A Multi-Task BETO-Based Framework with Synthetic Data Augmentation for Sentiment and Contextual Classification of Spanish Tourist Reviews

Alvaro Zaid Gallardo-Hernández[1,*], Ramón Aranda[2,3] and Angel Diaz-Pacheco[4]

[1]Departamento de Ciencias e Ingenieras, Universidad Iberoamericana Puebla, San Andrés Cholula, Puebla, México.
[2]Centro de Investigación en Matemáticas, Sede Mérida, Mérida, Yucatán, México.
[3]Consejo Nacional de Humanidades, Ciencias y Tecnologías, Ciudad de México, México
[4]Departamento de Ingeniería Electrónica, División de Ingenierías, Universidad de Guanajuato – Campus Irapuato-Salamanca, Yuriria, Mexico

Abstract

This paper presents our solution for the Rest-Mex 2025 shared task, which involves multilingual sentiment and contextual classification of Spanish-language tourist reviews. Given a review, the task is to determine its sentiment polarity (from 1 to 5), the type of destination (hotel, restaurant, or attraction), and the corresponding Magical Town from a predefined list. To address this, we developed a multi-task classification model based on the BETO transformer, incorporating three output heads to predict polarity, type, and town simultaneously. To mitigate class imbalance—especially for underrepresented sentiment classes—we implemented a data augmentation strategy that combines Jaccard distance–based sampling with synonym substitution using the Spanish WordNet. This approach allowed us to synthetically generate diverse reviews for minority classes. The model was trained using PyTorch with the AdamW optimizer and evaluated using macro F1 and accuracy across all tasks. Our system achieved competitive results, particularly excelling in the opinion type classification subtask. Additionally, we employed generative AI tools such as Gemini 2.5 to assist in code generation and experimentation, highlighting the emerging role of LLMs in reproducible NLP research.

Keywords: Sentiment Analysis, Multi-task Learning, Spanish NLP, BETO, Data Augmentation, Tourism, Transformers, WordNet, LLM-Assisted Research.

Keywords

Sentiment Analysis, Rest-Mex

1. Introduction

The rapid digitalization of the tourism industry has led to a proliferation of user-generated content, particularly online reviews, which now play a central role in shaping traveler decisions and influencing business strategies [1, 2, 3]. This paradigm shift has driven the need for automated systems capable of extracting actionable insights from natural language data, especially in low-resource and domain-specific contexts such as local tourism in Spanish-speaking regions [4].

Unlike to others editions [5, 6, 4], the Rest-Mex 2025 [7, 8] shared task poses a multi-faceted classification challenge that involves analyzing Spanish-language tourist reviews to predict three attributes: sentiment polarity (ranging from 1 to 5), the type of destination (hotel, restaurant, or attraction), and the specific town—typically drawn from a predefined list of Mexico's Pueblos Mágicos [5]. Addressing this task requires robust models that can handle multilingual input, unbalanced class distributions, and nuanced domain-specific expressions.

Recent advances in Natural Language Processing (NLP), particularly with the introduction of transformer-based models such as BERT [9] and its Spanish counterpart BETO [10], have significantly improved text classification performance in a wide range of tasks. BETO, trained specifically on large Spanish corpora, has shown strong results in sentiment and domain adaptation tasks. Multi-task

IberLEF 2025, September 2025, Jaén, Spain
*Corresponding author.
alvaro.gallardo@iberopuebla.mx (A. Z. Gallardo-Hernández); arac@cimat.mx (R. Aranda); angel.diaz@ugto.mx (A. Diaz-Pacheco)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

CEUR-WS.org/Vol-4098/RESTMEX2025_paper32.pdf

learning strategies further enhance model generalization by allowing simultaneous optimization over related objectives [11].

However, data imbalance remains a persistent challenge in real-world scenarios. In the tourism domain, positive reviews often dominate datasets, which can hinder model performance on minority sentiment classes. To mitigate this, data augmentation techniques have been explored, including lexical substitution using WordNet [12] and review recombination strategies based on lexical similarity measures such as Jaccard distance [13]. These techniques aim to synthetically increase diversity in underrepresented categories without compromising semantic plausibility.

