Lexical Data Augmentation for Tourism Sentiment Analysis Using FastText Embeddings

Andrea Bethsabe García-Gutiérrez[1,*]

[1]Centro de Investigación en Matemáticas A.C., Sede Monterrey

Abstract

This paper presents a lightweight lexical augmentation strategy for the REST-MEX 2026 Artificial Data Generation Track. The proposed approach generates synthetic tourist reviews by randomly selecting opinions from the original training corpus and replacing verbs and adjectives with semantically similar alternatives obtained from a FastText embedding model. Unlike prompt-based generation methods, the proposed technique preserves the original sentence structure while introducing lexical diversity through distributed word representations. The augmented corpus was incorporated into the official REST-MEX 2026 training set and evaluated using the competition protocol. Experimental results show that the proposed method outperformed the Majority Class baseline according to the official ranking metrics, demonstrating that simple embedding-based lexical transformations constitute a computationally efficient alternative for synthetic data generation in sentiment analysis.

Keywords

Sentiment Analysis, Lexical Data Augmentation, FastText, Word Embeddings, Tourism Reviews, REST-MEX 2026

1. Introduction

Distributed word representations have fundamentally changed the way Natural Language Processing (NLP) systems model lexical semantics. Rather than representing words as isolated symbols, embedding models encode lexical items into continuous vector spaces where semantic and syntactic similarities are reflected by geometric proximity. This representation has enabled numerous applications, including text classification, information retrieval, machine translation, question answering, and sentiment analysis. More importantly, vector representations provide an elegant mechanism for generating lexical variations while preserving the semantic content of a document.

Tourism has become one of the application domains that has benefited the most from recent advances in NLP. The continuous growth of online review platforms has produced enormous collections of user-generated content describing tourist experiences, service quality, destination image, and visitor satisfaction. These textual resources have supported the development of intelligent recommendation systems, hospitality assistants, destination image analysis, sentiment analysis, and digital reputation monitoring, contributing to more informed decision-making in tourism management [1, 2, 3, 4, 5, 6, 7, 8, 9].

Among these applications, sentiment analysis has emerged as one of the most extensively studied tasks due to its ability to automatically identify tourists' opinions at large scale. Recent deep learning architectures and transformer-based models have substantially improved sentiment classification performance in tourism corpora, enabling increasingly accurate analyses of customer experiences and destination perception [10, 11]. Nevertheless, regardless of the learning algorithm employed, the quality and diversity of the training corpus remain fundamental factors that directly influence classifier performance.

The REST-MEX evaluation campaign has established itself as one of the principal benchmarks for sentiment analysis in Mexican tourism reviews. Since its first edition, the competition has progressively evolved to address increasingly challenging problems involving recommendation systems, polarity

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

andrea.garcia@cimat.mx (A. B. García-Gutiérrez)

classification, domain adaptation, multitask learning, and synthetic data generation [12, 13, 14, 15]. The REST-MEX 2026 edition introduces the Artificial Data Generation Track, where participants are asked to enrich the original training corpus with automatically generated reviews that improve the performance of a predefined sentiment classification model [16, 17].

Although recent synthetic data generation methods are predominantly based on Large Language Models, lexical augmentation techniques remain attractive due to their simplicity, computational efficiency, and interpretability. Rather than generating completely new documents, lexical augmentation modifies existing sentences by replacing selected words with semantically related alternatives. When the replacements preserve the original sentiment orientation and grammatical structure, the resulting texts increase lexical variability without substantially altering the semantic meaning of the original review. Previous work on REST-MEX has shown that balancing strategies and controlled corpus enrichment can positively influence sentiment classification performance under imbalanced conditions [18, 19, 20, 21, 22].

In this work, we propose a lightweight lexical augmentation strategy based on distributed word representations. A subset of tourist reviews is randomly selected from the original REST-MEX 2026 training corpus, after which verbs and adjectives are identified and replaced by semantically similar alternatives obtained from a FastText embedding model. The underlying hypothesis is that introducing controlled lexical variability while preserving the original semantic content may increase the robustness of the sentiment classifier by exposing it to multiple lexical realizations of equivalent tourist experiences. Unlike prompt-based generation methods, the proposed approach relies exclusively on vector-space semantics, making it computationally inexpensive, easily reproducible, and independent of external generative language models.

