Gradiant at IberLEF 2026 REST-MEX: Enhancing Synthetic Text Generation with User and Travel Profile Simulation

Anthea Silvia Sasdelli[1,*,†], Rocío Mojica-Gómez[1,†], Héctor Cerezo-Costas[1,†], Pedro Alonso Doval[1,†] and Alberto Vilches-Mira[1,†]

[1]*Fundación Centro Tecnolóxico de Telecomunicacións de Galicia (GRADIANT), Vigo, Spain*

Abstract

This paper presents our proposal for the IberLEF Task Rest-Mex 2026: Research on Synthetic Text Modeling for Mexican Magical Towns. The task consists of generating a synthetic dataset of a fixed maximum size of 3K, the quality of which is verified through a subsequent text polarity classification task. Our approach combines the generation of realistic user profiles and travel itineraries to produce a result that is as close as possible to a genuine review written by a person. We explore three complementary generation strategies implemented through several pipelines, followed by filtering and selection processes that prioritize diversity and task performance. Our synthetic datasets achieved the highest overall ranking and occupied 8 of the top 10 positions in the competition, despite relying exclusively on small language models of 4B and 7B parameters.

Keywords

Synthetic Datasets, Text Generation, IberLEF, Rest-Mex

1. Introduction

The generation capabilities of LLMs have improved over time from pure text generation models, such as the early GPT models [1], to solving tasks using 'few-shot' strategies and fine-tuning through human alignment [2], producing LLMs capable of reproducing tasks that they have never been specifically trained for.

A common challenge in many practical machine learning problems is data scarcity. In many scenarios, data are difficult and costly to obtain, as human annotation is both time-consuming and labor-intensive. In addition, data availability is hindered by intellectual property or data protection policies. Therefore, the development or improvement of methods for building high-quality synthetic data could help overcome these limitations.

Thus, LLM models allow the generation of synthetic datasets at negligible costs, offering a promising alternative to manual annotations. Despite their theoretical prowess for this task, generating diverse datasets similar to real data is far from being solved. Several challenges remain open, including how to assess the expected performance of synthetic data prior to training, how to compare alternative synthetic datasets, and how to ensure that created data reflects real word conditions, not only in label distribution, but also in stylistic, lexical, and semantic properties.

In this context, this work presents our submission to the IberLEF task Rest-Mex 2026: Research on Synthetic Text Modeling for Mexican Magical Towns [3], [4]. In this task, participants were asked to generate synthetic sets of $3K$ size with the topic of travel reviews having a big set of $200K$ of real text as starting point. The text is distributed in five different polarities corresponding to the stars qualification given by customers of tourist packages, hotels, and restaurants.

Our approach solving this challenge includes a thoughtful process of analysis, generation, and selection. First, we performed a preliminary study of the real data and tested initial text to find the most

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

asasdelli@gradiant.org (A. S. Sasdelli); rmojica0@gradiant.org (R. Mojica-Gómez); hcerezo@gradiant.org (H. Cerezo-Costas); palonso@gradiant.org (P. A. Doval); avilches@gradiant.org (A. Vilches-Mira)

0009-0004-7687-3606 (A. S. Sasdelli); 0000-0001-5353-5947 (R. Mojica-Gómez); 0000-0003-2813-2462 (H. Cerezo-Costas); 0009-0000-8255-3466 (P. A. Doval); 0009-0005-0784-9503 (A. Vilches-Mira)

suitable synthetic text production methods. Once defined, three different data streams were utilized; they employ language models of $4B$ and $7B$ of parameters with strong and well-restricted conditions based on diverse user profiles and migration models, this aims to improve the variability of the reviews producing a large number of synthetic records for each category. Subsequently, these records are filtered with neighboring clustering heuristics to discard data far from the real distribution. Lastly, active learning [5] contributes to the final choice of the subsets presented during the submission task, where their quality is indirectly evaluated by training and measuring performance using text classifiers for each solution. The details of each of these stages are organized as follows: section 2 presents literature context related to this challenge, section 3 describes the pipeline for reviews generation, section 4 shows a preliminary data analysis, and section 5 illustrates the selection criteria of the datasets. Finally, sections 6, 7, and 8 explain the results obtained, discuss a lexical and semantic dataset analysis, and recaps the most important aspects about this task.

2. Background

2.1. Task

The task requests the generation of 3K synthetic tourist opinions about the Mexican villages that adhere to the **Pueblos Mágicos** program. Those opinions can be categorized in five levels of polarity (stars) ranging from 1 (very negative) to 5 (very positive opinion). Opinions are categorized in Restaurants, Hotels and Attractives.

$200K$ real records are provided to the participants in order to train or validate their solutions with an unbalanced polarity distribution 1. Positive reviews are much more frequent than negative reviews. Despite this clear imbalance in the real data, our datasets are generated balanced.

Solutions are indirectly evaluated using the subsidiary task of text classification. The metric used for evaluation in the task gives more priority to minority classes as compensation. The task is challenging because text must be generated following a similar distribution to the real input data and their labels. Furthermore, the number of texts to train the classifier are limited to $3K$ records, which is very limited taking into account that current state of the art classification models (which use for training in the range of $10K$ - $100K$ records depending on the task). Polarity classification is very subjective being extreme polarity classes more easily identified than the neutral and close to neutral ones.

Table 1
Distribution of polarity labels in the real dataset.

Polarity	Count
1	5441
2	5496
3	15519
4	45034
5	136561

2.2. Related work

The advancements in generative large language models in recent years have enabled the creation of synthetic data with increasingly high quality, focusing on different strategies to enhancing diversity and controllability.

Persona-driven synthesis methodologies have been proposed to generate diverse synthetic prompts and responses using LLMs. Persona prompting consists of instructing a model to respond from the perspective of an individual, demographic, or social group, described via a short textual profile [6]. This approach has been shown to increase lexical diversity compared to standard prompting, particularly when used with larger models [7].

Beyond persona-based approaches, several works explore diversity through explicit control of generation attributes. In [8], the authors investigate the use of prompts enriched with attributes such as length, style, or topic, demonstrating that attribute conditioning enables the creation of more diverse and structured datasets. Similarly, DATAGEN [9] is an LLM-powered framework designed to produce diverse and controllable datasets. It integrates an attribute-guided generation module with additional validation mechanisms, including code-based mathematical evaluation for label correctness and retrieval-augmented generation techniques for factual consistency.

To quantify such diversity gains, prior works rely on standardized evaluation metrics [10, 11] implemented in the *diversity* python package. These metrics include lexical diversity measures such as Compression ratio (CR), part-of-speech aware compression ratio (CR-POS), and N-gram diversity scores (NDS), as well as Self-repetition (SR), which captures the tendency of models to reuse long n-grams across outputs. Additionally, semantic homogenization is measured using HOM.VS, which leverages BERTScore embeddings. Beyond lexical aspects, these frameworks also assess readability diversity.

Another line of work, based on few shot methods, focuses on leveraging limited real data to guide synthetic data construction. GPT-3Mix [12] augments datasets by generating synthetic samples from mixtures of real examples, effectively increasing dataset diversity. In a similar direction, SynAlign [13] proposes a framework for few-shot synthetic data generation. It employs a Gaussian Process-based uncertainty tracker to select diverse and representative real-world samples. The method then follows a two-stage pipeline: first extracting linguistic attributes (e.g., style, tone, and length) from demonstrations, and subsequently generating new samples conditioned on these attributes to preserve diversity.

3. Datasets Generation Pipelines

In this section we explain our methodology to generate the synthetic datasets. We have three pipelines:**R5R_REGEN**, **HRP_UDGEN** and **R1R_UDGEN**.

3.1. R5R_REGEN

The first dataset generation pipeline, **R5R_REGEN** (**R**andom **5 R**eal_**RE**al **GEN**eration), was designed to exploit the large amount of available real-world data (figure 1). The main objective is to generate realistic reviews while preserving both the semantic coherence and the polarity distribution of the original data.

To achieve this, the original dataset was first reorganized and balanced according to the different review polarities, each with the same amount of reviews. This preprocessing step was necessary to avoid biases towards a specific polarity class and to ensure a more uniform generation process across all categories.

