# Multimodal Sentiment Analysis in Spanish Tourist Reviews: A Data Quality-Aware Approach

Tolulope Olalekan Abiola[1,*,†], Tewodros Achamaleh[1,‡], Olumide Ebenezer Ojo[1,†], Olaronke Oluwayemisi Adebanji[1,§], Oluwatobi Joseph Abiola[2,†], Temitope Dasola Ogunleye[3,‖] and Sidorov Grigori[1,†]

[1]*Centro de Investigación en Computación, Instituto Politécnico Nacional, CDMX, Mexico.*
[4]*Federal University Oye-Ekiti, Ekiti, Nigeria*
[3]*Ladoke Akintola University of Technology, Ogbomoso, Nigeria*

**Abstract**

This work presents baseline approaches for a Spanish-language classification task involving sentiment polarity, entity type, and town identification. We explore traditional machine learning models and pretrained language models like BERT, evaluating their performance across multiple metrics. Despite the linguistic challenges and class imbalances inherent in the dataset, our models provide competitive results and meaningful insights. The analysis highlights the strengths and limitations of each approach, offering a foundation for future improvements in Spanish-language text classification.

**Keywords**

Spanish NLP, User Review, Transformer Models, Sentiment Analysis, Rest-Mex 2025



## 1. Introduction

The exponential growth of user–generated reviews on platforms such as TripAdvisor, Google Reviews and regional booking sites has transformed the way travellers choose where to stay, dine and sight-see[1]. For destinations, especially those that depend on experience-based differentiation, such as Mexico's *Pueblos Mágicos* programme, understanding what visitors praise or criticise has become a strategic priority [2]. Prior research confirms that sentiments embedded in online narratives shape destination image and revisit intention [3, 4]. In the Mexican context, Monsalve-Pulido *et al.* demonstrated the promise of Spanish multimodal resources, yet their study focused on generic polarity and did not distinguish between attraction types or locations [5, 6, 7, 8]. Other scholars have delved into aspect-based or cross-cultural viewpoints from gastronomy tourism in Sarawak [9] to wellness retreats in the Algarve [10], but few address the compound task of simultaneously predicting sentiment strength *and* fine-grained tourism categories in Spanish.

Existing work also highlights two complementary research gaps. First, most tourism sentiment studies rely on classical machine-learning pipelines with handcrafted features [11, 12], which struggle to generalise across the varied Spanish found in Mexican reviews. Second, while the latest large language models (LLMs) achieve impressive accuracy, their computational footprint and cost can be prohibitive for research groups and tourism boards, as observed by Roumeliotis *et al.* when contrasting GPT-4 omni with BERT [13]. Consequently, there is a pressing need for a resource-efficient, Spanish-native model that can handle the tri-partite Rest-Mex 2025 challenge: (i) predict a five-point polarity score, (ii) classify the review target (*Restaurant*, *Hotel*, *Attractive*), and (iii) identify one of sixty *Pueblos Mágicos*.

*IberLEF 2025, September 2025, Zaragoza, Spain*
*Corresponding author.
†These authors contributed equally.
abiolato92@gmail.com (T. O. Abiola); teddymas97@gmail.com (T. Achamaleh); olumideoea@gmail.com (O. E. Ojo); oluwatobiabiola01@gmail.com (O. J. Abiola); ogunleyetemitopedasola@gmail.com (T. D. Ogunleye); sidorov@cic.ipn.mx (S. Grigori)
0009-0005-2817-6264 (T. O. Abiola)

In this paper, we present a single-encoder baseline that fine-tunes the Spanish whole-word-masking BERT model `dccuchile/bert-base-spanish-wwm-cased` for *all three subtasks*. We frame polarity prediction as ordinal classification and learn town and type labels in parallel via a shared `BertForSequenceClassification` head. Compared with pipelines that train separate models or stages, this unified architecture, like exploiting shared linguistic cues across tasks, reduces inference latency, critical for real-time dashboards used by destination managers, and remains computationally lighter than contemporary LLMs while retaining interpretability through attention analysis.

**Contributions** Our work offers:

- the first joint polarity–type–town benchmark on the Rest-Mex 2025 training corpus using a Spanish transformer;
- an efficient fine-tuning recipe and reproducible code that can serve as a strong baseline for future entrants; and
- an ablation study showing how a single multilingual-aware encoder outperforms classic TF–IDF + logistic regression pipelines [14, 15] while approaching the performance of substantially larger LLMs.

The remainder of the paper is organised as follows: Section 2 reviews related work; Section 3 describes the Rest-Mex dataset and preprocessing; Section and details our modelling pipeline; Section 4 reports experimental results and error analysis; and Section 6 and 7 concludes with future research directions and limitations.

