Multitask Classification of Mexican Tourist Reviews Using a Multi-Head Transformer Model Based on BETO

Juan Javier Monsivais Borjón [1,*], Miguel Ángel Álvarez-Carmona [1]

[1] *Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, Monterrey Nuevo León, 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

In recent years, online review platforms have become central to how travelers share their experiences, generating vast repositories of user-generated content [1, 2, 3, 4, 5]. This textual data, though unstructured, is a goldmine of insights into customer sentiments, service quality perceptions, and destination appeal [6, 2]. In tourism, automated analysis of these reviews enables businesses to refine their offerings while empowering travelers to make well-informed decisions—a dynamic especially relevant in Mexico [7], a nation with exceptional cultural, gastronomic, and geographic diversity and one of the country's largest economic pillars [8, 9, 10].

Since its inception in 2021, the Rest-Mex shared task has served as a leading benchmark in applying NLP to Mexican tourist texts. In 2021, the challenge focused on two tasks: predicting overall satisfaction scores (recommendation systems) and classifying sentiment polarity from TripAdvisor reviews [11]. The 2022 edition expanded to include a third track: predicting the federal "COVID-19 epidemiological semaphore" status [12, 13] from news texts [14]. By 2023, Rest-Mex had added text clustering as a new task and broadened its dataset to include reviews from Cuba and Colombia, though sentiment and type classification remained core [15]. Throughout these versions, the primary focus has remained on sentiment polarity, type of place, and in later years, country-level categorization.

The 2025 edition introduces a novel dimension by directly involving "pueblos mágicos" (magical towns): participants must now precisely identify the specific town mentioned among forty Mexican localities. This addition elevates the geographic granularity of the task, requiring models to discern fine-grained location cues—an essential step for geographically-aware tourism analytics [16, 17].

Transformer-based models such as BERT and its Spanish-adapted variant BETO have consistently excelled in such NLP challenges due to their strong contextual understanding. Their capacity to encode semantic subtleties makes them ideal for handling diverse and nuanced tourist reviews [18].

In this study, we propose an enhanced multilingual, multi-head architecture built atop BETO, aiming to simultaneously predict three outputs: sentiment polarity, town (now including magic towns), and establishment type. This multitask framework leverages shared representations and joint optimization to achieve computational efficiency and improve generalization across tasks.

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

juan.monsivais@cimat.mx (J. J. M. B.); miguel.alvarez@cimat.mx (M. : :)

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

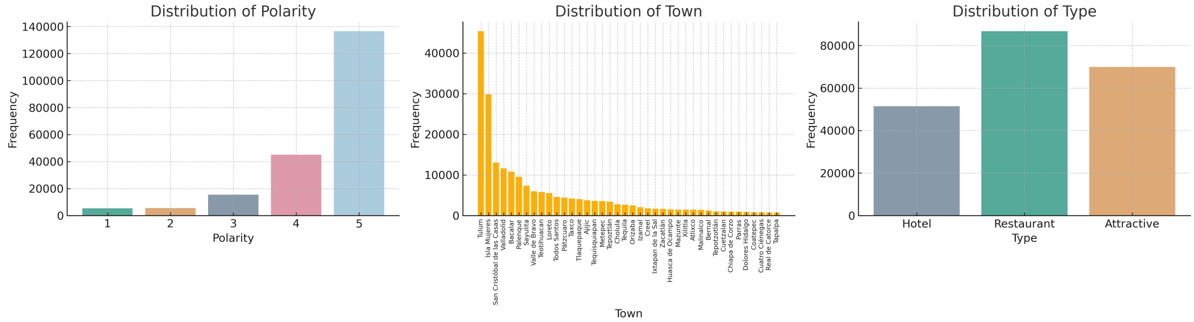

Figure 1: Distributions of polarity, type, and town in the Rest-Mex 2025 dataset.

2. Methodology

Before detailing the individual components—such as data collection, preprocessing, model design, and evaluation—we begin with a high-level overview of our methodological approach. Drawing on best practices in NLP, our pipeline adheres to a clear progression from data preparation to model deployment. [19]

First, we carry out comprehensive cleaning and normalization to ensure consistency and reduce noise in the input text. This step establishes a solid foundation, as preprocessing decisions can significantly influence downstream model performance [20].

