# Automatic Quality Assessment of Synthetic Tourism Reviews using Automatic Filtering Strategies

Angel Díaz-Pacheco[1], Ramón Aranda[2,*]

[1]Universidad de Guanajuato, Guanajuato, México.
[2]Centro de Investigación en Matemáticas, Mérida, México.

**Abstract**

The increasing adoption of Large Language Models (LLMs) has considerably simplified the generation of synthetic textual data. Nevertheless, automatically generated reviews frequently exhibit problems such as duplicated content, inconsistent sentiment, unrealistic descriptions, and low linguistic diversity. These issues become particularly relevant in shared tasks where the quality of the generated corpus directly influences the performance of downstream Natural Language Processing models. This paper presents an automatic quality assessment framework for synthetic tourism reviews generated for the REST-MEX 2026 shared task. Rather than proposing a new generation strategy, the objective is to validate the quality of the generated corpus before it is employed for sentiment classification. The proposed workflow combines multiple verification stages, including structural validation, duplicate detection, lexical diversity analysis, language verification, and sentiment consistency evaluation. Reviews satisfying all quality requirements are incorporated into the final corpus, while inconsistent samples are discarded and regenerated. The proposed framework provides a simple, reproducible, and model-independent quality assurance process that can be integrated into any synthetic review generation pipeline. Experimental evaluation demonstrates that the resulting corpus satisfies the quality requirements established by the REST-MEX 2026 evaluation protocol while preserving realistic tourism language and balanced sentiment distributions.

**Keywords**

Synthetic Reviews, Quality Assessment, Large Language Models, Tourism, REST-MEX, Natural Language Processing



## 1. Introduction

The tourism industry generates enormous amounts of textual information every day through online reservation platforms, social media, travel blogs, and customer review websites. These opinions constitute one of the principal information sources for understanding visitor satisfaction, destination image, service quality, and consumer behavior. Consequently, Natural Language Processing (NLP) has become an increasingly important research area within tourism analytics, supporting applications such as sentiment analysis, recommendation systems, topic discovery, conversational assistants, and destination management [1, 2, 3, 4, 5, 6, 7].

The continuous evolution of transformer architectures has significantly improved the quality of automatic language generation. More recently, Large Language Models (LLMs) have demonstrated the ability to generate coherent documents that closely resemble human-written text. These capabilities have motivated their adoption in numerous domains, including education, healthcare, finance, software engineering, and tourism.

Within tourism applications, synthetic review generation has attracted growing interest because it offers an alternative for mitigating class imbalance, preserving user privacy, and augmenting training datasets. Unlike manually written reviews, synthetic opinions can be generated under controlled conditions, allowing balanced datasets to be constructed while maintaining realistic linguistic characteristics.

The REST-MEX evaluation campaign has played an important role in promoting research on Spanish-language tourism corpora. Previous editions investigated recommendation systems, sentiment analysis,

*IberLEF 2026, September 2026, León, Spain*
*Corresponding author.
angel.diaz@ugto.mx (A. Díaz-Pacheco); arac@cimat.mx (R. Aranda)

and tourism-related text classification [8, 9, 10, 11]. The 2026 edition introduces a new challenge in which participants are required to generate a synthetic corpus capable of training sentiment classifiers evaluated over unseen authentic reviews [12, 13].

Among the different adaptation strategies available for Large Language Models, Supervised Fine-Tuning (SFT) has become one of the most widely adopted. Instead of modifying the architecture of a language model, SFT specializes an existing pre-trained model using domain-specific instruction-response pairs. This adaptation enables the model to better capture the vocabulary, writing style, and semantic characteristics of the target application while requiring relatively modest computational resources.

In this work we present a supervised fine-tuning framework for synthetic tourism review generation. Authentic reviews are transformed into instruction-following examples describing the desired sentiment polarity, attraction category, municipality, and approximate review length. These instruction-response pairs are subsequently used to adapt a pre-trained language model through supervised learning. Once fine-tuned, the model generates balanced synthetic reviews suitable for the REST-MEX 2026 shared task.

The main contributions of this paper are summarized as follows.

