# Contrastive Data-Driven Prompting for Synthetic Tourism Review Generation: FisBioUNAM at REST-MEX 2026

David Alexis García-Espinosa[1,*], Luis Eduardo Flores-Luna[1] and Angel Andrés Moreno-Sánchez[1]

[1]*Universidad Nacional Autónoma de México, Facultad de Ciencias, Ciudad de México, MX*

**Abstract**

This paper describes the FisBioUNAM system for REST-MEX 2026 at IberLEF 2026, where participants must generate a synthetic corpus of 3,000 Spanish tourism reviews labeled with sentiment polarity (1–5), evaluated indirectly by training a classifier on the synthetic data and testing on real reviews. We built and empirically tested a six-phase pipeline: corpus analysis with TF-IDF fingerprinting, a controlled mini-experiment comparing class distributions, contrastive prompting with Claude Sonnet, and three-stage filtering over 4,500 over-generated reviews. Each pipeline decision was validated before scaling: uniform distribution outperformed RP-weighted allocation by 4.9pp, and a specialized prompt for neutral reviews improved class 3 accuracy from 32.9% to 51%. The final corpus achieved a Track Score of 0.4366 (10th of 18 teams), below the organizer baseline (0.4721). Analysis reveals a persistent domain gap—12.5% vocabulary overlap and shorter synthetic texts (38 vs. 65 words)—pointing to LLM stylistic homogeneity as the main bottleneck.

**Keywords**

Synthetic Data Generation, Sentiment Analysis, Tourism NLP, Prompt Engineering, Spanish NLP, Large Language Models, Mexican Magical Towns



## 1. Introduction

REST-MEX 2026 [1], part of IberLEF 2026 [2], shifts the task from classifying tourist reviews to *generating* them: participants submit a synthetic corpus of 3,000 reviews and are evaluated by the performance of a classifier trained exclusively on that data against hidden real reviews. Previous editions focused on direct classification [3, 4, 5]; the 2026 edition asks whether LLM-generated reviews can replace real annotated data for sentiment analysis in Spanish.

The task has four concrete challenges. First, the real corpus exhibits severe class imbalance (65.6% class 5, 2.6% class 1). Second, the RP score penalizes systems that neglect minority classes via inverse-frequency weighting. Third, reviews span 40 Magical Towns with distinct characteristics. Fourth, the synthetic data must encode discriminative patterns—not just surface-level fluency—to enable classifier generalization.

The evaluation employs the RP score, a weighted macro F1 metric that gives higher weight to minority sentiment classes:

$$\mathrm{RP}(k) = \frac{\sum_{i \in C} \left(1 - \frac{TC_i}{TC}\right) F_i(k)}{\sum_{i \in C} \left(1 - \frac{TC_i}{TC}\right)}, \quad C = \{1, 2, 3, 4, 5\} \tag{1}$$

where $TC_i$ is the number of instances of class $i$ in the real test set, $TC$ is the total, and $F_i(k)$ is the F1 score for class $i$ obtained by system $k$.

*IberLEF 2026,September 2026, León, Spain*

*Corresponding author.

davale@ciencias.unam.mx (D. A. García-Espinosa); lefloresluna@ciencias.unam.mx (L. E. Flores-Luna); aams_fbm23@ciencias.unam.mx (A. A. Moreno-Sánchez)

0000-0001-6141-407X (D. A. García-Espinosa); 0009-0007-6087-8658 (L. E. Flores-Luna); 0009-0007-8978-6087 (A. A. Moreno-Sánchez)

### 1.1. Contributions

Our main contribution is a pipeline where each design decision—class distribution, prompt strategy, filtering thresholds—was tested empirically before scaling to the full 3,000-review corpus. Specifically:

1. A controlled mini-experiment (300 reviews, $0.10) showing uniform distribution outperforms RP-weighted allocation by 4.9pp.
2. Contrastive prompts informed by per-class TF-IDF fingerprints and prototypical examples, with a targeted variant for the class 3 bottleneck.
3. A three-stage filtering pipeline (consistency, length, diversity) over 1.5× over-generated reviews.
4. Quantitative analysis of the domain gap (12.5% vocabulary overlap, length mismatch) that explains the below-baseline result and identifies concrete failure modes.

## 2. Related Work

### 2.1. Synthetic Data Generation with LLMs

Ye et al. [6] showed that GPT-2/3 can generate competitive training data for text classification without labeled examples. Li et al. [7] explored label-conditioned generation for sentiment analysis, finding that prompt quality and example diversity are critical. Most prior work evaluates on English corpora and does not address Spanish tourism discourse.

