Targeted Minority-Class Augmentation with ChatGPT for Sentiment Analysis in REST-MEX 2026

Juan David Jurado-Buch[1,2,*]

[1]Universidad de Nariño, Colombia
[2]Universidad Nacional Abierta y a Distancia, Nariño, Colombia

Abstract

Class imbalance remains one of the main challenges in sentiment analysis, particularly in tourism-related corpora where positive opinions substantially outnumber negative reviews. The REST-MEX 2026 Artificial Data Generation Track addresses this problem by encouraging participants to generate synthetic instances capable of improving the performance of a predefined sentiment classification model. In this work, we present a targeted data augmentation strategy based on the generation of minority-class reviews using ChatGPT. Rather than generating synthetic examples for all sentiment categories, our approach focuses exclusively on the two most underrepresented classes of the corpus, corresponding to very negative and negative opinions. A prompt-engineering strategy was designed to produce realistic tourism reviews in Mexican Spanish, resulting in a synthetic dataset composed of 200 additional instances. These reviews were incorporated into the official training corpus and used during the fine-tuning process of the baseline classifier. Experimental results show that the highest F1-scores were obtained for the classes targeted during augmentation, suggesting that even a relatively small number of synthetic examples can influence the learning process of minority categories. The proposed system achieved an RP score of 0.0446, obtaining the 62nd position in the competition. Although the overall performance was modest, the results provide valuable insights into the potential and limitations of LLM-based synthetic data generation for addressing class imbalance in sentiment analysis.

Keywords

Sentiment Analysis, Data Augmentation, Large Language Models, ChatGPT, Class Imbalance, Tourism Analytics, Mexican Spanish, REST-MEX 2026.

1. Introduction

The rapid growth of online tourism platforms has transformed the way travelers share their experiences and evaluate destinations[1][2]. Websites such as TripAdvisor, Booking, and Google Reviews contain millions of user-generated opinions that provide valuable insights into customer satisfaction, service quality, and destination image [3]. These textual resources have attracted considerable attention from the Natural Language Processing (NLP) community, as they offer opportunities to automatically extract knowledge that can support both tourism stakeholders and decision makers[4][5].

Among the various NLP tasks applied to tourism analytics, sentiment analysis has become one of the most widely adopted approaches[6]. By automatically identifying the polarity expressed in textual reviews, sentiment analysis enables the large-scale assessment of visitor experiences that would otherwise require extensive manual effort[7]. In recent years, advances in deep learning and transformer-based architectures have significantly improved the performance of sentiment classification systems across multiple languages and domains[8][9][10]. Nevertheless, achieving robust performance remains challenging when the underlying data distribution is highly imbalanced[11][12].

Class imbalance is one of the most persistent problems in supervised machine learning[13]. In sentiment analysis datasets, positive opinions frequently outnumber negative ones by a substantial margin, causing learning algorithms to favor majority classes while underrepresenting minority categories. As a consequence, models often achieve high overall accuracy while exhibiting poor performance when identifying negative or dissatisfied customer experiences. From a practical perspective, this limitation is

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
jjuradobuch@gmail.com (J. D. Jurado-Buch)

particularly relevant in tourism applications, where negative reviews often contain the most actionable information for service improvement and quality assessment[14].

The REST-MEX evaluation campaign has emerged as an important benchmark [15][16][17][13] for studying sentiment analysis in Mexican tourism reviews. The 2026 edition [7][18] introduces a novel Artificial Data Generation Track, where participants are not asked to directly develop classification models, but instead generate synthetic reviews intended to improve the performance of a predefined sentiment classifier. This formulation transforms the challenge into a data-centric artificial intelligence problem, where the quality and usefulness of the generated data become more important than the classification architecture itself.

A distinctive characteristic of the REST-MEX 2026 corpus is its highly skewed polarity distribution. The training dataset contains 208,051 tourist reviews, of which more than 136,000 belong to the most positive category, while fewer than 6,000 instances are available for each of the two most negative classes. Such imbalance creates a substantial learning challenge, as minority classes represent only a small fraction of the available training data [19]. Consequently, generating additional examples for these underrepresented categories constitutes a promising strategy for improving classifier sensitivity toward negative opinions.

