Lexical-Only Synthetic Data Augmentation for Imbalanced Tourism Sentiment Analysis

Juan David Jurado-Buch[1,2,*], Lázaro Bustio-Martínez[3], Vitali Herrera-Semenets[4] and Miguel Angel Álvarez-Carmona[5]

[1]*Universidad de Nariño, Colombia*
[2]*Universidad Nacional Abierta y a Distancia, Nariño, Colombia*
[3]*Universidad Iberoamericana, Campus Santa Fe, Mexico*
[4]*OnDemand S.R.L, Ancona, Italy*
[5]*Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, Mexico*

Abstract

This paper investigates an extremely lightweight data augmentation strategy for the REST-MEX 2026 Artificial Data Generation Track. Instead of generating coherent synthetic reviews, the proposed methodology constructs artificial training instances by randomly sampling between 10 and 20 highly negative sentiment-bearing words extracted from authentic tourist opinions. These lexical collections are intentionally left without grammatical structure, allowing us to evaluate whether discriminative vocabulary alone can improve sentiment classification under severe class imbalance. The generated instances are incorporated into the official REST-MEX 2026 training corpus to fine-tune the baseline sentiment classifier provided by the organizers. Experimental results show that, despite the absence of coherent natural language, the proposed approach improves the Macro F1 score over the Majority Class baseline, demonstrating that isolated lexical information contributes to the learning process. Nevertheless, the obtained performance also indicates that lexical features alone are insufficient for accurately representing minority sentiment classes, highlighting the importance of combining discriminative vocabulary with coherent text generation in future synthetic data augmentation strategies.

Keywords

Sentiment Analysis, Data Augmentation, Lexical Features, Synthetic Text Generation, Class Imbalance, Tourism Reviews, REST-MEX

1. Introduction

The increasing availability of user-generated content has positioned Natural Language Processing (NLP) as one of the principal technologies supporting intelligent tourism systems [1, 2, 3, 4]. Online reviews published by travelers provide valuable information regarding customer satisfaction, service quality, destination image, and visitor experiences. Consequently, sentiment analysis has become an essential component of tourism analytics, enabling automatic opinion mining for recommendation systems, digital reputation monitoring, destination management, and decision-support applications [5, 6, 7, 8].

Recent advances in transformer architectures and Large Language Models (LLMs) have considerably improved the state of the art in sentiment classification. Nevertheless, the effectiveness of these models continues to depend heavily on the quality, diversity, and balance of the available training data[9, 10]. In practical tourism scenarios, user opinions exhibit a highly skewed sentiment distribution in which extremely positive reviews largely outnumber negative experiences. This imbalance frequently causes classifiers to become biased toward the dominant classes while underperforming on the minority categories that often contain the most valuable information for service improvement [11, 12, 13].

The REST-MEX benchmark has progressively addressed these challenges through successive shared tasks focused on Mexican tourism reviews. Since its introduction in 2021, the benchmark has evolved

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
jjuradobuch@gmail.com (J. D. Jurado-Buch); lazaro.bustio@ibero.mx (L. Bustio-Martínez); vherrera@cenatav.co.cu (V. Herrera-Semenets); miguel.alvarez@cimat.mx (M. A. Álvarez-Carmona)

from recommendation systems to fine-grained sentiment analysis, multitask learning, and, most recently, data-centric artificial intelligence through synthetic text generation [14, 15, 16, 17, 18, 19]. The Artificial Data Generation Track introduced in REST-MEX 2026 shifts the research focus from classifier optimization toward the construction of synthetic corpora capable of improving a common baseline model.

Most existing approaches rely on generating complete synthetic reviews using Large Language Models or applying sophisticated linguistic transformations to authentic opinions. Although these techniques often produce coherent and realistic texts, they also require considerable computational resources and carefully designed prompting strategies. An interesting question therefore arises: is it necessary to generate complete reviews to improve sentiment classification, or could a much simpler lexical representation already provide useful information to the classifier?