The pipeline is structured into two sequential generation stages. In the first stage, *five reviews* belonging to the same polarity class are randomly selected from the balanced dataset. These reviews are then provided as input to a generative language model, ***Qwen/Qwen3.5-4B***, which is instructed to synthesize a news-style article inspired by the content of the selected reviews. The purpose of this intermediate transformation is to abstract the information contained in the original reviews and reframe it into a different textual domain, namely a objective and newspaper-like narrative. This step introduces lexical and contextual variability while maintaining consistency with the original polarity.

In the second stage, the generated article is used together with an additional set of five reviews sharing the same polarity. These elements are passed to the same generative model whose task is to create a completely new review. In this way, the generated text inherits contextual information from the article and stylistic characteristics from real user reviews, producing outputs that are both diverse and realistic.

The two-step generation strategy was designed to increase variability in the synthetic data while reducing the risk of directly reproducing existing reviews from the original dataset. By introducing an

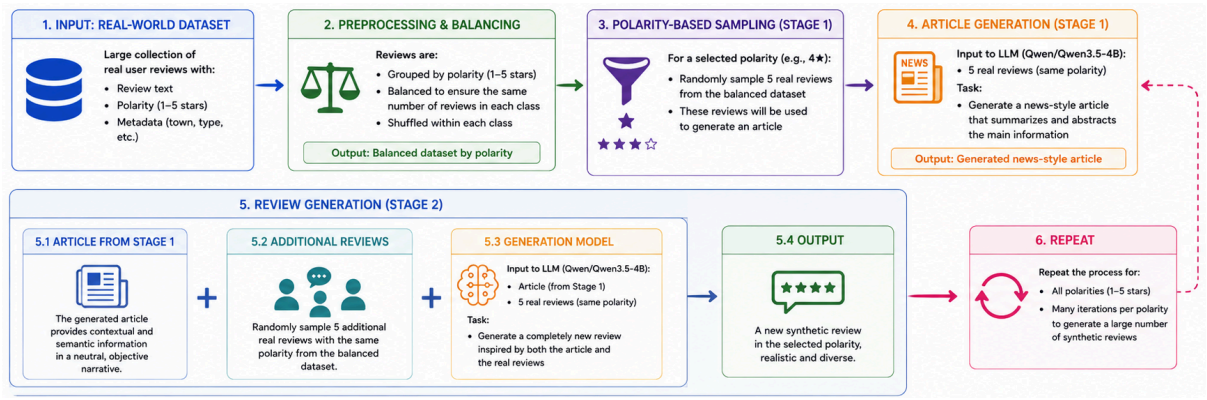

Figure 1: R5R_REGEN pipeline scheme.

intermediate article-generation phase, the pipeline encourages the model to reinterpret the information before producing the final review.

This pipeline was implemented with two react agents with access to the tools for paraphrasing (the conditions of the paraphrases are part of the parameters) and for adding typos. The first paraphrase is in charge of building the news article whereas the second agent generates the review using the article as the starting point. The last agent uses an additional tool to enforce that the final review follow ”first-person-writing-rules”.

3.2. HRP_UDGEN

The pipeline used to generate the HRP_UDGEN (**H**otel **R**estautant **P**ackage_h**U**man **D**riven **GEN**eration) dataset makes use of real-world data from a more qualitative perspective. In fact, the reviews were analysed in every respect:

- firstly, the texts contained in the *Review* column were analysed – the writing style, the tone used, but above all the length – in order to understand how a real review is structured;
- the *Polarity* column was used to check the balance of the dataset and to divide it into classes to be used separately;
- the *Town* column, which allowed us to extract a list of Mexican’s “Pueblos Mágicos”;
- the *Type* column, which allowed us to understand specifically what users were commenting on.

Once we had completed our analysis, we focused on five key elements: the *Users*, the *Language*, the *Polarity* of the reviews, *what* the users were supposed to be reviewing and *where*.

3.2.1. The User selection

There are countless possible combinations of characteristics that can define a user, but some are fundamental. To make the creation process repeatable and explainable, it was decided to create a file containing a set of characteristics that our system would then select:

- **Gender**: our user can be female, male or non-binary;
- **Family status**: our user can be single, in a relationship, married, divorced or widowed. Within these categories there are other subcategories that vary depending on the type of status chosen:
  - the **company**, i.e. the people with whom they shared the experience they are reviewing. It was necessary to create a list for each status because, for example, a single user could not travel with a spouse or partner. The selection is made from a list of possible combinations of individuals or groups of people;
  - the user’s **occupation**, which is indirectly linked to their status, as it is very rare to encounter a widowed university student. Here too, the choice is made from a list, but of occupations;

- the user's **age**, which, as in the previous case, requires a different range for each status. The choice is made from a list of numerical values tailored to the user's status and occupation (for example, a doctor must be aged between 30 and 65, as they must complete their studies before the age of 30 and should retire after 65);
- the **number of children** they are traveling with, again chosen from a list of numbers.

3.2.2. The Language selection

As the challenge involved reviews in Spanish, it was decided to make the text sound more authentic by using different varieties of Spanish or very close languages, ranging from Latin American dialects or contact varieties to those of the Iberian Peninsula. Specifically, the following were selected: **Andalusian**, **Bolivian**, **Castilian**, **Chilean**, **Cuban**, **Ecuadorian**, **Galician**, **Mexican**, **Miami Spanish** (which includes more English words), **Nicaraguan**, **Paisa**, **Peruvian**, **Porteño**.

A dedicated prompt was written for each variant of Spanish, describing the language's characteristics, such as style, commonly used words, slang and typical idioms. In this way, a review written in one of the selected styles would resemble, as closely as possible, a text written by someone who uses that specific type of Spanish.

3.2.3. The Polarity selection

When a user writes a review, the tone, language, and expressiveness are heavily influenced by the nature of their experience, whether positive or negative. For this reason, it was necessary to conduct a thorough analysis of the possible facets that could make the generated text as realistic as possible. Several lists were therefore drawn up for each of the following characteristics of the review:

- the **feelings**: depending on the polarity, the user may experience one of the emotions on the list (for example, with polarity 5, the list contains feelings such as ['happy', 'satisfied', 'calm', "excited", 'grateful', etc.], whilst with polarity 1, the list contains sentiments such as ['angry', 'upset', "stressed", 'unhappy', etc.]);
- the **text length**, to cover a different range of words used;
- the **tone** used by the user when writing the review, for example formal, informal or sarcastic ;
- the **expressiveness** used in describing the experience, which may be implicit, explicit, with few or many details;
- a list of **topics** is also provided from which the system can select between 1 and 5 of them to discuss (for example, the overall cleanliness of the accommodation, the safety of the area, or the speed of service from the staff)

3.2.4. The Town selection

As mentioned earlier, a list of Mexico's "Pueblos Mágicos" was extracted from the real-world dataset provided as training. To further increase the diversity of the reviews we intended to generate, we also added other towns that were not included in the original list. In the end, the towns used were the following: **Álamos**, **Baja California del Sur**, **Bernal**, **Chiapas**, **Chihuahua**, **Coahuila**, **Comala**, **Cosalá**, **Cuetzalan**, **Dolores Hidalgo**, **Guanajuato**, **Guerrero**, **Huasca Ocampo**, **Izamal**, **Jalisco**, **Mazamitla**, **Mexcaltitán**, **Michoacán**, **Mineral Monte**, **Morelos**, **Nayarit**, **Oaxaca**, **Parras**, **Pátzcuaro**, **Puebla**, **Querétaro**, **Quintana Roo**, **San Cristóbal**, **San Luis Potosí**, **San Miguel**, **Tapalpa**, **Taxco**, **Tequila**, **Tepoztlán**, **Tlalpujahua**, **Valle Bravo**, **Veracruz**, **Yucatán**.

For each selected city, a brief description was then retrieved from Wikipedia, which will subsequently be edited and expanded using the **Qwen/Qwen3-4B-Instruct-2507** generation model.

3.2.5. The HRP selection - Hotels Restaurants Packages

By analysing the topics in the real-world dataset provided to us, we were able to identify that the reviews covered three specific services: hotels, restaurants and activities (that will be classified as travel packages).

This helped us identify the topic on which users were supposed to write their reviews. However, after the first few rounds of unsupervised generation, we realised that the texts tended to be repetitive and predictable, making the dataset unusable.

