Balanced Text Classification for Tourism Data in Mexico Using NLP and Custom Sampling Techniques

Manuel Enrique Balan-Euan[1,†], Emmanuel Arturo Torres-Santana[2,*,†]

[1]*Instituto Tecnológico de Mérida, Ingeniería en Sistemas Computacionales, Mérida, Yucatán, México*

Abstract

This study addresses balanced text classification in the Mexican tourism context using Natural Language Processing (NLP) and custom sampling strategies. Based on the Rest-Mex 2025 corpus, which contains over 200,000 annotated reviews, supervised models were trained to predict sentiment polarity, business type (hotel, restaurant, or attraction), and geographic location (state and municipality). A two-phase class balancing strategy was implemented: redundancy reduction in overrepresented classes using similarity metrics (Jaccard and Fuzzy), and synthetic data generation for underrepresented classes using controlled domain-specific vocabularies. The text preprocessing pipeline included normalization, lemmatization, and formatting for FastText compatibility. Trained models demonstrated strong performance for business type classification (F1=0.9687), moderate performance for geographic location prediction (F1=0.6397), and more challenging results for sentiment analysis (F1=0.4403). This approach enhances fairness and applicability in multi-label, multilingual AI models for tourism-related text analysis and supports downstream tasks such as recommendation systems and public policy insights.

Keywords

Natural Language Processing, Data Imbalance, Tourism, Text Classification, Mexico

1. Introduction

In the field of artificial intelligence (AI), the performance of supervised learning models largely depends on the balance and quality of the training data. One of the most persistent challenges is data imbalance, where certain classes are significantly overrepresented or underrepresented compared to others. This issue is particularly critical when developing models intended for classification tasks across multiple categories and regions [1, 2, 3].

This work focuses on building a text classification model capable of analyzing tourist-related content and assigning it a polarity score from 1 to 5. The model also predicts whether the text refers to a restaurant, hotel, or tourist attraction, and determines its location among 40 Mexican states and their municipalities [4, 5]. Addressing the imbalance in the dataset was essential to improve the accuracy and generalization of the model.

To mitigate the imbalance, we employed a two-fold strategy. When overrepresented classes were identified, we applied a Jaccard distance metric to identify and randomly remove redundant samples until reaching the class mean. In contrast, for underrepresented classes, we leveraged preprocessed text data (already tokenized, lemmatized, and stripped of stop words, numbers, and punctuation) and expanded it by including contextually related words of similar category and length, supported by a custom vocabulary bank aimed at secure learning.

This problem is relevant in real-world AI applications where data collection does not always ensure class uniformity. In our case, the tourism domain suffers from such imbalance, potentially biasing models toward majority classes and limiting their usability in diverse environments[6, 7, 8, 9, 10].

Preliminary results show promising performance, including confusion matrices and predictions on a final test set of 89,000 records. Training was conducted on a dataset of approximately 200,000 entries, using an 80/20 split for training and validation purposes.

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

halomanue20@gmail.com (M. E. Balan-Euan); LE21081354@merida.tecnm.mx (E. A. Torres-Santana)

https://github.com/ManuBalan03/ (M. E. Balan-Euan); https://github.com/SystemTrabu (E. A. Torres-Santana)

© 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_paper12.pdf

This study contributes a practical methodology that combines advanced Spanish-language NLP preprocessing with FastText-based classification to build a multilingual, multi-label tourism sentiment model for Mexico.

2. Text Classification

Text classification is a fundamental task in natural language processing (NLP) that involves assigning predefined categories to textual data. In this project, the goal is to classify tourism-related user comments or descriptions according to three main aspects: sentiment polarity (from 1 to 5), category (restaurant, hotel, or attraction), and geographical location (states and municipalities of Mexico).

This multi-label classification problem requires a robust model capable of understanding contextual cues in the text and correctly mapping them to the corresponding labels. In tourism, such classification can support recommendation systems, improve customer experience, and assist government or private entities in analyzing public perception.

3. Natural Language Processing

To ensure high-quality input for training our classification models, a comprehensive text preprocessing pipeline was implemented. The process involved several Natural Language Processing (NLP) libraries and custom strategies to clean and normalize Spanish-language tourist reviews.

