Semantic Lexical Augmentation Using Embeddings for Synthetic Tourist Review Generation

Luisa Agudelo-Fuentes[1,*], Rodrigo Sebastián Rojas-Miranda[2]

[1]Universidad Iberoamericana, México
[2]Instituto Tecnológico de Toluca, México

Abstract

This paper presents a lightweight synthetic data generation approach for the REST-MEX 2026 Artificial Data Generation Track. The proposed methodology generates synthetic tourist reviews by applying controlled lexical transformations to authentic opinions from the training corpus. Specifically, verbs and adjectives are identified through morphological analysis and replaced with semantically similar alternatives retrieved from a FastText embedding model. Unlike approaches based on Large Language Models, the proposed method preserves the original sentence structure while increasing lexical diversity with minimal computational cost. Experimental results show that the augmented corpus improves the official evaluation metrics over the Majority Class baseline, demonstrating that semantic lexical substitution constitutes an effective and fully reproducible strategy for data augmentation in tourism sentiment analysis.

Keywords

Rest-Mex, Sentiment Analysis, Data Augmentation, Word Embeddings, Lexical Substitution

1. Introduction

The continuous growth of user-generated content has transformed Natural Language Processing (NLP) into one of the fundamental technologies supporting modern tourism research. Online reviews contain valuable information regarding visitors' experiences, service quality, destination image, and customer satisfaction, enabling both researchers and practitioners to understand tourist behavior at unprecedented scales. Consequently, sentiment analysis has become a key component of intelligent tourism systems, supporting applications such as recommendation engines, destination management, digital reputation analysis, conversational assistants, and decision-support platforms [1, 2, 3, 4, 5].

Recent advances in deep learning have substantially improved the automatic analysis of tourism-related text. Transformer-based architectures have demonstrated remarkable performance for sentiment classification, topic discovery, and multilingual opinion mining, significantly outperforming traditional machine learning approaches on numerous benchmark datasets [6, 7, 8, 9, 10]. Likewise, artificial intelligence techniques have expanded beyond textual analysis, contributing to destination image assessment, media analysis, multimodal tourism intelligence, gastronomic evaluation, and visual content understanding [11, 12, 13, 14, 15, 16].

The REST-MEX evaluation campaign has become one of the principal benchmarks for Spanish-language tourism NLP. Since its first edition, the shared task has progressively increased both the complexity of the datasets and the diversity of the proposed challenges. Earlier editions focused on recommendation systems, polarity classification, and service categorization, while subsequent competitions incorporated multitask learning and increasingly sophisticated transformer-based solutions [17, 18, 19, 20]. The 2026 edition introduces a different research perspective through the Artificial Data Generation Track, where participants are challenged to improve a predefined sentiment classifier by generating synthetic tourist reviews rather than modifying the classification architecture itself [21, 22].

Most current approaches to synthetic text generation rely on Large Language Models. Although these systems produce highly fluent and contextually coherent reviews, they often require considerable

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
A2617335@correo.uia.mx (L. Agudelo-Fuentes); rsrojasm@toluca.tecnm.mx (R. S. Rojas-Miranda)

computational resources and provide limited control over the generation process. Furthermore, producing entirely new opinions may unintentionally introduce factual inconsistencies or lexical distributions that differ from those observed in authentic tourist reviews. Consequently, alternative data augmentation strategies that preserve the linguistic characteristics of the original corpus remain attractive for sentiment classification tasks.

One of the simplest yet most effective mechanisms for increasing corpus diversity consists of generating new training instances through controlled transformations of existing documents. Rather than creating entirely novel reviews, the original opinions can be modified while preserving their semantic orientation, thereby enriching the lexical variability available during training. This philosophy has proven effective in several Natural Language Processing tasks, particularly when limited resources or highly specialized domains restrict the availability of manually annotated data [23, 24, 25, 26].

In this work, we propose a lightweight synthetic review generation strategy based on controlled linguistic transformations. Starting from authentic tourist reviews, the proposed methodology applies lexical filtering and systematic textual modifications to generate additional training instances that preserve the original sentiment labels while introducing lexical variability. Unlike prompt-based generation methods, the proposed approach requires no external generative models, making it computationally inexpensive, fully reproducible, and easily interpretable. The resulting synthetic corpus is incorporated into the official REST-MEX 2026 training dataset in order to evaluate its contribution under the competition's official evaluation protocol.

