Balanced Bag-of-Words Classification of Spanish Tourism Reviews Using Random Forests

Luisa Agudelo Fuentes[1,*], Rodrigo Sebastián Rojas Miranda[2]

[1]Universidad Iberoamericana, México
[2]Instituto tecnológico de toluca, México

Abstract

This paper presents a classical machine learning approach for sentiment and thematic classification of Spanish-language tourist reviews, using a bag-of-words representation filtered by part-of-speech and trained with Random Forest classifiers. To address the severe class imbalance in the Rest-Mex 2025 dataset, we applied random undersampling, equalizing the number of instances per class to that of the minority class. Additionally, we reduced linguistic noise by limiting the input features to nouns and verbs only. The resulting vectors were used to train independent models for polarity, type, and town classification. Despite the simplicity of the approach, it achieved a macro F1-score of 0.198 for sentiment polarity, 0.331 for type classification, and 0.025 for town classification. These results are considerably lower than those reported by Transformer-based models; however, our method offers transparency, interpretability, and computational efficiency. As such, it serves as a useful baseline for low-resource scenarios or educational settings where interpretability and deployment cost are prioritized over peak performance.

Keywords

Random Forest, Class Balancing, Bag-of-Words, Part-of-Speech Filtering, Sentiment Analysis, Spanish NLP, Low-Resource NLP, Rest-Mex 2025.

1. Introduction

The rise of user-generated content on platforms such as TripAdvisor has created new opportunities for understanding traveler preferences, satisfaction, and perceptions of destinations [1, 2]. These platforms capture vast amounts of textual data in natural language, offering a valuable resource for both tourism analytics and computational linguistics [3]. However, this data is often noisy, unstructured, and imbalanced, presenting considerable challenges for automated processing, especially in low-resource settings or in underrepresented languages like Spanish [4].

Sentiment analysis and thematic classification are essential tasks within this context. In Spanish-language tourism, the *Rest-Mex Shared Task* series has played a central role in defining benchmarks and stimulating research. Launched in 2021, Rest-Mex initially focused on polarity classification and satisfaction prediction using TripAdvisor reviews from Mexican destinations [5]. In 2022, the task expanded to include a third track on classifying COVID-19 epidemiological risk levels from news articles, further diversifying its application scenarios [6, 7]. The 2023 edition introduced data from Cuba and Colombia and incorporated unsupervised clustering as a fourth task, while maintaining polarity and service type classification as core components [8]. Now in its fourth edition, Rest-Mex 2025 presents a more complex challenge by including fine-grained town classification over 40 Mexican "Pueblos Mágicos," thereby emphasizing geographically-aware sentiment modeling [9, 10].

While recent winning systems at Rest-Mex have relied heavily on Transformer-based architectures, such as BETO [11], not all deployment contexts have access to high-end GPUs or sufficient infrastructure for training or inference with large-scale neural models. Moreover, in some applications—such as educational settings, embedded systems, or public sector deployments—model interpretability and computational efficiency may outweigh the need for state-of-the-art performance.

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
A2617335@correo.uia.mx (L. A. Fuentes); L24280633@toluca.tecnm.mx (R. S. R. Miranda)

In this paper, we revisit classical machine learning techniques to provide a lightweight and interpretable baseline for sentiment and thematic classification. Specifically, we train Random Forest classifiers using bag-of-words representations constructed from syntactically filtered texts[12]. By limiting the vocabulary to only nouns and verbs (extracted via part-of-speech tagging), we aim to capture content and action-oriented signals while reducing the noise associated with less informative word classes.

To address the class imbalance characteristic of the Rest-Mex dataset, we apply random undersampling so that each class has the same number of training instances as the smallest class. This uniform sampling is performed independently for each task—polarity (1–5), type (Hotel, Restaurant, Attractive), and town (40 categories).

Although this method cannot match the performance of modern Transformer-based approaches, it offers advantages in simplicity, explainability, and deployment efficiency. Our results demonstrate its viability as a baseline system, particularly in low-resource environments where transparency and accessibility are priorities.

2. Related works

Text classification has traditionally relied on linear models and decision tree ensembles trained over frequency-based representations such as bag-of-words (BoW) or term frequency-inverse document frequency (TF-IDF). These methods are computationally efficient, easy to interpret, and suitable for a wide range of text mining tasks. In particular, Random Forest classifiers [13] have proven robust in noisy domains due to their ensemble nature and ability to handle non-linear boundaries without extensive hyperparameter tuning.

