

Sentiment Motif Discovery through Smith–Waterman Local Alignment for Synthetic Review Generation

Estefania Lazcano-Díaz^{1,*}

¹*Centro de Investigación y de Estudios Avanzados (Cinvestav), Unidad Monterrey*

Abstract

This paper proposes a synthetic data generation methodology inspired by the Smith–Waterman local sequence alignment algorithm for the REST-MEX 2026 Artificial Data Generation Track. Instead of generating tourist reviews directly, the proposed approach identifies locally conserved sentiment motifs shared by authentic minority-class reviews through a semantic adaptation of the Smith–Waterman algorithm. These motifs are subsequently recombined and transformed into coherent tourist opinions using a Large Language Model, producing synthetic reviews that preserve the lexical and semantic characteristics of real negative experiences. The generated corpus is employed to fine-tune the official baseline classifier. Experimental results demonstrate that local sentiment alignment constitutes a promising mechanism for discovering informative linguistic patterns, although additional polarity-aware generation strategies are required to further improve minority-class recognition.

Keywords

Sentiment Analysis, Synthetic Text Generation, Smith–Waterman, REST-MEX 2026

1. Introduction

The exponential growth of user-generated content has transformed Natural Language Processing (NLP) into one of the most influential technologies for modern tourism analytics. Every day, millions of visitors publish reviews describing their experiences in hotels, restaurants, museums, and tourist attractions. These textual opinions provide a detailed description of customer satisfaction, service quality, destination image, and traveler expectations, allowing governments and tourism organizations to monitor visitor perception almost in real time. Consequently, sentiment analysis has become a fundamental component of intelligent tourism systems, supporting recommendation systems, reputation management, strategic planning, and destination competitiveness analysis [1, 2, 3].

Recent advances in Deep Learning and transformer-based language models have considerably improved the performance of sentiment classification systems. Domain-adapted BERT models, ensemble strategies, and Large Language Models have demonstrated remarkable capabilities for understanding the semantic complexity of tourist reviews and capturing subtle linguistic patterns associated with customer opinions [4, 5, 6]. Nevertheless, despite these architectural improvements, the quality of supervised sentiment classifiers remains fundamentally constrained by the characteristics of the available training data.

One of the principal difficulties encountered in tourism sentiment analysis is the severe imbalance naturally present in online review collections. Since satisfied visitors are considerably more likely to publish positive experiences than negative ones, most tourism corpora contain an overwhelming proportion of highly positive opinions. Consequently, minority sentiment classes receive comparatively little supervision during training, causing classification models to develop biased decision boundaries that primarily favor the majority categories [7, 8, 9]. This phenomenon becomes particularly problematic because the rare negative reviews frequently contain the most valuable information for identifying service deficiencies and improving customer satisfaction [10].

The REST-MEX benchmark has progressively evolved as an experimental platform for addressing these challenges. Beginning as a recommendation-oriented shared task, subsequent editions gradually

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

✉ estefania.lazcano@cimat.mx (E. Lazcano-Díaz)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

incorporated sentiment analysis, multitask learning, domain adaptation, and increasingly larger collections of Mexican tourism reviews. This evolution culminated in REST-MEX 2026, where the research focus shifted from designing increasingly sophisticated classifiers toward improving training corpora through synthetic data generation. In this new scenario, all participating systems employ exactly the same classifier architecture, while performance differences originate exclusively from the quality of the generated artificial reviews [11, 12, 13, 14, 15, 16].

Most synthetic review generation approaches proposed in recent years rely on prompt engineering, template-based generation, or unrestricted Large Language Model prompting. Although these methods produce fluent and grammatically correct text, they generally overlook an important characteristic shared by authentic negative opinions: discriminative linguistic information is usually concentrated in relatively short textual fragments rather than uniformly distributed throughout the complete review. Expressions such as "the room was dirty," "the service was extremely slow," or "the food arrived cold" constitute localized semantic patterns that largely determine the overall sentiment of the document.

A remarkably similar problem has been studied for decades in computational biology. Rather than comparing complete DNA or protein sequences, local sequence alignment algorithms such as Smith–Waterman identify highly conserved subsequences shared by different biological samples. These local regions frequently correspond to biologically meaningful motifs responsible for specific molecular functions. Interestingly, the underlying principles of biological sequence alignment have already demonstrated their usefulness in Natural Language Processing, where semantic similarity measures inspired by alignment techniques have successfully modeled relationships between textual sequences for applications such as plagiarism detection and semantic comparison [17].

