are most relevant to location information. By analyzing which parts of the text provide the strongest geographic signals, the model filters out less useful features, which simplifies the decision process and improves overall efficiency. This selective approach ensures that the classification is both fast and accurate. To further improve the performance of the system, we apply Bayesian hyperparameter optimization. This method carefully adjusts key parameters such as model complexity and regularization factors, helping to balance the trade-offs between accuracy and speed while also addressing potential class imbalances in the data. Overall, the hybrid framework not only separates the heavy lifting of context extraction from the prediction tasks but also achieves faster inference times compared to using a complete transformer model for every operation.

3.3. LLaMA-3 Instruction Tuning

In this approach, we adjust the LLaMA-3-8B model to enhance its understanding of geocultural contexts in tourism. We use a method called Low-Rank Adaptation (LoRA), which adds a small number of trainable components to specific layers of the model while keeping most of the original parameters unchanged. This allows the model to learn tourism-focused information without losing its general language abilities.

To incorporate tourism domain expertise, trainable adapters are inserted into the model's query and value projection layers. This selective tuning enables the model to effectively absorb and use domain knowledge while still relying on its robust pre-trained capabilities. A well-structured, three-part prompt strategy guides the model's response. To ensure the geographical information is accurate, two validation methods are applied. First, the system employs pattern matching to filter out any inputs that do not meet the expected format for town names. Second, a character-level similarity check is used as a fallback to correct minor errors in the town names by comparing them against an official list. This dual-check approach minimizes errors in geographic details and ensures the output remains precise.

Overall, this instruction tuning framework adapts the LLaMA-3 model to be both knowledgeable and reliable within the tourism domain. By combining targeted parameter tuning with a structured prompting and validation system, the model is capable of generating detailed, accurate responses while maintaining efficiency—a quality that is essential for deployment in resource-constrained environments.

3.4. Ensemble Strategies

Our ensemble approach is centered on two key voting techniques: soft voting and hard voting, each contributing uniquely to the final decision-making process.

In soft voting, the models provide probabilistic estimates that reflect the confidence of each prediction. These probabilities are combined in a way that gives higher influence to models with stronger performance. This method allows the ensemble to capture subtle distinctions in the data, effectively leveraging the context-aware abilities of transformer-based models when the differences between classes are not pronounced.

In contrast, hard voting involves each model casting a clear, discrete vote for its predicted outcome. The final prediction is determined by a majority rule—if the votes are tied, a simple tie-breaking procedure selects the outcome. This approach provides decisiveness and transparency, ensuring that the ensemble can deliver a clear prediction even when the individual model opinions diverge.

Category	Polarity	Type
1	5,441	
2	5,496	
3	15,519	
4	45,034	
5	136,561	
Hotel		51,410
Restaurant		86,720
Attractive		69,921
Total	**208,051**	**208,051**

Table 1: Distribution of Polarity and Type

4. Experiments

4.1. Dataset

Our experiments employ the competition dataset from Rest-Mex 2025, containing tourist reviews of Mexico's special "magical towns". The data includes user opinions with sentiment ratings and venue categories, showcasing real tourism feedback across diverse Mexican locations:

4.2. Experiment Setting

In this study, we employ the Rest-Mex 2025 dataset, a comprehensive corpus comprising 10,000 Spanish-language travel reviews collected from Mexican tourism destinations for the Rest-Mex 2025 challenge. Each review is systematically annotated for three distinct classification tasks: polarity (rated 1–5), destination type (restaurant, hotel, or attraction), and magical town (one of 40 distinct towns). The corpus exhibits moderate class imbalance for polarity—skewed toward positive ratings (4 and 5)—and for destination type, while the magical-town labels demonstrate high imbalance due to the rarity of some towns. Pre-processing removes entries with missing values and concatenates review titles and bodies using a `[SEP]` token to form a unified input sequence. Labels are numerically encoded (polarity: 0–4, destination type: 0–2, magical town: 0–39), and the data are partitioned through stratified sampling into 80% training, 10% validation, and 10% test splits to preserve class distributions.