2. Related Work

Natural Language Processing has become one of the principal technologies for analyzing large collections of user-generated tourism content. During the last decade, numerous studies have demonstrated that opinion mining techniques provide valuable insights into visitor satisfaction, destination image, service quality, and traveler behavior. Systematic reviews have documented the rapid adoption of NLP techniques within tourism research, highlighting sentiment analysis as one of the most active research topics in the field [1, 2]. Beyond textual analysis, recent studies have also incorporated computer vision and multimodal learning to analyze destination promotion, digital reputation, gastronomic evaluation, and visual perception of tourist destinations, illustrating the increasing integration of Artificial Intelligence into tourism research [11, 12, 13, 14, 15, 16].

Several intelligent tourism applications have emerged from these advances. Recommendation systems, conversational assistants, destination monitoring platforms, and hotel reputation analysis increasingly rely on NLP techniques to automatically process large volumes of textual information [3, 4, 27, 5]. These applications demonstrate that the quality of textual representations directly influences the effectiveness of downstream decision-support systems.

The REST-MEX shared task has played an important role in promoting research on Spanish-language tourism text mining. Since its first edition in 2021, the benchmark has progressively expanded both the complexity of its datasets and the diversity of the proposed tasks. Early editions focused on recommendation systems and polarity prediction, whereas subsequent competitions incorporated thematic classification, COVID-related challenges, multitask learning, and progressively larger tourism corpora [17, 18, 19]. The 2025 edition introduced fine-grained destination classification over forty Mexican *Pueblos Mágicos*, while REST-MEX 2026 shifted the research focus toward synthetic text generation, encouraging participants to improve sentiment classification through data augmentation instead of classifier modification [20, 21, 22].

Regarding sentiment classification, recent years have witnessed a clear transition from traditional machine learning methods toward Transformer-based architectures. Domain-adapted variants of BETO, RoBERTa, and multitask transformer models have consistently achieved state-of-the-art performance on Spanish tourism corpora [7, 28, 29, 10, 8, 9]. Other studies have explored ensemble learning, demonstrating that combining multiple classifiers often yields more robust predictions than relying on a single model [30]. Similar trends have also been observed in other Natural Language Processing tasks, where

deep neural architectures continue replacing conventional machine learning approaches [31, 32, 33].

While classifier architectures have received considerable attention, the problem of data augmentation has become increasingly relevant as modern models require large quantities of balanced training data. Several studies have investigated corpus balancing strategies for tourism sentiment analysis, including representative instance selection, controlled oversampling, and synthetic review generation guided by lexical information [23, 24]. These approaches seek to enrich minority sentiment classes while preserving the semantic characteristics of authentic reviews, thereby improving model generalization under highly imbalanced conditions.

Unlike recent approaches that rely primarily on Large Language Models for synthetic text generation, the methodology proposed in this work adopts a lightweight transformation-based strategy. Rather than generating completely new reviews, it constructs additional training instances by applying controlled linguistic modifications to authentic tourist opinions. This philosophy preserves the lexical and stylistic properties of the original corpus while maintaining low computational cost, making it particularly suitable for reproducible and resource-efficient sentiment analysis systems. More broadly, this direction aligns with current research trends that emphasize efficient machine learning algorithms, interpretable models, and scalable Artificial Intelligence solutions for Natural Language Processing [34, 35, 36, 37, 38].

3. Proposed Method

The proposed methodology generates synthetic tourist reviews by applying controlled lexical transformations to authentic opinions extracted from the REST-MEX 2026 training corpus. Unlike approaches based on Large Language Models, the proposed strategy does not create reviews from scratch. Instead, it preserves the semantic structure of existing opinions while introducing lexical variability through the replacement of selected words. The underlying hypothesis is that exposing the sentiment classifier to multiple lexical realizations of the same opinion may improve its generalization capability without modifying the original sentiment orientation.

Let $D = (x_i, y_i)_{i=1}^{N}$ denote the original training corpus, where x_i is a tourist review and $y_i \in 1, 2, 3, 4, 5$ is its corresponding polarity label. The objective is to construct an augmented corpus $D^* = D \cup D'$, where $D' = (x' * i, y_i) * i = 1^M$ contains the synthetic reviews and $M \leq 3000$ according to the competition rules.

3.1. Random Review Selection

The generation process begins by randomly selecting a subset $S = x_1, x_2, \ldots, x_M$ such that $S \subset D$. Since the reviews are sampled uniformly, the selected subset naturally preserves the diversity of destinations, tourism services, writing styles, and sentiment categories present in the original corpus.

3.2. Morphological Filtering

Each selected review is tokenized and morphologically analyzed before any lexical modification is performed. Given a review represented as $x = (w_1, w_2, \ldots, w_n)$, every token w_i is assigned a grammatical category through a part-of-speech tagging function $P(w_i)$.