- We present a complete supervised fine-tuning workflow for adapting Large Language Models to tourism review generation.
- We describe an automatic instruction generation strategy based on the metadata available in the REST-MEX corpus.
- We present a practical pipeline for producing balanced synthetic reviews using a specialized language model.
- We evaluate the generated corpus according to the official REST-MEX 2026 evaluation protocol.

## 2. Large Language Models for Tourism Text Generation

The rapid development of Large Language Models (LLMs) has significantly transformed Natural Language Processing by enabling high-quality text generation across a wide range of domains. Recent models are capable of producing coherent, context-aware, and linguistically fluent documents from relatively simple textual instructions. These capabilities have motivated their adoption in applications including education, healthcare, software development, scientific writing, and customer service.

Within tourism, Natural Language Processing has experienced remarkable growth during the last decade. Online reviews have become one of the primary sources of information for understanding customer satisfaction, destination image, travel behavior, and service quality. Consequently, numerous studies have applied machine learning and deep learning techniques to sentiment analysis, recommendation systems, destination image assessment, conversational agents, topic discovery, and digital reputation analysis [1, 2, 3, 4, 14, 15, 5, 6, 16, 7, 17].

The continuous evolution of transformer-based architectures has further expanded these applications. Instead of limiting their use to classification tasks, current research increasingly explores text generation, automatic summarization, question answering, and conversational assistants specialized in tourism. Language models are now capable of generating destination descriptions, travel itineraries, customer support responses, and synthetic reviews that closely resemble authentic user-generated content.

The REST-MEX shared task has played an important role in promoting research on Spanish-language tourism corpora. Since its first edition, the challenge has encouraged the development of Natural Language Processing techniques for recommendation systems, sentiment analysis, and tourism intelligence [8, 9, 10, 11]. The 2026 edition introduces a substantially different scenario in which participants are required to generate synthetic reviews instead of directly classifying existing documents [12, 13]. This new objective highlights the importance of controllable text generation techniques capable of producing realistic, balanced, and semantically consistent tourism reviews.

### 2.1. Supervised Fine-Tuning

Although pre-trained language models possess extensive linguistic knowledge, they are usually trained on general-purpose corpora collected from books, web pages, and online repositories. Consequently, they may not accurately reproduce the vocabulary, writing style, or discourse patterns observed in specialized domains such as tourism.

Supervised Fine-Tuning (SFT) has become one of the most widely adopted strategies for adapting Large Language Models to domain-specific applications. Instead of modifying the model architecture, SFT updates the model parameters using instruction-response pairs representative of the target domain. This adaptation enables the model to internalize specialized terminology while preserving the general language understanding acquired during pre-training.

Within tourism, supervised fine-tuning offers several advantages. The model learns how travelers naturally describe accommodations, restaurants, attractions, transportation, customer service, and overall travel experiences. In addition, metadata associated with each review—such as sentiment polarity, attraction category, municipality, or destination—can be naturally incorporated into the instruction, allowing the model to generate reviews conditioned on specific contextual information.

### 2.2. Instruction Construction

The quality of supervised fine-tuning strongly depends on the construction of instruction-response pairs. Rather than presenting isolated documents, each review is transformed into an instruction describing the characteristics that the generated text should satisfy.

For example, a training instance may request the generation of a positive review describing a restaurant located in a particular Mexican Magical Town. The expected response corresponds to the authentic review extracted from the REST-MEX corpus. Repeating this process for the entire collection produces thousands of supervised examples that teach the language model how to generate tourism reviews under different contextual conditions.

This instruction-based representation preserves both the textual content and the metadata available in the original corpus, facilitating the subsequent generation of balanced synthetic datasets.

### 2.3. Motivation

Recent advances in prompt engineering have demonstrated that carefully designed prompts can substantially improve the quality of generated text. Nevertheless, prompt engineering alone does not modify the internal knowledge of the language model. As a result, the generated reviews may still exhibit inconsistencies in terminology, writing style, or sentiment expression when applied to specialized domains.

In contrast, supervised fine-tuning directly adapts the language model to the linguistic characteristics of tourism reviews. By exposing the model to thousands of authentic examples during training, it gradually learns domain-specific vocabulary, sentence structure, and discourse patterns that are difficult to capture through prompting alone.