2. Methodology

The proposed data augmentation strategy is based on the hypothesis that lexical diversity can improve the generalization capability of sentiment classification models without modifying the semantic orientation of the original opinions. Instead of generating entirely new reviews, the method creates synthetic instances by performing controlled lexical substitutions over authentic tourist reviews. Specifically, verbs and adjectives are replaced by semantically similar alternatives obtained from a pre-trained FastText embedding model. Since these parts of speech usually convey the action and the sentiment expressed in a review, their substitution allows the construction of lexically different texts while preserving the overall meaning of the original opinion.

Formally, let

$$D = (x_i, y_i)_{i=1}^{N}$$

be the original REST-MEX 2026 training corpus, where x_i denotes a tourist review and $y_i \in 1, 2, 3, 4, 5$ represents its sentiment polarity. The proposed algorithm transforms a subset of D into a synthetic corpus

$$D' = (x'_i, y_i),$$

where each synthetic review x'_i preserves the original polarity label while introducing controlled lexical variations.

The complete procedure consists of four stages: random review selection, linguistic analysis, lexical substitution using FastText, and synthetic corpus construction.

2.1. Dataset

The experiments were carried out using the official REST-MEX 2026 training corpus, which contains 208,051 tourist reviews collected from Mexican *Pueblos Mágicos*. To generate the synthetic data, a subset

of 3,000 reviews was selected uniformly at random from the original corpus. Each selected opinion retained its original polarity label throughout the augmentation process.

2.2. Random Sampling

Instead of processing the complete dataset, the proposed method begins by randomly selecting a subset of reviews. Let

$$S = x_1, x_2, \ldots, x_n,$$

where

$$n = 3000$$

and

$$S \subset D.$$

Random sampling reduces computational cost while ensuring that the selected reviews cover different destinations, service types, writing styles, and sentiment categories present in the original corpus.

2.3. Part-of-Speech Identification

Each review is tokenized and morphologically analyzed to identify its grammatical structure. Let

$$x = (w_1, w_2, \ldots, w_m)$$

be a review composed of m tokens. A part-of-speech tagging function

$$P(w_i)$$

assigns a grammatical category to every token.
Only words satisfying

$$P(w_i) \in \text{Verb}, \text{Adjective}$$

are considered candidates for lexical substitution. Verbs and adjectives were selected because they are responsible for describing actions, events, and subjective evaluations, making them particularly relevant for sentiment expression while allowing lexical variability without drastically altering sentence structure.

2.4. FastText-Based Lexical Replacement

For each candidate word w, its vector representation is obtained from a pre-trained FastText embedding model.
Let

$$E : V \to \mathbb{R}^d$$

be the embedding function, where V is the vocabulary and d is the embedding dimension.
Semantic similarity between two words is computed using cosine similarity,

$$\text{sim}(w_i, w_j) = \frac{E(w_i) \cdot E(w_j)}{|E(w_i)||E(w_j)|}.$$

For every selected verb or adjective, the algorithm retrieves the set

$$N_k(w) = v_1, v_2, \ldots, v_k,$$

containing the k nearest neighbors according to cosine similarity.

A replacement word is then randomly selected from $N_k(w)$, provided that its similarity exceeds a predefined threshold,

$$\mathrm{sim}(w, v) \geq \tau.$$

This restriction avoids replacing words with semantically distant alternatives that could alter the sentiment orientation or produce unnatural sentences.

For example,

excellent	$\rightarrow$	outstanding
beautiful	$\rightarrow$	gorgeous
enjoyed	$\rightarrow$	liked
comfortable	$\rightarrow$	pleasant

Each replacement is performed independently, allowing multiple substitutions within the same review.

2.5. Synthetic Review Generation

Let

$$T(x)$$

denote the transformation operator that applies lexical substitutions over review x. The resulting synthetic review is defined as

$$x' = T(x).$$

Since only lexical substitutions are performed, the polarity label remains unchanged,

$$y(x') = y(x).$$

Applying the transformation operator to every review in S produces the augmented corpus

$$D' = (T(x_i), y_i)_{i=1}^{3000}.$$

Finally, the synthetic reviews are merged with the original REST-MEX 2026 training corpus,

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

and the resulting dataset is used during the fine-tuning stage of the official baseline sentiment classifier.