To improve text representation, linguistic preprocessing techniques such as part-of-speech (POS) filtering have been proposed. Selecting only nouns and verbs has been shown to preserve key semantic and action-related content while reducing dimensionality and noise [14]. This is particularly relevant in opinion mining, where sentiment-bearing words often correlate with content (nouns) and sentiment-laden actions (verbs).

However, classical models face limitations in handling polysemy, long-range dependencies, and contextual shifts. These issues motivated the emergence of deep learning approaches—first via word embeddings like Word2Vec and GloVe, and more recently with Transformer-based architectures such as BETO [4] or RoBERTa [15]. These models have demonstrated superior performance in virtually every NLP task, including sentiment analysis, named entity recognition, and thematic classification.

The *Rest-Mex Shared Task* has tracked this evolution since its inception. In 2021, participating systems primarily relied on BoW and embedding-based models for polarity classification and recommendation scoring [5]. By 2022, Transformer-based architectures had become dominant, especially in the new COVID-19 risk classification task [6, 7]. In 2023, the top-ranking system used a domain-adapted version of RoBERTa fine-tuned on tourism data, further confirming the importance of specialized pretraining [8].

Despite their accuracy, these models require substantial computational resources and infrastructure for training and deployment. In contrast, classical models remain relevant in scenarios where efficiency, interpretability, or hardware limitations are critical—particularly in educational, governmental, or embedded system contexts.

Our work contributes to this ongoing conversation by revisiting a traditional pipeline—POS-filtered bag-of-words with Random Forests—and testing its viability on the challenging Rest-Mex 2025 benchmark. Although not competitive with neural models in terms of raw performance, this approach offers clarity, speed, and fairness through class-balanced training, making it a useful alternative or baseline in constrained environments.

3. Methodology

Our classification pipeline is designed to be simple, transparent, and computationally lightweight. It consists of the following stages: dataset balancing, part-of-speech-based preprocessing, feature extraction via bag-of-words, and model training using Random Forest classifiers. This section details each component.

3.1. Dataset and Class Distribution

We use the official training split of the Rest-Mex 2025 dataset, which contains over 200,000 Spanish-language tourist reviews labeled for three classification tasks: sentiment polarity (five levels), service type (three categories), and geographic destination (forty towns). Table 1 presents the class distribution for each task in raw counts and relative frequencies.

Label	Class	Count	Percent (%)
Sentiment Polarity (1–5)	1	5,441	2.61
	2	5,496	2.64
	3	15,519	7.46
	4	45,034	21.64
	5	136,561	65.65
Service Type	Hotel	51,410	24.71
	Restaurant	86,720	41.68
	Attractive	69,921	33.61
Town Labels (top 3)	Tulum	45,345	21.8
	Isla Mujeres	29,826	14.3
	San Cristóbal de las Casas	13,060	6.3

Table 1: Rest-Mex 2025 dataset class distribution before balancing

As shown above, all three classification axes are imbalanced, particularly sentiment polarity and town labels. To ensure fair learning, we apply random undersampling independently for each task so that every class contains the same number of instances as the least frequent one. For example, all polarity classes are subsampled to 5,441 reviews (equal to the number in class 1).

3.2. Text Preprocessing and POS Filtering

Each review is processed using a Spanish part-of-speech tagger. We retain only tokens tagged as nouns or verbs, removing adjectives, adverbs, function words, and punctuation. This linguistic filtering reduces vocabulary size and focuses on content-bearing words and action verbs, which are often more informative for tourism sentiment and topic classification.

The filtered tokens are lowercased and lemmatized to normalize inflections (e.g., *comer*, *comió*, *comiendo* $\rightarrow$ *comer*). No additional stopword removal or stemming is applied, to avoid discarding potentially meaningful tourism-related terms.

3.3. Bag-of-Words Feature Representation

We use a simple binary bag-of-words (BoW) model to convert the filtered text into feature vectors. Each document is represented as a sparse vector indicating the presence or absence of a word in a fixed vocabulary. The vocabulary is constructed from the training split after POS filtering, keeping the top 5,000 most frequent lemmas across the corpus.

This BoW representation is interpretable and compatible with tree-based models. It also allows us to inspect which nouns and verbs are most relevant for each class prediction through feature importance scores.