Initially, each review was transformed to lowercase and stripped of trailing whitespace. We used regular expressions to replace date patterns and long numerical sequences with placeholder characters to prevent misleading token frequency patterns. Then, we applied two normalization strategies from the `spanlp` library: `NumbersToVowelsInLowerCase` and `NumbersToConsonantsInLowerCase`. These strategies replaced digits with semantically plausible letters, which helped maintain syntactic structure without numerical noise.

Accents and diacritics were removed using Unicode normalization (`unicodedata`), followed by filtering all non-alphabetical characters. Tokenization was performed using the NLTK library, and Spanish stopwords were removed based on an extended list. After this step, we applied lemmatization using the `spaCy` model `es_core_news_sm`, which converts each token to its base form, improving the semantic consistency of inputs.

This preprocessing logic was encapsulated in a class named `Preprocesador`, which was used to transform each row in the dataset before being written to three FastText-compatible training files: one for sentiment polarity, one for business type (restaurant, hotel, or attraction), and one for geographic location (state and municipality). Only entries with valid and non-empty reviews and labels were included.

The resulting processed datasets were saved in plain text files following the FastText input format, with lines like:

```
__label__positivo excelente comida y servicio atencion rapida
__label__hotel lugar tranquilo y limpio en el centro
__label__yucatan-merida paseo cultural interesante y economico
```

These datasets were then used to train three independent classification models using the FastText library, achieving efficient training with n-gram features and dimensional embeddings. All these steps were crucial in preparing data that is both clean and semantically rich for downstream classification tasks.

This FastText-compatible structure allowed efficient supervised training using n-gram word representations and low-dimensional embeddings, while remaining computationally lightweight and suitable for large-scale experimentation.

4. Data Imbalance and Bias

One of the main challenges in designing supervised classification models is class imbalance, a condition where some labels have significantly more samples than others. In this project, we observed this phenomenon especially in classes representing small municipalities or less common types of businesses within the tourism sector, such as "attraction."

Imbalance can generate biases in machine learning, favoring dominant classes during training and reducing the model's ability to generalize to minority classes. To mitigate this problem, we developed a two-phase balancing strategy: redundancy reduction in overrepresented classes and data augmentation for underrepresented classes.

First, for classes with samples above the global average (calculated as the mean of the number of instances per label), we removed redundant examples using a hybrid text similarity approach. Both Jaccard similarity, based on sets of tokens, and fuzzy similarity using Python's `SequenceMatcher` algorithm were used. These metrics allowed us to detect highly similar or near-duplicate phrases, which were eliminated to retain only unique and representative examples.

Subsequently, for underrepresented minority classes, new synthetic examples were generated. Using a JSON dictionary with representative vocabulary per category, random phrases were synthesized through controlled word sampling, ensuring semantic consistency. These phrases were tokenized and formatted in the style required by FastText, preserving the structure of the original corpus.

Finally, all classes were truncated or augmented until they equaled the average number of instances per class, thus generating a balanced dataset. This procedure helped reduce class bias during training, improving prediction fairness for lower-frequency categories.

This approach not only improves overall model performance but also increases the interpretability and reliability of predictions, especially in sensitive applications such as tourism and localized services.

In total, X redundant examples were removed and Y synthetic examples were generated, resulting in Z uniformly distributed examples per class.

5. Tourism Context

This study is based on the **Rest-Mex 2025** corpus, a large-scale dataset of tourist reviews focused on the most iconic and visited towns across Mexico. The dataset consists of 208,051 annotated entries, each representing a tourist's opinion and metadata collected from multiple sources. It was released as part of the "Sentiment Analysis Magical Mexican Towns" research initiative, and is intended exclusively for academic and research purposes.

Each review includes a title, a textual review, and three key labels:

- **Polarity**: A sentiment score from 1 (very dissatisfied) to 5 (very satisfied).
- **Type**: The category of place described, labeled as `Hotel`, `Restaurant`, or `Attractive`.
- **Geographic Location**: The name of the town and its corresponding region (state) in Mexico.

The corpus spans opinions from 40 carefully selected Mexican towns, such as Tulum, Isla Mujeres, San Cristóbal de las Casas, and Valladolid — places known for their cultural, historical, or ecological significance. These locations are distributed across 40 different states of Mexico, reflecting a wide geographic and touristic diversity. Tulum alone accounts for over 45,000 reviews, while towns like Tapalpa and Real de Catorce have under 1,000 reviews each, illustrating a natural class imbalance in the distribution of the data.

