Exploring Diffusion Language Models for Review Generation on REST-MEX 2026

Iván Reyes-Hernández[1,*], Ivan Meza[2]

[1]Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México, México
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México

Abstract

This paper describes our participation in REST-MEX 2026, a shared task on generating synthetic tourist reviews for *Pueblos Mágicos* from México. We explored MDLM and BD3LM using diffusion-converted Qwen 3 0.6B models, employing a two-stage training pipeline with continual pretraining and supervised fine-tuning. Our system generated 2,880 synthetic reviews, which were evaluated through a downstream sentiment classification task on real reviews. Results show strong performance on positive reviews (F1=0.7778 for class 5) but poor performance on minority classes (macro F1=0.4360). We attribute these limitations to model size and training data imbalance.

Keywords

Language Modeling, Diffusion Language Models, Sentiment Analysis, Autoregressive to Diffusion, Data-constrained

1. Introduction

Generative modeling remains a central challenge in machine learning as it serves as a metric for evaluating the ability of models to approximate underlying data distributions. Autoregressive models has dominated the field of Natural Language Processing but in recent years some novel approaches have been proposed in order to improve inherent limitations of this family of models. Among the most promising alternatives are discrete diffusion models [1].

In this paper, we explore the efficacy of discrete diffusion for controlled text generation by participating in the Generative Language Modeling shared-task at REST-MEX, held as part of IberLEF 2026 [2, 3, 4, 5].

The objective of this task is the generation of a dataset consisting of Spanish-language reviews for Mexican "Magic Towns" (Pueblos Mágicos de México). The quality of the generated data is evaluated based on its utility in a sentiment analysis classification task performed on real, human-written reviews.

We considered two distinct modeling frameworks: Masked Diffusion Language Models [6] (MDLM), a non-autoregressive approach, and Block Diffusion Language Models [7] (BD3LM), a semi-autoregressive approach. These were implemented within a two-step training framework.

The remainder of this paper is organized as follows: section 2 describes the shared-task and the provided dataset; section 3 details our proposed solution; section 4 presents out experimental results; the paper concludes discussing identified limitations and directions for future work in section 5.

2. Task and Data Description

2.1. Task description

Unlike previous editions, the shared task for REST-MEX 2026 does not consist of classifying tourist reviews; instead, it introduces a fully generative evaluation paradigm. So, the task involves generating synthetic tourist opinions and their corresponding polarities. The problem is defined as follows [8]:

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
ivannpy@ciencias.unam.mx (I. Reyes-Hernández); ivanvladimir@turing.iimas.unam.mx (I. Meza)

Table 1
Polarity statistics. There are about 25 times more 5-polarity reviews than 1-polarity ones.

Polarity	Instances
1	$5,441$
2	$5,496$
3	$15,519$
4	$45,034$
5	$136,561$
Total	$208,051$

> Given a limited collection of real tourist reviews written in Spanish, each participating team must develop a generative model capable of producing a synthetic corpus of 3,000 original tourist opinions related to Mexican Pueblos Mágicos.

The task also seeks that the generated reviews be original, resemble authentic opinions in style, content, and sentiment expression, and include an explicit sentiment polarity (from 1 (very bad) to 5 (very good)) assigned by the system itself. It is intended as an experimental testbed to explore how generative artificial intelligence can be used to replace or complement real annotated data.

The evaluation is indirect. A synthetic corpus is considered effective if a sentiment classifier model trained on it generalizes well to real-world data. Once the model is trained, it is evaluated on a held-out test partition consisting solely of real reviews, and its performance on this real test set determines the final score. The performance is measured by the impact of the generated synthetic data on downstream sentiment classification. Qualitatively, it is measured by the Polarity Score, which is designed to fairly assess sentiment classification performance under severe class imbalance, a common characteristic of real tourist reviews.

2.2. Dataset

The corpus consists of $208,051$ real reviews, each containing 6 fields: title, review, polarity, town, region, and type. The polarity is a score given by the review's author about their experience at the place; it's an integer from 1 to 5, where 5 means "very good" and 1 means "very bad". Type refers to the place, it can be a hotel, a restaurant, or an attraction. Town and Region refer to the city and state in México, respectively.

