Generating Synthetic Tourist Reviews for Sentiment Classification Using a Multi-LLM Dual-Stream Pipeline

Antonio Zempoaltecatl Navarro[1,*], Diana Coyotl Carcamo[1] and Mauricio Castro Valencia[1]

[1]*Benemérita Universidad Autónoma de Puebla, Puebla, México*

Abstract

The REST-MEX 2026 shared task challenges participating teams to generate a synthetic corpus of Spanish-language tourist reviews rather than submit classification predictions. Team *Modern Data* from the Benemérita Universidad Autónoma de Puebla (BUAP) addresses this challenge with a multi-LLM dual-stream generation pipeline deployed via OpenRouter. Stream 1 conditions generation on implicit sentiment key-points using Llama 3.1 70B, while Stream 2 anchors generation to real-data seeds using Claude Sonnet 4.6, GPT-4o-mini, and Llama 3.1 70B. Generated reviews undergo agentic label verification by Claude 3 Haiku (99.9% acceptance rate) followed by BETO-based perplexity filtering. The submitted corpus of 3,000 reviews intentionally overrepresents low-frequency negative classes to improve performance under the RP(k) metric. The official evaluation yielded an RP Score of 0.4663 and a Macro F1 of 0.5044, with strong performance on class 5 (F1 = 0.7697) but lower scores on class 2 (F1 = 0.3486) and class 3 (F1 = 0.3830).

Keywords

synthetic data generation, sentiment analysis, tourist reviews, large language models, Spanish NLP, TSTR evaluation

1. Introduction

Sentiment analysis of user-generated content from tourism platforms enables service providers and destination managers to track quality trends across hotels, restaurants, and cultural attractions. In the context of Mexican tourism, the REST-MEX series at IberLEF has established the *Pueblos Mágicos* review collection as a challenging benchmark for Spanish-language opinion mining [1, 2, 3].

REST-MEX 2026 [4], organized within IberLEF 2026 [5], introduces a paradigm shift from prior editions: instead of evaluating classifiers on provided data, participating teams are asked to submit a synthetic corpus of 3,000 labeled tourist reviews in Spanish. The official evaluators then train a standard sentiment classifier exclusively on each submitted corpus and evaluate it on a hidden set of real reviews. This indirect evaluation scheme, known as Train on Synthetic, Test on Real (TSTR), directly measures the downstream utility of generated data for sentiment classification rather than surface-level text quality.

The task presents several challenges. First, real reviews are strongly skewed toward positive polarities (65.6 % of the provided training set is rated 5/5), so a corpus that mirrors this distribution would perform poorly on minority classes under the RP(k) weighted metric. Second, models must produce colloquial Mexican Spanish that is stylistically consistent with authentic tourist reviews. Third, each review must express a specific polarity without containing an explicit numerical rating, preventing trivially learned lexical cues.

In this paper we describe the approach of team *Modern Data* from BUAP. Our main contributions are: (1) a dual-stream multi-LLM generation pipeline that combines key-point-conditioned and seed-anchored generation strategies; (2) an intentional class-rebalancing strategy that overrepresents negative polarities to improve RP(k) on minority classes; and (3) a multi-stage quality-filtering pipeline consisting of agentic label verification and BETO-based perplexity filtering.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

anthonydevxp@gmail.com (A. Z. Navarro); diana.coyotlc@alumno.buap.mx (D. C. Carcamo); castrovalencia844@gmail.com (M. C. Valencia)

2. Related Work

2.1. LLM-Based Synthetic Data Generation

Generating synthetic training corpora with large language models (LLMs) has emerged as a practical strategy for low-resource and class-imbalanced NLP tasks. Few-shot prompting and chain-of-thought conditioning have been shown to produce labeled text that is useful for downstream classification even when the generating model is not fine-tuned on the target domain. A key challenge is *distribution shift*: LLM-generated text often occupies a different manifold in the feature space compared to human-generated text, which directly penalizes TSTR-style evaluations. MAUVE [6], a divergence-frontier metric based on feature-space comparison, quantifies this shift and has been used as a go/no-go criterion in data generation pipelines.

2.2. Dual-Stream Generation

The dual-stream generation paradigm uses two complementary strategies to improve both diversity and label fidelity of synthetic corpora. Stream 1 generates text conditioned on implicit key-points or attributes, ensuring topical coverage without explicitly encoding sentiment labels in the surface form. Stream 2 uses real instances as few-shot anchors, grounding the generation in the authentic register of the target domain. Our pipeline adapts this paradigm for the tourist review setting.

