Prototype-Guided Synthetic Review Generation for Minority Sentiment Augmentation

Juan-Luis García-Mendoza[1,*]

[1]*Pangeanic BI Europe, Spain*

Abstract

This paper presents a prototype-guided synthetic review generation methodology for the REST-MEX 2026 Artificial Data Generation Track. Unlike conventional prompting strategies that generate reviews from scratch, the proposed approach first identifies representative reviews from the minority sentiment classes and subsequently employs them as semantic prototypes to guide a Large Language Model during text generation. Prototype selection is performed through semantic similarity in the embedding space, allowing the generated reviews to preserve the lexical, semantic, and stylistic characteristics of authentic tourist opinions while introducing controlled contextual variations. The resulting synthetic corpus is incorporated into the original REST-MEX 2026 training dataset to fine-tune the official baseline sentiment classifier. Experimental results demonstrate that prototype-guided generation substantially improves the Macro F1-score compared with the official Majority Class baseline, suggesting that representative seed examples constitute an effective mechanism for producing high-quality synthetic reviews for imbalanced tourism sentiment analysis.

Keywords

Sentiment Analysis, Prototype-Guided Generation, Synthetic Text Generation

1. Introduction

The widespread adoption of online tourism platforms has generated an unprecedented volume of user-generated content describing travelers' experiences in hotels, restaurants, and tourist attractions. These reviews constitute an invaluable source of information for understanding customer satisfaction, destination image, service quality, and visitor expectations. Consequently, Natural Language Processing (NLP) has become one of the principal technologies supporting intelligent tourism systems, enabling the automatic extraction of knowledge from textual opinions at a scale that would be impossible through manual analysis [1, 2].

Among the numerous NLP applications in tourism, sentiment analysis has emerged as one of the most extensively studied research topics. The automatic identification of positive and negative opinions provides useful insights for destination management organizations, tourism businesses, and governmental agencies seeking to improve visitor experiences and monitor public perception. Recent advances in transformer-based architectures have significantly increased classification performance, allowing modern models to capture subtle semantic relationships and contextual information within tourism reviews [3, 4, 5].

Despite these advances, sentiment analysis systems remain highly dependent on the quality and representativeness of the available training corpora. Real-world tourism datasets naturally exhibit severe class imbalance, where highly positive reviews largely outnumber negative opinions [6]. As a result, supervised learning algorithms tend to optimize their decision boundaries toward the majority classes, reducing their ability to correctly identify dissatisfied visitors. This limitation has motivated the development of numerous balancing strategies, including resampling techniques, data augmentation, ensemble learning, and more recently, synthetic text generation using Large Language Models [7, 8, 9].

The recent emergence of generative artificial intelligence has shifted the attention from model-centric approaches toward data-centric machine learning [9]. Instead of modifying classifier architectures,

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

garciamendoza@lipn.univ-paris13.fr (J. García-Mendoza)

researchers increasingly seek to improve model performance by constructing richer and more representative training corpora. Large Language Models have demonstrated an extraordinary capacity for generating coherent and contextually appropriate text, making them attractive tools for creating artificial training instances capable of mitigating class imbalance while preserving semantic consistency [10, 11].

Nevertheless, unrestricted synthetic generation presents important challenges. Reviews generated without considering the characteristics of authentic minority opinions frequently fail to reproduce the lexical, semantic, and stylistic properties observed in real tourist experiences. Consequently, many synthetic instances contribute little useful information during the fine-tuning process. This observation suggests that effective data augmentation should not only increase the number of minority examples but also preserve the linguistic patterns that define them.

Motivated by this idea, the present work proposes a prototype-guided synthetic review generation strategy for the REST-MEX 2026 Artificial Data Generation Track. Rather than generating reviews completely from scratch, the proposed methodology first identifies representative minority-class opinions within the original corpus and subsequently uses them as semantic templates for controlled generation. The objective is to preserve the linguistic characteristics of authentic negative reviews while simultaneously increasing corpus diversity through the generation of new synthetic instances.

2. REST-MEX Through Time

The REST-MEX benchmark has become one of the most important evaluation campaigns devoted to Natural Language Processing for tourism in Spanish. Since its creation, the benchmark has provided a common evaluation framework that promotes reproducible research while encouraging the development of increasingly sophisticated methods for analyzing Mexican tourism reviews [12].

The first edition of REST-MEX focused primarily on recommendation systems for Mexican tourist destinations, establishing the initial benchmark and introducing the first publicly available corpus for tourism recommendation research in Spanish. This collection laid the foundations for subsequent editions by demonstrating the potential of user-generated reviews as a valuable source of tourism intelligence [12, 13, 7].