For this reason, we decided to generate, for each town, a list of descriptions for ten local restaurants (made up), ten local hotels (also made up) and ten suggested travel packages.

This was made possible thanks to the descriptions of the cities generated previously and a tailor-made prompt designed to provide a comprehensive overview of the topics, including the advantages, disadvantages, staff members characteristics, what makes it distinctive, and other features that could make it compatible with the location.

3.2.6. Generation pipeline

Once all the sections described above had been defined, it was possible to proceed with the generation of the dataset described in figure 2.

First of all, the system must select the generation model it will use; in our case, we have decided to utilize two models:

- **Qwen/Qwen3-4B-Instruct-2507**: This model was selected because it represents a strong balance between generation quality, instruction-following capabilities, and computational efficiency. As an instruction-tuned model, it is particularly effective at understanding complex prompts and consistently generating coherent and well-structured reviews. Its ability to follow stylistic constraints, maintain contextual consistency, and adapt to different tones and sentiments made it especially suitable for the controlled generation pipeline proposed in this work. Furthermore, the relatively compact size of the model allows the generation of large-scale datasets while maintaining reasonable inference times and hardware requirements.
- **Orion-zhen/Qwen2.5-7B-Instruct-Uncensored**: The uncensored variant was chosen to complement the more controlled behavior of the standard instruct model. Since this model applies fewer safety and alignment restrictions, it is capable of generating more spontaneous, emotionally intense, and sometimes more provocative text. This characteristic is particularly valuable when simulating realistic online reviews, as real users often express opinions in informal, exaggerated, emotional, or highly critical ways that heavily aligned models tend to suppress or sanitize. The reduced filtering mechanisms therefore help produce reviews that appear more authentic and closer to real user-generated content commonly found on platforms such as TripAdvisor or social media environments.

Another fundamental step of the pipeline concerns the loading and organization of all the resources generated during the previous stages. At this point, the system imports the synthetic user profiles, the language configurations containing regional Spanish variants and expressions, the polarity settings associated with the different review ratings, the HRP descriptions (*Hotels, Restaurants, and Packages*), and the collections of real reviews that will be used as stylistic references.

They play a particularly important role within the generation process. Before being used, they are divided according to both their polarity and their semantic category, separating reviews related to hotels, restaurants, and tourist attractions or travel packages. This organization allows the language model to receive examples that are coherent not only with the desired sentiment but also with the specific domain being described, significantly improving the realism and variability of the generated content.

For each selected *Pueblo Mágico*, the pipeline loads the three JSONL files containing the descriptions of hotels, restaurants, and tourist packages associated with that location. These descriptions are then stored and later used as contextual information during the prompt construction phase.

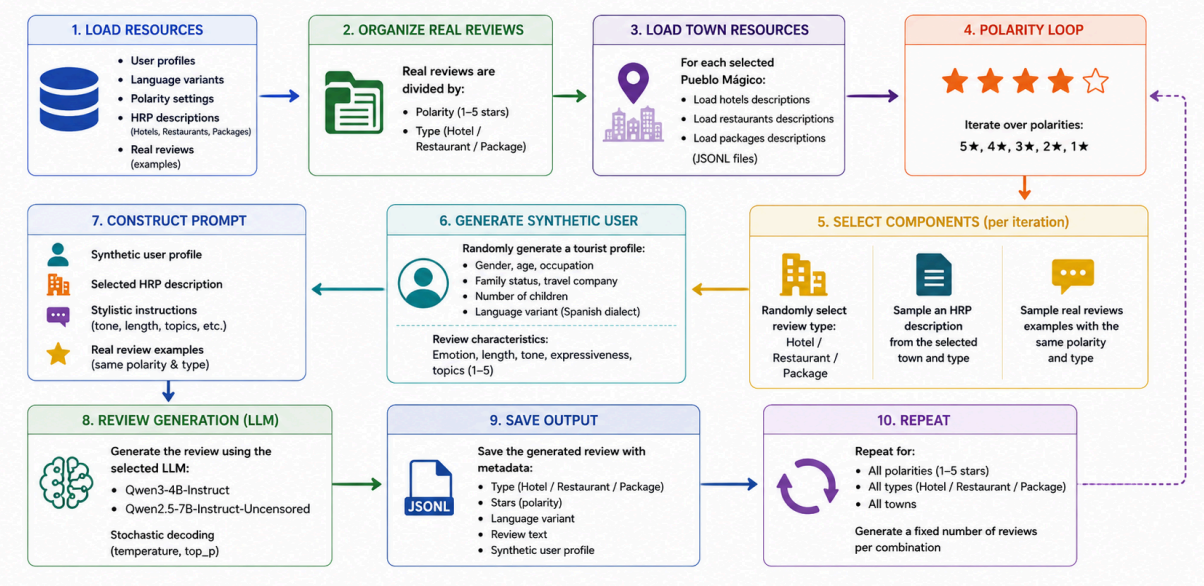

Figure 2: HRP_UDGEN pipeline scheme.

The generation process is performed iteratively for each review polarity, ranging from one-star to five-star reviews. For every polarity level, the system retrieves the corresponding subset of real reviews previously filtered by sentiment and category. These reviews are not copied directly; instead, they serve as stylistic inspiration that guides the model toward producing coherent emotional tones, realistic phrasing, and diverse writing styles.

For each town and polarity combination, the pipeline generates a fixed number of synthetic reviews. During every iteration, the system randomly selects the type of entity to review, which can be a hotel, a restaurant, or a trip package. Afterward, a corresponding HRP description is sampled together with a small group of real reviews sharing the same sentiment. At the same time, a completely synthetic tourist profile is dynamically generated.

Once all these components have been selected, the final prompt is constructed by combining the HRP description, the synthetic tourist profile, the stylistic instructions, and a set of real review examples coherent with the target polarity. The language model is further guided through a dedicated system prompt specifically designed to encourage natural and human-like writing. Particular attention is given to reducing repetitive patterns, increasing lexical variability, and generating reviews that resemble authentic user-generated content. The prompt also explicitly encourages the use of realistic typos and small grammatical imperfections in order to simulate the writing style commonly found on platforms such as TripAdvisor.

3.3. R1R_UDGEN

The pipeline used to generate the R1R_UDGEN (**R**andom **1** **R**eal_h**U**man **D**riven **GEN**eration) dataset closely follows the architecture and methodology previously described for the main synthetic review generation process, with only a small but significant modification introduced during the final generation stage (figure 3).

More specifically, all the preliminary phases remain unchanged. The system still performs the same polarity-driven review generation process, the same synthetic user creation procedure, and the same selection of linguistic, emotional, and stylistic characteristics. Similarly, the overall prompt construction strategy is preserved: the model receives detailed information regarding the user profile, the desired sentiment, the writing style, the language variant, and the behavioral constraints that the generated review must satisfy.

The main difference lies in the source of contextual information used during the generation step. In

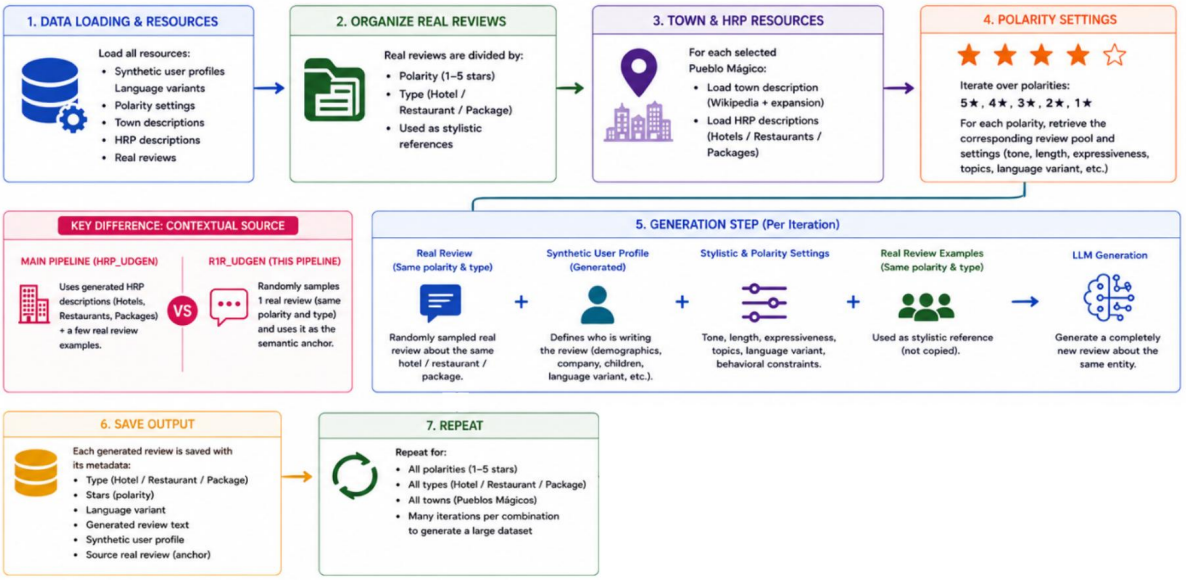

Figure 3: R1R_UDGEN pipeline scheme.

the original pipeline, the review is generated starting from previously created HRP descriptions related to hotels, restaurants, or tourist packages with the help of a few example from the original real dataset. In contrast, the R1R_UDGEN pipeline directly samples a real review belonging to the selected polarity and semantic category.