The polarity distribution among the reviews is very imbalanced. There are about 25 times as many 5-polarity reviews as 1-polarity reviews. This is an important factor to take into account since a system trained on these data could generate imbalanced samples [9].

The review types are not as imbalanced as their polarities. There are $6,720$ reviews about restaurants, $69,921$ about attractive and $51,410$ about hotels. The reviews are about popular Mexican places such as Tulum, Isla Mujeres, San Cristobal de las Casas, Valladolid, Balacar, etc.

The corpus is a subset of the original REST-MEX dataset, containing only 70% of the data. The remaining 30% is used for testing the classifier models, as described in 2.1.

3. System Description

As described in 1, we explored the novel language modeling paradigm based on discrete diffusion models. We considered two different approaches in order to apply discrete diffusion: MDLM [6] and BD3LM [7].

Both methods were applied in the framework Tiny-A2D using the *dllm* library [10]. We used the diffusion-converted [11] versions of Qwen 3 0.6B, released by Z. Zhou [10], on Hugging Face [12]. There are checkpoint versions for MDLM and BD3LM. Both were adapted from the autoregressive Qwen 3 version using `tulu-3-sft-mixture, smoltalk, opc-sft-stage1 and opc-sft-stage2` datasets.

Table 2
The original training dataset was partitioned in order to apply continual pretraining and supervised fine-tuning.

SFT dataset		CPT dataset	
Polarity	Instances	Polarity	Instances
1	5,441	1	0
2	5,496	2	0
3	5,441	3	10,078
4	5,441	4	39,593
5	5,441	5	131,120

We applied minimal pre-processing to the original data. We just fixed some encoding errors and cleaned the data, removing HTML entities, URLs, email addresses, and reducing repeated characters, if any.

We divided the training into two steps: continual pretraining and supervised fine-tuning. To accomplish this, we divided the training dataset into two partitions, each of which was used for each training step. The SFT partition consists of a polarity-balanced corpus, with as many samples for classes 3, 4, and 5 as the minimum number for classes 1 and 2 in the original dataset. The CPT partition contains the remaining instances. Table 2 shows the details.

The intention was for the model to learn to write reviews during the CPT step without using polarity labels.

We also aimed to make the model generate reviews on demand while explicitly producing their polarity. For this reason, we performed SFT on the balanced corpus, explicitly passing the polarity.

For CPT, we used text in the following format:

> TYPE en TOWN, REGION. TITLE: REVIEW.
>
> *TYPE in TOWN, REGION. TITLE: REVIEW.*

For SFT, we used the following instruction:

> Instrucción: Escribe una reseña sobre un TYPE en TOWN, REGION. Respuesta: TITLE. Calificación: POLARITY. REVIEW.
>
> *Instruction: Write a review about a TYPE in TOWN, REGION.*
>
> *Response: TITLE Score: POLARITY. REVIEW.*

3.1. Training and experimental configurations

For training, we used the following pipeline. Figure 1 summarize the process.

1. Data cleaning.
2. Data splitting into CPT and SFT partitions.
3. Template application.
   For CPT
   - Tokenize using the diffusion converted version of Qwen 3 0.6 B with sequence length equal to 1024 and insert the special EOS token.
   - Train using MDLM method on a 24 GB vRAM 4090 RTX NVIDIA GPU using maximum length 1024, 10 epochs, learning rate 3×10^{-5}, warmup ratio 0.05, and weight decay 0.1.
   - For BD3LM training, we used the same parameters and block size 32.

   For SFT
   - Tokenize using the diffusion converted version of Qwen 3 0.6 B using mask prompt loss.
   - Train using MDLM method on a 12 GB vRAM 3060 RTX NVIDIA GPU using maximum length 512, 5 epochs, learning rate 3×10^{-5}, warmup ratio 0.05 and weight decay 0.1.