From a classification perspective, this dataset presents a multi-label challenge that involves:

1. Assigning a sentiment polarity (ordinal classification).
2. Identifying the type of business or site (nominal classification).
3. Predicting the corresponding municipality and state (geographic classification).

Such a setting closely simulates real-world tourism dynamics where both subjective experiences (e.g., satisfaction) and structured information (e.g., location and service type) coexist. The dataset not only enables experiments in multilingual NLP and sentiment analysis but also provides a practical foundation for regional tourism recommendation systems or public policy insights.

All preprocessing and model training in this research strictly adhere to the terms of academic use specified by the Rest-Mex 2025 initiative.

6. Methodology

Our methodology consists of three main phases: text preprocessing, class balancing, and model training. Each phase is designed to prepare the dataset and mitigate issues such as noise, imbalance, and redundancy in tourist-related textual reviews.

6.1. Text Preprocessing

To ensure semantic clarity and consistency across the dataset, we implemented a custom preprocessing class named `Preprocesador`, designed specifically for Spanish-language tourist reviews. The main steps of the normalization pipeline are illustrated in Algorithm 1.

Algorithm 1 Normalization Pipeline

1: Remove dates and long numerical patterns using regex
2: Replace digits using `NumbersToVowelsInLowerCase`
3: Normalize Unicode accents (NFKD form)
4: Tokenize and lemmatize using spaCy (`es_core_news_sm`)
5: Filter out Spanish stopwords using an extended list

This pipeline ensures that the resulting tokens retain only meaningful, context-rich elements suitable for classification.

6.2. Class Balancing Strategy

Let n_i be the number of samples in class c_i, and let $\bar{n}$ denote the global average number of samples across all classes. The balancing procedure is defined by:

$$\text{Action}(c_i) = \begin{cases} \text{Redundancy Reduction}, & \text{if } n_i > \bar{n} \\ \\ \text{Data Augmentation}, & \text{if } n_i < \bar{n} \end{cases} \tag{1}$$

Redundancy was addressed by comparing intra-class instances using a hybrid similarity metric:

$$\text{Sim}(d_i, d_j) = \alpha \cdot \text{Jaccard}(d_i, d_j) + (1 - \alpha) \cdot \text{FuzzyMatch}(d_i, d_j) \tag{2}$$

Algorithm 2 describes the pruning process applied to high-similarity entries.

Algorithm 2 Redundancy Pruning

for each pair (d_i, d_j) in c_i **do**
    Compute hybrid similarity $\text{Sim}(d_i, d_j)$
    if $\text{Sim} > 0.8$ **then**
        Remove instance d_j
    end if
end for

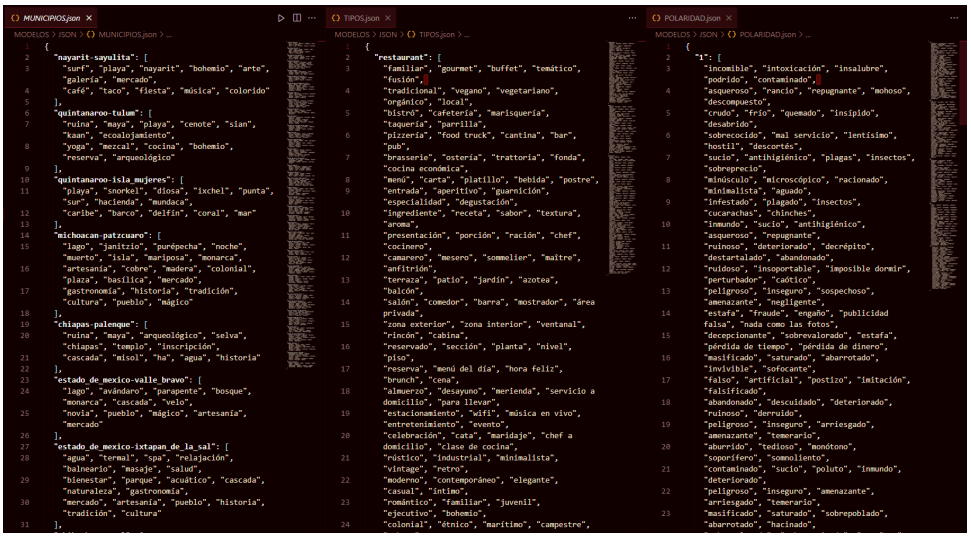

Figure 1: Categorized vocabulary used for controlled augmentation

6.3. Synthetic Data Generation

For underrepresented classes, new synthetic samples were created using a domain-specific vocabulary stored in structured JSON files. Random sequences of words were generated while preserving semantic consistency, as illustrated in Figure 1.