Inspired by these observations, this paper proposes a novel synthetic review generation methodology that transfers the philosophy of Smith–Waterman local alignment to tourism sentiment analysis. Instead of generating complete reviews directly from prompts, minority-class opinions are first interpreted as linguistic sequences. A local alignment process subsequently discovers recurrent sentiment motifs shared among authentic negative reviews. These motifs are then recombined and verbalized through a Large Language Model to construct new synthetic opinions that preserve the lexical, semantic, and stylistic characteristics of real minority-class reviews while introducing sufficient contextual diversity.

The central hypothesis of this work is that locally conserved linguistic patterns constitute more informative generation units than complete reviews. By explicitly extracting and recombining these sentiment motifs before invoking the language model, the proposed methodology seeks to produce synthetic reviews that remain closer to the original data distribution and provide more effective supervision during classifier fine-tuning.

2. Proposed Method

The proposed methodology is inspired by the Smith–Waterman local alignment algorithm originally developed for biological sequence analysis. In molecular biology, local alignment identifies highly conserved subsequences that frequently correspond to functional DNA or protein regions. Instead of comparing complete sequences globally, Smith–Waterman searches for local similarities capable of revealing biologically meaningful motifs. Similar alignment principles have previously been adapted to Natural Language Processing for semantic similarity estimation, demonstrating that biological sequence analysis can successfully inspire textual comparison algorithms [17]. In this work, the same philosophy is transferred to synthetic review generation.

Rather than considering each tourist review as an indivisible document, the proposed framework models every opinion as a sequence of linguistic elements. The objective is to identify locally conserved sentiment fragments shared by authentic minority-class reviews and subsequently employ these fragments as semantic building blocks for generating new synthetic opinions.

Let $D = \{(x_i, y_i)\}_{i=1}^N$ denote the REST-MEX 2026 training corpus, where x_i represents a tourist review and $y_i \in \{1, 2, 3, 4, 5\}$ its corresponding sentiment polarity. Since the objective is to reinforce the underrepresented classes, only reviews belonging to polarities 1 and 2 are employed during motif

discovery.

2.1. Linguistic Sequence Representation

Each review is transformed into an ordered linguistic sequence

$$x = (u_1, u_2, \dots, u_n),$$

where every element u_i corresponds to a normalized lexical unit obtained after tokenization, lemmatization, and stop-word removal. Unlike Bag-of-Words representations, this sequential formulation preserves the contextual relationships among neighboring words, allowing local alignment algorithms to identify coherent linguistic fragments rather than isolated vocabulary items.

Representing reviews as ordered sequences enables the direct adaptation of biological alignment techniques without modifying their fundamental dynamic programming structure.

2.2. Local Sentiment Alignment

Given two minority-class reviews

$$x = (u_1, \dots, u_n)$$

and

$$z = (v_1, \dots, v_m),$$

their local similarity is computed through a Smith–Waterman-inspired dynamic programming procedure. The alignment matrix is recursively defined as

$$H(i, j) = \max \{0, H(i-1, j-1) + s(u_i, v_j), H(i-1, j) - g, H(i, j-1) - g\},$$

where $s(u_i, v_j)$ denotes the semantic similarity between two lexical units and g represents the gap penalty.

Unlike biological alignment, where symbols correspond to nucleotides or amino acids, the similarity function combines exact lexical coincidence with semantic similarity computed from distributed word representations. Consequently, semantically related expressions such as *dirty*, *filthy*, and *unclean* contribute positively to the alignment score even when they are not lexically identical.

The highest-scoring local alignments correspond to textual fragments repeatedly observed among authentic negative reviews. These fragments are interpreted as *sentiment motifs*, representing localized linguistic patterns that characterize dissatisfaction in tourism experiences.

2.3. Motif Extraction and Repository Construction

After aligning every pair of minority reviews, the resulting high-scoring local fragments are extracted and stored in a motif repository

$$M = \{m_1, m_2, \dots, m_k\}.$$

Each motif generally represents a short semantic structure describing a specific negative situation, including complaints regarding cleanliness, customer service, waiting times, food quality, pricing, accessibility, room conditions, or personnel behavior.

Since every motif originates from authentic tourist opinions rather than manually designed templates, the repository naturally captures the lexical richness, syntactic variability, and discourse style characteristic of genuine customer complaints.

To avoid redundancy, motifs exhibiting extremely high semantic similarity are merged into a single representative pattern. This produces a compact repository containing only the most informative local sentiment structures.