Recent advances in Large Language Models (LLMs) have enabled the generation of coherent, diverse, and contextually relevant synthetic texts. Models such as ChatGPT have demonstrated remarkable capabilities for producing human-like content that can be leveraged for data augmentation tasks. Several studies have shown that carefully designed prompts can guide these models to generate domain-specific examples while preserving semantic consistency and label fidelity. This capability has opened new research opportunities for addressing class imbalance through targeted synthetic data generation [20].

In this work, we investigate a simple yet focused augmentation strategy based on the generation of minority-class reviews using ChatGPT. Rather than distributing synthetic instances across all sentiment categories, our approach concentrates exclusively on the two most underrepresented classes of the REST-MEX 2026 dataset. To achieve this objective, a prompt-engineering strategy was designed to encourage the generation of realistic tourism reviews expressing strong dissatisfaction and dissatisfaction, corresponding to polarity classes 1 and 2, respectively. A total of 200 synthetic instances were generated and incorporated into the training corpus. The central hypothesis of this work is that even a relatively small number of carefully generated minority-class examples may contribute to mitigating the effects of class imbalance and improving sentiment classification performance on negative reviews.

2. Related Work

Sentiment analysis has become one of the most extensively studied tasks within Natural Language Processing, with applications spanning domains such as e-commerce, social media, healthcare, and tourism. Early approaches relied primarily on lexicon-based methods and traditional machine learning algorithms using handcrafted features, including Bag-of-Words representations, term frequency statistics, and sentiment lexicons. Although these methods achieved reasonable performance in specific scenarios, their ability to capture contextual and semantic information was limited.

The emergence of deep learning significantly transformed the field. Neural architectures based on recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and convolutional neural networks (CNNs) demonstrated substantial improvements by learning distributed representations directly from text. More recently, transformer-based models such as BERT and its multilingual variants have established new state-of-the-art results across numerous sentiment analysis benchmarks. In the context of Spanish-language sentiment analysis, models such as BETO and RoBERTa-BNE have shown particularly strong performance due to their pretraining on large Spanish corpora [21].

Within the tourism domain, sentiment analysis has been widely employed to analyze online reviews, evaluate service quality, and understand visitor perceptions of destinations. Several studies have leveraged machine learning and deep learning techniques to classify tourist opinions collected from platforms such as TripAdvisor and Booking, demonstrating the value of automated sentiment analysis for

supporting tourism management and decision-making processes. The REST-MEX evaluation campaigns have further contributed to this area by providing benchmark datasets specifically focused on Mexican tourism reviews, encouraging the development and comparison of specialized approaches for Spanish-language sentiment classification [22].

A recurring challenge in sentiment classification is the presence of class imbalance, where positive opinions substantially outnumber negative ones. Numerous strategies have been proposed to address this issue, including cost-sensitive learning, class reweighting, oversampling techniques, and synthetic data generation. Among these, data augmentation has gained increasing attention due to its ability to enrich minority classes without requiring additional manual annotation [19].

Recent advances in Large Language Models (LLMs) have opened new possibilities for synthetic data generation. Models such as GPT, Gemini, Claude, and LLaMA are capable of producing coherent and contextually appropriate text that can be used to augment training datasets. Several recent studies have reported improvements in downstream classification tasks when synthetic examples generated by LLMs are incorporated into the training process, particularly in low-resource and imbalanced settings. However, the effectiveness of such approaches depends heavily on the quality of the prompts used, the diversity of the generated instances, and the degree to which synthetic examples accurately reflect the characteristics of the target domain [23].

The REST-MEX 2026 Artificial Data Generation Track provides an ideal framework for studying these questions. By evaluating the contribution of synthetic reviews to the performance of a fixed sentiment classifier, the shared task shifts the focus from model architecture design to data-centric artificial intelligence. In this context, our work explores whether a small collection of carefully generated minority-class reviews can help alleviate class imbalance and improve sentiment classification performance in Mexican Spanish tourism reviews.