Three transformer models are fine-tuned for each task: BETO (dccuchile/bert-base-spanish-wwm-cased), RoBERTa (bertin-project/ bertin-roberta-base-spanish), and XLM-RoBERTa (xlm-roberta-large). For polarity and destination type classification, models predict 5 and 3 classes respectively, with a maximum sequence length of 128 tokens; the magical-town task utilizes 256 tokens and predicts 40 classes. Fine-tuning is conducted for three epochs with a learning rate of `2e-5`, the AdamW optimizer, a batch size of 16, and gradient accumulation (4 steps for BETO and RoBERTa, 2 for XLM-RoBERTa). Training employs mixed precision (FP16) for enhanced memory efficiency. Additionally, a domain-adapted BETO is created by pre-training on the corpus with masked-language modeling for two epochs before task-specific fine-tuning. To explore complementary methodologies, final-layer `[CLS]` embeddings from BETO, RoBERTa, and XLM-RoBERTa are fed to XGBoost classifiers. XGBoost is configured with depth 6, learning rate 0.1, and 1,000 boosting rounds, utilizing early stopping on validation loss and class weights, particularly for the highly imbalanced magical-town task. Ensemble strategies include soft voting (probability averaging with weights tuned for macro-F_1 on the validation set) and hard voting (equal-weight majority vote). A LLaMA-3.2-3B-Instruct model is additionally fine-tuned via LoRA ($r = 16$, $\alpha = 32$) in a multi-task setting to generate structured outputs for all three labels, though this remains exploratory due to output-parsing challenges. Transformer models provide robust baselines for Spanish NLP, while XGBoost and ensemble methods leverage complementary inductive biases to offset individual weaknesses.

Evaluation relies on accuracy, macro-F_1, and weighted F_1 metrics to reflect performance across imbalanced classes. These experimental settings balance computational feasibility with rigorous analy-

Method	F1 Score	Accuracy	Precision	Recall
BETO	0.7535	0.7630	0.7421	0.7654
RoBERTa	0.7590	0.7700	0.7834	0.7360
XLM-RoBERTa	0.7674	0.7776	0.7512	0.7847
BETO + XGBoost	0.7529	0.7524	0.7689	0.7378
RoBERTa + XGBoost	0.7589	0.7588	0.7423	0.7767
DA-BETO	0.7622	0.7659	0.7756	0.7494
Ensemble (Soft)	0.7663	0.7778	0.7534	0.7801
Ensemble (Hard)	0.7649	0.7759	0.7612	0.7687
LLM-FT	0.7659	0.7761	0.7498	0.7832
XLM-RoBERTa + XGBoost	**0.7690**	**0.7689**	**0.7845**	**0.7543**

Table 2: Results for Polarity Classification

Method	F1 Score	Accuracy	Precision	Recall
BETO	0.9801	0.9801	0.9823	0.9779
RoBERTa	0.9814	0.9813	0.9789	0.9839
XLM-RoBERTa	0.9830	0.9830	0.9845	0.9815
BETO + XGBoost	0.9800	0.9800	0.9834	0.9766
RoBERTa + XGBoost	0.9813	0.9813	0.9801	0.9825
XLM-RoBERTa + XGBoost	0.9829	0.9829	0.9812	0.9846
DA-BETO	0.9812	0.9812	0.9798	0.9826
Ensemble (Hard)	0.9836	0.9836	0.9849	0.9823
LLM-FT	0.9703	0.9704	0.9745	0.9662
Ensemble (Soft)	**0.9838**	**0.9837**	**0.9821**	**0.9855**

Table 3: Results for Type Classification

Method	F1 Score	Accuracy	Precision	Recall
BETO	0.8912	0.8923	0.8834	0.8991
RoBERTa	**0.9305**	**0.9311**	**0.9234**	**0.9377**
XLM-RoBERTa	0.9173	0.9183	0.9098	0.9249
BETO + XGBoost	0.8933	0.8935	0.8967	0.8899
RoBERTa + XGBoost	0.9291	0.9291	0.9345	0.9238
XLM-RoBERTa + XGBoost	0.9156	0.9157	0.9089	0.9224
DA-BETO	0.9036	0.9051	0.8945	0.9129
Ensemble (Soft)	0.9293	0.9303	0.9187	0.9401
Ensemble (Hard)	0.9253	0.9264	0.9156	0.9352
LLM-FT	0.8396	0.8457	0.8523	0.8271

Table 4: Results for Town Classification

sis, enabling comprehensive comparison across models while respecting the dataset's linguistic and distributional characteristics.