2.3. Spanish Sentiment Analysis and BETO

Pre-trained language models for Spanish, particularly BETO [7] — a BERT-style [8] model trained on the Spanish Wikipedia and other corpora using whole-word masking — have become the standard baseline for Spanish NLP classification benchmarks. In the REST-MEX task series [4], BETO-based classifiers serve as the reference evaluator, making it the natural choice for both local TSTR evaluation and the official evaluation pipeline.

2.4. TSTR Evaluation

The TSTR paradigm [9] evaluates synthetic text by training a classifier on the generated corpus and measuring its performance on real held-out data. This approach avoids the shortcomings of intrinsic metrics (perplexity, BLEU) that do not correlate reliably with downstream utility. REST-MEX 2026 adopts TSTR as its sole evaluation protocol, using the RP(k) metric — a class-frequency-weighted Macro F1 — to reward performance on rare polarities.

3. Dataset

The REST-MEX 2026 organizers provided a publicly available training set of 208,051 tourist reviews collected from Mexican Pueblos Mágicos. Each review is labeled with a polarity score from 1 (very negative) to 5 (very positive) and includes metadata: venue type (Hotel, Restaurant, Attractive), geographic region, and town name.

Table 1 shows the class distribution of the provided training corpus. The dataset is heavily skewed: 65.6 % of reviews carry polarity 5, while fewer than 3 % are negative (polarity 1 or 2 combined). This severe imbalance is characteristic of online review platforms and directly motivates our corpus design strategy.

Reviews span 19 Mexican states. The three largest regional contributors are Quintana Roo (41.3 %), Chiapas (11.3 %), and Estado de México (9.3 %). By venue type: Restaurant (41.7 %), Attractive (33.6 %), and Hotel (24.7 %).

Table 1
Class distribution of the real training data (208,051 reviews).

Polarity	Label	Count	%
1	Very negative	5,441	2.6%
2	Negative	5,496	2.6%
3	Neutral	15,519	7.5%
4	Positive	45,034	21.6%
5	Very positive	136,561	65.6%
Total		208,051	100%

4. Methodology

4.1. Initial Corpus and Diagnostic

An initial corpus of 3,000 reviews was generated using Gemini 2.5 Flash-Lite via OpenRouter. The quality of this corpus was evaluated using MAUVE [6], which compares the embedding-space distributions of real and synthetic texts. The resulting MAUVE score was 0.0698 — well below the 0.4 threshold for acceptable distribution shift. Supporting diagnostics confirmed the issue: 30 % bigram overlap with real reviews and a vocabulary of 14 artificial phrases not observed in the real corpus. Based on this analysis, the initial corpus was discarded and the generation pipeline was redesigned from scratch.

4.2. Dual-Stream Generation Pipeline

The final corpus was generated using a dual-stream approach: Stream 1 (key-point driven) and Stream 2 (instance driven), run concurrently across three large language models accessed via OpenRouter using the OpenAI Python SDK [10]. The combined pipeline produced 5,000 reviews, from which 3,000 were selected for submission. Generation spanned 18 Pueblos Mágicos across three venue types (Hotel, Restaurant, Attractive) and required approximately 404.7 minutes of wall-clock time.

Target distribution. Rather than mirroring the real corpus distribution (65.6 % class 5), the pipeline targeted an intentionally rebalanced distribution that overrepresents minority polarities: class 1 = 25 %, class 2 = 20 %, class 3 = 20 %, class 4 = 18 %, class 5 = 17 %. This design was motivated by the RP(k) metric, which assigns higher weight to correctly classifying rare polarity classes.

Stream 1: Key-Point Driven. For each review in Stream 1, a *key-point generation* step first produced six concrete, observable aspects for the target polarity and venue type using Llama 3.1 70B (temperature 0.7, max_tokens 300). Polarity was communicated through a natural-language descriptor (e.g., *"experiencia muy por debajo de las expectativas; múltiples fallas graves que arruinaron la visita"* for class 1) rather than a numeric label, preventing stereotyped lexical patterns. A random subset of 3–5 key-points was then used to condition a second call to Llama 3.1 70B (temperature 0.9, max_tokens 350) that generated the review text as a JSON object with `titulo` and `resena` fields. Stream 1 produced 2,000 reviews.

Stream 2: Instance Driven. A seed bank of 8 real reviews per polarity class was sampled from the training corpus and stored for few-shot conditioning. For each Stream 2 generation call, 2–3 seeds were randomly selected and provided as in-context examples; the model was instructed to capture the tone and register of the seeds while producing an entirely new review for a different Pueblos Mágico. An explicit polarity reinforcement clause was appended to prompts for extreme polarities (1 and 5) to counteract drift observed in smoke tests. Three models shared Stream 2 in equal parts: Claude Sonnet 4.6 (Anthropic), GPT-4o-mini (OpenAI), and Llama 3.1 70B (Meta), each contributing approximately 1,000 reviews (temperature 0.85, max_tokens 350). Stream 2 produced 3,000 reviews.