## 2. Recent Work

Sentiment analysis in the tourism domain faces challenges due to the multimodal nature of user-generated content, which often combines text, images, and ratings. Monsalve et al. [5] propose a multimodal sentiment analysis model tailored to Spanish-language tourism data, comprising extraction, classification, fusion, and visualisation phases, alongside a quality evaluation component. Their approach integrates an adapted version of *SenticNet 5* for Spanish and demonstrates that classifiers like Random Forest and SVM perform better on an automatically generated dataset than on manually labelled data. Complementing this, Chen et al. [3] explore cultural differences in tourist perceptions by applying Hofstede's cultural dimensions theory and a two-stage LDA and BERT-BiLSTM framework to analyse sentiment in Chinese and English reviews of Xiamen. Their findings highlight notable cultural variations, offering strategic insights for destination marketing.

The gastronomy tourism industry significantly contributes to local economies, enriches travel experiences, and preserves regional culinary heritage. Understanding customer sentiments is essential for informed business decisions, marketing, and service enhancements. Traditional sentiment analysis methods, however, are often subjective and lack visual clarity. To address this, Razali et al.[9] propose a hybrid framework combining lexicon-based sentiment and emotion analysis with data augmentation and feature engineering to better classify minority sentiment classes. Applied in Sarawak's gastronomy tourism, their system integrates real-time business intelligence visualisation and achieves high performance metrics (accuracy of 0.98; F1/ROC-AUC of 0.99). Complementing this, Sánchez et al.[16] utilise NLP, data mining, and machine learning to analyse unstructured guest reviews, extracting key terms and latent topics to support strategic decisions. Their visual and predictive approaches help managers better understand guest preferences, ultimately improving service and competitiveness.

Sentiment analysis of user-generated content on social media offers valuable insights into wellness tourism destinations. George et al.[10] examined 1,294 TripAdvisor reviews to evaluate tourists' sentiments toward various wellness components in the Algarve. Using text mining techniques, the study identified key motivators, such as spa and massage services, and revealed commonly used positive terms like "great," "love," and "visit," highlighting satisfaction with wellness infrastructure. Similarly, Wang et al.[4] explored destination image in the context of panda tourism at the Chengdu Research Base using Latent Dirichlet Allocation and topic-based sentiment analysis. By comparing English and Chinese

reviews, the study revealed cross-cultural perception differences and emphasised the importance of topic-specific sentiment analysis for shaping targeted destination marketing strategies.

Urban green spaces and cultural heritage sites increasingly attract scholarly interest due to their significance in sustainable tourism. Saoualih et al.[17] investigate visitor sentiment toward the Majorelle Garden in Marrakech, Morocco, using VADER sentiment analysis and LDA topic modelling on TripAdvisor reviews from 2006 to 2023. Their findings highlight the garden's symbolic and cultural value, with predominantly positive sentiments reflecting its role in heritage preservation and sustainable tourism. Similarly, JADS et al.[14] analyse English-language reviews of Indonesia's Nusantara Temples using machine learning classifiers—SGD, Logistic Regression, and SVM—finding LR most effective (91.66% accuracy). While offering insights into international tourists' views, the study notes limitations in local representation and language diversity. Together, these works illustrate how sentiment analysis and text mining can inform strategic management of heritage and cultural tourism attractions in diverse contexts.

Tourism translation and hospitality service analysis both require specialised approaches to address the evolving demands of the global tourism industry. Nasution [18] explores translation methods used in the English version of the brochure Borobudur: Inspiring Historical Heritage, identifying equivalence (28%) as the most frequently applied method, followed by adaptation, transposition, modulation, calque, and borrowing. These methods reflect the need to balance linguistic accuracy with cultural sensitivity. The study underscores the importance of equipping English students with translation strategies that integrate language, culture, and technology. Complementing this, Nawawi [15] employs aspect-based sentiment analysis and zero-shot learning with models like RoBERTa to analyse TripAdvisor reviews of Central Java's tourism services. This approach identifies key satisfaction drivers—such as food, accommodation, and cultural experiences—and demonstrates the value of advanced NLP techniques in enhancing service quality and understanding tourist expectations. Together, these studies emphasise the role of linguistic and technological competencies in improving tourism communication and visitor satisfaction.

Understanding tourist sentiment and satisfaction is vital for enhancing tourism experiences and informing destination management. Fu et al.[19] analysed online reviews of Wuyishan National Park using LDA and SnowNLP to explore human–nature interactions. Their findings highlight predominantly positive emotions, with key themes including natural and cultural heritage, tourism activities, and service infrastructure. The study recommends transforming the park into a multifunctional space that combines recreation with environmental education. Similarly, Huang et al.[20] applied LDA and SnowNLP to 26,966 online reviews of ethnic tourism attractions in Guizhou Province, identifying six critical satisfaction factors: experience, transportation, management, commercialisation, natural scenery, and ticket price. The study contributes to a deeper understanding of tourist satisfaction through data-driven insights and highlights the value of online reviews over traditional survey methods. Together, these works demonstrate the effectiveness of sentiment analysis and topic modelling in capturing tourists' perceptions, guiding service improvements, and promoting sustainable tourism development.