Motivated by this question, this paper investigates an intentionally minimalist data augmentation strategy. Instead of generating fluent sentences, the proposed methodology constructs synthetic instances by randomly selecting a small set of highly negative sentiment-bearing words extracted from the original corpus. These lexical collections do not form grammatically correct sentences and are not intended for human interpretation. Rather, they function as artificial feature vectors that increase the representation of strongly negative vocabulary during fine-tuning.

This work evaluates whether increasing the frequency of highly discriminative lexical items alone is sufficient to improve sentiment classification under severe class imbalance. Although the experimental results indicate that this simplified representation is considerably less effective than generating coherent reviews, the study provides useful insights into the role played by linguistic structure during synthetic data generation. The findings suggest that lexical information alone is insufficient and reinforce the importance of combining discriminative vocabulary with coherent natural language generation, motivating future research based on Large Language Models capable of transforming sentiment-bearing lexical sets into realistic tourist reviews.

2. Proposed Method

Unlike most synthetic data generation approaches, the proposed methodology does not attempt to generate grammatically correct tourist reviews. Instead, it is based on the hypothesis that sentiment classifiers can benefit from increasing the frequency of highly discriminative lexical items, even if these words are not organized into coherent natural language. The proposed strategy therefore generates synthetic instances composed exclusively of strongly negative sentiment-bearing words extracted from authentic tourist reviews.

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the REST-MEX 2026 training corpus, where x_i is a tourist review and $y_i \in \{1, 2, 3, 4, 5\}$ its corresponding sentiment polarity. Since the objective is to reinforce the minority sentiment class, only reviews belonging to polarity level 1 are considered during the lexical extraction stage.

2.1. Extraction of Polarized Vocabulary

The first stage consists of identifying the lexical units that most strongly characterize highly negative tourist opinions. Let

$D_1 = \{x_i \mid y_i = 1\}$

be the subset of reviews assigned to the most negative sentiment category.

After tokenization and basic text preprocessing, every token is assigned a sentiment relevance score according to its frequency within D_1. The resulting vocabulary is represented as

$V = \{w_1, w_2, \ldots, w_m\}$,

where each word is associated with a relevance value indicating its representative power for the negative sentiment class.

Examples of extracted terms include words expressing dissatisfaction, poor service quality, unpleasant experiences, or negative evaluations frequently observed in authentic tourist complaints.

2.2. Synthetic Instance Construction

Instead of generating complete sentences, synthetic documents are created by randomly sampling words from the polarized vocabulary.

For every artificial instance, a subset

$S = \{w_{i_1}, w_{i_2}, \ldots, w_{i_k}\}$

is constructed, where k is randomly selected between 10 and 20 words. Sampling is performed uniformly over the extracted vocabulary while allowing different combinations across synthetic documents.

A typical synthetic instance therefore consists of an unordered collection of highly negative lexical items such as

> *dirty, terrible, disgusting, broken, expensive, horrible, disappointing, rude, uncomfortable, unacceptable, unsafe, noisy*

These lexical collections are intentionally left unmodified. No attempt is made to impose grammatical structure, preserve word order, or generate meaningful sentences.

2.3. Corpus Augmentation

Each generated lexical collection is assigned the polarity label corresponding to class 1. The synthetic dataset is therefore defined as

$D' = \{(S_i, 1)\}_{i=1}^{M}$,

where M denotes the total number of generated instances, respecting the maximum number established by the REST-MEX 2026 competition.

Finally, the augmented corpus is obtained as

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

which is subsequently employed to fine-tune the official baseline classifier provided by the competition organizers.

2.4. Underlying Hypothesis

The proposed methodology is motivated by the assumption that transformer-based classifiers may benefit from repeated exposure to highly discriminative lexical features, even in the absence of coherent syntactic structure. If the contextual encoder successfully captures semantic information from isolated sentiment-bearing words, these synthetic instances could reinforce the internal representation of the minority class during fine-tuning.

Unlike conventional data augmentation techniques that seek to maximize linguistic realism, the proposed strategy deliberately ignores grammatical correctness in order to isolate the contribution of lexical information alone. Consequently, this methodology can be interpreted as a controlled ablation experiment designed to evaluate whether coherent sentence structure is truly necessary for effective synthetic data generation in tourism sentiment analysis.