4.3. Quality Filtering

Agentic label verification (Phase 9). Each of the 5,000 generated reviews was submitted to Claude 3 Haiku as an independent evaluator via OpenRouter. The model assigned a predicted polarity score in the range 1–5 (temperature 0.0, max_tokens 5); reviews differing from the assigned polarity by more than one point were rejected. This step yielded a 99.9 % acceptance rate: 4,997 of 5,000 reviews passed. Only three reviews were rejected (two from class 1, one from class 2).

Length filter (Phase 10). Reviews outside the range of 30–300 words were discarded.

BETO perplexity filter (Phase 10). Perplexity was computed for each review using BETO [7] (`dccuchile/bert-base-spanish-wwm-cased`, batch size 16, max_length 128 tokens). Reviews in the bottom 2nd percentile (overly generic text with anomalously low perplexity) and in the top 98th percentile (incoherent text) were removed.

4.4. Corpus Composition

After filtering, 3,000 reviews were selected via stratified sampling according to the target distribution. Table 2 shows the final composition of the submitted corpus.

Table 2
Distribution of the submitted synthetic corpus (3,000 reviews).

Polarity	Label	Count	%	Avg words
1	Very negative	750	25.0%	91.8
2	Negative	600	20.0%	90.7
3	Neutral	600	20.0%	93.3
4	Positive	540	18.0%	90.0
5	Very positive	510	17.0%	88.0
Total		3,000	100%	90.8

By source model: Llama 3.1 70B contributed 1,801 reviews (60.0%), Claude Sonnet 4.6 contributed 607 (20.2%), and GPT-4o-mini contributed 592 (19.7%). The average review length was approximately 90.8 words, within the 40–120-word target range specified in the generation prompts.

5. Experiments

5.1. Classifier Setup

Following the REST-MEX 2026 TSTR protocol, BETO [7] (`dccuchile/bert-base-spanish-wwm-cased`) was fine-tuned exclusively on the synthetic corpus using HuggingFace Transformers [10]. The classifier was implemented with `BertForSequenceClassification` (109.8M parameters, 5 output classes). Training details: batch size 16, learning rate 2×10^{-5}, 4 epochs, maximum sequence length 128 tokens, AdamW optimizer with weight decay 0.01 and a 10 % linear warm-up schedule. Cross-entropy loss was weighted by the inverse class frequency of the real corpus to compensate for the known evaluation-set imbalance (class weights: $w_1 = w_2 \approx 1.995$, $w_3 = 0.691$, $w_4 = 0.240$, $w_5 = 0.079$). Training ran on an NVIDIA GeForce RTX 3060 Ti (8.3 GB VRAM) for approximately 26 seconds per epoch, reaching a validation Macro F1 of 0.9983 on the synthetic validation split.

5.2. Local Evaluation

Prior to submission, a local TSTR evaluation was conducted to estimate the expected performance on the hidden real test set. A stratified 10 % split of the real training corpus (20,806 reviews) was held

out as a proxy test set. BETO was fine-tuned on a development synthetic corpus and evaluated on this split using Macro F1 as the primary metric, yielding Macro F1 = 0.24 as a development estimate. The proxy test set preserves the natural class distribution (class 5 accounts for 65.7 % of the proxy set), representative of the evaluation conditions expected in the official submission.

6. Results

Table 3 presents the aggregate official evaluation results for team *Modern Data* in REST-MEX 2026.

Table 3
Official evaluation results — REST-MEX 2026 (Team: *Modern Data*, submission `Team_BUAP_RESTMEX2026_submission`).

Metric	Score
RP Score (official)	0.4663
Macro F1	0.5044
Accuracy	62.83%
MAE	0.4279

Table 4 shows the per-class precision, recall, and F1-score from the official evaluation.

Table 4
Per-class results — official REST-MEX 2026 evaluation.

Polarity	Label	Precision	Recall	F1
1	Very negative	0.5531	0.5686	0.5607
2	Negative	0.6348	0.2403	0.3486
3	Neutral	0.4086	0.3604	0.3830
4	Positive	0.5828	0.3799	0.4600
5	Very positive	0.6715	0.9016	0.7697

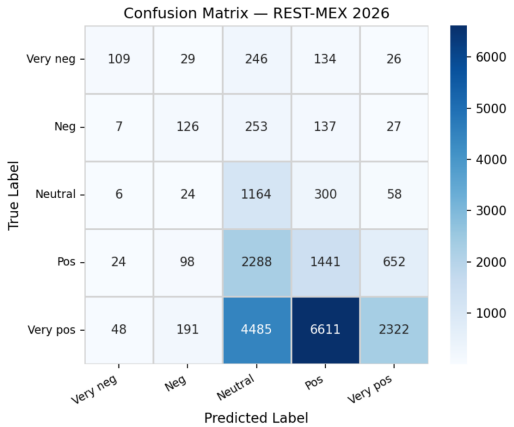

Figure 1: Confusion matrix from local TSTR evaluation (proxy test set, 20,806 real reviews). This preliminary experiment, conducted on a development corpus prior to submission, illustrates the class-confusion patterns inherent to TSTR on imbalanced real data. Official per-class results are reported in Table 4.