The growing influence of social media and digital storytelling has transformed tourism marketing and sentiment analysis. Kumar et al.[21] conducted sentiment classification on user-generated reviews from various tourism platforms, integrating TF-IDF and GloVe embeddings for feature extraction and evaluating multiple machine learning models. Random Forest achieved the highest accuracy (89.54%), demonstrating the effectiveness of combining traditional and semantic features with ensemble learning for robust sentiment analysis. Similarly, Singgalen[12] applied the CRISP-DM framework and VADER sentiment analysis to travel vlog reviews of Gili Trawangan, using an SVM model enhanced by SMOTE. The model reached 88.57% accuracy and 90.95% precision, with results revealing a strong positive sentiment toward vlog content. These studies underscore the power of both textual and visual user-generated content in shaping tourist perceptions and emphasise the importance of advanced data mining techniques in tourism research.

The analysis of online tourism reviews offers valuable insights into visitor satisfaction and preferences, though their unstructured nature poses analytical challenges. Iswari et al.[11] address this by integrating

semantic similarity techniques into aspect-based sentiment analysis, focusing on key aspects such as scenery, dusk, surf, amenities, and sanitation. Their approach enhances F-Measure scores, particularly for aspects with high lexical variability, demonstrating improved sentiment classification through nuanced semantic understanding. In parallel, Roumeliotis et al.[13] compare sentiment classification performance between traditional models like BERT and advanced large language models (LLMs), including GPT-4o and GPT-4o mini, using few-shot learning on hotel reviews. GPT-4o achieved superior accuracy (67%) over BERT (60.6%), highlighting LLMs' capacity for deeper contextual comprehension. Together, these studies underscore the value of semantic-aware NLP and LLMs in refining sentiment analysis, offering powerful tools for data-driven decision-making in tourism and hospitality.

As sentiment analysis increasingly supports decision-making across diverse domains such as tourism, law, and healthcare, concerns over model interpretability have become more pronounced. Hinojosa-Cardenas et al. [22] propose an explainable sentiment analysis approach for restaurant reviews, leveraging an evolutionary algorithm to address transparency and trust in sentiment classification. The method involves preprocessing reviews using natural language processing, determining initial sentiment scores through a lexical dictionary, and then applying an evolutionary model to classify sentiment. Crucially, the model highlights influential words based on polarity scores, offering users clear insights into the basis of sentiment predictions. When evaluated against traditional sentiment analysis models, the proposed approach demonstrated competitive performance and improved explainability, making it especially useful in user-centric applications where understanding the rationale behind decisions is essential.

## 3. Methodology

This section outlines the methodological pipeline adopted for the sentiment classification of tourist reviews in the Rest-Mex 2025 dataset. The workflow integrates natural language preprocessing, transformer-based model architecture, and evaluation of sentiment predictions by level, town and tourism type. All experiments were conducted using the Logistic Regression (LR) BERT-based multilingual (BERT) model fine-tuned for Spanish, with downstream tasks framed as multiclass classification.

### 3.1. Dataset Description

In this study, we utilise the official **Rest-Mex 2025** [23] dataset in [24], a large-scale corpus specifically designed to support sentiment analysis research in the Mexican tourism domain. The dataset comprises user-generated content written in Spanish, collected from tourism platforms where tourists express their experiences in restaurants, hotels, and tourist attractions across designated *Pueblos Mágicos* of Mexico. The goal of this research is to classify sentiment polarity using the textual content of reviews and analyse variations across towns and tourism types.

The training dataset consists of 208,051 entries. Each entry includes a review title, the full review text, a numerical polarity label (ranging from 1.0 to 5.0), and metadata about the location and type of tourism service. The polarity score reflects the sentiment of the review, with higher values indicating more positive sentiment. Reviews are labelled with the town, region, and one of three types: *Restaurant*, *Hotel*, or *Attractive* (i.e., tourist attraction).

Table 1 presents an excerpt from the training dataset to illustrate the structure and nature of the data. The textual content of the review has been truncated for formatting purposes.

The test dataset comprises 89,166 unlabeled reviews, representing 30% of the original corpus. These entries include a unique identifier (ID), title, and review text. While sentiment labels are not provided, the structure mirrors the training data, enabling consistent preprocessing and classification.

Both datasets serve as the foundation for model development and evaluation. The classifier used in this study is based on the `BertForSequenceClassification` architecture, specifically the pretrained `dccuchile/bert-base-spanish-wwm-cased` model, selected for its suitability in Spanish-language processing tasks. The review polarity predictions are examined not only globally but also by town and tourism type to derive localised insights on sentiment trends.