### 2.2. Data Augmentation for Sentiment Analysis

Traditional augmentation techniques (back-translation, synonym replacement, EDA [8]) preserve labels but add limited diversity. LLM-based generation offers greater diversity but risks modal collapse—superficially varied but semantically homogeneous outputs [9]. Our over-generation and filtering pipeline targets this problem.

### 2.3. Indirect Evaluation of Synthetic Data

The train-on-synthetic, test-on-real paradigm has been explored in computer vision [10] and recently in NLP. The central obstacle is the domain gap between synthetic and real distributions, addressed through statistical alignment, consistency filtering, or adversarial refinement.

## 3. Methodology

Our system is structured as a six-phase iterative pipeline, illustrated in Figure 1.

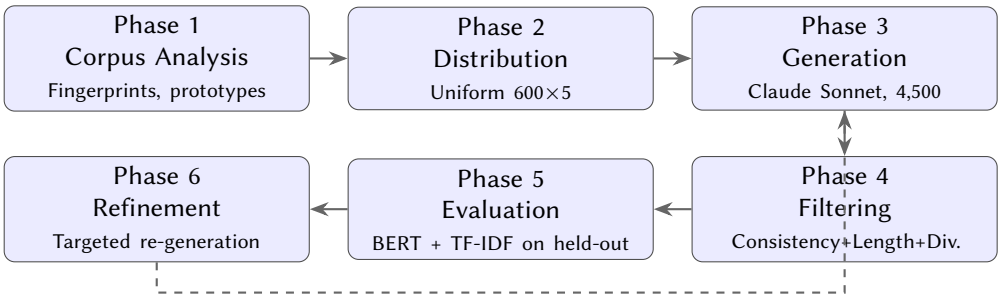


**Figure 1:** Six-phase pipeline for synthetic review generation. Dashed arrow indicates iterative refinement.

### 3.1. Data Preprocessing and Corpus Analysis

The organizers provided 208,051 real tourist reviews in Spanish. The corpus exhibits severe class imbalance: polarity 5 accounts for 65.6% of reviews, while polarities 1 and 2 each represent only 2.6%. We removed 210 duplicate reviews and normalized polarity labels from float to integer.

For each polarity class, we extracted discriminative n-grams using TF-IDF with chi-squared feature selection. These *linguistic fingerprints* reveal the vocabulary that best separates each class:

- **Class 1:** anger and indignation ("pésimo", "nunca vuelvo", "robo", "asco")
- **Class 2:** moderate disappointment ("esperaba más", "no cumplió", "decepcionante")
- **Class 3:** balanced/neutral ("está bien pero", "nada especial", "regular")
- **Class 4:** satisfaction with nuance ("buen lugar", "recomendable", "bonito pero")
- **Class 5:** superlative enthusiasm ("increíble", "espectacular", "100% recomendado")

We also identified *prototypical examples*—reviews classified with highest confidence by a baseline Logistic Regression + TF-IDF model—to serve as few-shot exemplars in prompts. A confusion matrix analysis confirmed that adjacent-class confusion (3↔4, 3↔2, 4↔5, 1↔2) dominates classification errors, directly informing our contrastive prompt design.

### 3.2. Distribution Validation

We hypothesized that over-representing minority classes (inversely proportional to RP weights) would maximize the score. To test this without committing budget, we conducted a controlled mini-experiment with 300 reviews ($0.10 USD) comparing two distributions:

**Table 1**
Distribution experiment: Uniform vs. RP-weighted (300 reviews).

| Config | P1 | P2 | P3 | P4 | P5 | Total | RP |
|---|---|---|---|---|---|---|---|
| Uniform | 60 | 60 | 60 | 60 | 60 | 300 | **0.2590** |
| RP-weighted | 70 | 85 | 60 | 50 | 35 | 300 | 0.2100 |

The uniform distribution outperformed RP-weighted by 4.9 points. Analysis revealed that RP-weighted improved class 2 (+14.0pp) but destroyed class 3 (−18.6pp) and class 5 (−9.0pp), because the RP formula weights *all* minority classes, not just the most extreme ones. We adopted uniform 600×5 for the final corpus.

Counter-intuitively, uniform distribution outperforms RP-weighted allocation because reducing class 5 examples to boost classes 1–2 destroys the classifier's ability to learn the majority class, which despite its low RP weight ($w_5 \approx 0.344$) still contributes to the score and affects decision boundaries for adjacent classes. Notably, classes 1–2 together account for ∼48.5% of the total RP weight despite representing only 5.2% of the data.