Only tokens satisfying $P(w_i) \in$ Verb, Adjective are considered candidates for replacement. These grammatical categories were selected because they largely determine the subjective content of tourist reviews while allowing lexical substitutions that minimally affect sentence structure.

3.3. Semantic Lexical Substitution

Candidate words are replaced using semantic information extracted from a pre-trained FastText embedding model. Let $E : V \rightarrow \mathbb{R}^d$ denote the embedding function, where V is the vocabulary and d is the embedding dimension.

The semantic similarity between two words is computed using cosine similarity,

$$\text{sim}(u, v) = \frac{E(u) \cdot E(v)}{|E(u)||E(v)|}.$$

For each candidate word w, the algorithm retrieves its neighborhood $N_k(w) = v_1, v_2, \ldots, v_k$ composed of the k nearest words in the embedding space. A replacement word is randomly selected from this neighborhood provided that it satisfies the similarity constraint $\text{sim}(w, v) \geq \tau$, where τ is a predefined threshold.

3.4. Synthetic Review Generation

The lexical transformation is represented by the operator $T : X \to X$, where each original review x is converted into a synthetic review $x' = T(x)$. Since the transformation only modifies semantically related words, the sentiment polarity remains unchanged, that is, $y(x') = y(x)$.

Applying the transformation operator to every review in the selected subset produces the synthetic corpus $D' = (T(x_i), y_i)_{i=1}^{M}$. Finally, the augmented dataset $D^* = D \cup D'$ is used during the fine-tuning stage of the official baseline classifier released by the REST-MEX 2026 organizers.

The proposed approach combines three desirable properties for data augmentation. First, every synthetic instance originates from a real tourist review, preserving the linguistic characteristics of the corpus. Second, semantic substitutions introduce lexical diversity without altering the original polarity annotation. Finally, the methodology is lightweight, fully reproducible, and independent of external generative models, making it suitable for efficient synthetic data generation in sentiment analysis tasks.

4. Proposal Results

The proposed lexical transformation strategy was evaluated following the official protocol of the REST-MEX 2026 Artificial Data Generation Track. The competition employs the *Ranking Performance* (RP) score as its primary evaluation metric, since it assigns greater importance to minority sentiment classes and provides a more representative assessment under highly imbalanced class distributions. In addition to RP, the evaluation includes Macro F1, overall Accuracy, and the F1-score for each individual sentiment category, allowing a detailed analysis of the classifier's behavior across the five polarity levels.

The proposed approach obtained an RP score of 0.0142 and a Macro F1 score of 0.0257. These results indicate that the generated synthetic reviews contributed to the learning process by enabling the classifier to recognize sentiment categories beyond the trivial majority-class prediction. Although the improvement remains modest, the obtained scores demonstrate that controlled lexical transformations constitute a viable data augmentation strategy despite their simplicity.

Analyzing the individual sentiment categories reveals that the best performance was achieved for polarity class 5, reaching an F1-score of 0.0812, followed by class 2 with an F1-score of 0.0470. In contrast, classes 1 and 3 obtained an F1-score of 0, while class 4 achieved only 0.0003. This behavior reflects the inherent difficulty of learning the minority classes in the REST-MEX dataset, where extremely positive reviews dominate the training corpus. Consequently, lexical transformations appear to benefit primarily those sentiment categories that already possess a relatively richer lexical representation.

Compared with the Majority Class baseline, the proposed method produced positive values for both RP and Macro F1, confirming that the generated reviews introduced useful variability into the training process. Unlike the baseline, which predicts a single sentiment category for every instance, the augmented corpus enabled the classifier to identify multiple polarity levels, resulting in measurable improvements under the official competition metrics.

An important advantage of the proposed methodology lies in its simplicity. The generation process requires neither external Large Language Models nor expensive inference procedures, relying exclusively on semantic relationships captured by distributed word representations. Consequently, the entire augmentation pipeline is computationally efficient, fully reproducible, and straightforward to implement. Furthermore, because every synthetic review originates from an authentic tourist opinion, the generated corpus naturally preserves the lexical style and domain-specific vocabulary of the original dataset.

Nevertheless, the experimental results also reveal several limitations. Lexical substitutions modify only a relatively small portion of each review while preserving the original syntactic structure. As a consequence, although the generated opinions introduce additional lexical diversity, they may not provide sufficient semantic variability to substantially improve the generalization capability of the sentiment classifier. In addition, semantic proximity in the embedding space does not always imply complete equivalence in contextual usage, meaning that some substitutions may produce expressions that are grammatically correct but less natural in authentic tourism discourse.