**Table 1**
Rest-Mex 2025 Training Dataset

| Title | Review (Truncated) | Polarity | Town | Region | Type |
|---|---|---|---|---|---|
| Mi Lugar Favorito!!!! | Excelente lugar para comer y pasar una buena noche... | 5.0 | Sayulita | Nayarit | Restaurant |
| Lugares interesantes para visitar | andar mucho, así que un poco difícil para personas mayores... | 4.0 | Tulum | Quintana Roo | Attractive |
| No es el mismo Dreams | Es nuestra cuarta visita a Dreams Tulum, elegimos volver por... | 3.0 | Tulum | Quintana Roo | Hotel |
| Un buen panorama cerca de Cancún | Estando en Cancún, fuimos al puerto y tomamos un ferry... | 4.0 | Isla Mujeres | Quintana Roo | Attractive |
| El mejor | Es un lugar antiguo y por eso me encantó, tiene mucha historia... | 5.0 | Pátzcuaro | Michoacán | Hotel |

### 3.2. Preprocessing

The reviews in the dataset are written in Spanish and vary in quality and length. To ensure consistent input to the classification model, a structured preprocessing pipeline was implemented. Each review underwent tokenisation using the Spanish tokeniser associated with the BERT model (`dccuchile/bert-base-spanish-wwm-cased`). In addition, titles and reviews were concatenated to form a single textual input per instance, allowing the model to consider both elements jointly. Basic cleaning operations included the removal of redundant white spaces and character normalisation. Since the model operates on fixed-length inputs, each review was truncated or padded to a maximum token length of 512 using the Hugging Face tokeniser framework.

### 3.3. Model Architecture

We adopted the model built on the `BertForSequenceClassification` architecture from the Transformers library. This variant of BERT includes a classification head on top of the pooled output vector corresponding to the [CLS] token. The model used, `dccuchile/bert-base-spanish-wwm-cased`, is a case-sensitive Spanish model pretrained on whole-word masking, making it well-suited for sentiment tasks in the tourism domain. The output layer is a softmax classifier configured for five classes, corresponding to the five-point sentiment polarity scale in the training data.

### 3.4. Training and Evaluation

We fine-tuned the model on the labelled training dataset of 208,051 instances. Our training process involved stratified splitting of the data into training and validation subsets, with the AdamW optimiser and a learning rate of $2e^{-5}$ over 4 epochs. The loss function used was cross-entropy, standard for multiclass classification tasks. Evaluation metrics include overall accuracy and macro-averaged F1-score to account for class imbalance. We experimented on an Intel Core i9 11th Gen CPU with 32GB of RAM and an RTX3080 16GB NVIDIA GPU to accelerate the experiment with the use of the GPU, and early stopping was employed based on validation performance.

### 3.5. Prediction and Analysis

Following model training, we generated the predictions on the official test dataset comprising 89,166 entries. Each review instance included a unique identifier, title, and review body, but no associated polarity label. We used the trained BERT and LR models to infer sentiment scores, with the class of

highest predicted probability selected as the final label. Results were aggregated at three levels: (1) by Sentiment scores/level (rating 1-5) (2) by tourism type *restaurant*, *hotel*, or *attractive*, and (3) by town, to analyse geographical trends in sentiment distribution. This enabled a multi-perspective evaluation of sentiment across local tourism sectors.

Our outputs from this stage form the basis for downstream visualisations and exploratory data analysis. The sentiment scores, paired with town and type metadata, enable a deeper understanding of how perceptions vary across tourism destinations, further aiding policymakers and stakeholders in improving service quality.

## 4. Results

In this section, we present a comprehensive evaluation of the model variants across three key subtasks: sentiment polarity classification, attraction type classification, and town classification. We compare two fine-tuned models (LR, BERT) and a baseline. Results are reported using standard classification metrics: accuracy, macro F1, precision, and recall.

### 4.1. Sentiment Polarity Prediction Results

We evaluated three models: Baseline, Logistic Regression (LR), and BERT on the 5-class sentiment polarity classification task. Table 2 summarises the macro-level and per-class metrics, including F1 score, precision, and recall.

**Table 2**
Sentiment polarity classification performance across five sentiment classes.

| Model | Macro F1 | Accuracy (%) | MAE | $F1_5$ | $Prec_5$ | $Recall_5$ |
|---|---|---|---|---|---|---|
| Baseline | 0.1584 | **65.54** | **0.551** | 0.792 | 1.000 | 0.655 |
| LR | 0.2004 | 53.48 | 0.753 | 0.703 | 0.758 | 0.656 |
| BERT | 0.2004 | 49.43 | 0.813 | 0.666 | 0.677 | **0.676** |

