Binary Fine-Tuning with Centroid-Based Sampling and Transformer Fusion: The Top-Scoring System at REST-MEX 2025

Juan David Jurado-Buch[1,*]

[1]Universidad de Nariño, Pasto, Colombia

Abstract

In this work, we propose a novel architecture that achieved the best overall performance in the REST-MEX 2025 shared task for Spanish-language sentiment and thematic classification. Our method transforms the multiclass classification task into a series of binary classification problems—one for each class across three axes: polarity (1–5), type (Hotel, Restaurant, Attractive), and town (40 selected locations in Mexico). For each class, we fine-tune a RoBERTa-based Transformer using all positive examples of the target class and a balanced number of negative examples from the remaining classes. Two strategies for negative sampling are explored: 1) random selection and 2) centroid-based selection using contextual embeddings derived from the [CLS] token of a pre-trained model roberta-base-bne-finetuned-sentiment_analysis_RestMex2023. In the centroid-based approach, we compute the geometric center of each class in the 768-dimensional embedding space and select the most representative samples based on cosine proximity. The final classification decision is obtained by concatenating the binary outputs from each class-specific model and feeding them into a 10-layer multilayer perceptron. This ensemble strategy significantly outperforms traditional multiclass models, achieving macro F1-scores 0.644 (polarity), an outstanding 0.987 (type) and 0.691 (town), setting a new benchmark in the task. These results confirm the efficacy of task decomposition, embedding-based sampling, and transformer specialization for fine-grained sentiment analysis in low-resource languages.

Keywords

, RoBERTa, REST-MEX 2025, Contextual Embeddings, Centroid Sampling, Spanish NLP

1. Introduction

Tourism plays a critical role in the economic, cultural, and social development of many countries. In Mexico, the promotion and management of tourist destinations—particularly the nationally recognized "Pueblos Mágicos"—relies increasingly on the analysis of user-generated content such as reviews, ratings, and narratives. These texts contain valuable insights into visitor satisfaction, service quality, and destination image, but extracting actionable information at scale remains a complex task [1, 2, 3].

In recent years, artificial intelligence (AI) and natural language processing (NLP) have emerged as transformative tools in tourism analytics. Sentiment analysis, topic modeling, and classification systems allow tourism boards, local businesses, and policymakers to better understand traveler perceptions and adapt strategies accordingly[4, 5, 6]. However, despite advances in multilingual language models, high-quality tools and datasets tailored to Spanish-language tourism content are still limited[7, 8, 9, 6].

The REST-MEX shared task series has addressed this gap since 2021, providing a benchmark framework for Spanish sentiment and thematic classification. The first edition (REST-MEX 2021) [10] focused on polarity classification for tourist reviews. The 2022 and 2023 editions introduced new classification axes, including type of establishment and town identification, while fostering methodological innovation in the Spanish NLP community[11, 12]. In 2025, the fourth edition of REST-MEX presents the most comprehensive dataset to date, with over 200,000 labeled reviews annotated across three classification tasks: polarity (from 1 to 5), service type (Hotel, Restaurant, Attractive), and 40 carefully selected Mexican towns[13, 14].

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

jjuradobuch@gmail.com (J. D. Jurado-Buch)

While many participating systems have adopted multiclass architectures based on fine-tuned Transformers, limitations remain when classes are highly imbalanced or semantically overlapping. In this work, we introduce an alternative approach grounded in binary decomposition and embedding-based data selection. Instead of building a single multiclass model per task, we train independent binary classifiers for each class. Each classifier specializes in detecting the presence or absence of a single category, using balanced datasets curated via centroid-aware selection strategies in embedding space. These classifiers are then fused using a lightweight multilayer perceptron, allowing for a final softmax decision over all classes[15].

Our method, based on the best RoBERTa variant from REST-MEX 2023 [16], not only simplifies the training process for each class but also enables better control over data balance and representation. The experimental results demonstrate the effectiveness of this approach, setting a new state-of-the-art across all three REST-MEX 2025 tasks. This work contributes a scalable and interpretable architecture applicable to other multilingual or imbalanced-class scenarios in tourism and beyond.

2. Methodology

The REST-MEX 2025 dataset comprises 208,051 tourist reviews annotated along three classification tasks: sentiment polarity (five classes), type of establishment (three classes), and town (40 selected locations). One major challenge is the significant class imbalance—especially in the polarity task, where class 5 dominates the distribution.

To address this, we reformulate each multiclass classification problem as a set of binary classification subproblems. For each class c in a given task, we train a Transformer-based binary classifier f_c that distinguishes between class c and "not-c".

2.1. Binary Classifier Construction

Each binary classifier is trained independently using a balanced subset of the dataset:

- For minority classes (e.g., polarity class 1), we select all available positive instances and randomly sample an equal number of negative instances from the remaining classes.
- For majority classes (e.g., polarity class 5), all negative instances (i.e., those not labeled as class 5) are included as the negative class.

To improve representativeness, we employ two selection strategies for negative examples:

1. **Random Sampling**: Standard uniform sampling from the non-target classes.
2. **Centroid-based Sampling**: We compute contextual embeddings for each instance using the [CLS] token output from the `vroberta-base-bne-finetuned-sentiment_analysis_RestMex2023`[1] model. For each class, we define a centroid vector μ_c, and select the closest non-class-c embeddings to form a balanced dataset.