Once a real review has been selected, the language model is instructed to generate a completely new review that refers to the same hotel, restaurant, or travel package described in the original text, while still adhering to all the synthetic characteristics defined during the previous stages of the pipeline. In practice, the real review acts as a semantic anchor that provides contextual information about the reviewed entity, whereas the generated output must remain original in terms of wording, structure, tone, and narrative style.

This modification introduces a higher degree of realism into the dataset generation process. Since the model is conditioned on authentic user-generated content rather than artificially generated HRP descriptions, the resulting reviews tend to preserve more realistic details, situations, and contextual references associated with real tourism experiences. At the same time, the synthetic user profile and the stylistic constraints ensure variability across the generated samples, preventing simple paraphrasing or direct rewriting of the original review.

As a consequence, the R1R_UDGEN dataset represents a hybrid generation strategy that combines the authenticity of real reviews with the controllability and scalability of synthetic text generation.

4. Data analysis

This section describes the experiments designed to analyze the linguistic patterns of the synthetically generated texts, which structure is presented in annex A, with the aim of assessing the level of diversity and identifying recurring structures in natural language generation. This analysis was essential to understand which parts of the generation pipeline were working properly before providing an automatic strategy to filter the data.

4.1. Preliminary Analysis

The first analysis consists of a qualitative overview of the semantic diversity of the data, where we plotted a graphical representation of the data, using the embeddings obtained with a Sentence transformers

models, flatten to two dimensions with the clustering UMAP algorithm [14]. Thus, we ensure that texts of all polarities cover similar topics and are not semantically far, as well as they are close to real data from the training set. Figure 4 and 5 show examples of these analysis.

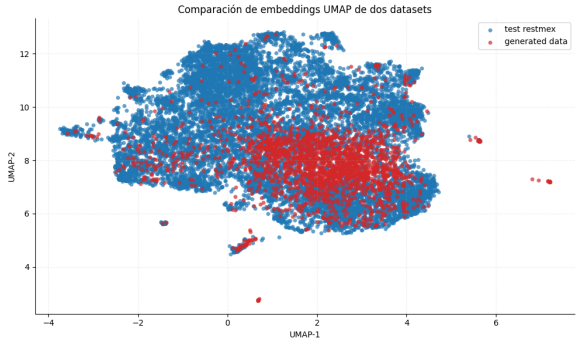

Figure 4: 2D representation of synthetic data vs real training data

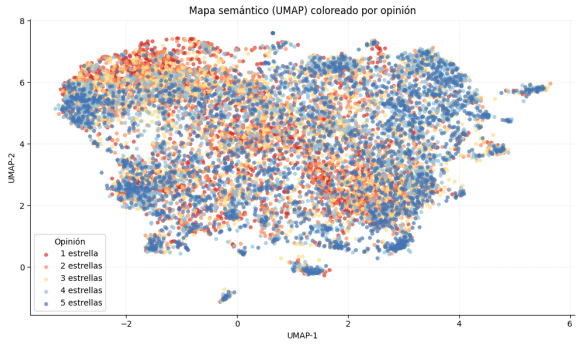

Figure 5: 2D representation of data per polarity level

4.2. Analysis Methodology

In order to identify recurring patterns in the original and generated texts, the different datasets were analyzed using the Patterns tool available at Text Diversity [11], [11] [1]. This tool allows for the extraction of frequently occurring textual sequences (e.g., n-grams or repeated expressions) along with their frequency of appearance across documents.

Our aim in conducting this analysis was to determine the variability of the different reviews in each of the four datasets (the original and the three synthetic ones). This is of paramount relevance as we are using very small language models for generation, which tend to overuse some linguistic constructions if they appear in the instruction prompt. Once we had carried out this analysis of the datasets, we compiled the results into several JSON files in order to draw conclusions and modify the instructions accordingly to avoid repetitive constructions. The patterns or expressions that appeared most frequently in each dataset are shown in the table 2.

In the original dataset (REST-MEX_2026_train), the most frequently repeated phrases appear with very low percentages (e.g., *"vale la pena"* 0.02%, *"la comida es"* 0.01%, *"es un lugar"* 0.01%). This is a strong indication of greater lexical and syntactic diversity and less reuse of templates: even the most frequent n-grams barely 'count' in the total. The conclusion here is that the original serves as a benchmark for 'natural variety' (many different ways of saying similar things).

In the synthetic data, the top patterns rise to significantly higher levels:

- R5R_REGEN: *"la falta de"* 0.28%, *"y el personal"* 0.14%, *"y la comida"* 0.12%, *"en el restaurante"* 0.09%, *"y el servicio"* 0.08%.
- HRP_UDGEN: *"verdad es que"* 0.29%, *"la verdad es"* 0.22%, *"como si el"* 0.13%, *"vale la pena"* 0.06%, *"el tiempo se"* 0.06%.
- R1R_UDGEN: *"verdad es que"* 0.16%, *"como si el"* 0.05%, *"pensé que un"* 0.05%.

Synthetic datasets show template markers (connectives and filler words) that recur with greater frequency than in human texts.

4.2.1. Conclusion by Polarity

From a polarity perspective, synthetic pipelines are designed to generate text by star rating and control tone and expressiveness, which should preserve sentiment consistency. However, the aggregated

[1] https://ai-templates.app/

Table 2
N-gram pattern analysis. Most frequent trigram expressions per dataset as a percentage of the total number of records in each dataset

Dataset	Expression / Pattern	Dataset %
REST-MEX_2026_train (28107 docs)	vale la pena	0.02
	la comida es	0.01
	es un lugar	0.01
	además de que	0.007
	de la habitación	0.007
	de los mejores	0.007
R5R_REGEN (12078 docs)	la falta de	0.28
	y el personal	0.14
	y la comida	0.12
	en el restaurante	0.09
	y el servicio	0.08
HRP_UDGEN (17240 docs)	verdad es que	0.29
	la verdad es	0.22
	como si el	0.13
	vale la pena	0.06
	el tiempo se	0.06
R1R_UDGEN (19000 docs)	verdad es que	0.16
	como si el	0.05
	pensé que un	0.05

analysis by dataset suggests that, regardless of polarity, cross-cutting discourse markers (*"la verdad es"*, *"verdad es que"*, *"como si el"*) emerge, which may homogenise the style across stars. Expressions such as *"vale la pena"* and *"la falta de"* are consistent with recommendations and complaints respectively.

4.2.2. Conclusion by Linguistic Variety

As for linguistic variety, although the HRP_UDGEN pipeline incorporates specific prompts for multiple varieties of Spanish (Andalusian, Castilian, Chilean, Porteño, etc.), the most frequently observed patterns consist mainly of neutral expressions. This suggests that dialectal variation, if it exists, is not dominating the most frequent expressions or is overshadowed by the model's high-probability fillers and connectors.

5. Datasets Selection

In this section, two methods for the final selection of the dataset are explained. The first method uses the coordinates of the real data as baseline to clean the synthetic data from non relevant text. The second method selects data implementing active learning or using random sampling, which yields the final output datasets. One part of the results employs a combination of the first and the second method meanwhile, the other part only takes the second one.