All classes were finally normalized to a balanced state of approximately $\bar{n}$ samples.

7. Mathematical Approach

This section formalizes the key mathematical and algorithmic components of the proposed methodology, from text preprocessing to data balancing and model training using FastText.

7.1. Text Normalization and Preprocessing

Tourist reviews were preprocessed using a custom class `Preprocesador`, which implemented the following steps:

1. Conversion to lowercase and trimming of whitespace.
2. Replacement of dates and long numerical sequences using regular expressions.
3. Normalization via Unicode NFD to remove accents and non-ASCII characters.
4. Digit substitution using two transformation strategies: `NumbersToVowelsInLowerCase` and `NumbersToConsonantsInLowerCase`.
5. Tokenization using `nltk.tokenize.word_tokenize`.
6. Stopword filtering using an extended list of Spanish stopwords.
7. Lemmatization using SpaCy's `es_core_news_sm`.

7.2. Data Balancing with Similarity Metrics

Let $D = \{d_1, d_2, \ldots, d_n\}$ be the set of reviews in class C. For each class C_i, we define its cardinality $n_i = |C_i|$ and compute the global mean $\bar{n} = \frac{1}{K}\sum_{i=1}^{K} n_i$ where K is the total number of classes.

Each class was processed as follows:

$$\text{Action}(C_i) = \begin{cases} \text{PruneRedundancy}(C_i), & \text{if } n_i > \bar{n} \\ \text{AugmentSynthetic}(C_i), & \text{if } n_i < \bar{n} \end{cases}$$

Redundancy Detection: A hybrid similarity score was computed between pairs of sentences using:

$$\text{Sim}(d_i, d_j) = \alpha \cdot \text{Jaccard}(d_i, d_j) + (1 - \alpha) \cdot \text{Fuzzy}(d_i, d_j)$$

where:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad \text{Fuzzy}(s_1, s_2) = \frac{2 \cdot \text{match}(s_1, s_2)}{|s_1| + |s_2|} \tag{3}$$

Synthetic Augmentation: For underrepresented classes, phrases were generated from a domain-specific vocabulary V_c associated with category c, where:

$$\text{frase}_i = \text{random\_sample}(V_c, k), \quad 5 \leq k \leq 10$$

This ensured semantic relevance while compensating for data scarcity.

7.3. Supervised Classification with FastText

We trained three separate supervised classifiers using FastText:

- **Sentiment Polarity Model** on a 5-point ordinal scale.
- **Business Type Model** (Hotel, Restaurant, Attraction).
- **Geographic Model** (State and Municipality).

FastText's softmax formulation estimates the probability of class y given input vector x as:

$$P(y|x) = \frac{e^{s(x,y)}}{\sum_{y'} e^{s(x,y')}}$$

where:

$$s(x, y) = \sum_{i=1}^{n} w_{iy} \cdot x_i + b_y$$

with w_{iy} representing the weight of feature x_i for class y, and b_y the bias.

7.4. Prediction Pipeline

Once trained, the models were used to predict each of the three labels per review using:

1. **Model Loading:** FastText models were restored using `fasttext.load_model()`.
2. **Prediction:** Each preprocessed comment was passed through the three classifiers, outputting both labels and confidence probabilities.
3. **Integration:** Final results were saved as an Excel file including predicted labels and their associated probabilities for analysis and visualization.

This mathematically grounded and computationally optimized pipeline enabled accurate and scalable analysis of tourism-related sentiment across multiple dimensions.