3. Methodology

The main objective of this work is to mitigate the class imbalance present in the REST-MEX 2026 training corpus through the generation of synthetic reviews belonging exclusively to the minority sentiment categories. To achieve this, we employed a prompt-engineering strategy using ChatGPT to generate artificial tourist opinions that resemble authentic reviews while maintaining the sentiment characteristics associated with the target classes.

The proposed methodology consists of three stages: (i) analysis of the sentiment distribution in the original dataset, (ii) generation of synthetic reviews for the minority classes, and (iii) integration of the generated instances into the training corpus used for model fine-tuning.

3.1. Dataset

The experiments were conducted using the training corpus provided by the REST-MEX 2026 Artificial Data Generation Track. The dataset contains 208,051 tourist reviews collected from Mexican *Pueblos Mágicos*, each annotated with a sentiment polarity label ranging from 1 (very bad) to 5 (very good) [7].

One of the most notable characteristics of the corpus is its highly imbalanced distribution. While the most positive category contains 136,561 instances, the two most negative classes contain only 5,441 and 5,496 examples, respectively. This imbalance motivated our decision to focus exclusively on generating synthetic data for polarity classes 1 and 2.

3.2. Minority Class Augmentation

Rather than generating synthetic examples for all sentiment categories, we adopted a targeted augmentation strategy centered on the two most underrepresented classes. The underlying hypothesis is that improving the representation of negative opinions may help the classifier learn more robust decision boundaries for these difficult categories.

A total of 100 synthetic reviews were generated for polarity class 1 (very bad) and 100 synthetic reviews for polarity class 2 (bad), resulting in a synthetic corpus of 200 additional instances. The generated opinions were designed to emulate realistic tourist experiences involving hotels, restaurants, and attractions commonly found in Mexican tourist destinations.

3.3. Prompt Design

The quality of synthetic data generated by Large Language Models depends heavily on the instructions provided to the model. Therefore, a detailed prompt was designed to encourage the generation of realistic, diverse, and domain-specific tourism reviews.

The prompt instructed ChatGPT to assume the role of a dissatisfied tourist and generate short reviews that could plausibly appear on online review platforms. Additionally, the model was explicitly requested to avoid repetitive expressions, vary the described situations, and maintain consistency with the target sentiment label.

An abbreviated version of the prompt is shown below:

> Generate a realistic tourist review written in Mexican Spanish. The review must correspond to sentiment class [1 or 2], expressing either extreme dissatisfaction or dissatisfaction. The opinion should resemble a genuine online review related to a hotel, restaurant, or tourist attraction in Mexico. Use natural language, avoid repetitive phrases, and describe specific aspects such as service quality, cleanliness, staff behavior, facilities, food quality, waiting times, pricing, or overall experience. The generated text must be coherent, realistic, and clearly aligned with the assigned sentiment category.

This prompt was repeatedly executed until the desired number of instances for each minority class was obtained.

3.4. Synthetic Data Integration

Once generated, the synthetic reviews were assigned their corresponding polarity labels and merged with the original REST-MEX 2026 training corpus. No modifications were performed on the original data, and all synthetic instances were incorporated as additional training examples.

The resulting augmented dataset was then used by the organizers to fine-tune the baseline sentiment classification model. The effectiveness of the proposed augmentation strategy was subsequently evaluated on the hidden test set according to the official REST-MEX 2026 evaluation protocol.

4. Results

The performance of the proposed approach was evaluated using the official metrics defined by the REST-MEX 2026 organizers. Given the strong class imbalance present in the dataset, the competition employed a specialized metric called *Ranking Performance* (RP), which serves as the official leaderboard score. Unlike traditional evaluation measures, RP assigns greater importance to minority classes, reducing the influence of the dominant positive categories and providing a more balanced assessment of sentiment classification performance.

In addition to RP, the evaluation included Macro F1, overall Accuracy, and the F1-score obtained for each sentiment class individually. Macro F1 computes the arithmetic mean of the F1-scores across all classes, giving equal importance to minority and majority categories. Accuracy measures the proportion of correctly classified instances, while the per-class F1-scores provide a detailed view of the classifier's behavior for each sentiment polarity.

Table 1 presents the results obtained by the proposed synthetic data generation strategy.