2.4. Motif Recombination

Once the repository has been constructed, several compatible motifs are randomly selected to compose the semantic skeleton of a new review.

Let $M' = \{m_{i_1}, m_{i_2}, \dots, m_{i_r}\} \subseteq M$ denote the selected motif subset. Compatibility constraints ensure that motifs describe coherent tourism scenarios and do not introduce semantic contradictions.

For example, fragments referring to hotel rooms are combined with hotel-related service complaints, whereas restaurant motifs are combined with food and service descriptions.

This recombination stage is conceptually analogous to biological sequence recombination, where functional subsequences are combined to produce new genetic configurations while preserving meaningful structural properties.

2.5. Large Language Model Generation

The selected motif collection is subsequently provided to a Large Language Model responsible for transforming the fragmented semantic skeleton into a coherent tourist review.

Formally, every synthetic opinion is generated according to $x' = G(M', y)$, where $G(\cdot)$ denotes the language model and y represents the target sentiment polarity.

Instead of generating reviews from unrestricted prompts, the language model receives explicit instructions to preserve the semantic content of every motif while creating a fluent narrative that naturally integrates all selected fragments. This strategy guarantees that every generated opinion contains authentic negative linguistic patterns inherited from real minority-class reviews while simultaneously introducing sufficient lexical and contextual diversity.

3. Results

The proposed local-alignment-guided generation strategy was evaluated following the official protocol of the REST-MEX 2026 Artificial Data Generation Track. The synthetic reviews generated from sentiment motifs were incorporated into the original training corpus and used to fine-tune the official baseline classifier. The evaluation considered the official Ranking Performance (RP) metric, Macro F1, the overall competition score, and the individual F1-scores for each polarity class.

The proposed methodology obtained an RP score of 0.0104 and a Macro F1-score of 0.1228, reaching an overall competition score of 49.8203. Regarding the individual sentiment categories, the F1-scores for classes 1 to 5 were 0.0005, 0.0133, 0.0005, 0.0006, and 0.5990, respectively. The highest performance was clearly obtained for polarity class 5, while the remaining classes continued to show very limited classification performance.

Although the RP score remained low, the Macro F1-score was notably higher than in several simpler augmentation strategies. This suggests that extracting and recombining local sentiment motifs provides useful linguistic information to the classifier. In particular, the strong performance observed for class 5 indicates that the synthetic reviews generated through motif recombination may reinforce general sentiment patterns, even if they do not sufficiently improve the recognition of minority categories.

The limited performance for classes 1 and 2 shows that local alignment alone is not enough to fully address the severe imbalance of the REST-MEX corpus. One possible explanation is that the identified motifs capture short discriminative fragments, but the final generated reviews may still resemble majority-class discourse or may not provide enough diversity in strongly negative contexts. Therefore, while the Smith–Waterman-inspired strategy is useful for identifying relevant textual fragments, additional mechanisms are needed to ensure that these motifs produce more effective minority-class supervision.

Overall, the results indicate that local alignment is a promising mechanism for discovering sentiment-bearing fragments in tourist reviews. However, the generation stage must be further controlled to preserve the intended polarity and to increase the diversity of negative synthetic opinions.

4. Conclusions

This paper presented a synthetic review generation strategy inspired by the Smith–Waterman local alignment algorithm. The proposed method identifies sentiment motifs in minority-class reviews and uses them as semantic building blocks for generating new tourist opinions.

The results show that the approach improves Macro F1, but it still struggles to improve minority-class recognition. As future work, we plan to combine local alignment with polarity-aware filtering and LLM-based validation to generate more coherent and discriminative negative 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. Díaz-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-Rodríguez, 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] 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.
- [6] 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).
- [7] 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.
- [8] 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.
- [9] O. G. Toledano-López, M. Á. Álvarez-Carmona, J. Madera, A. Simón-Cuevas, Y. A. López-Rodríguez, H. González Díez, 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.
- [10] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancan case, seen from the usa, canada, and mexico, *International Journal of Tourism Cities* 10 (2024) 639–661.
- [11] 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).
- [12] 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: Recommen-

dation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, *Procesamiento del Lenguaje Natural* 69 (2022) 289–299.

- [13] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, V. Muñoz-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.
- [14] 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.
- [15] 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).
- [16] 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.
- [17] M. A. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, L. Villaseñor-Pineda, Semantically-informed distance and similarity measures for paraphrase plagiarism identification, *Journal of Intelligent & Fuzzy Systems* 34 (2018) 2983–2990.