8. Evaluation and Results

8.1. Experimental Setup

The models were evaluated on a held-out test set (20% of 208,051 samples) with the following class distributions:

- Business Type: 3 classes (Hotels, Restaurants, Attractions)
- Polarity: 5 classes (Rating scales 1-5)
- Location: 40 classes (Mexican tourist destinations)

8.2. Overall Performance

Table 1 shows macro-averaged results across tasks:

Table 1
Overall Model Performance

Task	Accuracy	F1	Precision	Recall
Business Type	0.9696	0.9687	0.9690	0.9684
Location	0.6657	0.6397	0.6172	0.6771
Polarity	0.5905	0.4403	0.4824	0.4402

8.3. Per-Task Analysis

8.3.1. Business Type Classification

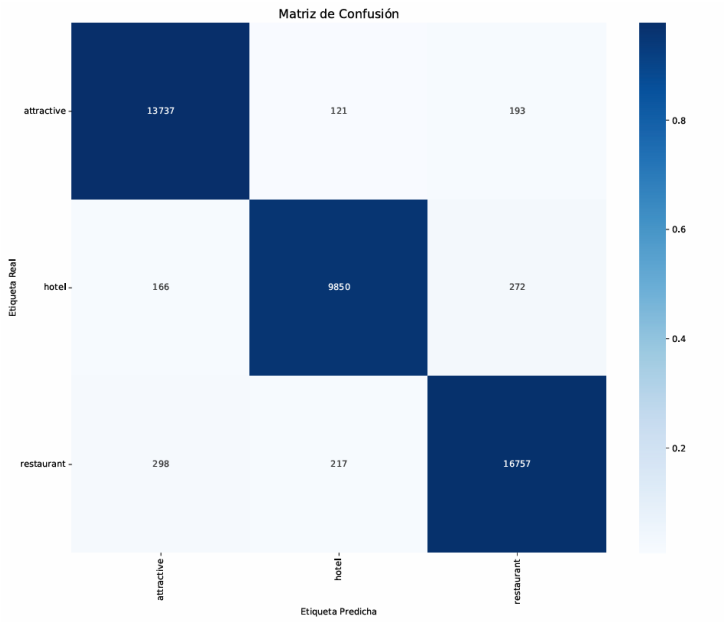

Figure 2: Confusion matrix for business type classification showing near-diagonal dominance

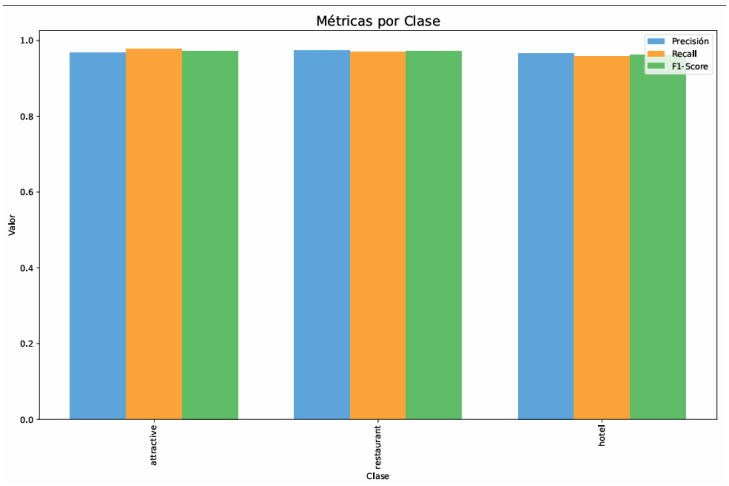

Figure 3: Per-class metrics for business type classification (Precision, Recall, F1)

Key observations:

- Uniform performance across all classes (F1 > 0.96)
- Minimal confusion between categories (Fig. 2)
- Balanced precision-recall tradeoff (Fig. 3)