### 3.3. Contrastive Data-Driven Generation

We used Claude Sonnet (`claude-sonnet-4-20250514`) [11] via the Anthropic API, generating reviews in batches of 10 per API call. Each prompt incorporated:

1. **Linguistic fingerprint:** The discriminative n-grams for the target polarity class, anchoring vocabulary in real-data patterns.
2. **Prototypical examples:** 2–3 high-confidence exemplars from the real corpus as few-shot references.
3. **Contrastive boundary instructions:** Explicit descriptions of how each class differs from its neighbors (e.g., "Class 2 expresses clear disappointment but not extreme anger like Class 1; it may acknowledge some positive aspect").

4. **Geographic specificity:** A randomly assigned Magical Town with its region, encouraging location-specific details.
5. **Type specification:** Hotel, Restaurant, or Attractive, following the real corpus distribution.

A specialized prompt variant (v3) was developed for class 3 (neutral), which was identified as the main bottleneck—its accuracy improved from 32.9% to 51% between the validation-600 and final-3000 experiments.

Unlike generic sentiment-conditioned generation ("write a negative review"), which produces ambiguous text straddling class boundaries, this approach explicitly defines what each class is *not*, yielding more discriminative training examples.

Generation used 5 parallel workers via `ThreadPoolExecutor`, completing 4,500 reviews in approximately 8 minutes at a total API cost of $2.70 USD.

### 3.4. Filtering Pipeline

From 4,500 over-generated reviews, we applied three filtering stages:

1. **Consistency filter:** A baseline classifier trained on real data predicts the polarity of each synthetic review. Reviews with predicted polarity deviating more than $\pm 1$ from the assigned label are discarded. Result: only 1 review removed (0.0% inconsistency rate).
2. **Length filter:** Reviews shorter than a minimum threshold are removed. Result: 94 reviews removed.
3. **Diversity filter:** Cosine similarity between TF-IDF representations of same-class reviews; near-duplicates ($>0.85$) are reduced. Result: 0 duplicates found.

After filtering, 600 reviews per class were selected for the final 3,000-review corpus. Generating $1.5\times$ the target and filtering to 3,000 is more robust than generating exactly 3,000, as the length filter removes degenerate outputs while the near-zero inconsistency rate (1/4,500) confirms high label fidelity from the contrastive prompting strategy.

## 4. Implementation Details

**Table 2**
Hyperparameter configuration.

| Parameter | Value | Variable |
|---|---|---|
| Generative model | claude-sonnet-4-20250514 | API model ID |
| Reviews generated (raw) | 4,500 | $1.5\times$ overgeneration |
| Reviews final | 3,000 ($600\times5$) | Uniform distribution |
| Magical Towns | 40 | `random.choice()` |
| Types | R:1350, H:1017, A:633 | Real-data proportional |
| Batch per API call | 10 reviews | Title + Review |
| Parallel workers | 5 | `ThreadPoolExecutor` |
| Internal eval model | BERT Spanish (BETO) [12] | `dccuchile/bert-base-spanish-wwm-cased` |
| Eval split | 70/30 | `random_state=2026` |

The implementation uses Python with the Anthropic SDK for generation and HuggingFace Transformers [13] with PyTorch [14] for BERT-based evaluation. All experiments ran on Google Colab with a Tesla T4 GPU. The output format is a UTF-8 CSV with columns: Title, Review, Polarity, Town, Region, Type.

### 4.1. Code Availability

Our implementation is publicly available at https://github.com/Ironsss/IBERLEF_2026_FisBioUNAM.

# 5. Results and Evaluation

## 5.1. Internal Validation Results

We evaluated the synthetic corpus internally by training classifiers on the 3,000 synthetic reviews and testing on a held-out partition of 62K real reviews (70/30 split, `random_state=2026`).

**Table 3**
Internal validation results on held-out real reviews.

| Classifier | RP Score |
|---|---|
| BERT Spanish (GPU) | **0.3659** |
| LinearSVC + TF-IDF | 0.2778 |
| Naïve Bayes + TF-IDF | 0.2540 |
| Logistic Regression (balanced) | 0.2454 |

BERT [15] outperformed all TF-IDF-based classifiers (Table 3), suggesting that pretrained semantic representations partially compensate for low vocabulary overlap between synthetic and real reviews.

## 5.2. Official Competition Results

Table 4 presents the official REST-MEX 2026 results for the top teams and baselines.