Overall, the experiments demonstrate that controlled lexical transformation constitutes an effective lightweight baseline for synthetic review generation. While its performance remains below that of more sophisticated neural generation approaches, the methodology successfully enriches the training corpus with minimal computational cost and without requiring external generative models. These findings suggest that lexical augmentation can serve as a complementary component within larger synthetic data generation pipelines, particularly when computational efficiency, reproducibility, and interpretability are desirable characteristics.

5. Conclusions

This paper presented a lightweight synthetic data generation strategy based on controlled lexical transformations for the REST-MEX 2026 Artificial Data Generation Track. By replacing selected verbs and adjectives with semantically similar alternatives obtained from FastText embeddings, the proposed method generated additional training instances while preserving the original sentiment labels and the linguistic characteristics of authentic tourist reviews.

The experimental results demonstrated that the augmented corpus improved the official evaluation metrics over the Majority Class baseline, confirming that lexical augmentation can provide useful variability for sentiment classification despite its simplicity. Although the overall performance remains below that of more sophisticated generation approaches, the proposed methodology offers an efficient, interpretable, and fully reproducible alternative that does not rely on Large Language Models.

As future work, we plan to incorporate contextual embedding models for more accurate lexical substitutions, explore syntactic and semantic paraphrasing techniques, and investigate hybrid augmentation strategies that combine lexical transformations with neural text generation to produce more diverse and informative synthetic tourist 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] 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] 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).
[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] 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.
[7] 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).
[8] D. S. L. Guilarte, V. Herrera-Semenets, L. Bustio-Martínez, M. Á. Álvarez-Carmona, Bert-based models for joint sentiment, type, and location classification of spanish tourist reviews (2025).
[9] J. J. M. Borjón, M. Á. Álvarez-Carmona, Multitask classification of mexican tourist reviews using a multi-head transformer model based on beto (2025).
[10] I. Siliceo-Guzmán, R. Aranda, M. Alvarez-Carmona, Hybrid prompt engineering and transfer learning for sentiment analysis in mexican tourism reviews, IberLEF@ SEPLN (2025).
[11] 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.
[12] 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).
[13] A. D. Pacheco, M. A. Á. Carmona, A. Y. R. González, H. Carlos, R. Aranda, Measuring the difference between pictures from controlled and uncontrolled sources to promote a destination. a deep learning approach., International Journal of Interactive Multimedia and Artificial Intelligence 9 (2025) 18–31.
[14] 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.
[15] A. Díaz-Pacheco, M. Á. A. Carmona, R. Aranda, R. Guerrero-Rodríguez, An advanced deep learning framework for discerning relevant and irrelevant content in image analysis for tourism studies (2025).
[16] 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.
[17] 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).
[18] 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.
[19] 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.

[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.
[23] 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.
[24] 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).
[25] J. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, Lexical complexity assessment of spanish in ecuadorian public documents, Procesamiento del Lenguaje Natural 74 (2025) 291–303.
[26] D. Fajardo-Delgado, L. Rodriguez-Coayahuitl, M. G. Sánchez-Cervantes, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, Readability formulas for elementary school texts in mexican spanish, Applied Sciences 15 (2025) 7259.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] M. Á. Alvarez-Carmona, R. Aranda, Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022).
[34] A. Y. Rodríguez-González, M. Á. Álvarez-Carmona, Machine learning and soft computing: Current trends and applications, 2026.
[35] A. Y. Rodríguez-González, R. M. Valdovinos-Rosas, G. Bernal Baró, R. Aranda, A. Díaz-Pacheco, M. Á. Álvarez-Carmona, Pso-fspminer: A metaheuristic approach for mining a representative subset of frequent similar patterns, Algorithms 19 (2026) 229.
[36] L. Bustio-Martínez, V. Herrera-Semenets, J. A. González-Ordiano, Y. Pérez-Guadarrama, L. N. Zú niga Morales, D. Montoya-Godínez, J. Van Den Berg, et al., Enhanced phishing detection using multimodal data, Knowledge-Based Systems (2025) 115105.
[37] V. Herrera-Semenets, L. Bustio-Martínez, J. v. d. Berg, M. Á. Álvarez-Carmona, Detecting economic vulnerability via multi-agent llm architecture and context-aware cluster analysis, in: International

Congress on Artificial Intelligence and Pattern Recognition, Springer, 2025, pp. 255–267.
[38] M. Á. Álvarez-Carmona, I. S. Guzmán[1], 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.