8.3.2. Sentiment Analysis (Polarity)

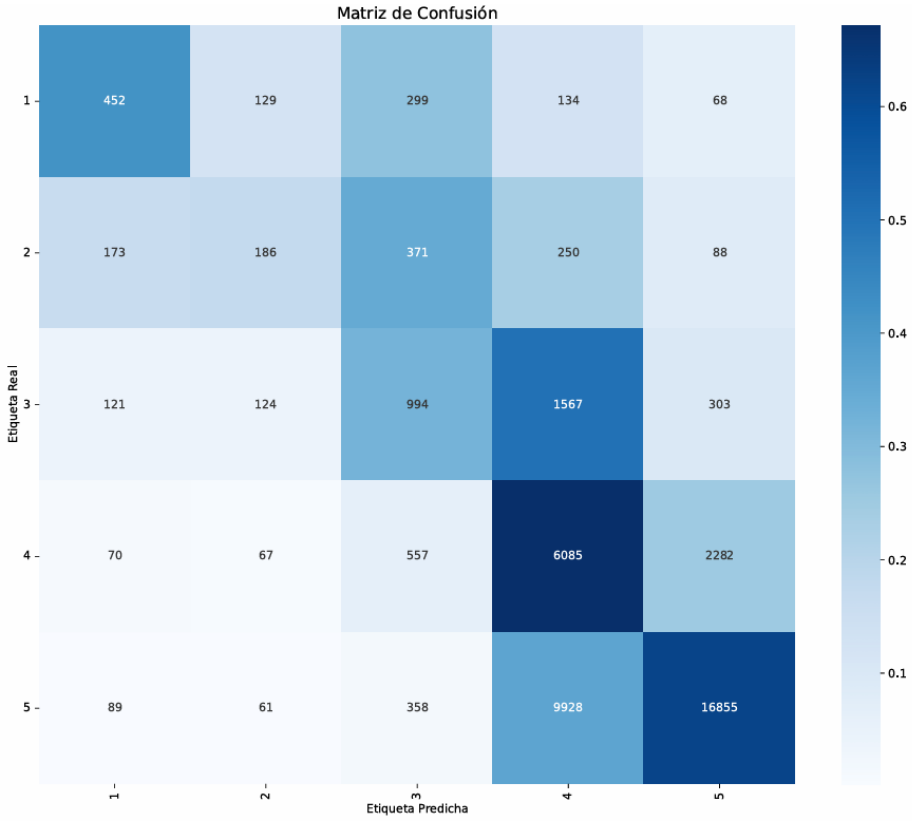

Figure 4: Confusion matrix showing ordinal bias between adjacent classes (3-4)

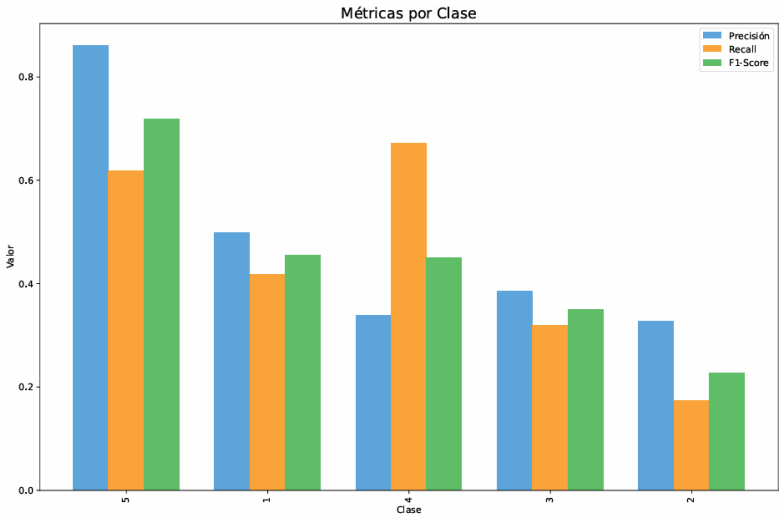

Figure 5: Class-wise metrics revealing sentiment analysis challenges

Notable patterns:

- Strong performance bias toward Class 5 (Fig. 5)
- Frequent 3↔4 misclassifications (Fig. 4)
- Macro F1 depressed by minority classes (2-4)