Motivated by these observations, this work presents a complete supervised fine-tuning workflow for synthetic tourism review generation. The proposed methodology emphasizes simplicity and reproducibility, describing each stage of the adaptation process from corpus preparation to the generation of the final synthetic dataset. Rather than proposing a new optimization algorithm, the objective is to provide a practical framework that can be easily replicated and extended for future REST-MEX editions or other domain-specific text generation tasks.

## 3. Proposed Methodology

The proposed workflow adapts a pre-trained Large Language Model to the tourism domain through supervised fine-tuning. Rather than modifying the model architecture, the methodology focuses on preparing domain-specific instruction-response pairs that allow the model to learn the linguistic characteristics of authentic tourism reviews.

Figure ?? summarizes the complete workflow, which consists of five stages: corpus preparation, instruction construction, supervised fine-tuning, synthetic review generation, and quality verification.

### 3.1. Corpus Preparation

The official REST-MEX 2026 training corpus was used as the source of authentic reviews. Each review contains textual content together with metadata describing its sentiment polarity, attraction type, municipality, and state.

Before fine-tuning, the corpus was preprocessed using a lightweight normalization procedure consisting of:

- Removal of duplicated white spaces.
- Unicode normalization.
- Elimination of incomplete reviews.
- Standardization of punctuation.
- Preservation of the original capitalization.

Unlike traditional Natural Language Processing pipelines, no stop-word removal or lemmatization was applied because the objective was to preserve the original writing style of authentic users.

### 3.2. Instruction Construction

Each review was automatically transformed into an instruction-response example.

The instruction contains the contextual information required for generation, while the response corresponds to the original review extracted from the dataset.

An example instruction is shown below.

> *Generate a four-star review describing a restaurant located in the Mexican Magical Town of Taxco. The review should contain approximately 80 words and describe a realistic customer experience.*

The corresponding response is the authentic review associated with the selected metadata.

Repeating this process for every review produces a collection of instruction-response pairs suitable for supervised fine-tuning.

### 3.3. Supervised Fine-Tuning

After constructing the instruction dataset, the language model is adapted using supervised learning. During this stage, the model receives an instruction $I_i$ and predicts the corresponding review $R_i$. Model parameters are updated by minimizing the prediction loss over the complete instruction dataset.

Unlike prompt engineering, where the model remains unchanged, supervised fine-tuning gradually adapts the internal representations of the language model to the vocabulary and writing style characteristic of tourism reviews.

### 3.4. Synthetic Review Generation

Once fine-tuning is completed, the adapted model is used to generate the synthetic corpus required by REST-MEX 2026.

Generation is performed by creating instructions describing the desired characteristics of each review.

For every generated instance, the following information is specified:

- Sentiment polarity.

- Attraction type.
- Municipality.
- Approximate review length.

The model then generates a complete review satisfying these constraints.
Generation continues until obtaining the required number of reviews for every sentiment category.

### 3.5. Quality Verification

Every generated review is automatically verified before being incorporated into the final corpus.
Three validation criteria are considered.

1. Minimum review length.
2. Duplicate detection.
3. Verification of the requested sentiment polarity.

Reviews satisfying all criteria are accepted, whereas rejected instances are regenerated using the same instruction.

### 3.6. Implementation Details

The complete workflow was implemented using the Hugging Face Transformers library. Training examples were stored in JSON format following the instruction-response structure commonly adopted by instruction-tuned language models.

Generation was performed using stochastic decoding to encourage lexical diversity while maintaining coherent sentence structure. The same instruction template was employed throughout the generation process, modifying only the metadata describing the desired review.

The proposed methodology is independent of the specific language model employed during fine-tuning and can therefore be applied to different open-source instruction-tuned architectures with minimal modifications.

## 4. Experimental Evaluation

### 4.1. Experimental Configuration

The proposed workflow was evaluated using the official REST-MEX 2026 training corpus. The authentic reviews were automatically converted into instruction-response pairs and used to adapt an instruction-tuned Large Language Model through supervised fine-tuning.

After training, the adapted model generated a balanced synthetic corpus containing 3,000 reviews distributed across the five sentiment levels and the three attraction categories according to the task requirements.

The generated corpus was subsequently employed to train the sentiment classifier required by the official REST-MEX evaluation protocol.

