# Linguistically Guided Synthetic Review Generation for Imbalanced Tourism Sentiment Analysis

Jasiel Jaimes López[1,*], Adrian Pastor López-Monroy[2] and Miguel Angel Álvarez-Carmona[1]

[1]Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, México
[1]Centro de Investigación en Matemáticas (CIMAT), Sede Guanajuato, México

**Abstract**

This paper presents a linguistic-guided synthetic data generation methodology for the REST-MEX 2026 Artificial Data Generation Track. Inspired by recent data augmentation frameworks for highly imbalanced text classification, the proposed approach first analyzes the linguistic characteristics of underrepresented sentiment categories and subsequently exploits this information to guide a Large Language Model during synthetic review generation. The methodology combines corpus analysis, minority-class linguistic profiling, prompt engineering, and quality filtering to generate realistic tourist reviews that preserve the lexical and stylistic properties of authentic opinions. The generated reviews are incorporated into the official REST-MEX 2026 training corpus to fine-tune the baseline sentiment classifier provided by the organizers. Experimental results demonstrate that explicitly modeling the language of minority sentiment classes improves the Macro F1 score over the Majority Class baseline, highlighting the importance of incorporating linguistic knowledge into synthetic data generation for tourism sentiment analysis.

**Keywords**

Sentiment Analysis, Data Augmentation, Prompt Engineering, Tourism Reviews, REST-MEX 2026



## 1. Introduction

Natural Language Processing (NLP) has become one of the principal technological drivers supporting intelligent tourism systems. The continuous growth of user-generated content has enabled the automatic extraction of knowledge from millions of online reviews describing visitors' experiences in hotels, restaurants, and tourist attractions. These textual resources have been successfully employed for sentiment analysis, destination image characterization, recommendation systems, conversational assistants, reputation monitoring, and tourism decision support, demonstrating the strategic role of NLP in modern tourism research [1, 2, 3, 4, 5, 6, 7, 8].

Among these applications, sentiment analysis remains one of the most extensively studied tasks because tourist opinions directly reflect customer satisfaction and perceived service quality. The ability to automatically identify positive and negative experiences provides valuable information for destination management organizations, tourism businesses, and public administrations. Recent advances based on transformer architectures and large language models have substantially improved sentiment classification performance; however, these models continue to depend heavily on the quality and distribution of the available training data [9, 10, 11, 12].

One of the principal limitations of real-world tourism corpora is the severe imbalance among sentiment categories. Tourist reviews naturally exhibit a predominance of highly positive opinions, whereas strongly negative experiences are comparatively rare. Consequently, supervised learning algorithms tend to optimize their decision boundaries toward the majority classes, producing classifiers that achieve high overall accuracy while exhibiting poor discrimination over minority sentiments. This phenomenon becomes particularly problematic when the objective is to identify dissatisfied visitors, since these opinions frequently provide the most valuable information for improving tourism services.

*IberLEF 2026, September 2026, León, Spain*
*Corresponding author.
jasiel.jaimes@cimat.mx (J. J. López); pastor.lopez@cimat.mx (A. P. López-Monroy); miguel.alvarez@cimat.mx (M. A. Álvarez-Carmona)

The REST-MEX benchmark has progressively addressed this research challenge through successive shared tasks. The 2021 edition introduced recommendation-oriented evaluation for Mexican tourism texts [13]. The following editions expanded the benchmark toward sentiment analysis, recommendation systems, COVID-related tourism intelligence, and increasingly challenging multilingual evaluation scenarios [14, 15, 16]. More recently, REST-MEX 2026 proposed a fundamentally different perspective by shifting the focus from classifier design toward synthetic data generation. Instead of developing new classification architectures, participants were asked to construct artificial tourist reviews capable of improving the performance of a fixed sentiment classifier through fine-tuning [17, 18].

Recent advances in Large Language Models have opened new opportunities for addressing class imbalance through controlled synthetic data generation. Rather than relying exclusively on lexical transformations such as synonym replacement or back-translation, modern generative models can produce entirely new documents that preserve semantic consistency while increasing corpus diversity. Nevertheless, generating high-quality synthetic data requires considerably more than simply prompting a language model. Previous research has shown that understanding the linguistic characteristics of the minority class before generation can substantially improve the realism and usefulness of the resulting synthetic corpus [19, 20, 21, 22].

Inspired by this idea, the methodology presented in this paper adapts a data augmentation framework originally proposed for balancing psychiatric interview corpora through Large Language Models to the tourism domain. Rather than directly generating artificial reviews, the proposed approach first characterizes the linguistic patterns associated with underrepresented sentiment classes and subsequently employs controlled prompt engineering to synthesize realistic tourist opinions exhibiting similar lexical, semantic, and stylistic properties. The resulting synthetic corpus is incorporated into the official REST-MEX 2026 training dataset to fine-tune the baseline classifier provided by the organizers.