8.3.3. Location Identification

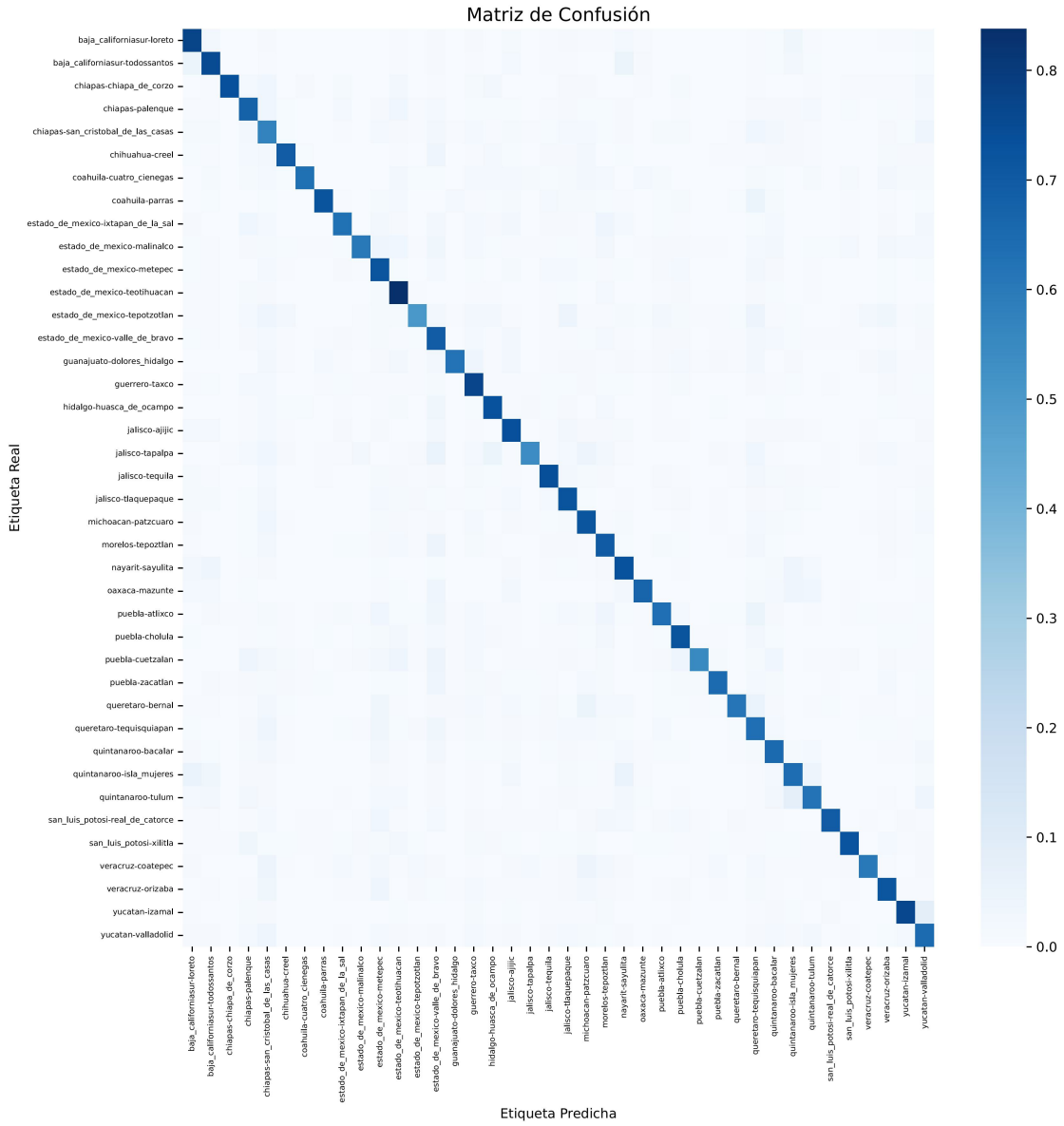

Figure 6: Sparse confusion matrix for 40 location classes

Figure 7: F1-scores for top 15 performing locations

Critical findings:

- High variance in class performance (Fig. 7)
- Phonetic confusion visible in off-diagonals (Fig. 6)
- Top performers: Chiapas-chiapa_de_corzo (F1=0.7577), QuintanaRoo-tulum (F1=0.7413)

8.4. Key Insights

The analysis reveals:

- **Class imbalance** significantly impacts minority classes
- **Ordinal bias** affects sentiment analysis
- **Geographic complexity** requires specialized handling for location names

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, 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.

[2] 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).

[3] 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).

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A. Diaz-Pacheco, M. A. 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 0 (2022) 1–31. doi:10.1080/0952813X.2022.2153276.

[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, International Journal of Interactive Multimedia and Artificial Intelligence In Press (2023) 1–14. URL: http://dx.doi.org/10.9781/ijimai.2023.10.003. doi:10.9781/ijimai.2023.10.003.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.