The proposed methodology intentionally avoids complex generation models and instead exploits the semantic structure captured by distributed word representations. Consequently, the augmentation process is computationally efficient, deterministic once the sampling seed is fixed, and fully reproducible, while preserving the lexical and stylistic characteristics of authentic tourist reviews.

3. Results and Discussion

The proposed lexical augmentation strategy was evaluated following the official protocol of the REST-MEX 2026 Artificial Data Generation Track. Since the dataset exhibits a highly imbalanced sentiment distribution, the competition adopts the *Ranking Performance* (RP) score as the primary evaluation metric. RP emphasizes the contribution of minority sentiment categories by assigning them greater importance than majority classes, thereby providing a more informative assessment than conventional accuracy. In addition to RP, the organizers report Macro F1, overall Accuracy, and the F1-score obtained for each individual sentiment category.

The experimental results indicate that the proposed lexical augmentation strategy improves upon the Majority Class baseline according to both the official RP metric and Macro F1. Although the numerical improvements are relatively small, they demonstrate that introducing lexical variability through controlled semantic substitutions allows the classifier to recognize sentiment categories beyond the dominant class. Consequently, the generated corpus contributes useful information during fine-tuning despite relying on a remarkably simple generation procedure.

The highest F1-score was obtained for polarity class 2, reaching 0.0505, whereas polarity class 5 achieved an F1-score of 0.0295. A marginal improvement was also observed for class 4. These results suggest that lexical substitutions based on distributed word representations are capable of preserving the semantic orientation of the original reviews while exposing the classifier to alternative lexical realizations. Since FastText places semantically related words close to each other in the embedding space, replacing verbs and adjectives with neighboring terms introduces lexical diversity without substantially modifying the underlying sentiment.

Nevertheless, the generated corpus was not sufficient to outperform the original baseline model released by the organizers. One possible explanation is that lexical substitution alone modifies only a limited portion of each review. Although different surface forms are produced, the global syntactic structure and discourse organization remain practically unchanged. Consequently, the augmented corpus may still exhibit a relatively low level of structural diversity compared with naturally occurring reviews or texts generated by modern Large Language Models.

Another important consideration concerns the embedding space itself. Semantic proximity does not necessarily imply sentiment equivalence. Two neighboring words in the FastText vector space may differ subtly in intensity, connotation, or contextual usage. Although the similarity threshold reduces the probability of inappropriate substitutions, some generated reviews may still contain lexical choices that are uncommon or less natural in the tourism domain, limiting their contribution during classifier training.

Despite these limitations, the proposed methodology presents several practical advantages. The entire augmentation process is computationally inexpensive, requires neither additional model training nor access to external generative models, and is fully reproducible once the random seed and embedding model are fixed. Furthermore, unlike prompt-based generation, every synthetic review can be directly traced back to an authentic opinion, facilitating the interpretation of the generated corpus and reducing the likelihood of introducing hallucinated or factually inconsistent content.

Overall, the experimental results suggest that lexical augmentation using distributed word representations constitutes a simple yet effective baseline for synthetic data generation. While its impact on classification performance remains modest, the approach demonstrates that semantic-preserving lexical transformations can enrich the training corpus and provide measurable improvements over trivial baselines. Future work could combine lexical substitution with contextual generation methods, syntactic paraphrasing, or sentence-level semantic transformations to produce richer and more diverse synthetic reviews capable of further improving sentiment classification performance.

4. Conclusions and Future Work

This paper presented a simple lexical augmentation strategy based on FastText embeddings for the REST-MEX 2026 Artificial Data Generation Track. By replacing verbs and adjectives with semantically similar words, the proposed method generated synthetic reviews that increased the lexical diversity of the training corpus while preserving the original sentiment labels. Although the approach did not outperform the official baseline model, it achieved better performance than the Majority Class baseline, demonstrating that lightweight lexical transformations can provide useful additional training instances.

As future work, we plan to incorporate contextual constraints during word replacement, explore sentence-level paraphrasing techniques, and combine lexical augmentation with Large Language Models to generate more diverse and semantically rich synthetic 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] 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.

[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] 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.

[7] 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).

[8] 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.

[9] 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.

[10] 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.
[11] 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.
[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] 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.
[14] 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.
[15] 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.
[16] 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).
[17] 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.
[18] 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.
[19] A. 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 16 (2026) 7398.
[20] 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).
[21] 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.
[22] 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.