The proposed system achieved an RP score of 0.0446, which positioned the team in the 62nd place of the competition leaderboard. Although the overall performance was modest, the results provide interesting insights regarding the effectiveness of targeted synthetic augmentation.

Table 1
Official results obtained in the REST-MEX 2026 evaluation.

Metric	Value
RP Score	0.0446
Macro F1	0.0415
Accuracy	0.0513
F1(1)	0.0672
F1(2)	0.0950
F1(3)	0.0000
F1(4)	0.0108
F1(5)	0.0346

The highest per-class performance was obtained for polarity class 2, reaching an F1-score of 0.0950, followed by class 1 with an F1-score of 0.0672. This behavior is particularly noteworthy because these were precisely the classes targeted during the synthetic data generation process. The results suggest that the additional examples contributed to improving the classifier's ability to recognize negative opinions, supporting the hypothesis that focused augmentation can influence learning in underrepresented categories.

However, improvements in the minority classes were not sufficient to substantially increase overall performance. The classifier exhibited limited effectiveness on the remaining sentiment categories, especially class 3, for which no correct predictions were achieved according to the F1-score. This outcome indicates that concentrating exclusively on the most negative classes may reduce the model's ability to adequately represent the complete sentiment spectrum.

Several factors may explain these results. First, only 200 synthetic instances were generated, representing less than 0.1 % of the original training corpus. Second, the generated reviews were produced using a single prompting strategy and a single language model, potentially limiting the diversity of the synthetic examples. Finally, while the generated texts were designed to reinforce minority categories, they may not have fully captured the linguistic variability present in real tourist reviews.

Overall, the results highlight both the potential and limitations of small-scale targeted data augmentation. While the approach appears to have positively influenced the performance of the minority classes, future work could explore larger synthetic datasets, more diverse prompting strategies, and the simultaneous generation of examples across multiple sentiment categories to achieve a more balanced improvement in classification performance.

5. Conclusions

This paper presented a simple data augmentation strategy for the REST-MEX 2026 Artificial Data Generation Track based on the generation of synthetic reviews using ChatGPT. The proposed approach focused exclusively on the two most underrepresented sentiment categories of the training corpus, namely polarity classes 1 and 2, with the objective of mitigating the severe class imbalance present in the dataset.

A total of 200 synthetic reviews were generated through prompt engineering and incorporated into the original training set. The evaluation results showed that the best F1-scores were obtained precisely for the two classes targeted during the augmentation process, suggesting that even a relatively small number of synthetic instances can influence the learning process of minority categories. This observation supports the idea that targeted data generation may be a viable strategy for improving the representation of scarce classes in sentiment analysis tasks.

Despite these encouraging findings, the overall performance remained limited, indicating that small-scale augmentation alone is insufficient to substantially improve classification performance in highly imbalanced datasets. The results suggest that additional factors such as synthetic data diversity, prompt design, generation volume, and class coverage play a critical role in determining the effectiveness of LLM-based augmentation approaches.

Future work could explore the generation of larger synthetic datasets, the use of multiple Large Language Models, more sophisticated prompting strategies, and quality-control mechanisms to increase the diversity and realism of generated reviews. Furthermore, combining targeted minority-class augmentation with examples from intermediate sentiment categories may help achieve a more balanced representation of the sentiment spectrum and ultimately improve overall classification performance.

The REST-MEX 2026 challenge demonstrates the growing importance of data-centric artificial intelligence approaches, where improvements are achieved not only through better models but also through the creation of higher-quality training data. In this context, synthetic data generation using Large Language Models represents a promising research direction for addressing class imbalance in tourism-related sentiment analysis and other low-resource NLP applications.

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] 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.

[2] 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.

[3] 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.

[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] 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.

[6] 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.

[7] 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).

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

[9] 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).

[10] 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.
[11] 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.
[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] 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.
[14] 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.
[15] 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).
[16] 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.
[17] 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.
[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] A. Y. Rodríguez-González, M. Á. Álvarez-Carmona, Machine learning and soft computing: Current trends and applications, 2026.
[21] 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).
[22] 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.
[23] 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.