5.1. First Filter: Cluster Filter

In order to find the best final datasets, a filtering procedure is applied. The idea is simple; we take each generated dataset independently and compute similar reviews using the sentence transformers **all-MiniLM-L6-v2** [15] and **MiniLM-L12-v2** [16], as they generate texts with semantic content similar to the real text. An illustrative example can be seen in Figure 6, where real data (12500 texts) and synthetic data (12078 texts) are plotted. Then, we collect real and synthetic reviews with shared coordinates, using the maximum and minimum (x, y)—location of the real text, we delimit the synthetic outputs leaving a total of 8330 texts; this avoids taking reviews semantically away from the real text and eliminates

noisy data, as can be checked in Figure 7. This process is utilized for 10 of the total submissions; which subsequently were delivered to perform the computations with Active Learning or Random Sampling.

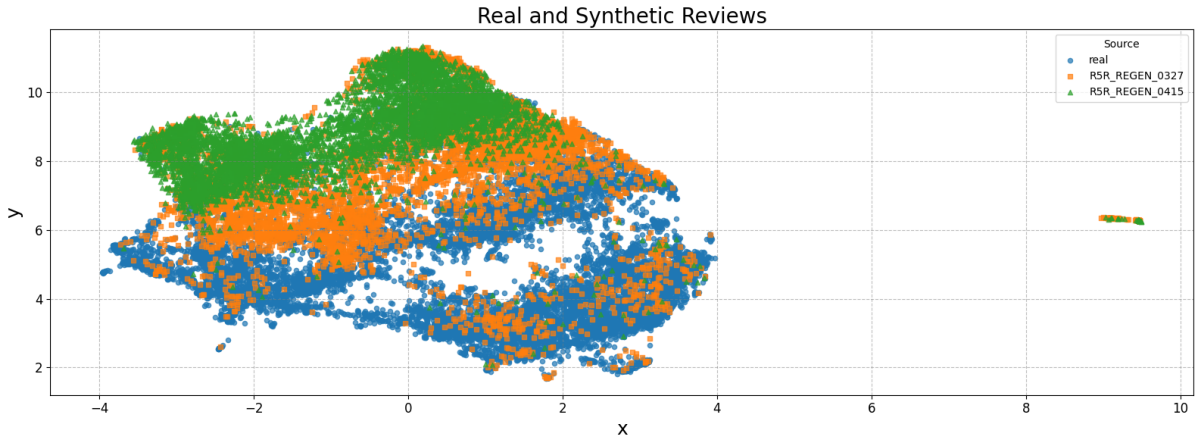

Figure 6: Coordinate distribution of real reviews and synthetic generated text before filtering for the R5R dataset with Mini-L12 transformer.

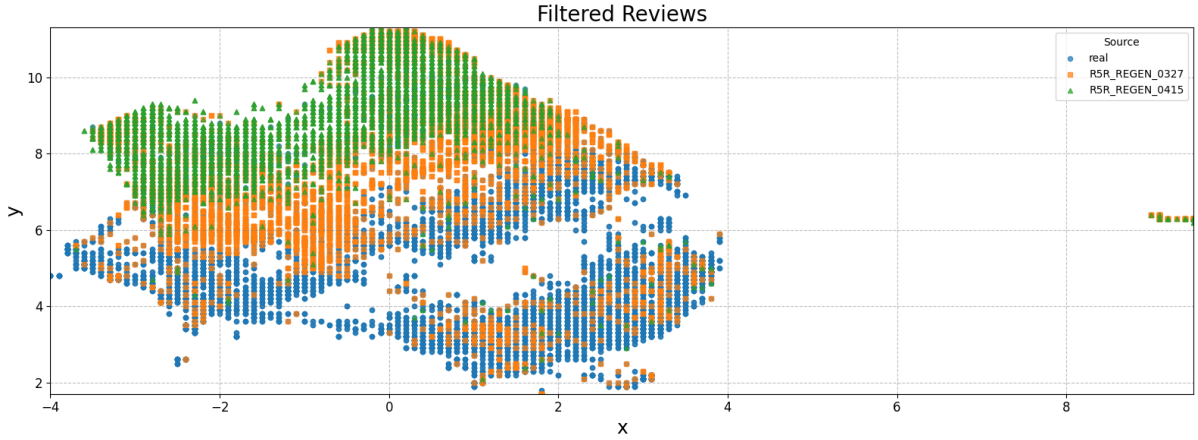

Figure 7: Coordinate distribution of real reviews and synthetic generated text after filtering for the R5R dataset with Mini-L12 transformer.

5.2. Second Filter: Active Learning and Random Sampling

The final step in the construction of the various datasets was carried out using two alternative selection strategies: **Active Learning** (AL) and the naive **Random Sampling**.

The Active Learning strategy follows an iterative selection pipeline using the entropy of the label distribution to guide the selection. Instead of picking all the samples with low or high entropy, the heuristic splits the classification results in low, mid and high entropy per class selecting 20% from the low entropy, 30% from the medium entropy and 50% from the high entropy samples. The full process can be seen in figure 8. During the cold start in the first step, the complete synthetic dataset is provided as input to a classifier previously fine-tuned on the real dataset. After inference, 1,000 reviews are selected from the classifier outputs and stored in the output repository using the entropy selection criteria. For the next step, the remaining samples, not selected in the previous step, are then processed by a new classifier, fine-tuned on the selected 1,000 synthetic texts. From its predictions, an additional 100 samples are selected, consisting of 20 low-entropy, 30 medium-entropy, and 50 high-entropy examples.

This procedure is repeated iteratively: at each step, a new classifier is fine-tuned on the progressively expanded labeled set (i.e., $1{,}000 + 100 + \dots$), and further samples are selected following the same entropy-

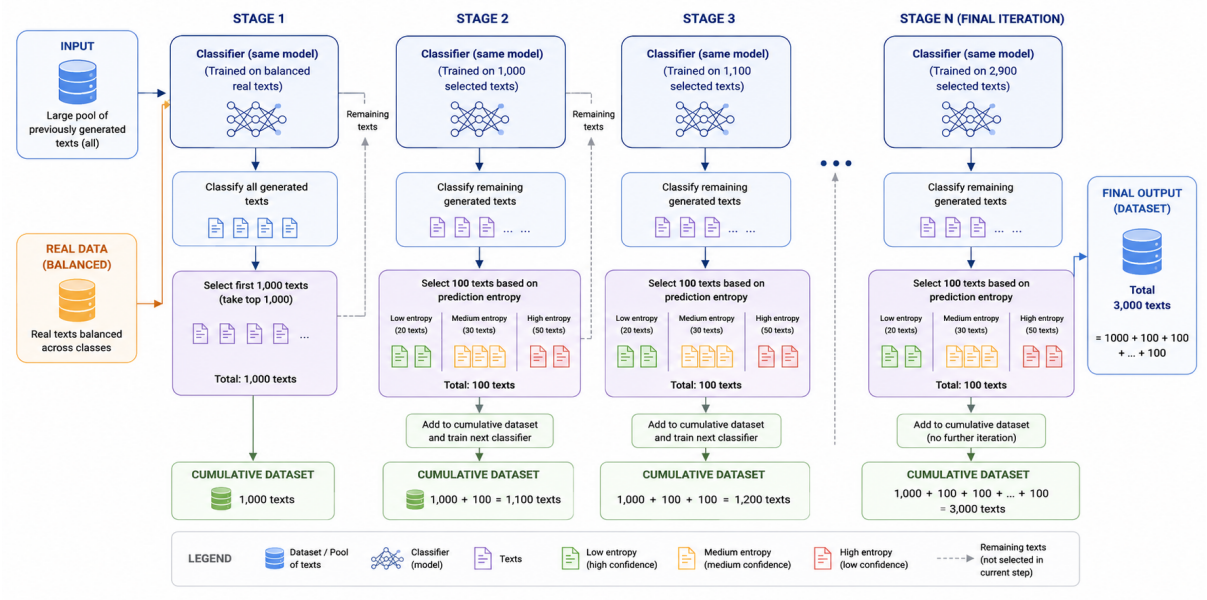

Figure 8: Active Learning workflow.

based criterion. The process continues until a total of 3,000 reviews is reached. This final set constitutes the dataset obtained under the Active Learning strategy. As the AL strategy is performed per class, we always obtain a balanced dataset with 600 samples per class.