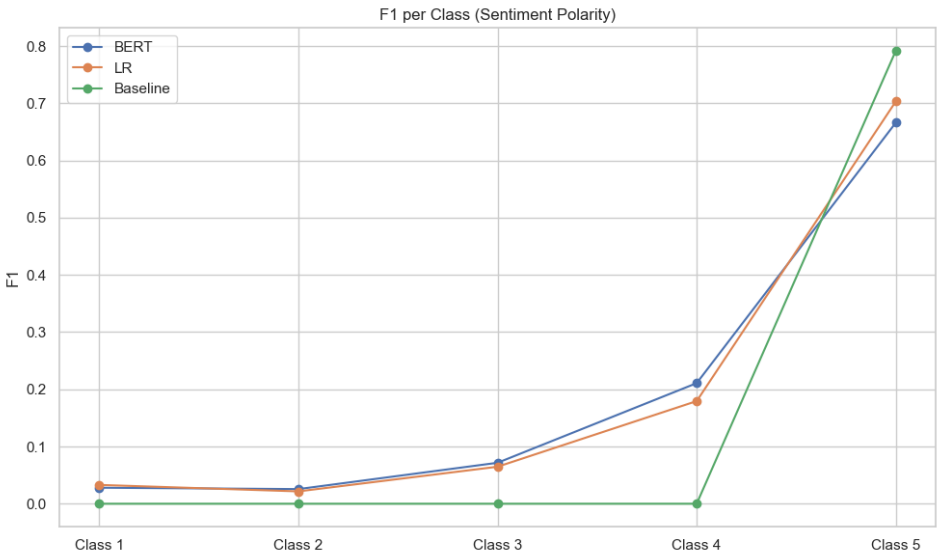


**Figure 1:** F1 score per sentiment class across models.

Although both BERT and LR yield a Macro F1 score of approximately 0.20, their accuracy and MAE differ. Surprisingly, the Baseline model (which always predicts class 5) achieves the highest accuracy (65.54%) and lowest MAE (0.551), but this performance is misleading: it is driven entirely by the dominant class. The macro F1 score reveals its inability to generalise to minority classes.

Figure 1 shows F1 scores per sentiment class. All models exhibit highly skewed performance, with substantial gains on class 5 (strongly positive sentiment) but near-zero performance on minority classes (1–3). BERT marginally outperforms LR on class 4, while LR shows higher recall on class 2.

The figure (Figure 1) further illustrates that both BERT and LR are highly biased toward the majority class. The Baseline model trivially achieves perfect precision by always predicting class 5. BERT, while showing slightly better recall, fails to learn effective boundaries for low-frequency classes. Overall, none of the models generalises well beyond class 5.

## 4.2. Type Classification Performance

This section evaluates the models' performance on the task of type classification, where the goal is to classify review mentions into one of three types: *Attractive*, *Hotel*, or *Restaurant*. We compare the Logistic Regression (LR), BERT-based model, and a rule-based Baseline using multiple evaluation metrics, including F1 Score, Precision, Recall, and overall Accuracy.

**Table 3**
Type Classification Performance (F1, Precision, Recall per class and Accuracy overall)

| Model | Accuracy | F1 Score | | | Precision | | | Recall | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (%) | Attractive | Hotel | Restaurt | Attractive | Hotel | Restaurt | Attractive | Hotel | Restaurant |
| Baseline | 41.87 | 0.000 | 0.000 | 0.590 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.419 |
| BERT | 34.71 | 0.334 | 0.250 | 0.415 | 0.333 | 0.252 | 0.414 | 0.335 | 0.248 | 0.416 |
| LR | 34.91 | 0.334 | 0.247 | 0.420 | 0.333 | 0.246 | 0.422 | 0.335 | 0.249 | 0.417 |

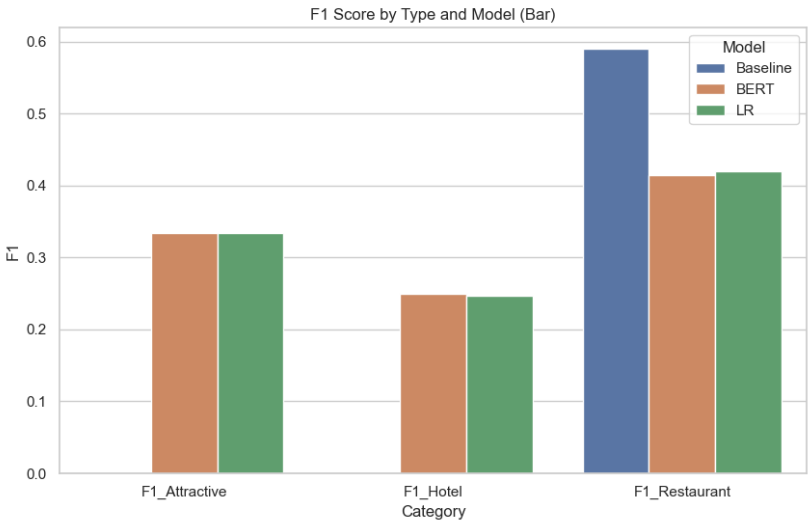


**Figure 2:** F1 score per sentiment class across models.

The baseline system performed relatively well only on the *Restaurant* class, with a high precision of 1.0, but failed on *Attractive* and *Hotel*, leading to an overall moderate accuracy of 41.87%. This suggests the rule-based approach likely defaulted to predicting Restaurant in many cases, skewing the results.

Both the BERT and LR models performed consistently across all three classes. The BERT model achieved F1 scores of 0.334 (Attractive), 0.250 (Hotel), and 0.415 (Restaurant), closely matched by LR with 0.334, 0.247, and 0.420, respectively. Precision and Recall metrics followed similar trends, with BERT slightly favouring recall and LR slightly favouring precision for the Restaurant class.