REST-MEX 2022 considerably expanded the scope of the shared task by incorporating sentiment analysis together with recommendation systems and the prediction of the Mexican COVID-19 epidemiological semaphore from textual information [14]. This edition also introduced more challenging evaluation metrics and encouraged the development of domain-adapted transformer models specifically designed for tourism applications [15, 16].

The 2023 edition represented another important milestone by concentrating the competition on tourism sentiment analysis. The benchmark attracted numerous participants proposing transformer-based architectures, ensemble learning strategies, balancing techniques, and domain adaptation methods. These contributions demonstrated that specialized language models substantially outperform traditional machine learning approaches while simultaneously highlighting the persistent challenges imposed by severe sentiment imbalance [17, 18, 19].

Building upon these advances, REST-MEX 2025 substantially increased the scale of the benchmark by collecting more than 250,000 tourist opinions covering hotels, restaurants, and attractions across forty Mexican Magical Towns. The evaluation also evolved into a multitask scenario involving sentiment polarity, place type, and destination classification, providing one of the largest publicly available tourism corpora in Spanish for Natural Language Processing research [20].

The most recent edition, REST-MEX 2026, represents a significant paradigm shift toward data-centric artificial intelligence. Instead of requiring participants to design increasingly sophisticated classifiers, the competition fixes the classification architecture and challenges participants to improve its performance exclusively through synthetic data generation. This formulation encourages the exploration of novel corpus augmentation strategies capable of enriching minority sentiment classes while preserving the linguistic characteristics of authentic tourist reviews [21, 22].

The methodology proposed in this paper is aligned with this new research direction. Rather than emphasizing architectural innovations, it investigates how representative minority-class reviews can guide the generation of realistic synthetic opinions. In this way, the work contributes to the growing body of research that views data quality as a central component for improving Natural Language Processing systems in tourism.

3. Proposed Method

The proposed methodology is based on the assumption that the quality of synthetic reviews depends not only on the capabilities of the Large Language Model but also on the quality of the examples used to guide the generation process. Instead of prompting the language model with generic instructions, the proposed framework first identifies the most representative reviews belonging to the minority sentiment classes and subsequently employs them as semantic prototypes for generating new synthetic instances.

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the 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. Since the objective is to reinforce the minority categories, only reviews whose labels belong to classes 1 and 2 are considered during the prototype selection stage.

3.1. Prototype Extraction

Each minority review is transformed into a semantic embedding using a sentence representation model. Let $E = \{e_1, e_2, \ldots, e_m\}$ denote the embedding vectors associated with every minority review. A semantic centroid is then computed as $c = \frac{1}{m}\sum_{i=1}^{m} e_i$, representing the average semantic profile of the minority corpus. Subsequently, the cosine similarity between every embedding and the centroid is calculated as $sim(e_i, c) = \frac{e_i \cdot c}{\|e_i\|\|c\|}$. Reviews exhibiting the highest similarity values are selected as semantic prototypes because they best represent the lexical, semantic, and stylistic characteristics shared by authentic minority-class opinions.

Unlike random sampling strategies, selecting reviews close to the semantic centroid reduces the probability of choosing atypical opinions while preserving the linguistic properties that characterize the minority sentiment categories.

3.2. Prototype-Guided Generation

Each prototype is subsequently employed as guidance for a Large Language Model. Instead of copying the original review, the language model receives instructions to preserve the sentiment polarity, writing style, and overall discourse organization while modifying contextual information such as tourist attractions, hotel names, restaurants, locations, prices, services, and specific events.

Formally, every synthetic review is generated according to $x' = G(x_p, y)$, where x_p denotes the selected prototype review, y is the target sentiment polarity, and $G(\cdot)$ represents the language model. Multiple synthetic variants are produced from each prototype, increasing the representation of the minority classes without losing their linguistic characteristics.

3.3. Semantic Filtering

Although prototype-guided generation naturally preserves semantic consistency, excessive similarity between synthetic reviews and their corresponding prototypes may reduce the diversity of the augmented corpus. To avoid this situation, every generated review is compared against both the original corpus and the previously accepted synthetic reviews using semantic similarity. Reviews whose similarity exceeds a predefined threshold are discarded and regenerated.

This filtering stage seeks to maximize lexical diversity while maintaining the semantic characteristics inherited from the original prototype.