We use the same classifier in all the steps, an XLM-Roberta [17] and hyperparameters specified in Table 3.

Table 3
Hyperparameters used during XLM-Roberta finetuning

Parameters	r (LORA)	α (LORA)	Dropout (LORA)	Learning Rate	Epochs	Batch Size
Value	8	16	0.1	$5e^{-5}$	12	16

As a baseline, the Random Sampling strategy selects 3,000 reviews uniformly at random from the synthetic filtered data set. We do not impose any restriction in number of samples per class but the original datasets were already balanced.

6. Results

The table 4 shows the results for the different combinations of generation, filtering and selection. The validation accuracy results were obtained by executing the text classifier XLM-Roberta with the synthetic samples and validated using in a held-out balanced data extracted from the training data with 2500 samples per polarity for a total of 12500 records. As can be seen, the validation accuracy is higher when using Active Learning in comparison with a naive random selection. However, despite being close to a classifier trained with a $3K$ random subset of real data, none of the synthetic solutions surpassed this random selection. Thus, there are still room for improvement with synthetic data generation and selection.

There is not a clear winner in the generation strategies in validation although $R1R$ tend to perform better in the tests over $R5R$ and HRH. This insight is reassured with the task test results in which 3 of the best 5 solutions are $R1R$ subsets in contrast to just one for the HRP and $R5R$. Active learning for subset selection is clearly beneficial as it is applied in 7 from the 8 best of the internal ranking of our solutions.

All in all, our pipelines obtained 8 of the first 10 solutions in RESTMEX with more than 60 submissions,

including the top-3 submissions. Our best solution is a *R1R* with clustering filter and AL for the final selection. Our worse submission was awarded with the 24 position, above the half of the submissions.

Table 4
Validation and Test results of different combinations of generation, filtering and selection. Validation accuracy is performed over a held-out balanced data from the training distribution. Test results were derived from the organization using the Track Score evaluation metric. Rank shows the position of the submission among themselves and all the competitors - in parenthesis - in the RESTMEX challenge.

Dataset Descriptions						
ID	**Gen. Pipeline**	**Cluster Filter**	**Selection**	**Accuracy Val.**	**Score Test**	**Rank(Position)**
1	R1R	Mini-L12	AL	0.519	0.512	4 (4)
2	HRP	Mini-L12	AL	0.490	0.507	6 (8)
3	R1R	–	AL	0.514	0.518	3 (3)
4	HRP	–	AL	0.505	0.518	2 (2)
5	R5R	–	AL	0.498	0.507	7 (9)
6	R1R	Mini-L6	AL	0.490	**0.522**	**1 (1)**
7	HRP	Mini-L6	AL	0.497	0.504	11 (13)
8	R5R	–	Random	0.509	0.511	5 (6)
9	HRP	–	Random	0.470	0.492	15 (21)
10	R1R	–	Random	0.503	0.506	9 (11)
11	R5R	Mini-L6	Random	0.487	0.495	14 (17)
12	HRP	Mini-L6	Random	0.486	0.476	12 (24)
13	R1R	Mini-L6	Random	0.503	0.507	8 (10)
14	R5R	Mini-L12	Random	0.485	0.500	12 (15)
15	HRP	Mini-L12	Random	0.461	0.498	13 (16)
16	R1R	Mini-L12	Random	0.485	0.505	10 (12)
-	Real	–	Random	**0.535**	–	

7. Discussion

7.1. Diversity Analysis of the Submitted Datasets

In this section, we present an analysis of the text diversity from two complementary perspectives: lexical and semantic. Specifically, the goal is to compare the variability of vocabulary (lexical diversity) and to examine the underlying contextual meaning of the submitted texts with respect to a subset of the real data.

7.1.1. Lexical Diversity

The lexical analysis is performed using two concepts. First, compression ratio-part of speech (CR-POS) [18], which is proposed to detect text repetition. This metric builds on the concept of compression ratio, which uses gZip to compress the concatenated text and is specified as the ratio between the size of the compressed text and the original text. The CR-POS is given by Equation 1; thus, a higher CR implies greater redundancy. As a consequence, the POS diversity is lower.

$$Diversity(POS) = \frac{1}{CompressionRatio} \tag{1}$$

Second, we explore the normalized n-gram entropy. N-gram entropy[19] computes how uniformly n-grams are distributed and provides insights into the richness and variability of a text at different sequence levels, (unigrams, bigramss, trigrams, and fourgrams). Since the n-gram entropy does not have a fixed score, normalization generates values between 0 and 1, facilitating the comparison between

the real text and synthetically generated text. In this metric, maximum diversity is achieved when all the n-grams have the same frequency. Thus, higher scores indicate more diversity.

These results are presented in table 5, where the real data at the top is highlighted as a reference, meanwhile bold values in each column correspond to those closest to the reference metric for that column. Among the submitted solutions, ID 16 (R1R) shows the closest score to the real data, for both in terms of compression ratio and for most of the normalized n-grams entropy metrics.

Table 5
Lexical analysis for each submitted solution with respect to real data. Compression ratio (Ratio) and normalized n-gram entropy where n=1,2,3, and 4 are presented.

Lexical Analysis						
ID	**Gen. Pipeline**	**CR-POS**	**n=1**	**n=2**	**n=3**	**n=4**
	Real	0.368	0.583	0.873	0.972	0.994
1	R1R	0.342	0.573	**0.845**	**0.952**	0.952
2	HRP	0.338	0.559	0.829	0.941	0.978
3	R1R	0.342	0.573	**0.845**	0.951	0.984
4	HRP	0.336	0.558	0.826	0.939	0.977
5	R5R	0.337	0.554	0.827	0.942	0.978
6	R1R	0.340	0.568	0.835	0.940	0.972
7	HRP	0.336	0.557	0.823	0.933	0.969
8	R5R	0.341	0.555	0.829	0.945	0.981
9	>	0.339	0.559	0.829	0.942	0.980
10	R1R	0.342	0.571	0.844	**0.952**	**0.986**
11	R5R	**0.343**	0.552	0.828	0.946	0.946
12	HRP	0.337	0.560	0.826	0.937	0.974
13	R1R	0.342	**0.574**	0.845	0.951	0.984
14	R5R	0.340	0.552	0.827	0.944	0.981
15	HRP	0.337	0.557	0.825	0.938	0.975
16	R1R	**0.343**	0.572	**0.845**	**0.952**	**0.986**

7.1.2. Semantic diversity

Semantic diversity is explored using two methods: pairwise diversity evaluation and DCScore [20]. The pairwise method uses sentence transformers to generate sentence embeddings, capture their semantic meaning, and perform similarity-based semantic comparisons. Since each polarity exhibits semantic differences, this computation is carried out separately for each polarity value (P). On the other hand, DCScore takes the pairwise similarity of the sentence embeddings and computes the overall diversity by taking the mean value of the diagonal of the pairwise similarity matrix.

The pairwise similarity is calculated for all the solutions submitted and compared with the reference values of the real data, registered in the top row of Tables 6 and 7. The semantic analysis scores use the previously selected sentence transformers Mini-L6 and Mini-L12, they are applied to their respective combinations given in Table 4.

These results are further complemented by the DCScore metric. Note that for the all-MiniLM-L6-v2 model, the closest value with transformer-based reference corresponds to ID 6, while the non-transformer-based setting, the closest match is ID 3. In contrast, for the MiniML-L12-v2 model, the closer transformer-based score is obtained with ID 16, whereas ID 8 gives the closest match in the opposite scenario. Although, for the latter case, the scores are closer to the real values, there is no clear

or consistent pattern indicating that this parameter provides a strong basis for selecting one method or another. This observation is supported by the behavior illustrated in Figures 9, 10, and 11.