Interestingly, despite slightly lower accuracy scores ( 34.9%) compared to the baseline, the learned models demonstrated better balance across all classes and avoided the extreme class bias seen in the

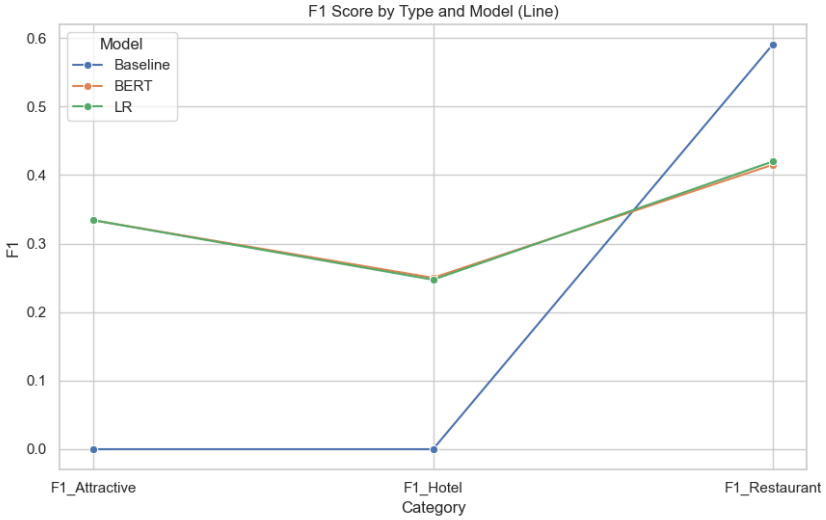


**Figure 3:** F1 score per sentiment class across models.

baseline.

While rule-based methods may yield high metrics in skewed or repetitive cases, machine learning models such as BERT and Logistic Regression provide more reliable performance across diverse review types. Their ability to generalise to multiple classes makes them more suitable for fine-grained type classification in real-world review datasets.

## 4.3. Town Classification Performance

In this subsection, we evaluate the models' performance in classifying reviews by town. The task involves predicting the town associated with each review across more than 60 different locations. Given the granularity and large class count, this is a challenging multi-class classification problem.

Table 4 summarises the key evaluation metrics for each model, including Accuracy, Average Precision, and Average Recall. The Baseline model, which relies on majority class prediction or random guess, achieved a relatively higher accuracy of 21.80% due to class imbalance. However, its precision and recall were very low, indicating its inability to generalise well beyond the dominant classes.

Both BERT and Logistic Regression (LR) models significantly underperformed on accuracy, achieving only 8.84% and 10.29% respectively. Nevertheless, they slightly improved average precision and recall metrics, suggesting better distributional sensitivity despite overall weak performance.

We also evaluated the models on selected towns to observe finer-grained performance. As shown in Figure 4, BERT and LR provided marginal improvements over the Baseline in towns such as *isla_mujeres*, *san_cristobal_de_las_casas*, and *valladolid*. Yet, most towns, including *bacalar* and *tulum*, showed relatively low F1 scores across all models.

The line charts depict the trends in F1 scores across towns. While BERT slightly outperforms the Baseline in nuanced class handling, the overall low F1 values highlight a need for model enhancements, better data balancing, or town-specific feature engineering.

**Table 4**
Summary of Town Classification Metrics

| Model | Accuracy (%) | Avg Precision | Avg Recall |
|---|---|---|---|
| Baseline | 21.80 | 0.0250 | 0.0054 |
| BERT | 8.84 | 0.0250 | 0.0250 |
| Logistic Regression | 10.29 | 0.0252 | 0.0253 |