3.4. Dataset Augmentation

The accepted synthetic reviews compose the artificial dataset $D' = \{(x'_i, y_i)\}_{i=1}^{M}$, where M denotes the total number of generated reviews. Finally, the augmented corpus is obtained as $D^* = D \cup D'$, which is subsequently employed to fine-tune the official baseline classifier released by the REST-MEX 2026 organizers.

The central hypothesis of the proposed methodology is that reviews generated from representative semantic prototypes remain closer to the original data distribution than reviews generated through unrestricted prompting, thereby providing more informative training examples for the minority sentiment classes.

4. Results and Discussion

The proposed methodology was evaluated according to the official protocol of the REST-MEX 2026 Artificial Data Generation Track. Following the competition guidelines, the generated synthetic reviews were merged with the original training corpus and employed to fine-tune the official baseline sentiment classifier. The evaluation was performed using the official Ranking Performance (RP) metric together with Macro F1 and the individual F1-score for each sentiment category.

The official evaluation metric is defined as $RP(k) = \frac{\sum_{i \in C}\left(1-\frac{TC_i}{TC}\right)F_i(k)}{\sum_{i \in C}\left(1-\frac{TC_i}{TC}\right)}$, where TC_i denotes the number of training instances belonging to class i, TC is the total number of training examples, and $F_i(k)$ corresponds to the F1-score obtained for sentiment class i. In addition, the Macro F1-score is computed as $MacroF1 = \frac{1}{|C|}\sum_{i \in C} F1_i$, assigning identical importance to every sentiment category independently of its frequency.

The proposed prototype-guided strategy obtained an RP score of 0.0120 and a Macro F1-score of 0.0837, achieving an overall competition score of 32.9930. The best individual performance corresponded to polarity class 5, with an F1-score of 0.3916, while class 2 obtained an F1-score of 0.0265. Although the remaining minority classes continued to be difficult to identify, the overall Macro F1 considerably exceeded the official Majority Class baseline.

Compared with the baseline, the proposed methodology nearly tripled the Macro F1-score while substantially increasing the performance of the dominant sentiment category. These results indicate that selecting representative reviews before generation allows the language model to produce synthetic opinions that remain closer to the linguistic distribution of authentic tourist reviews. Consequently, the augmented corpus provides more informative supervision than synthetic reviews generated from unrestricted prompts.

Another interesting observation is that the semantic consistency of the generated reviews was noticeably higher than that obtained using purely prompt-based strategies. Since every synthetic opinion originates from an authentic prototype, the generated reviews naturally preserve realistic discourse organization, sentiment intensity, and tourism-related vocabulary while introducing sufficient contextual variation to avoid simple duplication of the original corpus.

Although the improvements obtained for the minority classes remain relatively modest, the proposed methodology demonstrates that prototype selection constitutes an important stage in synthetic data generation. Rather than increasing the number of generated reviews indiscriminately, identifying representative seed examples before generation appears to produce synthetic corpora with greater usefulness for downstream sentiment classification.

5. Conclusions

This paper presented a prototype-guided synthetic review generation strategy for the REST-MEX 2026 Artificial Data Generation Track. Instead of relying on unrestricted prompting, the proposed

methodology first identifies representative minority-class reviews and subsequently employs them as semantic prototypes to guide the generation of new synthetic opinions.

The experimental results demonstrate that prototype-guided generation substantially improves the Macro F1-score compared with the official Majority Class baseline, suggesting that the representativeness of the seed examples plays a fundamental role in the quality of the generated corpus. These findings reinforce the importance of combining semantic retrieval with Large Language Models for data-centric sentiment analysis.

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] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism, Journal of King Saud University-Computer and Information Sciences 35 (2023) 101746.

[4] 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.

[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] M. Á. Álvarez-Carmona, I. S. Guzmán, R. Aranda, Towards accurate and legible scene text generation in spanish; a text-to-image, in: Progress in Artificial Intelligence and Pattern Recognition: 9th International Congress, IWAIPR 2025, Varadero, Cuba, October 14–17, 2025, Proceedings, Springer Nature, 2026, p. 152.

[7] 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.

[8] 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.

[9] 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).

[10] 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.

[11] A. Y. Rodríguez-González, M. Á. Álvarez-Carmona, Machine learning and soft computing: Current trends and applications, 2026.
[12] 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).
[13] 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.
[14] M. Á. Alvarez-Carmona, R. Aranda, Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022).
[15] 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.
[16] 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.
[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] 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).
[19] 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.
[20] 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.
[21] 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).
[22] 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.