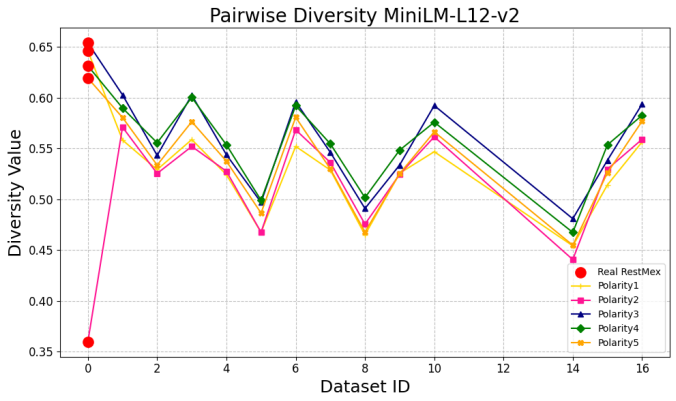

Figure 9: Pairwise diversity values by polarity for the sentence transformer Mini-L12. Real values are represented by the red dot.

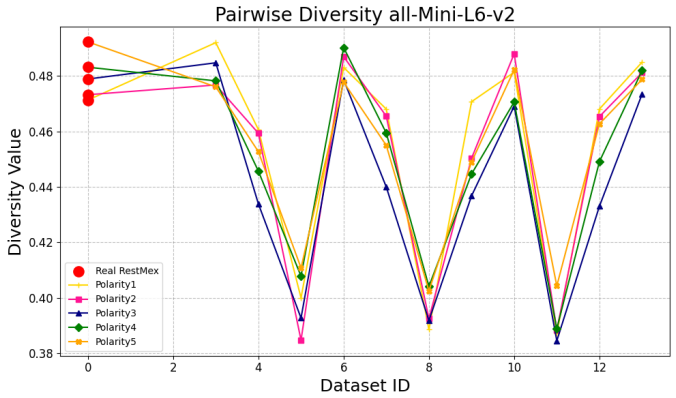

Figure 10: Pairwise diversity values by polarity for the sentence transformer Mini-L6. Real values are represented by the red dot.

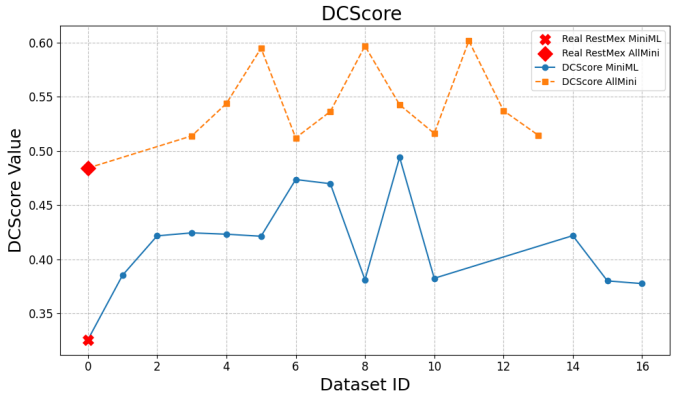

Figure 11: DCScore. Mini-L6 transformer results are represented by the orange dotted line where the red diamond represents the real text value. Similarly, Mini-L12 results follow blue line with reference value represented by a red X.

Table 6
Semantic analysis. Pairwise diversity evaluation for each polarity value with sentence transformer Mini-L6.

Semantic Analysis - Mini-L6							
ID	**Gen. Pipeline**	**P=1**	**P=2**	**P=3**	**P=4**	**P=5**	**DC-Score**
	Real	0.471	0.473	0.478	0.483	0.492	0.484
3	R1R	0.492	0.476	0.484	0.478	0.476	0.513
4	HRP	0.460	0.459	0.433	0.445	0.452	0.543
5	R5R	0.399	0.384	0.392	0.392	0.410	0.594
6	R1R	0.483	0.486	0.478	0.490	0.477	0.512
7	HRP	0.468	0.465	0.439	0.459	0.454	0.536
8	R5R	0.388	0.392	0.391	0.404	0.402	0.596
9	HRP	0.470	0.450	0.436	0.444	0.448	0.542
10	R1R	0.481	0.488	0.469	0.470	0.482	0.516
11	R5R	0.386	0.386	0.384	0.388	0.404	0.601
12	HRP	0.468	0.465	0.433	0.449	0.462	0.537
13	R1R	0.485	0.481	0.473	0.481	0.478	0.514

Table 7
Semantic analysis. Pairwise diversity evaluation for each polarity value with sentence transformer Mini-L12.

Semantic Analysis - Mini-L12							
ID	**Gen. Pipeline**	**P=1**	**P=2**	**P=3**	**P=4**	**P=5**	**DC-Score**
	Real	0.646	0.359	0.653	0.631	0.619	0.325
1	R1R	0.558	0.570	0.602	0.589	0.580	0.385
2	HRP	0.529	0.525	0.543	0.555	0.533	0.421
3	R1R	0.558	0.552	0.602	0.600	0.576	0.424
4	HRP	0.524	0.527	0.543	0.553	0.537	0.423
5	R5R	0.467	0.467	0.496	0.499	0.486	0.421
8	R5R	0.465	0.475	0.490	0.501	0.467	0.381
9	HRP	0.525	0.524	0.533	0.548	0.525	0.494
10	R1R	0.546	0.561	0.592	0.575	0.566	0.382
14	R5R	0.454	0.440	0.480	0.467	0.454	0.421
15	HRP	0.514	0.529	0.538	0.553	0.526	0.380
16	R1R	0.556	0.559	0.593	0.582	0.582	0.377

8. Conclusion

In this paper, we presented an approach to generate diverse and useful synthetic data for review generation. Our datasets are constructed through a multi-stage pipeline that combines generation, profiling, clustering, and selection.

In the generation stage, diversity is encouraged through extensive user and travel profiles that varies across executions to produce unique and informative instructions. Among the explored configurations, the **R1R_UDGEN** pipeline achieved the best results by combining stylistic profiles with real data for topic grounding. All the strategies rely on relatively small language models ($7B$ models and below), producing datasets of good quality.

Additionally, we show that data selection plays a crucial role alongside generation. Clustering was

used to retain samples closer to the real data distribution, whilst active learning was employed to prioritize the most useful instances for the task. Thanks to this setting, active learning consistently outperformed random selection, especially when initialized with real data.

Overall, our approach achieved top results in the competition, with the best performance obtained when the full pipeline is applied. However, randomly selected real data still outperformed purely synthetic datasets in our internal validation benchmark, highlighting that there is still significant room for improving synthetic data generation methods in future work.

Acknowledgments

This work was carried out in the context of the PRESERVE project, funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101168309. In this project, we deal with the generation of synthetic training data facing the same issues present in this task: Data quality, diversity, and royal representation of real word data.

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4o version) [2] for the generation of images with diagrams for figures 1, 2, 3 and 8 by giving natural language descriptions and instructions of the desired content. Authors also used DeepL[3] translator to ensure the correct use of English across different sections.

After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.

[3] 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.

[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] C. Schröder, A. Niekler, A survey of active learning for text classification using deep neural networks, arXiv preprint arXiv:2008.07267 (2020).

[6] T. Ge, X. Chan, X. Wang, D. Yu, H. Mi, D. Yu, Scaling synthetic data creation with 1,000,000,000 personas, arXiv preprint arXiv:2406.20094 (2024).

[7] G. Kambhatla, C. Shaib, V. S. Govindarajan, Measuring lexical diversity of synthetic data generated through fine-grained persona prompting, Findings of the Association for Computational Linguistics: EMNLP 2025 (2025) 21024–21033.

[2] https://chat.openai.com

[3] https://www.deepl.com/es/translator

[8] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. J. Ratner, R. Krishna, J. Shen, C. Zhang, Large language model as attributed training data generator: A tale of diversity and bias, Advances in neural information processing systems 36 (2023) 55734–55784.

[9] Y. Huang, S. Wu, C. Gao, D. Chen, Q. Zhang, Y. Wan, T. Zhou, J. Gao, C. Xiao, L. Sun, et al., Datagen: Unified synthetic dataset generation via large language models, arXiv preprint arXiv:2406.18966 (2024).

[10] C. Shaib, Y. Elazar, J. J. Li, B. C. Wallace, Detection and measurement of syntactic templates in generated text, in: Conference on Empirical Methods in Natural Language Processing, 2024. URL: https://api.semanticscholar.org/CorpusID:270869797.

[11] C. Shaib, V. S. Govindarajan, J. Barrow, J. Sun, A. F. Siu, B. C. Wallace, A. Nenkova, Standardizing the measurement of text diversity: A tool and a comparative analysis of scores, arXiv preprint arXiv:2403.00553 (2024).