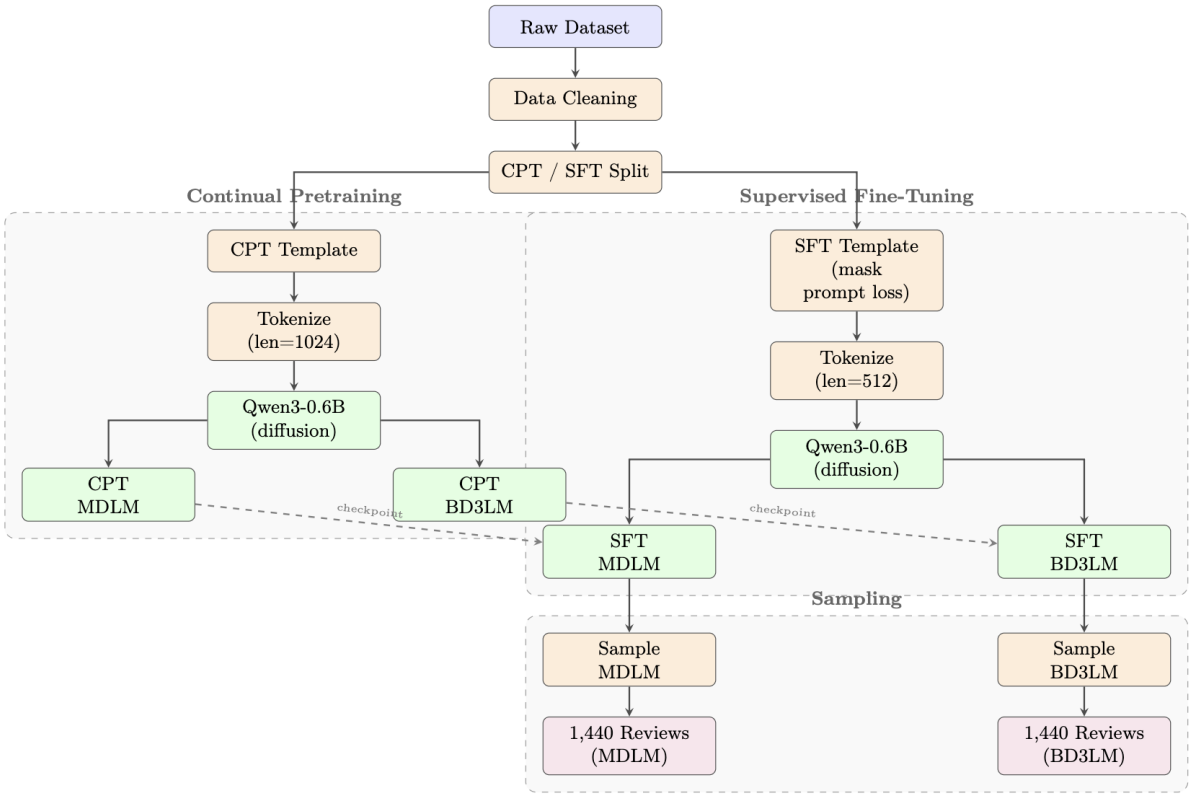

Figure 1: Training and inference pipeline for discrete diffusion language models (MDLM and BD3LM). The CPT dataset contains reviews with polarities 3, 4 and 5; SFT dataset contains an almost balanced dataset. The tokenizer used was the same for Qwen 3 0.6B. We used the diffusion converted framework [11] and model checkpoints released by [10].

- For BD3LM training we used the same parameters and block size 32 over 3 epochs.

Once the models were trained, we automated the sampling process using a basic prompt template

> Escribe una reseña P sobre T en L
>
> *Write a review P about T in L*

where P were replaced by "", "negativa/*negative*" or "positiva/*positive*", T were replaced by "un hotel/*a hotel*", "un restaurante/*a restaurant*" or "una atracción turística/*a touristic attraction*" and L were replaced by each of the Town, Region pairs in the original dataset.