Furthermore, the integration of large language models (LLMs) into research workflows is reshaping the NLP development landscape. Tools such as Gemini [14], ChatGPT, and Claude have shown potential not only for code generation but also for experiment design and rapid prototyping [15]. As such, their use introduces new opportunities for accelerating research while raising important questions about reproducibility and human oversight.

In this paper, we present a multi-task classification framework based on BETO for the Rest-Mex 2025 challenge. Our contributions include: (1) a synthetic data generation strategy combining Jaccard sampling and WordNet-based synonym substitution, (2) a multi-head neural architecture leveraging shared representations, and (3) a discussion on the use of LLMs to support NLP research. We evaluate our approach using official task metrics and report competitive results, particularly in opinion type classification.

2. Task Description: Sentiment Analysis

The goal of this task is to analyze TripAdvisor reviews and classify them based on three main aspects: sentiment polarity, type of site, and associated Pueblo Mágico (Magical Town).

Each review provides useful insights into a tourist's experience. First, the task involves predicting the sentiment polarity of the review, assigning a score from 1 (very negative) to 5 (very positive) based on the original rating given by the user.

Next, participants must classify the type of site being reviewed. A review can refer to a *hotel*, *restaurant*, or *attraction*, based on context and metadata.

Finally, the third sub-task is to predict the Pueblo Mágico (Magical Town) associated with each review. This classification is based on location metadata and aims to ensure that the review is accurately linked to the correct town.

Evaluation:

Systems are evaluated using standard metrics such as precision, recall, and F1-score. The evaluation is divided into three sub-tasks:

1. **Polarity classification:** The macro-F1 score over all polarity classes (1–5) is computed as:

$$Res_P(k) = \frac{\sum_{i=1}^{|C|} F_i(k)}{|C|}$$

   where $F_i(k)$ is the F1-score for class i by system k, and $C = \{1, 2, 3, 4, 5\}$.
2. **Type classification:** The macro-average F1-score over three categories (Attraction, Hotel, Restaurant):

$$Res_T(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3}$$

   where $F_A(k)$ is the F1-score for the Attraction class, $F_H(k)$ for Hotel, and $F_R(k)$ for Restaurant.
3. **Pueblo Mágico classification:** A macro-F1 score is calculated over all towns in the Magical Towns list (MTL):

$$Res_{MT}(k) = \frac{\sum_{i=1}^{|MTL|} F_{MTL_i}(k)}{|MTL|}$$

   where $F_{MTL_i}(k)$ is the F1-score for the i-th Magical Town.

Final Score:

The final score for a system k is computed as:

$$Sentiment(k) = \frac{2 \cdot Res_P(k) + Res_T(k) + 3 \cdot Res_{MT}(k)}{6}$$

This formula gives more weight to the polarity and Pueblo Mágico sub-tasks, reflecting their greater importance in the evaluation.

3. Methodology

3.1. Data Preparation and Advanced Text Cleaning

The original dataset, comprising Spanish-language reviews of Mexican tourist destinations, underwent rigorous preprocessing to ensure data quality and consistency. This included handling missing values (replacing NaN with empty strings), Unicode normalization, emoji removal, conversion to lowercase, and substitution of URLs, user mentions, and hashtags with standardized placeholders (`_URL_`, `_MENCION_`, `_HASHTAG_`). Extraneous punctuation and redundant whitespace were removed. Linguistic processing with spaCy was applied for tokenization and lemmatization, as well as stopword and number removal. Generic tags were preserved during processing.

3.2. Synthetic Data Generation and Class Balancing

To address the pronounced class imbalance in sentiment polarity, we implemented a mixed strategy based on both downsampling and synthetic oversampling.