2.2. System Architecture

Each binary classifier is implemented as a `roberta-base-bne-finetuned-sentiment_analysis_RestMex2023`-based Transformer fine-tuned for two epochs. Once all binary classifiers for a task are trained, their outputs (logits or soft probabilities) are concatenated and passed to a multilayer perceptron (MLP) with the following configuration:

- 10 fully connected hidden layers
- 10 neurons per layer
- ReLU activation for hidden layers
- Softmax activation in the output layer
- Trained for 10 epochs using cross-entropy loss

[1] https://huggingface.co/vg055/roberta-base-bne-finetuned-sentiment_analysis_RestMex2023

2.3. Classifier Fusion and Final Decision

Let $P_1, P_2, \ldots, P_K$ be the outputs of the K binary classifiers for a given task. We concatenate these outputs into a feature vector $\mathbf{x} \in \mathbb{R}^K$, and feed it into the MLP, which outputs the final class prediction. This decoupled architecture allows each binary classifier to specialize while still contributing to a global multiclass decision.

2.4. Dataset Summary (Balanced Sampling)

For clarity, Table 1 presents the number of instances used per class during binary training using centroid-based sampling:

Class	Positive Examples	Negative Examples
Polarity 1	5,441	5,441 (from classes 2–5)
Polarity 2	5,496	5,496 (from classes 1, 3–5)
Polarity 3	15,519	15,519 (from classes 1, 2, 4, 5)
Polarity 4	45,034	45,034 (from classes 1–3, 5)
Polarity 5	136,561	71,490 (all other classes combined)

Table 1: Balanced instance count per binary classifier (Polarity task)

Similar balancing strategies were applied to the Type and Town classification tasks. For the Town task, which includes 40 classes, training and inference are parallelized to optimize runtime.

3. Results

To evaluate the effectiveness of the proposed per-class binary classification strategy, we compared two methods for constructing balanced datasets: (1) **random sampling** of negative instances and (2) **centroid-based selection** using contextual embeddings. Both strategies were applied independently for each class in the three tasks: polarity, type, and town. After training individual RoBERTa classifiers and integrating their outputs through a multilayer perceptron, we computed macro-averaged F1 scores and other standard metrics on the official REST-MEX 2025 test set.

3.1. Overall Performance

Table 2 summarizes the performance obtained using each strategy. The centroid-based approach outperformed random sampling in every task and metric, particularly for the more challenging polarity and town classifications.

3.2. Performance by Class

A more detailed per-class analysis further supports the advantage of centroid-based selection. For example, in the polarity task, the F1 score for class 2 increased from 0.443 to 0.472, and for class 5 from 0.875 to 0.888. Similarly, in town classification, hard-to-discriminate locations such as `Sayulita` and `Tequisquiapan` showed improvements of over 2 percentage points in F1.

3.3. Discussion of Gains

These results validate the hypothesis that using contextual embeddings to identify representative negative instances leads to better decision boundaries and improved generalization. By selecting negatives that are closer to the semantic center of their class, the classifiers learn to distinguish finer nuances rather than overfitting to outliers or noise.

Task	Metric	Random	Centroid-based	Gain
Polarity	Macro F1	0.7012	**0.7254**	+0.0242
	Accuracy	0.6691	**0.6919**	+0.0228
	MAE	0.2493	**0.2262**	-0.0231
	Precision	0.6021	**0.6275**	+0.0254
Type	Macro F1	0.6215	**0.6445**	+0.0230
	Accuracy	0.9831	**0.9882**	+0.0051
	Precision	0.9729	**0.9876**	+0.0147
Town	Macro F1	0.9735	**0.9877**	+0.0142
	Accuracy	0.7538	**0.7703**	+0.0165
	Precision	0.6381	**0.6575**	+0.0194

Table 2: Comparison between instance selection strategies on the test set

Overall, the centroid-based strategy not only achieved the best macro F1 scores across all tasks in REST-MEX 2025 but also proved computationally feasible due to the modest size of the final fine-tuning subsets. It constitutes a principled and scalable method for class balancing in low-resource multiclass scenarios.

4. Conclusions

In this paper, we proposed a class-specific binary classification framework using fine-tuned RoBERTa models, trained independently for each class of polarity, service type, and town in the REST-MEX 2025 dataset. By transforming each multiclass task into a set of balanced binary problems and combining their outputs through a multilayer perceptron, the system achieved outstanding results.

The centroid-based selection of negative examples proved consistently superior to random sampling. This strategy leverages contextual embeddings to identify representative samples, allowing the classifiers to learn more discriminative and generalizable boundaries. As a result, our approach achieved the highest scores among all participating systems in the 2025 edition of the shared task.

Despite these achievements, the method incurs significant computational costs. Training one binary classifier per class (a total of over 50 models), followed by the final MLP stage, required approximately **64 GPU-hours** using an **NVIDIA A100 (40GB)**. This makes the approach less suitable for rapid iteration or deployment in low-resource environments.

Future work will explore methods to reduce the number of required classifiers, such as parameter sharing, knowledge distillation, or sparse training strategies. Nonetheless, our findings demonstrate the effectiveness of decomposing multiclass NLP problems into interpretable, balanced binary decisions—especially when supported by high-quality pre-trained language models and semantically meaningful instance selection.

Declaration on Generative AI

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

References

[1] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico, International Journal of Tourism Cities 10 (2024) 639–661.

[2] I. Castillo-Ortiz, M. Á. Álvarez-Carmona, R. Aranda, Á. Díaz-Pacheco, 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 (2024) 101070.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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).
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[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] 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.
[15] 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.
[16] 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).