4.3. Main Results

Our experimental evaluation reveals distinct performance patterns across the three classification tasks. For polarity classification, XLM-RoBERTa combined with XGBoost achieved the best overall performance, demonstrating the effectiveness of hybrid approaches. The type classification task showed consistently high performance across all methods, with the soft-ensemble approach slightly outperforming individual models. Town classification exhibited significant variation, with the standalone RoBERTa model surprisingly outperforming more complex ensemble approaches, highlighting that optimal model selection is highly task-dependent. Overall, these findings stress the importance of matching model complexity to task characteristics rather than adopting a one-size-fits-all solution.

Place	Run	Track Score	Macro F1(Polarity)	Macro F1(Type)	Macro F1(Town)
1st	UDENAR_1	0.725420071	0.644452173	0.987723640	0.691964146
2nd	Axolotux_E_T3	0.722583909	0.639542301	0.987621480	0.689599124
3rd	Pandas_Rojos_1	0.686457349	0.616301540	0.981814631	0.634775461
HM	**lephuquy_3 (Ours)**	**0.607854464**	**0.628981369**	**0.981954095**	**0.469069983**

Table 5: Rest-Mex 2025 Final Ranking

5. Error Analysis and Discussion

The most prominent weakness in our approach is observed in the Town Classification task, where our model achieved a macro-F1 score of 0.4690, significantly lower than the top-performing system's score of 0.6919. This performance gap stems primarily from our use of metadata (Region) during training to distinguish between towns. While this metadata enhanced the model's ability to differentiate towns in the training data, it was not available in the test set. As a result, the model failed to generalize effectively, particularly for less frequent towns, leading to poor classification performance.

This issue highlights the critical importance of maintaining data consistency between training and testing phases. The absence of the Region metadata during inference created a mismatch that undermined the model's predictive capability. To address this, potential solutions include eliminating reliance on metadata entirely or integrating external knowledge sources, such as geographical or cultural databases, to provide contextual cues independent of the training data. Additionally, improving data balance—perhaps through oversampling or synthetic data generation—and enhancing the model's ability to handle linguistic variations, such as slang or dialects, could further increase its reliability and robustness for future applications.

5.1. Conclusion

This study examines transformer-based and hybrid approaches for multi-task tourism review analysis, focusing on the classification of magical towns in the Rest-Mex 2025 dataset. While the overall method shows promise, the town classification task achieved a macro-F1 score of 0.4690—significantly lower than the leading 0.6919—primarily due to using Region metadata during training that was not available at testing, resulting in poor generalizability. These findings underscore the importance of consistent feature availability across training and testing, suggesting that future models should avoid reliance on such metadata by incorporating external knowledge, advanced data augmentation, and improved handling of linguistic diversity like regional slang to ensure robust real-world performance..

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Declaration on Generative AI

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

References

[1] R. Guerrero-Rodriguez, M. A. Álvarez Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, Current Issues in Tourism 26 (2023) 289–304. URL:

https://doi.org/10.1080/13683500.2021.2007227. doi:10.1080/13683500.2021.2007227. arXiv:https://doi.org/10.1080/13683500.2021.2007227.

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

[3] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-GonzÁlez, R. Aranda, A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism, Journal of King Saud University - Computer and Information Sciences 35 (2023) 101746. URL: http://dx.doi.org/10.1016/j.jksuci.2023.101746. doi:10.1016/j.jksuci.2023.101746.

[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] 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.

[6] 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).

[7] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

[8] 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.

[9] 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.

[10] B. A. Muritala, M.-V. Sánchez-Rebull, A.-B. Hernández-Lara, A bibliometric analysis of online reviews research in tourism and hospitality, Sustainability 12 (2020) 9977.

[11] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).

[12] M. Faseeh, A. Jaleel, N. Iqbal, A. Ghani, A. Abdusalomov, A. Mehmood, Y.-I. Cho, Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy, Mathematics 12 (2024) 3416.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[14] S. M. Bhatt, S. Agarwal, O. Gurjar, M. Gupta, M. Shrivastava, Tourismnlg: a multi-lingual generative benchmark for the tourism domain, in: European Conference on Information Retrieval, Springer, 2023, pp. 150–166.

[15] O. O. Awe, G. O. Opateye, C. A. G. Johnson, O. T. Tayo, R. Dias, Weighted hard and soft voting ensemble machine learning classifiers: Application to anaemia diagnosis, in: Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, Ghana, 2022, Springer, 2024, pp. 351–374.