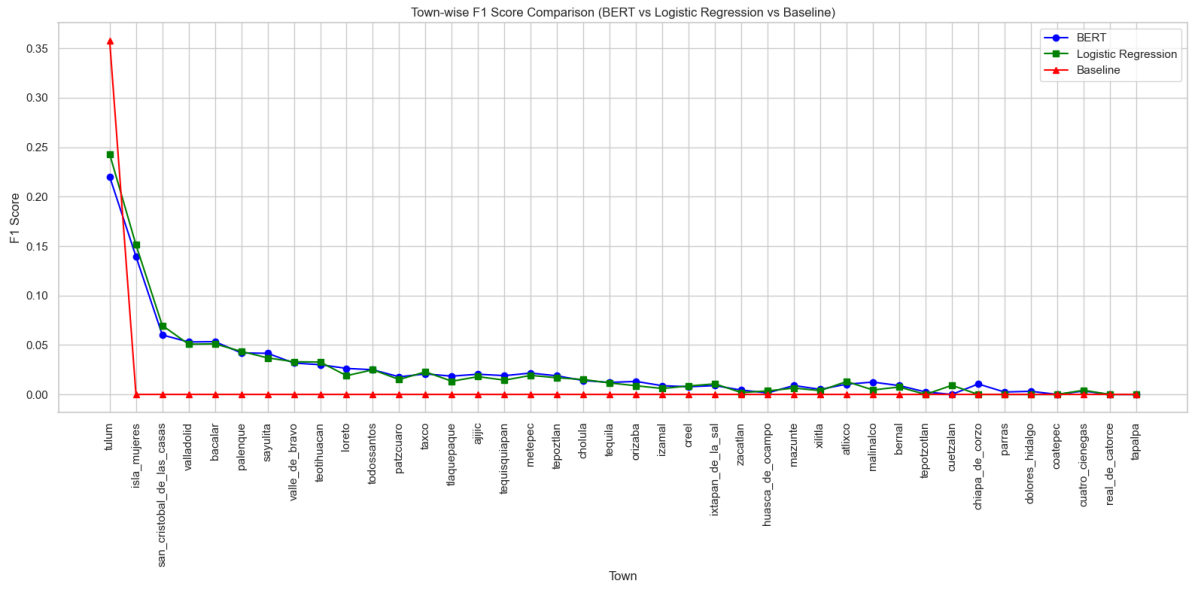


**Figure 4:** F1 score per sentiment class across models.

## 5. Comparison with Other Submissions

To evaluate the performance of our models (BERT and Logistic Regression), we compare them against the full set of submissions using the summary statistics provided. These include minimum, maximum, mean, and standard deviation values for each evaluation metric across all submitted systems. This comparison offers insight into where our models stand relative to the broader range of approaches.

### 5.1. Polarity Classification

Our models (BERT and LR) achieved **Macro F1 scores** of 0.2004 and 0.2004, respectively, which are significantly below the mean of 0.4494 and well outside the one standard deviation range ($\mu \pm \sigma = [0.2717, 0.6270]$). Similarly, their **Accuracy** scores were 49.43% and 53.48%, both below the average accuracy of 61.97%. This suggests that while our approaches performed better than the minimum submitted systems and Baseline, they lag behind most other participants in terms of polarity prediction.

### 5.2. Type Classification

In terms of **Macro F1 (Type)**, both BERT and LR achieved scores around 0.333, which again fall below the average of 0.7962 and outside the standard deviation band ($[0.5095, 1.0830]$). Precision and Recall metrics for the individual types (Attractive, Hotel, Restaurant) followed similar trends, showing underperformance compared to the leaderboard average.

### 5.3. Town Classification

Town classification posed a particularly difficult challenge for our models. The BERT model achieved a **Macro F1** of just 0.0249 and LR similarly scored 0.0248, which are far below the average of 0.4027 and nearly at the minimum end of the distribution (min = 0.0089). This highlights that our current approach struggles significantly in recognising town-level distinctions.

### 5.4. Overall Performance and Observations

Across the board, our models fell short of the average and typical performance (mean $\pm$ std) observed in the shared task. Particularly in the Polarity and Town subtasks, both models operated well below the central range, suggesting limitations in feature representation. In contrast, other submissions on the leaderboard have demonstrated better generalisation, likely due to more sophisticated architectures, domain adaptation, or data augmentation strategies.

In future work, enhancing model complexity (e.g., domain-tuned transformers), integrating external knowledge sources, or performing task-specific fine-tuning could help improve these metrics and bridge the performance gap with top submissions.

## 6. Conclusion

In this paper, we presented our approaches for tackling the multilingual sentiment, type, and town classification tasks. Using baseline models such as BERT and Logistic Regression, we evaluated their effectiveness on the provided dataset and compared our results against the overall leaderboard statistics.

While our models demonstrated moderate performance in certain areas and well above the Baseline, such as polarity classification, they underperformed significantly in more complex tasks like town identification and fine-grained type classification. The comparison with other submissions highlighted the limitations of our models in capturing the diverse linguistic and semantic nuances required by this task.

These findings suggest several directions for future improvement. In particular, leveraging domain-adapted transformer models, incorporating multilingual pretraining, and employing advanced fine-tuning strategies may substantially boost performance. Additionally, enriching the models with external knowledge sources or context-aware embeddings could help bridge the gap observed between our results and those of top-performing systems.

Overall, our participation offered valuable insights into the challenges of multilingual classification and the importance of robust modelling techniques in real-world sentiment and content understanding tasks.

## 7. Limitations

While our models provided a solid starting point for the classification tasks, several limitations constrained their overall performance. The class imbalance observed in both the sentiment and town classification tasks led to biased predictions despite handling it with the AdamW optimiser, with the models favouring dominant classes and struggling to correctly identify less frequent ones. Our approach also relied exclusively on raw textual inputs, without incorporating additional features such as syntactic structures, named entities, or domain-specific signals that could have improved model discrimination. Moreover, due to computational constraints, we limited our experiments to a small set of models and epochs and did not perform extensive hyperparameter optimisation or ensemble learning, which might have enhanced robustness and accuracy. These limitations suggest that while our models establish useful baselines, future work could benefit from more targeted fine-tuning, richer feature representations, and better handling of class imbalance to advance performance in Spanish-language classification tasks.

## 8. Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías

del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

## 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] 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.

[2] 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.

[3] Q. Chen, R. Liu, Q. Jiang, S. Xu, Exploring cross-cultural disparities in tourists' perceived images: a text mining and sentiment analysis study using lda and bert-bilstm models, Data Technologies and Applications 58 (2024) 669–690. doi:10.1108/DTA-10-2023-0645.

[4] Z. Wang, P. Udomwong, J. Fu, P. Onpium, Destination image analysis and marketing strategies in emerging panda tourism: a cross-cultural perspective, Cogent Business & Management 11 (2024). doi:10.1080/23311975.2024.2364837.