Synthetic Review Generation. For minority classes, we designed an algorithmic pipeline to create lexically diverse and semantically plausible synthetic reviews, which operates as follows:

1. **Selection of Base and Candidate Reviews:** For each synthetic sample, a base review r_1 is randomly selected from the minority class. Ten additional candidate reviews $\{r_{c,i}\}$ from the same class are sampled at random.
2. **Lexical Similarity via Jaccard Distance:** For each candidate, we compute the Jaccard distance between the base review and the candidate as follows:

   $$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

   where A and B are the sets of alphabetic tokens (after lowercasing and stopword removal) from the two reviews. This metric quantifies the lexical dissimilarity between reviews.
3. **Candidate Selection:** Based on the computed distances, three key candidates are identified: the most similar (r_c), the most distant (r_a), and an intermediate candidate (r_i).
4. **Fragment Extraction and Synonym Substitution:** From r_1, r_c, r_i, and r_a, a random subset (typically 25%) of their tokens is selected. To further diversify the generated text, each selected token is optionally replaced with a synonym using the Spanish WordNet corpus (via NLTK and OMW 1.4). If a suitable synonym is available and distinct from the original token, it is used; otherwise, the original token is retained.
5. **Synthetic Review Assembly:** The selected fragments (enriched with synonyms) are concatenated to synthesize a new artificial review, which inherits the label of the target class.

This process is repeated iteratively until the minority class reaches the desired number of samples.

Class Balancing. To address the pronounced class imbalance across all three classification tasks, we implemented a generalized balancing strategy involving both random downsampling and synthetic oversampling. This procedure was applied independently to the sets of labels for polarity, destination type, and Magical Town, with the aim of creating a more uniform data distribution for the model.

Let $\mathcal{C}_t$ be the set of classes for a given task $t \in \{\text{polarity, type, town}\}$. Our objective is to transform the original sample distribution, where each class $i \in \mathcal{C}_t$ has a size of $|C_i|$, into a balanced distribution with a new size $|C_i'|$. This is achieved by defining a target sample size, N_{target}, and applying the following transformation:

$$|C_i'| = \begin{cases} N_{\text{target}} & \text{if } |C_i| > N_{\text{target}} \quad \text{(Downsampling)} \\ N_{\text{target}} & \text{if } |C_i| < N_{\text{target}} \quad \text{(Oversampling)} \\ |C_i| & \text{if } |C_i| \approx N_{\text{target}} \quad \text{(Invariant)} \end{cases}$$

where downsampling is performed by random selection without replacement, and oversampling is achieved using the synthetic review generation method described previously.

For the polarity task, this strategy was applied with a target size of $N_{\text{target}} = 15519$. The majority classes (polarities 4.0 and 5.0) were downsampled, while the minority classes (1.0 and 2.0) were synthetically oversampled to reach N_{target}. The neutral class (polarity 3.0), whose size was close to the target, was kept invariant, preserving its original data.

A similar principle was applied to the destination type and Magical Town tasks. For each, we identified majority and minority classes relative to a target size and applied downsampling or synthetic oversampling, respectively, to mitigate the skew in their distributions.

The final comprehensive dataset, composed of a mix of original, downsampled, and synthetic reviews for all tasks, is then shuffled and prepared for model training.

3.3. Data Preparation for PyTorch

The balanced dataset is further processed for compatibility with PyTorch. Categorical textual labels (e.g., sentiment polarity, destination type, and town) are mapped to integer values via label encoding. Each review is tokenized using the pre-trained BETO tokenizer, with sequences padded or truncated to a fixed maximum length $L_{\max}$. The dataset is then stratified into training and validation sets. PyTorch `DataLoader` objects are instantiated to manage mini-batch sampling and data shuffling.

3.4. Neural Network Architecture: MultiTaskBETO

For classification, we employ a multi-task neural network architecture based on the BETO transformer model, specialized for Spanish. The architecture includes:

- **BETO Encoder:** The core of the model is the pre-trained BETO encoder (`AutoModel.from_pretrained`), which generates contextualized vector representations for each input review.
- **Dropout Regularization:** A dropout layer is applied to the pooled output to reduce overfitting.
- **Multi-task Classification Heads:** Three independent linear layers receive the pooled output (corresponding to the `[CLS]` token):
  1. Sentiment polarity (5 classes)
  2. Destination type (hotel, restaurant, attraction)
  3. Magical Town (up to 60 classes)

  Each head outputs unnormalized logits for its corresponding task.