Next, we design a multitask Transformer-based architecture that leverages shared representations to handle simultaneous predictions of sentiment, location, and type. Training proceeds under a unified objective, combining losses for each task head. Finally, we evaluate the system using established metrics like accuracy and macro F1-score, which are particularly suitable for handling class imbalances [21].

In the following subsections, we detail each phase—starting with the dataset and preprocessing steps—before exploring the model architecture, training setup, and evaluation strategies.

2.1. Dataset and Preprocessing

The foundation of any robust NLP model is high-quality, well-prepared training data. For this reason, we began our methodology using the Rest-Mex 2025 training dataset, which comprises over 208,000 authentic Spanish-language tourist reviews. Figure shows the distribution for the three labels.

To ensure reliable downstream performance, we implemented an extensive preprocessing pipeline informed by established best practices in text classification and sentiment analysis:

1. **Encoding Correction:** We detect and remedy common encoding anomalies such as misrepresented accents or fragmented characters—prevalent in user-submitted web text.
2. **Pattern Removal:** Employing custom regular expressions, we filtered out HTML remnants and encoding artifacts (e.g., `_xABCD_`), preventing non-linguistic tokens from polluting the vocabulary.
3. **Whitespace & Line Break Normalization:** Multi-space sequences, tabs, and inconsistent newlines were standardized to single spaces, improving tokenization stability without altering semantics.
4. **Unicode Normalization:** Utilizing NFKC normalization via `unicodedata.normalize`, we harmonized composite Unicode characters into canonical forms, reducing variation in accented characters and special symbols.

Empirical studies show that even simple cleaning decisions—such as normalization and encoding repair—can significantly affect model performance, especially in sentiment tasks [4].

Following preprocessing, we extracted the key columns: `Review`, `Polarity`, `Town`, and `Type`. We discarded less relevant fields like `Title` (present in only 2 samples) and `Region` (not used for optimization). Target labels were then converted into numerical indices using `LabelEncoder`, enabling efficient mapping during model training.

This thorough approach to dataset cleaning and label encoding established a consistent and noise-reduced foundation, crucial for the effective representation learning that underpins our multi-head Transformer model.

2.2. Model Architecture: MultiHeadBETO

Inspired by hard parameter sharing in multitask learning:contentReference[oaicite:2]index=2, we designed a model named `MultiHeadBETO`, based on the BETO Transformer encoder [22]. The model architecture consists of:

- A shared BETO encoder (`dccuchile/bert-base-spanish-wwm-uncased` [23]).
- Three task-specific classification heads branching from the shared [CLS] token:
  1. **Polarity Head:** 5-class linear layer.
  2. **Town Head:** 40-class linear layer.
  3. **Type Head:** 3-class linear layer.

Training is guided by a joint loss function combining the cross-entropy of each head:

$$L_{\text{total}} = L_{\text{polarity}} + L_{\text{town}} + L_{\text{type}}$$

This hard-sharing setup encourages shared contextual understanding while allowing task-specific discrimination.

2.3. Training Setup

The cleaned dataset was split 80/20 into training and validation sets using a fixed random seed. Tokenization leveraged the BETO tokenizer with a maximum sequence length of 128.

Training was performed using HuggingFace's `Trainer` API with the following configuration:

- Epochs: 3
- Batch size: 16
- Mixed precision (FP16): Enabled
- Evaluation: Once per epoch
- Checkpoints: Max 2 saved
- Early stopping was optionally implemented through learning rate scheduling.

This setup balances computational efficiency with model robustness.

2.4. Evaluation Metrics

Model performance was evaluated on the validation set using:

- **Accuracy:** Fraction of correct predictions.
- **F1-score (macro):** Averaged harmonic mean of precision and recall to balance class-level performance across imbalanced categories.

These metrics are standard in multitask NLP evaluation:contentReference[oaicite:3]index=3 and are particularly sensitive to the class imbalance present in Town and Polarity labels.

Metric	Polarity	Town	Type	General
F1-score (macro)	0.551	0.623	0.976	-
Accuracy	0.742	0.713	0.977	-
Evaluation Loss				1.786
Eval Runtime				134.48 s
Samples/sec				309.43
Steps/sec				38.68

Table 1: Performance of `MultiHeadBETO` on the validation set

3. Results

Performance metrics on the validation set are summarized in Table 1.

Type classification achieved near-perfect performance, likely due to distinguishable lexical patterns across the three classes.

Town classification was more challenging, given the 40-class imbalance and lexical overlap between towns.