**Table 4**
Official REST-MEX 2026 competition results (best run per team).

| Rank | Team | Track Score | Macro F1(Pol.) | Acc.(Pol.) |
|---|---|---|---|---|
| – | Combination MLP† | 0.6717 | 0.6471 | 82.78% |
| 1 | Gradiant | 0.5223 | 0.5564 | 67.26% |
| 2 | AxoloTux | 0.5114 | 0.5451 | 65.51% |
| 3 | LynxesAI-LKEBUAP | 0.5085 | 0.5515 | 71.85% |
| 4 | Erick_Barrios | 0.5032 | 0.5456 | 71.26% |
| – | Baseline† | 0.4721 | 0.5168 | 70.00% |
| 5 | Tata-VQ_UMSNH | 0.4933 | 0.5190 | 58.86% |
| 6 | UCOUCCICESEPlenitas | 0.4898 | 0.5238 | 62.97% |
| 7 | Jose_Rodriguez | 0.4747 | 0.5178 | 67.26% |
| 8 | Team_BUAP | 0.4663 | 0.5044 | 62.83% |
| 9 | hijos_del_gradiente | 0.4408 | 0.4701 | 54.35% |
| **10** | **FisBioUNAM** | **0.4366** | **0.4746** | **59.43%** |
| 11 | REST_MEX_LATE_iimas | 0.4242 | 0.4448 | 48.02% |
| … | … | … | … | … |
| – | Majority class† | 0.0090 | 0.1584 | 65.54% |

† Organizer-provided systems (not participant submissions).

Our submission placed 10th of 18 teams with a Track Score of 0.4366, below the organizer baseline (0.4721) and 0.0857 points behind the top team (Gradiant, 0.5223).

## 5.3. Per-Class Analysis

Figure 2 shows our per-class F1 scores compared to the top team and baseline.

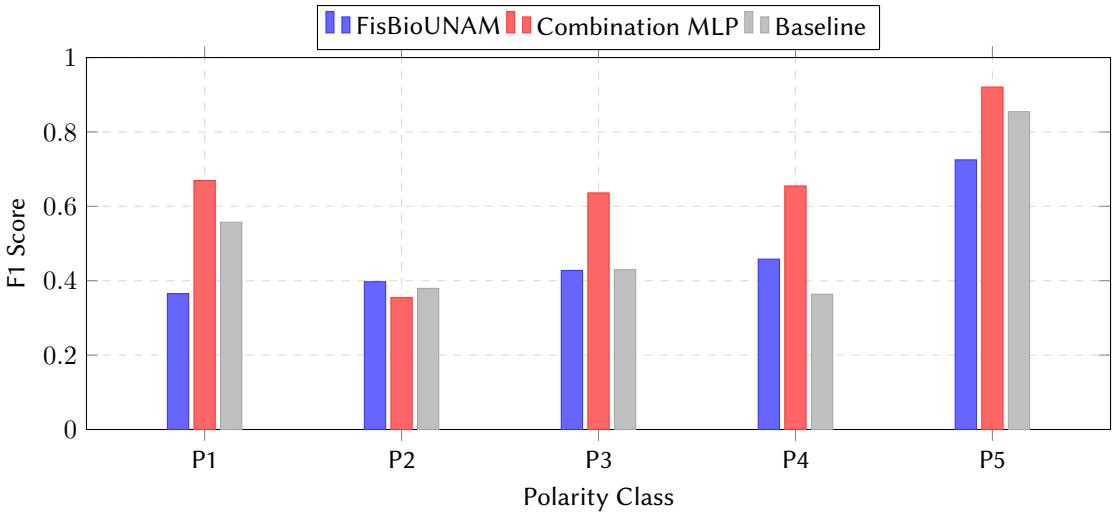


**Figure 2:** Per-class F1 scores: FisBioUNAM vs. organizer upper bound (Combination MLP) and baseline.

**Table 5**
Detailed per-class metrics for FisBioUNAM.

| Metric | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| Precision | 0.235 | 0.766 | 0.504 | 0.647 | 0.595 |
| Recall | 0.821 | 0.268 | 0.371 | 0.354 | 0.927 |
| F1 | 0.365 | 0.397 | 0.428 | 0.458 | 0.725 |

Two patterns stand out: (1) the classifier over-predicts extreme polarities—class 1 has 82.1% recall but only 23.5% precision, class 5 shows 92.7% recall and 59.5% precision; (2) mid-range classes (2–4) have low recall (26.8%–37.1%), meaning the classifier defaults to extreme predictions when uncertain. The synthetic reviews encode strong sentiment markers well but lack the nuanced vocabulary of moderate polarities.