3.4. Classification with Random Forests

We train a separate Random Forest classifier for each task (polarity, type, town). Each model uses 100 decision trees, Gini impurity as the splitting criterion, and no depth restriction. The balanced class distribution allows us to skip class weighting and directly optimize accuracy and macro-averaged F1-score.

All classifiers are trained on 80% of the balanced dataset and evaluated on the remaining 20% using stratified splits. No external resources or embeddings are used, keeping the method fully self-contained and computationally efficient.

4. Results

We evaluated our approach across the three classification tasks defined in the Rest-Mex 2025 challenge: sentiment polarity (5 classes), service type (3 classes), and tourist town identification (40 classes). All models were trained on class-balanced subsets using only nouns and verbs as input features in a bag-of-words representation, and evaluated on held-out validation data.

Table 2 summarizes the macro-averaged F1-scores obtained for each task. While the performance remains modest compared to Transformer-based baselines, the results demonstrate the viability of interpretable, resource-light pipelines in multilingual tourism opinion classification.

Task	**Macro F1-score**
Polarity Classification (5 classes)	0.198
Service Type Classification (3 classes)	0.331
Town Identification (40 classes)	0.025

Table 2: Macro F1-scores across tasks using Random Forest on POS-filtered BoW

4.1. Polarity Classification

Despite the use of balanced training data, the polarity classifier struggled to distinguish between nuanced sentiment levels. Performance was particularly low on the negative and neutral classes (F1 $\approx$ 0.02–0.07), while the model showed slightly better ability in detecting positive (F1 $\approx$ 0.21) and very positive (F1 $\approx$ 0.65) opinions. This asymmetry may reflect the limitations of BoW in capturing subtle linguistic signals like irony, intensifiers, or negation.

4.2. Service Type Classification

Service type classification achieved the highest F1-score among the three tasks. The model was moderately effective in distinguishing between Hotel, Restaurant, and Attractive categories (macro F1 = 0.331). This is likely due to the presence of characteristic nouns (e.g., *habitación*, *comida*, *vista*) that provide strong lexical cues.

4.3. Town Classification

Identifying the town mentioned in the review proved to be the most difficult task, with a macro F1-score of just 0.025. Although some highly frequent towns (e.g., Tulum, Isla Mujeres) were recognized better than others, most low-frequency classes exhibited near-zero performance. This highlights a fundamental limitation of BoW models: their inability to capture entity-level semantics or geographical relationships.

4.4. Efficiency and Practicality

From a computational perspective, the entire training and evaluation pipeline completed in under five minutes on a standard laptop (Intel i7, 16 GB RAM). This reinforces the main advantage of the proposed method: efficiency and ease of deployment. While accuracy is sacrificed, the cost of experimentation is significantly reduced.

5. Conclusions

In this study, we explored a resource-efficient approach to multilingual text classification in the tourism domain, using a classical machine learning pipeline based on bag-of-words representations filtered by part-of-speech and Random Forest classifiers. By balancing the dataset through random undersampling and restricting input features to content-bearing words (nouns and verbs), we aimed to reduce noise and emphasize interpretability.

Our results on the Rest-Mex 2025 dataset show that, while such classical approaches are far from achieving state-of-the-art performance—particularly in complex tasks like fine-grained sentiment analysis or town identification—they offer a lightweight and transparent alternative in scenarios with limited computational resources. The best results were obtained in service type classification, suggesting that even simple lexical patterns can be sufficient for detecting domain-specific categories.

Nonetheless, the substantial gap between our method and Transformer-based architectures highlights the inherent limitations of bag-of-words models, including their inability to capture context, semantics, or syntactic dependencies. Future work may explore hybrid strategies that combine classical interpretability with lightweight neural embeddings, or evaluate the trade-offs between model complexity and portability in real-world tourism applications.

Ultimately, this work contributes a strong, interpretable baseline for Spanish-language opinion classification, and underscores the value of accessible methods in multilingual NLP, especially for educational and low-resource settings.

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. 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.

[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] 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.

[4] 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.

[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) 289–299.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[14] M. Gamon, A. Aue, S. Corston-Oliver, E. Ringger, Pulse: Mining customer opinions from free text, in: Advances in Intelligent Data Analysis VI: 6th International Symposium on Intelligent Data Analysis, IDA 2005, Madrid, Spain, September 8-10, 2005. Proceedings 6, Springer, 2005, pp. 121–132.

[15] 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).