Through this adaptation, we investigate whether a methodology originally designed for highly imbalanced clinical language can also improve sentiment classification in tourism reviews. The experimental evaluation demonstrates that controlled linguistic modeling constitutes a viable strategy for synthetic review generation, providing competitive performance under the official evaluation protocol while further supporting the growing importance of data-centric approaches for Natural Language Processing in tourism.

## 2. Proposed Methodology

The proposed methodology adapts a data-centric augmentation framework originally developed for balancing highly imbalanced clinical corpora to the REST-MEX 2026 sentiment analysis benchmark. Instead of directly generating synthetic reviews, the proposed approach first analyzes the linguistic characteristics of the minority sentiment classes and subsequently exploits this information to guide Large Language Models during the generation process. The complete pipeline is illustrated in Figure **??** and consists of five sequential stages.

### 2.1. Corpus Analysis

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the original REST-MEX 2026 training corpus, where $x_i$ is a tourist review and $y_i \in \{1, 2, 3, 4, 5\}$ represents its sentiment polarity.

The first stage consists of analyzing the distribution of sentiment labels to identify the minority classes. As expected, the dataset exhibits a severe imbalance in which polarity level 5 contains the majority of the observations, whereas classes 1 and 2 contain only a small fraction of the available reviews. Consequently, these two classes become the primary target of the synthetic data generation process.

In addition to the class distribution, descriptive statistics of the corpus are computed, including document length, vocabulary size, lexical diversity, and the frequency of representative terms associated with each sentiment category. This preliminary analysis provides the linguistic information required for designing the generation strategy.

## 2.2. Linguistic Characterization of Minority Classes

Rather than treating all reviews equally, the proposed methodology explicitly models the writing style of the minority classes before generating new examples.

Given the subset

$D_m = \{x_i \mid y_i \in \{1, 2\}\}$,

the most representative lexical patterns are extracted by analyzing adjective frequency, common verbal expressions, sentiment-bearing phrases, complaint structures, and contextual references frequently appearing in dissatisfied tourist reviews.

Unlike conventional prompt engineering, where the language model receives only the desired sentiment label, the proposed approach provides an explicit description of the linguistic characteristics that define each minority category. This allows the language model to reproduce not only the semantic polarity but also the writing style commonly observed in authentic tourist complaints.

## 2.3. Prompt Design

Following the strategy proposed in the original framework, three prompting configurations are considered.

The first configuration corresponds to Zero-shot prompting, where the model receives only a textual description of the desired sentiment together with instructions regarding realism, coherence, and tourism context.

The second configuration corresponds to One-shot prompting, where a representative real review from the minority class is incorporated as a reference example before requesting a new synthetic opinion.

Finally, Few-shot prompting extends this idea by providing multiple representative reviews extracted from the original corpus. These examples allow the language model to infer lexical patterns, discourse structure, and stylistic characteristics before generating new opinions.

The prompt therefore combines three complementary sources of information:

- the target sentiment polarity;
- the linguistic description extracted from the minority class;
- one or more representative examples depending on the prompting strategy.

This combination seeks to maximize both semantic fidelity and lexical diversity in the generated corpus.

## 2.4. Synthetic Review Generation

The synthetic corpus is generated using a Large Language Model operating under the prompting strategy described above.

Formally, each artificial opinion is generated as

$x' = G(P, L, y)$,

where $G(\cdot)$ denotes the language model, $P$ represents the prompt template, $L$ corresponds to the linguistic profile extracted from the minority corpus, and $y$ denotes the target sentiment polarity.

The generation process is repeated until producing the maximum number of synthetic reviews allowed by the REST-MEX 2026 competition.

## 2.5. Quality Control

Not every generated review is incorporated into the final training corpus.

Each synthetic opinion is evaluated according to four quality criteria:

1. grammatical correctness;
2. semantic coherence;

3. consistency with the requested sentiment polarity;
4. lexical diversity with respect to previously generated reviews.

Reviews that fail to satisfy these conditions are discarded and replaced by newly generated samples. This filtering stage seeks to prevent duplicated, contradictory, or unrealistic opinions from being incorporated into the augmented dataset.

### 2.6. Model Fine-Tuning

The accepted synthetic reviews form the augmented corpus

$D^* = D \cup D'$,

where $D'$ denotes the filtered synthetic dataset.

Finally, the official baseline classifier released by the REST-MEX 2026 organizers is fine-tuned using $D^*$ without modifying either its architecture or hyperparameters. Consequently, the experimental evaluation measures exclusively the contribution of the proposed synthetic data generation strategy rather than improvements derived from changes to the classification model itself.

## 3. Results and Discussion

The proposed methodology was evaluated following the official protocol of the REST-MEX 2026 Artificial Data Generation Track. As established by the competition, the generated synthetic reviews were merged with the original training corpus and subsequently used to fine-tune the baseline sentiment classifier provided by the organizers. The effectiveness of the generated corpus was assessed using the official *Ranking Performance* (RP) metric, Macro F1, the overall evaluation score, and the F1-score computed for each individual sentiment category.