3. Results

The proposed lexical augmentation strategy was evaluated under the official protocol of the REST-MEX 2026 Artificial Data Generation Track. Following the competition guidelines, the synthetic lexical instances were incorporated into the original training corpus and used to fine-tune the baseline sentiment classifier released by the organizers. Performance was assessed using the official Ranking Performance (RP) metric, Macro F1, the overall competition score, and the F1-score computed for each sentiment category.

Table 1 summarizes the results obtained by the proposed methodology together with the official Majority Class baseline.

Table 1
Performance comparison between the Majority Class baseline and the proposed lexical augmentation strategy.

Method	RP	Macro F1	Score	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)
Proposed Method	0.0127	0.0648	24.5802	0.0005	0.0338	0.0005	0.0007	0.2886

The proposed methodology obtained an RP score of 0.0127 and a Macro F1 score of 0.0648, while reaching an overall competition score of 24.5802. Regarding the individual sentiment categories, the highest performance was achieved for polarity class 5 with an F1-score of 0.2886, followed by polarity class 2 with an F1-score of 0.0338. Classes 1 and 3 obtained F1-scores of 0.0005, whereas class 4 reached 0.0007.

Compared with the Majority Class baseline, the proposed approach substantially increased the Macro F1 score, more than doubling the baseline performance (0.0648 versus 0.0291). Likewise, the F1-score for the dominant polarity class increased from 0.1455 to 0.2886, indicating that reinforcing the training corpus with highly polarized lexical items contributes useful discriminative information to the classifier. These improvements suggest that the lexical content itself carries a considerable amount of sentiment information, even when presented outside a coherent linguistic context.

Nevertheless, the proposed strategy was considerably less effective for the minority sentiment categories. Despite explicitly generating synthetic instances for the most negative class, the improvements obtained for polarity classes 1 and 2 remained modest. This behavior indicates that simply increasing the occurrence of highly negative vocabulary is insufficient for accurately modeling the complex semantic and contextual patterns that characterize authentic negative tourist reviews.

An interesting finding of this work is that transformer-based classifiers appear capable of extracting useful sentiment information from unordered collections of words. Although the generated instances lack grammatical structure and cannot be interpreted as natural language by humans, they still provide additional lexical evidence that positively influences the learning process. However, the absence of syntactic relationships, discourse structure, and contextual dependencies limits the representational richness of these synthetic examples.

Overall, the experiments suggest that lexical information alone constitutes only one component of effective synthetic data generation. While the proposed methodology provides a computationally inexpensive mechanism for increasing the representation of discriminative vocabulary, realistic sentence structure appears to remain an essential factor for producing synthetic instances capable of substantially improving sentiment classification performance.

4. Conclusions

This paper presented a minimalist lexical augmentation strategy in which synthetic instances are constructed exclusively from randomly sampled highly negative sentiment-bearing words rather than coherent tourist reviews. The experimental results demonstrate that isolated lexical information can contribute to sentiment classification, yielding improvements over the Majority Class baseline in terms of Macro F1.

However, the overall performance indicates that lexical information alone is insufficient to accurately represent the complexity of authentic tourist opinions. As future work, we plan to exploit Large Language Models to transform these sentiment-bearing lexical sets into grammatically correct and semantically coherent reviews, preserving the discriminative vocabulary while introducing the linguistic structure required for more effective synthetic data generation.

Acknowledgments

This research was supported by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico, through the Humanistic Research Project H-2025-G-283, entitled *"Procesamiento*

de lenguaje natural adaptativo con enfoque humanístico para apoyo comunicacional y diagnóstico en pacientes con enfermedades neurodegenerativas". The authors gratefully acknowledge the financial support provided by this project, which made this research possible.

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-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.

[2] 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).

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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).

[11] 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.

[12] 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.
[13] 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.
[14] 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).
[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. 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.
[17] 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.
[18] 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).
[19] 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.