We sampled $1,440$ reviews per method, allowing random and confidence remasking, with temperature 0.6, block size 32 (for BD3LM) and max-tokens 64. The sampling denoising process ran for 128 steps.

4. Results

We generated 2880 synthetic reviews. The following reviews were generated using MDLM:

> Polarity: 1. "Muy malo servicio. Visitamos este lugar al ver el calor y en la ciudad, cuando visitamos atlixco. la comida está bien adecuado y los precios bien, pero el servicio es demasiado malmano de lo que es la atención local"
>
> Polarity: 2. "Las ruinas son viejas. La historia de las ruinas es muy mala, la representación de Santa Victoria es un poco de colorado comparado a las otras ruinas que están a la entrada, recomendaría que vayan a las otras ruinas".

Polarity: 3. "Caro pero vale la visita. Es un lugar visitado, es buena, pero no como se ve desde la ciudad, por dentro de los lugares mayas, la vista muy buena y la comida de buen precio. visite encarecidamente, vale la pena visita"

Polarity: 4. "Agradable. Es un lugar agradable, con una vista de la playa, muy pintoresco, la gente es buena y amable, es muy pequeño, la comida es rica y la atención es buena, desde la recepción hasta el entorno del sitio

Polarity: 5. "Excelente hotel en todo. Excelente hotel en todo, relación calidad precio servicio, la comida es excelente. Excelente ubicacion en el hotel. Excelente servicio a todo el personal. Excelente la atención. Mia, Eli, Tom, Anel y Elia"

The following were generated using BD3LM:

Polarity: 1. "Muy mal ubicado. Muy mal ubicado. la habitación estaba sucia, la cama estaba sucia, el baño estaba sucia, el restaurante estaba sucia. la habitación estaba sucia, la cama estaba sucia, el baño estaba sucia, el restaurante estaba sucia."

Polarity: 2. "No es un lugar espectacular. El lugar es bonito pero la vista es muy linda, el lugar es muy bonito"

Polarity: 3. "Comida y mal servicio. El lugar es muy bonito, la comida es buena, pero el servicio es lento"

Polarity 4: "Muy bonito el lugar. El lugar es muy bonito, la decoración es muy bonita, el arte es muy interesante"

Polarity 5: "Muy bonito lugar. El lugar es muy bonito, la vista es espectacular, las ruinas son muy interesantes"

As can be seen, some generated samples look very simple, even repetitive, whereas others are more elaborate. This likely had an impact on the official results released.

Table 3 shows the distribution of the polarities of the generated text.

Table 3: Polarity distribution of the generated text.

Polarity	**Instances**
1	825
2	390
3	400
4	698
5	567

We also measured the cosine similarity between the training and generated datasets. The mean cosine similarity was 0.81 ± 0.02 with minimum and maximum values 0.60 and 0.85, respectively.

The following table shows the shared task results for our dataset.

Table 4: Overall Performance Metrics

Overall Polarity Metrics	
Metric	**Value**
Track Score	0.3862
Macro F1	0.4360
Accuracy	60.0173
Avg Precision	0.5503
Avg Recall	0.4020
MAE	0.5543

Table 5: Comparison of Per-Class Performance Metrics

Class	F1 Score	Precision	Recall
1	0.4467	**0.7528**	0.3176
2	0.2696	0.4860	0.1865
3	0.2970	0.4274	0.2275
4	0.3890	0.4012	0.3774
5	**0.7778**	0.6843	**0.9001**

Table 5 shows that the classifier model trained on our synthetic data had exceptional recall for class 5, but it performes poorly on classes 1-4. On the other hand, class 1 has higher precision, but low recall. This means that the trained model usually predicts class 1 instances correctly, but this rarely happens. Also, as accuracy is 60% but Macro F1 is 0.436, the model is classifying the majority class correctly, but fails on minority classes. There is an obvious bias toward class 5, as expected due to the review's nature.

The overall metrics suggest that the methodology was not as effective as expected. We hypothesize that the limited performance was mainly due to the small model size and the absence of classes 1 and 2 during continual pretraining.