The official RP Score of 0.4663 falls slightly below the organizers' *baseline_original* benchmark (0.4720). The per-class breakdown in Table 4 explains this gap. Class 2 (negative) achieved the lowest F1 at 0.3486, driven by a poor recall of 0.2403: the model rarely predicts polarity 2, likely because its linguistic features overlap strongly with class 1 and class 3 — both of which are better represented

in our synthetic corpus. Class 3 (neutral) similarly underperformed at F1 = 0.3830 (precision 0.4086, recall 0.3604), reflecting the inherent ambiguity of neutral reviews and the difficulty of anchoring generation at a midpoint between positive and negative sentiment.

In contrast, the corpus design strategy proved effective for the extreme classes. Class 1 (very negative) achieved F1 = 0.5607, a strong result for a class that represents only 2.6 % of the real training corpus, validating the decision to allocate 25 % of the synthetic corpus to this polarity. Class 5 (very positive) achieved the highest F1 at 0.7697, supported by a recall of 0.9016: the model correctly identified almost all very-positive examples, reflecting the linguistic salience and consistency of this category. Class 4 (positive) landed at F1 = 0.4600 with balanced precision and recall (0.5828 vs. 0.3799), indicating moderate coverage.

The accuracy of 62.83 % and MAE of 0.4279 suggest that when the model errs, it tends to misclassify by a single ordinal step rather than making large-scale polarity inversions — a finding consistent with the confusion patterns shown in Figure 1.

7. Conclusions

We presented a multi-LLM dual-stream pipeline for generating synthetic Spanish tourist reviews for the REST-MEX 2026 shared task. The pipeline combined key-point-driven and instance-driven generation across Llama 3.1 70B, Claude Sonnet 4.6, and GPT-4o-mini, with agentic label verification (99.9 % pass rate) and BETO-based perplexity filtering as quality gates.

The official evaluation yielded an RP Score of 0.4663 and Macro F1 of 0.5044, falling marginally below the organizers' baseline of 0.4720. The corpus design strategy was partially successful: overrepresenting class 1 (very negative, 25 % of the corpus) yielded a strong F1 of 0.5607, and class 5 achieved the best per-class score (0.7697). However, classes 2 and 3 remained weak points, with F1 scores of 0.3486 and 0.3830 respectively. The low recall on class 2 (0.2403) suggests that negative but not extreme reviews are the hardest to generate distinctively, as their surface features overlap with both very-negative and neutral language.

Future work should address four directions: (1) Targeted generation improvements for minority classes 2 and 3, including harder constraints on lexical and syntactic features that differentiate negative from neutral sentiment in Spanish; (2) Class-balanced sampling strategies during corpus assembly that account not only for polarity frequency but also for inter-class linguistic similarity; (3) Retrieval-augmented generation anchored in dense embeddings of real reviews to reduce distributional shift; and (4) Intermediate proxy metrics (MAUVE per polarity class, vocabulary overlap) that correlate more reliably with official TSTR performance and can guide corpus design before submission.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with grammar checking, phrasing suggestions, and structural editing of specific sections of this paper. For *corpus generation*, Llama 3.1 70B (Meta), Claude Sonnet 4.6 (Anthropic), GPT-4o-mini (OpenAI), and Claude 3 Haiku (Anthropic) were used via OpenRouter to produce and verify the synthetic tourist reviews that constitute the submission; this usage is fully described in Section 4. After using AI-assisted tools, the authors reviewed and edited all content as needed and take full responsibility for the publication.

References

[1] 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.

[2] 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.

[3] 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.

[4] M. A. Álvarez Carmona, A. Díaz-Pacheco, R. Aranda, M.-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).

[5] A. Bonet-Jover, J. A. 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.

[6] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, Z. Harchaoui, MAUVE: Measuring the gap between neural text and human text using divergence frontiers, in: Advances in Neural Information Processing Systems (NeurIPS), volume 34, 2021.

[7] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: PML4DC at ICLR 2020, 2020.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, ACL, 2019, pp. 4171–4186.

[9] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, S. Bengio, Generating sentences from a continuous space, in: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 10–21.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.