3.5. Training Pipeline

The model is trained using the AdamW optimizer and categorical cross-entropy loss for each classification head. Let $\mathcal{L}_{\text{pol}}$, $\mathcal{L}_{\text{type}}$, and $\mathcal{L}_{\text{town}}$ denote the losses for sentiment, type, and town prediction, respectively; the total loss for each batch is computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pol}} + \mathcal{L}_{\text{type}} + \mathcal{L}_{\text{town}}$$

The model is trained for E epochs, where in each epoch the model iterates through all training batches, computes forward and backward passes, updates model weights, and evaluates performance on the validation set.

3.6. Iterative Training and Evaluation

For each epoch, training and validation loss, as well as macro-F1 and accuracy for each task, are monitored and recorded. Upon completion, the best model (according to validation performance) is selected for further analysis and reporting.

4. Results

We obtained an overall **Track Score of 0.6145**. In particular, our system excelled in the **opinion type classification** subtask, where we achieved a **Macro F1 score of 0.9688**. This indicates that our model was highly effective in differentiating between reviews of hotels, restaurants, and attractions.

For the **polarity classification** task, we reached a **classification accuracy of 71.52%**, which was among the highest in the evaluation. However, the **Macro F1 score** for this subtask was lower, at **0.4381**, highlighting performance imbalances across sentiment classes. Specifically, our system performed best on **class 4** with an **F1-score of 0.8525**, while performance was notably weaker on minority classes such as **class 1**, which yielded an **F1-score of 0.1856**.

In the **town classification** task, we obtained a **Macro F1 score of 0.5902**. Although slightly below the scores achieved by the top-performing systems, this result still reflects a reasonable performance given the large number of distinct town labels in the dataset.

Our system demonstrated strong capabilities in the opinion type classification task and achieved high accuracy in polarity classification. Nonetheless, the results suggest that future work should focus on improving the balance of predictions across all sentiment classes and enhancing performance on the more granular town classification subtask.

5. Conclusions

In this paper, we presented a multi-task classification approach for the Rest-Mex 2025 shared task, focusing on the automatic prediction of sentiment polarity, opinion type, and associated town in Spanish-language tourist reviews. Our system was built upon the BETO transformer and demonstrated competitive performance—particularly in the opinion type classification task, where it achieved near-perfect macro F1 results.

To mitigate the strong class imbalance present in the original dataset, we applied a synthetic data generation pipeline using Jaccard-based sampling and synonym substitution via Spanish WordNet. This approach increased lexical diversity and helped improve performance in minority classes, although challenges remained—especially in sentiment classes with few training examples.

For future work, we intend to explore and compare alternative architectures, including encoder-decoder models and large language models (LLMs) such as LLaMA, Mistral, and other multilingual transformers. We also recognize the increasing relevance of generative AI tools in research workflows. While this work leveraged Google's Gemini 2.5 Pro for code generation and experiment support, further experimentation is needed to systematically compare the contributions of various LLM-based

assistants—such as ChatGPT, Claude, or open-source alternatives—on reproducibility, model design, and code quality.

By incorporating these tools into a reproducible benchmarking framework, we aim to assess not only model performance but also the qualitative impact of AI-assisted research, ultimately enhancing the effectiveness and transparency of sentiment analysis in multilingual tourism datasets.

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] F. Calderón, M. Blanco, Impacto de internet en el sector turístico, Revista UNIANDES Episteme 4 (2017) 477–490.

[2] 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 (2023) 639–661. URL: http://dx.doi.org/10.1108/IJTC-09-2022-0223. doi:10.1108/ijtc-09-2022-0223.

[3] 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.

[4] M. A. Alvarez-Carmona, A. Díaz-Pacheco, R. Aranda, et al., Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

[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. Herrera-Semenets, Overview of rest-mex at iberlef 2025: Researching sentiment evaluation in text for mexican magical towns, volume 75, 2025.

[8] 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.

[9] 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.
[10] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, D. Rogers, Spanish pre-trained bert model and evaluation data, 2020. ArXiv:2003.12171.
[11] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).
[12] G. A. Miller, Wordnet: A lexical database for english, Communications of the ACM 38 (1995) 39–41.
[13] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson, 2005.
[14] G. DeepMind, Gemini 1.5 technical report, 2024. Https://deepmind.google/technologies/gemini/.
[15] OpenAI, Gpt-4 technical report, 2023. Https://openai.com/research/gpt-4.