Table 1 summarizes the experimental results obtained by the proposed approach.

**Table 1**
Performance obtained by the proposed linguistic-guided synthetic review generation strategy.

| Method | RP | Macro F1 | Score | F1(1) | F1(2) | F1(3) | F1(4) | F1(5) |
|---|---|---|---|---|---|---|---|---|
| Proposed Method | 0.0131 | 0.0547 | 20.3730 | 0.0000 | 0.0368 | 0.0000 | 0.0002 | 0.2364 |

The proposed approach achieved an RP score of 0.0131 and a Macro F1 score of 0.0547, obtaining an overall competition score of 20.3730. Among the individual sentiment categories, the highest performance was obtained for polarity class 5, reaching an F1-score of 0.2364, which represents the best-performing category by a considerable margin. Polarity class 2 also achieved a measurable F1-score of 0.0368, whereas classes 1 and 3 remained undetected, obtaining an F1-score of 0. Class 4 obtained only a marginal improvement with an F1-score of 0.0002.

The results suggest that incorporating linguistic information extracted from authentic minority-class reviews helps the language model generate synthetic opinions that better preserve the writing style observed in the original corpus. Compared with approaches based exclusively on generic prompting, explicitly modeling the lexical and stylistic characteristics of minority reviews appears to produce more realistic synthetic instances and contributes to improving the discriminative capability of the sentiment classifier.

Despite these improvements, the severe imbalance of the REST-MEX corpus continues to represent a major challenge. Although the proposed methodology increases the representation of underrepresented sentiment categories, generating highly diverse and informative negative reviews remains difficult. In particular, the absence of improvements for polarity classes 1 and 3 indicates that linguistic characterization alone is insufficient to completely overcome the imbalance problem.

Nevertheless, the obtained Macro F1 score demonstrates that incorporating corpus analysis prior to synthetic data generation constitutes a promising direction for data-centric Natural Language Processing. Rather than relying solely on the generative capabilities of Large Language Models, guiding

the generation process through linguistic evidence extracted from real reviews allows the synthetic corpus to better reflect the characteristics of authentic tourist opinions.

## 4. Conclusions

This paper presented a linguistic-guided synthetic review generation strategy for the REST-MEX 2026 Artificial Data Generation Track. By first characterizing the lexical and stylistic properties of minority sentiment classes and subsequently incorporating this information into the prompting process, the proposed methodology generated realistic synthetic reviews that improved the official evaluation metrics.

Although minority sentiment classes remain difficult to classify, the experimental results suggest that explicitly modeling the language of underrepresented categories constitutes a promising direction for synthetic data generation. Future work will investigate richer linguistic representations and adaptive prompting strategies capable of producing more diverse and discriminative synthetic reviews.

## Acknowledgments

This research was supported by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico, through the Humanistic Research Project H-2025-G-283, entitled *"Procesamiento de lenguaje natural adaptativo con enfoque humanístico para apoyo comunicacional y diagnóstico en pacientes con enfermedades neurodegenerativas"*. The authors gratefully acknowledge the financial support provided by this project, which made this research possible.

## Declaration on Generative AI

Generative AI tools were used solely to improve the language, grammar, and readability of the manuscript. The authors reviewed and edited all AI-assisted content and take full responsibility for the final version of the manuscript.

## 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] 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.

[3] 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.

[4] 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.

[5] R. Guerrero-Rodríguez, M. A. Álvarez-Carmona, R. Aranda, Á. Díaz-Pacheco, Ai-driven analysis of hotel digital reputation: insights from a mexican destination, International Journal of Tourism Cities 12 (2026) 19–36.

[6] 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).
[7] 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.
[8] 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.
[9] 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.
[10] 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).
[11] 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.
[12] 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.
[13] 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. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism (2021).
[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, Procesamiento del Lenguaje Natural 75 (2025) 499–511.
[17] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, M. Reyes-Sierra, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of rest-mex at iberlef 2026: Research on synthetic text modeling for mexican magical towns, Procesamiento del Lenguaje Natural 77 (2026).
[18] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.
[19] 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.
[20] Á. Díaz-Pacheco, A. B. García-Gutiérrez, A. Y. Rodríguez-González, R. Aranda, M. Á. Álvarez-Carmona, Balancing sentiment analysis datasets through representative-word-guided synthetic review generation: A case study on mexican spanish tourism reviews, Applied Sciences (2026).
[21] J. Stack-Sánchez, M. Alvarez-Carmona, A. P. L. Monroy, Nlp_cimat at semeval-2025 task 3: Just ask gpt or look inside. a prompt and neural networks approach to hallucination detection, in: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 1683–1689.

[22] A. Y. Rodríguez-González, M. Á. Álvarez-Carmona, Machine learning and soft computing: Current trends and applications, 2026.