5. Conclusions

In this work, we presented our exploration of discrete diffusion language models (MDLM and BD3LM) for the generative task of REST-MEX 2026. Our approach used a two-stage training strategy on a 0.6B-parameter Qwen model converted to a diffusion model, separating the learning of review syntax (via continual pretraining) from sentiment control (via supervised fine-tuning on a balanced subset).

Our results indicate that while the evaluated diffusion models were able to generate coherent Spanish reviews, they struggled to capture the nuances of minority sentiment classes when trained on highly imbalanced real-world human-written data. The downstream classifier achieved good recall for the majority class (polarity 5) but failed to generalize to negative and neutral sentiments, resulting in a low macro F1 score (0.4360).

We identified two primary limitations in our methodology. First, the exclusion of classes 1 and 2 during the continual pretraining phase likely affected the model's ability to learn the linguistic patterns associated with negative experiences, forcing it to rely solely on the fine-tuning stage to acquire this capability. Second, the relatively small model size (0.6B parameters) may have limited its capacity to disentangle style from sentiment effectively compared to larger autoregressive counterparts.

Future work will focus on three directions: (1) experimenting with larger diffusion-converted models to improve semantic understanding; (2) modifying the training pipeline to eliminate the pretraining phase and focus only on the supervised fine-tuning phase; and (3) training over more epochs to test the findings of [13] and [14], and compare with a autoregressive base model. Despite the current limitations, this experiment demonstrates the viability of diffusion models as a tool for synthetic data generation in low-resource sentiment analysis scenarios.

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] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, R. van den Berg, Structured denoising diffusion models in discrete state-spaces, in: Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 17981–17993.

[2] 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.

[3] 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).

[4] 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.

[5] 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.

[6] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, V. Kuleshov, Simple and Effective Masked Diffusion Language Models, 2024. doi:10.48550/arXiv.2406.07524. arXiv:2406.07524.

[7] M. Arriola, A. Gokaslan, J. T. Chiu, Z. Yang, Z. Qi, J. Han, S. S. Sahoo, V. Kuleshov, Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models, 2025. doi:10.48550/arXiv.2503.09573. arXiv:2503.09573.

[8] REST-MEX 2026 Workshop Organizers, REST-MEX 2026: Research on Synthetic Text Modeling for Mexican Magical Towns At IberLEF 2026, https://sites.google.com/cimat.mx/rest-mex-2026/home, 2026. Last accessed: 2026/05/22.

[9] A. Díaz-Pacheco, A. B. García-Gutiérrez, A. Y. Rodríguez-González, R. Aranda, M. Á. Álvarez-Carmona, Balancing sentiment analysis datasets through representative-word-guided synthetic review generation: A case study on mexican spanish tourism reviews, Applied Sciences 16 (2026) 7398.

[10] Z. Zhou, L. Chen, H. Tong, D. Song, dLLM: Simple Diffusion Language Modeling, 2026. doi:10.48550/arXiv.2602.22661. arXiv:2602.22661.

[11] S. Gong, S. Agarwal, Y. Zhang, J. Ye, L. Zheng, M. Li, C. An, P. Zhao, W. Bi, J. Han, H. Peng, L. Kong, Scaling Diffusion Language Models via Adaptation from Autoregressive Models, 2024. doi:10.48550/ARXIV.2410.17891.

[12] Hugging Face Hub - dllm-hub, Tiny A2D checkpoints, https://huggingface.co/collections/dllm-hub/tiny-a2d, 2026. Last accessed: 2026/05/23.

[13] M. Prabhudesai, M. Wu, A. Zadeh, K. Fragkiadaki, D. Pathak, Diffusion Beats Autoregressive in Data-Constrained Settings, 2025. doi:10.48550/ARXIV.2507.15857.

[14] X. Pan, E. Hahami, J. Fan, Z. Xie, H. Sompolinsky, Diffusion-Inspired Masked Fine-Tuning for Knowledge Injection in Autoregressive LLMs, 2025. doi:10.48550/ARXIV.2510.09885.