### 4.2. Evaluation Metrics

The generated corpus was evaluated using the official metrics provided by the shared task.
The reported measures include:

- Track Score
- Macro Precision
- Macro Recall
- Macro F1-score

Additionally, two complementary corpus statistics were computed:

- Average review length.
- Vocabulary size.

These measurements provide an overview of both the predictive performance and the linguistic characteristics of the generated corpus.

## 5. Results

Table 1 summarizes the evaluation results obtained using the synthetic corpus generated by the proposed workflow.

**Table 1**
Evaluation results.

| Metric | Value |
|---|---|
| Track Score | 0.516 |
| Macro Precision | 0.541 |
| Macro Recall | 0.538 |
| Macro F1-score | 0.539 |

The obtained results indicate that the generated corpus satisfies the requirements of the REST-MEX 2026 evaluation framework. The specialized language model successfully generated reviews covering the complete range of sentiment categories while preserving coherent tourism-related language.

Table 2 summarizes several descriptive statistics of the generated corpus.

**Table 2**
Characteristics of the generated corpus.

| Characteristic | Value |
|---|---|
| Generated reviews | 3000 |
| Average words per review | 81 |
| Vocabulary size | 14562 |
| Average sentences per review | 4.2 |
| Duplicate reviews | 0 |

The resulting corpus presents a balanced distribution of reviews across all sentiment levels while maintaining a vocabulary representative of tourism-related language. Manual inspection of randomly selected reviews also indicated that the generated texts preserved coherent descriptions of accommodations, restaurants, tourist attractions, and travel experiences.

## 6. Discussion

The proposed workflow demonstrates that supervised fine-tuning constitutes a practical strategy for adapting Large Language Models to tourism review generation. By learning directly from authentic reviews, the model naturally acquires the vocabulary, writing style, and discourse patterns commonly observed in user-generated tourism content.

An important advantage of the proposed workflow is its simplicity. The complete adaptation process only requires authentic reviews together with their associated metadata, making it straightforward to reproduce for other tourism datasets or related application domains.

The instruction construction stage also facilitates the generation of balanced corpora because each generated review is explicitly conditioned on the desired sentiment polarity, attraction type, and municipality. Consequently, corpus generation becomes systematic and reproducible without requiring manually designed prompts for every individual review.

Although this work focuses on tourism reviews, the same workflow can be extended to other domains where labeled textual data are available. Examples include customer feedback, healthcare narratives, educational comments, legal documents, and product reviews.

## 7. Conclusions

This paper presented a supervised fine-tuning workflow for adapting Large Language Models to synthetic tourism review generation under the REST-MEX 2026 shared task.

The proposed methodology transforms authentic tourism reviews into instruction-response pairs that enable domain adaptation through supervised learning. After fine-tuning, the specialized language model generates balanced synthetic reviews conditioned on sentiment polarity, attraction type, municipality, and review length.

The resulting workflow is simple to implement, reproducible, and independent of the specific language model employed during adaptation. Because the methodology relies exclusively on authentic labeled reviews, it can be readily transferred to other specialized domains requiring synthetic text generation.

Future work will investigate parameter-efficient fine-tuning techniques, multilingual adaptation strategies, automatic instruction diversification, and the incorporation of additional contextual information such as traveler profiles, seasonal characteristics, and destination-specific knowledge to further enrich the generated reviews.

## 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] 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).

[6] 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.

[7] 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.
[8] 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.
[9] 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).
[10] 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).
[11] 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.
[12] 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).
[13] 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.
[14] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, Quantifying differences between ugc and dmo's image content on instagram using deep learning, Information Technology & Tourism 26 (2024) 293–329.
[15] A. Diaz-Pacheco, M. A. Álvarez-Carmona, A. Y. Rodríguez-González, H. Carlos, R. Aranda, Measuring the difference between pictures from controlled and uncontrolled sources to promote a destination. a deep learning approach (2023).
[16] I. Castillo-Ortiz, M. Á. Álvarez-Carmona, R. Aranda, Á. Díaz-Pacheco, Evaluating culinary skill transfer: A deep learning approach to comparing student and chef dishes using image analysis, International Journal of Gastronomy and Food Science 38 (2024) 101070.
[17] 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). doi:https://doi.org/10.13053/CyS-26-2-4055.