# 6. Error Analysis

## 6.1. Domain Gap Analysis

The below-baseline result traces to the *domain gap* between synthetic and real reviews, which we measured along three dimensions:

**Table 6**
Domain gap analysis: synthetic vs. real reviews.

| Property | Synthetic | Real | Gap |
|---|---|---|---|
| Mean length (words) | $38 \pm 11$ | $65 \pm 52$ | $-41.5\%$ |
| Length std. dev. | 11 | 52 | $-78.8\%$ |
| Vocabulary overlap | 12.5% | | Low |

**Length distribution mismatch.** Synthetic reviews are shorter and more uniform ($38 \pm 11$ words vs. $65 \pm 52$). Real reviews range from one-word reactions to multi-paragraph narratives; the LLM produces consistently mid-length text, missing both extremes.

**Low vocabulary overlap (12.5%).** Despite prompts incorporating discriminative n-grams, the LLM favors formal Spanish over the colloquial, error-prone, and regionally marked language in real reviews—abbreviations, misspellings, code-switching, and platform conventions.

**Stylistic homogeneity.** Generated reviews converge on a uniform register, lacking the diversity of real user content (single-word reviews, lists, complaints addressed to management, humorous remarks).

### 6.2. Internal vs. Official Score Discrepancy

Our internal BERT evaluation yielded RP = 0.3659, while the official Track Score was 0.4366. The difference likely reflects a different classifier or training configuration used by the organizers. This gap means internal metrics were conservative predictors of official performance.

### 6.3. Classification Bias Patterns

The precision-recall asymmetry (Table 5) shows the classifier over-relies on strong sentiment markers and defaults to extreme predictions when uncertain. This follows from our prompting strategy: extreme classes received the most distinctive vocabulary, while moderate classes share overlapping features.

## 7. Conclusion

We presented a pipeline for REST-MEX 2026 where each decision was tested at small scale before committing resources. The system achieved a Track Score of 0.4366 (10th of 18 teams), below the organizer baseline (0.4721).

The empirical tests within the pipeline produced concrete findings: (1) uniform distribution outperforms RP-weighted allocation despite the intuition that over-representing minorities should help; (2) contrastive prompting improves class separability but does not address stylistic homogeneity; (3) the domain gap—12.5% vocabulary overlap, length mismatch—is the dominant bottleneck, not class distribution or prompt design.

Future work should target the domain gap directly: incorporating real review fragments via style transfer, generating reviews with controlled length distributions, and using adversarial filtering to break LLM stylistic uniformity.

## Acknowledgments

We thank the REST-MEX 2026 organizers at CIMAT for providing the training corpus and evaluation infrastructure. Computational resources were provided by Google Colab. Generation used the Claude API (Anthropic).

## Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: synthetic review generation (core task component) and grammar review. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

[1] L. A. Medina-Guerrero, et al., Overview of REST-MEX 2026: Research on synthetic text modeling for Mexican magical towns, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), 2026. To appear.

[2] IberLEF Organizers, Overview of IberLEF 2026: Natural language processing challenges for Iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), 2026. To appear.

[3] M. A. Alvarez-Carmona, et al., Overview of REST-MEX at IberLEF 2022: Recommendation system, sentiment analysis and prediction of the type of tourism for Mexican tourist texts, in: Procesamiento del Lenguaje Natural, volume 69, 2022, pp. 289–300.
[4] M. A. Alvarez-Carmona, et al., Overview of REST-MEX at IberLEF 2023: Research on sentiment analysis task for Mexican tourist texts, in: Procesamiento del Lenguaje Natural, volume 71, 2023, pp. 365–376.
[5] M. A. Alvarez-Carmona, et al., Overview of REST-MEX at IberLEF 2025, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), 2025.
[6] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, L. Kong, ZeroGen: Efficient zero-shot learning via dataset generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 11653–11669.
[7] Z. Li, H. Zhu, Z. Lu, M. Yin, Synthetic data generation with large language models for text classification: Potential and limitations, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 10443–10461.
[8] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6382–6388.
[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.
[10] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, X. Qi, Is synthetic data from generative models ready for image recognition?, arXiv preprint arXiv:2210.07574 (2022).
[11] Anthropic, Claude: A family of large language models, https://www.anthropic.com/claude, 2025.
[12] 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.
[13] 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.
[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019, pp. 8026–8037.
[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, Proceedings of NAACL-HLT (2019) 4171–4186.