[5] J. Monsalve-Pulido, C. A. Parra, J. Aguilar, Multimodal model for the spanish sentiment analysis in a tourism domain, Social Network Analysis and Mining 14 (2024) 46. doi:10.1007/s13278-024-01202-3.

[6] 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.

[7] 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).

[8] 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).

[9] M. N. Razali, S. A. Manaf, R. B. Hanapi, M. R. Salji, L. W. Chiat, K. Nisar, Enhancing minority sentiment classification in gastronomy tourism: A hybrid sentiment analysis framework with data augmentation, feature engineering and business intelligence, IEEE Access 12 (2024) 49387–49407. doi:10.1109/ACCESS.2024.3362730.

[10] O. A. George, C. M. Q. Ramos, Sentiment analysis applied to tourism: exploring tourist-generated content in the case of a wellness tourism destination, International Journal of Spa and Wellness 7 (2024) 139–161. doi:10.1080/24721735.2024.2352979.

[11] N. Iswari, N. Afriliana, E. Dharma, N. Yuniari, Enhancing aspect-based sentiment analysis in

visitor review using semantic similarity, Journal of Applied Data Sciences 5 (2024) 724–735. URL: https://doi.org/10.47738/jads.v5i2.249. doi:10.47738/jads.v5i2.249.

[12] Y. A. Singgalen, Understanding digital engagement through sentiment analysis of tourism destination through travel vlog reviews, KLIK: Kajian Ilmiah Informatika dan Komputer 4 (2024) 2992–3004. URL: https://djournals.com/klik. doi:10.30865/klik.v4i6.1947, open access article under Creative Commons Attribution License.

[13] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, Leveraging large language models in tourism: A comparative study of the latest gpt omni models and bert nlp for customer review classification and sentiment analysis, Information 15 (2024) 792. URL: https://doi.org/10.3390/info15120792. doi:10.3390/info15120792.

[14] H. Hariyono, A. Wibawa, E. Noviani, G. Lauretta, H. Citra, A. Utama, F. Dwiyanto, Exploring visitor sentiments: A study of nusantara temple reviews on tripadvisor using machine learning, Journal of Applied Data Sciences 5 (2024).

[15] I. Nawawi, K. F. Ilmawan, M. R. Maarif, M. Syafrudin, Exploring tourist experience through online reviews using aspect-based sentiment analysis with zero-shot learning for hospitality service enhancement, Information 15 (2024) 499. doi:10.3390/info15080499.

[16] M. J. Sánchez-Franco, S. Rey-Tienda, The role of user-generated content in tourism decision-making: an exemplary study of andalusia, spain, Management Decision 62 (2024) 2292–2328. doi:10.1108/MD-06-2023-0966.

[17] A. Saoualih, L. Safaa, A. Bouhatous, M. Bidan, D. Perkumienė, M. Aleinikovas, B. Šilinskas, A. Perkumas, Exploring the tourist experience of the majorelle garden using vader-based sentiment analysis and the latent dirichlet allocation algorithm: The case of tripadvisor reviews, Sustainability 16 (2024) 6378. doi:10.3390/su16156378.

[18] D. K. Nasution, A. J. Kharisma, Learning translation methods: Study analysis of tourism brochure "inspiration of historical heritage", Jurnal As-Salam 8 (2024). doi:10.37249/assalam.v8i1.680.

[19] W. Fu, B. Zhou, Theme exploration and sentiment analysis of online reviews of wuyishan national park, Land 13 (2024) 629. doi:10.3390/land13050629.

[20] X. Huang, S. Chelliah, Attributes influencing tourist satisfaction: Sentiment analysis and topic modeling of online reviews, Journal of China Tourism Research (2024) 1–20. doi:10.1080/19388160.2024.2440323.

[21] B. S. K. Kumar, M. L. Prajwal, Nivedita, Sentiment analysis of indian tourist place reviews: A machine learning-based exploration, in: 2024 4th International Conference on Intelligent Technologies (CONIT), Bangalore, India, 2024, pp. 1–5. doi:10.1109/CONIT61985.2024.10626602.

[22] E. Hinojosa-Cardenas, L. Laura-Ochoa, E. Sarmiento-Calisaya, Explainable sentiment analysis on restaurant reviews using an evolutionary algorithm, in: J. Lossio-Ventura, E. Ceh-Varela, E. Díaz, F. Paz Espinoza, C. Tadonki, H. Alatrista-Salas (Eds.), Information Management and Big Data. SIMBig 2024, volume 2496 of *Communications in Computer and Information Science*, Springer, Cham, 2025. URL: https://doi.org/10.1007/978-3-031-91428-7_22. doi:10.1007/978-3-031-91428-7_22.

[23] M. A. Alvarez-Carmona, A. Díaz-Pacheco, A. Y. Aranda, L. Bustio-Mart, V. Herrera-Semenets, Overview of rest-mex to pass 2025: Research sentiment evaluation in text for mexican magical town, 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), volume 75 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[24] 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.