Polarity classification yielded moderate results, reflecting the difficulty of detecting nuanced sentiment in natural language.

4. Discussion

The results demonstrate that our multi-head architecture based on BETO effectively handles the multitask challenges posed by Rest-Mex 2025. By sharing a common Transformer encoder and using task-specific classification heads, the model benefits from cross-task representation learning—a strategy well-supported by prior studies in multitask learning. For instance, trade-offs between sentiment and type classification arise naturally, yet the shared layers amplify performance across related tasks by extracting mutually informative features.

Notably, classification of establishment type achieved near-perfect performance, suggesting that lexical cues and domain-specific vocabulary are highly discriminative. However, sentiment polarity remains the most demanding task, as it requires detecting nuanced tone, sarcasm, and implicit opinions within user-generated text. Enhancements such as task-specific attention modules or dynamic loss weighting could improve model focus on sentiment subtleties. Similarly, town classification accuracy could benefit from incorporating geographical priors or embeddings that capture spatial adjacency and regional similarity—techniques proven effective in geo-aware NLP research.

Implementing alternative sampling strategies during training, such as square-root or proportional scheduling, might also balance gradient contributions across tasks and reduce overfitting in underrepresented classes:contentReference[oaicite:4]index=4. Additionally, external linguistic and geographic knowledge sources—such as knowledge graphs or gazetteers—could provide disambiguation support, especially for town names with overlapping feature sets.

Overall, while the current architecture offers a strong baseline, our findings indicate clear avenues for refinement through modular enhancements and richer contextual grounding.

5. Conclusion

We introduced MultiHeadBETO, a multitask Transformer model designed to jointly predict sentiment polarity, magical-town affiliation, and establishment type from Spanish-language tourist reviews. The model achieved outstanding classification results, especially for establishment type ($F1_{\text{macro}} = 0.976$), while delivering competitive performance in town and sentiment tasks. These outcomes confirm that the combination of shared Transformer representations and task-specific heads is a powerful strategy for multi-dimension text classification in specialized domains like tourism.

These findings also highlight the versatility of multitask Transformer architectures in leveraging the interdependence between tasks, thereby improving overall robustness and generalization—an advantage underscored in previous studies:contentReference[oaicite:5]index=5. Moreover, incorporating "pueblos mágicos" in the classification schema demonstrates the model's adaptability to incorporate more granular geographic tasks without disproportionately affecting performance.

Our work affirms the potential of end-to-end models in extracting valuable insights from real-world tourist review datasets. At the same time, persistent challenges in sentiment and fine-grained geographic classification point to promising directions involving enriched linguistic and spatial modeling.

6. Future Directions

Beyond the current scope, several research avenues warrant exploration. One promising line involves the introduction of task-specific attention layers or dynamic weighting mechanisms aimed at improving task calibration—especially for sentiment prediction, which still lags behind in F1-score. Another frontier is the integration of geographic embeddings or knowledge graphs to provide richer spatial context, which could bolster performance in town identification and reduce confusion among similar locales.

Furthermore, improvements in class balance through data augmentation or sophisticated sampling techniques may mitigate skewed distributions in polarity and town labels. A comprehensive error analysis leveraging attention visualization or misclassification diagnostics could guide targeted refinements. Finally, experimenting with more expressive or hierarchical classification heads—potentially involving non-linear or deeper architectural blocks—may enhance the model's capability to capture complex dependencies and nuanced expression across tasks.

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] 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.

[3] 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.

[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] 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.

[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] 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.

[9] 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).

[10] 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).

[11] 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.

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

[13] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Science 50 (2024) 568–589.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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).

[19] 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.

[20] O. G. Toledano-López, M. Á. Álvarez-Carmona, J. Madera, A. Simón-Cuevas, Y. A. López-Rodríguez, H. González Diéz, Polarity prediction in tourism cuban reviews using transformer with estimation of distribution algorithms, in: International Workshop on Artificial Intelligence and Pattern Recognition, Springer, 2023, pp. 335–346.

[21] J. D. Jurado-Buch, S. Minayo-Díaz, J. Tello, K. Chaucanes, L. Salazar, M. Oquendo-Coral, M. Á. Álvarez-Carmona, A single model based on beto to classify spanish tourist opinions through the random instances selection, 2023.

[22] 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.

[23] J. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, Lexical complexity assessment of spanish in ecuadorian public documents, Procesamiento del Lenguaje Natural 74 (2025) 291–303.