[12] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, W. Park, Gpt3mix: Leveraging large-scale language models for text augmentation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2225–2239.

[13] J. Ren, Z. Du, Z. Wen, Q. Jia, S. Dai, C. Wu, Z. Dong, Few-shot llm synthetic data with distribution matching, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 432–441.

[14] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[15] Sentence-Transformers, all-minilm-l6-v2 sentence transformer model, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2020. Accessed: 2026-05-22.

[16] Sentence-Transformers, sentence-transformers/paraphrase-multilingual-minilm-l12-v2, https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, 2020. Accessed: 2026-05-22.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8440–8451.

[18] C. Shaib, Y. Elazar, J. J. Li, B. C. Wallace, Detection and measurement of syntactic templates in generated text, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 6416–6431.

[19] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, B. Dolan, Generating informative and diverse conversational responses via adversarial information maximization, Advances in Neural Information Processing Systems 31 (2018).

[20] Y. Zhu, H. Zhang, B. Wu, J. Li, Z. Zheng, P. Zhao, L. Chen, Y. Bian, Evaluating diversity of llm-generated datasets: A classification perspective (2024).

A. Appendix: Dataset Details

Below are some examples with a polarity of 5 generated for the HRP_UDGEN dataset.


```
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Michoacan",
"type": "Package",
"stars": "5 stars",
"language": "Ecuatoriano",
"review": "¡Qué chévere! Este hotel es de ley. La seguridad está asegurada, pues siempre hay un cuidador de la memoria rondando. El traslado a la Sierra Madre Oriental estuvo perfecto, hasta que llovió un poquito. El río Balsas, o sea, ¡qué belleza! El precio incluye todo, menos los caprichos. Unas comidas más caras de ley, pero ¡qué sabrosas! La casa flotante fue lo máximo.",
"user_profile": {"person": {"gender": "female", "family_status": "divorced", "company": "trip to escape the routine and find oneself", "work": "taxi driver", "age": 52, "sons": 2}, "review": {"stars": "5 stars", "emotion": "satisfied", "length": "between 0 and 15 words", "tone": "critical", "expressivity": "neutral", "topics": ["Perceived safety", "Transparency regarding additional expenses", "How the experience is affected (rain, heat, visibility)", "Common areas"]}, "language": {"variant": "Ecuatoriano"}}
}
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Puebla",
"type": "Hotel",
"stars": "5 stars",
"language": "Cubano",
"review": "Asere, me alojé en El Sol de Puebla y fue una fiesta. La organización, la verdad, estuvo en el nivel superior. El personal fue super atento y nos hicieron sentir como en casa. La cena, la verdad, fue un festín para los ojos y el paladar. El único inconveniente es que a veces el wifi en las habitaciones más altas es un poco lento. Pero todo en general, perfecto. Dale, definitivamente te recomiendo este lugar.",
"user_profile": {"person": {"gender": "male", "family_status": "divorced", "company": "trip with just a backpack and a sense of adventure", "work": "athlete", "age": 32, "sons": 1}, "review": {"stars": "5 stars", "emotion": "satisfied", "length": "between 15 and 50 words", "tone": "constructive", "expressivity": "neutral", "topics": ["Organization", "Wi-Fi"]}, "language": {"variant": "Cubano"}}
}
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Yucatan",
"type": "Restaurant",
"stars": "5 stars",
"language": "Peruano",
"review": "Reservas facilísimas, servicio del equipo increíble, comida fresquísima, ambiente tranquilo y acogedor, ideal para la familia. ¡Pucha, qué buen lugar!",
"user_profile": {"person": {"gender": "female", "family_status": "married", "company": "with the entire family", "work": "salesperson", "age": 51, "sons": 3}, "review": {"stars": "5 stars", "emotion": "calm", "length": "between 0 and 15 words", "tone": "
```

```
    expressivity": "neutral", "topics": ["Ease and clarity of booking", "Speed of
    service from the staff", "Quality and freshness of food", "Local or low-impact
    options"]}, "language": {"variant": "Peruano"}}
}
```


Below are some examples with a polarity of 4 generated for the R1R_UDGEN dataset.


```
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "DoloresHidalgo",
"type": "Restaurant",
"stars": "4 stars",
"language": "Mexicano",
"review": "Me tocó cenar aquí, güey, es una locura de lugar, qué hermoso, la verdad, no
    me lo esperaba al entrar. La comida está buena, recomendable, pero la música en vivo
     me molestó un poquito, chido que tuvieran ese detalle, pero se escuchaba bastante
    alto, no manches. El personal fue super amable, eso sí, la variedad y calidad de la
    comida son lo mejor, no me canso de decirlo.",
"user_profile": {"person": {"gender": "male", "family_status": "dating", "company": "
    trip to reflect on life and set new goals", "work": "architect", "age": 30, "sons":
    1}, "review": {"stars": "4 stars", "emotion": "happy", "length": "more than 50 words
     but less than 100 words", "tone": "constructive", "expressivity": "implicit", "
    topics": ["Friendliness of the staff", "Variety and quality of food"]}, "language":
    {"variant": "Mexicano"}}
}
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Pátzcuaro",
"type": "Package",
"stars": "4 stars",
"language": "Chileno",
"review": "¡Vaya, Tulum y Coba son lugares bacán! El problema fue el transporte: la
    entrada estaba llena de vendedores que te engañan. Pero la vista de las ruinas fue
    increíble. Y los kayaks... ¡eran una mierda!",
"user_profile": {"person": {"gender": "no-binary", "family_status": "dating", "company":
     "with the family dog", "work": "employee", "age": 40, "sons": 0}, "review": {"stars
    ": "4 stars", "emotion": "pleased", "length": "between 0 and 15 words", "tone": "
    sarcastic", "expressivity": "explicit", "topics": ["Condition of equipment (boats,
    bicycles, kayaks, etc.)", "Available plugs", "Gentler alternatives"]}, "language": {
    "variant": "Chileno"}}
}
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "SanCristóbal",
"type": "Hotel",
"stars": "4 stars",
"language": "Porteño",
"review": "¡Qué quilombo! Reservamos este hotel hace meses, pensamos que estaríamos
    cómodos y todo estaría perfecto. Pero al llegar, ni siquiera tenían habitaciones
    disponibles. Nos dieron una suite enorme, ¡incluso más barato que una habitación
    normal! La verdad es que estuvo re bueno. El desayuno era una masa, con todo lo que
    uno necesita. Y cuando hicimos un tour, nos ofrecieron una copada comida que vale un
     billete de avión. En general, el lugar estaba muy limpio y cómodo. Si bien hubo un
```

```
    bajón por la falta de habitaciones, el precio parecía justificado por la experiencia
    . En resumen, un hotel que recomendaría para una estadía especial.",
"user_profile": {"person": {"gender": "no-binary", "family_status": "married", "company"
    : "with the partner", "work": "self-employed", "age": 40, "sons": 3}, "review": {"
    stars": "4 stars", "emotion": "pleased", "length": "more than 50 words but less than
     100 words", "tone": "professional", "expressivity": "detailed but not too much", "
    topics": ["Accuracy of descriptions", "Perception of whether the price is justified
    by the experience", "Overall comfort", "Quality of tastings, samplings, or meals"]},
     "language": {"variant": "Porteño"}}
}
```


Below are some examples with a polarity of 3 generated for the R1R_UDGEN dataset.


```
{
"model": "Qwen3-4B-Instruct-2507",
"town": "Chihuahua",
"type": "Restaurant",
"stars": "3 stars",
"language": "Miami",
"review": "Servicio OK, pero la salsa was too salty, y el ruido en la mesa era too much,
     no es mi vibe.",
"user_profile": {"person": {"gender": "male", "family_status": "divorced", "company": "
    with the family cat", "work": "designer", "age": 54, "sons": 2}, "review": {"stars":
     "3 stars", "emotion": "indifferent", "length": "between 0 and 15 words", "tone": "
    informal", "expressivity": "explicit", "topics": ["Correspondence between what was
    advertised and what was received", "Level of tranquility or noise"]}, "language": {"
    variant": "Miami"}}
}
{
"model": "Qwen3-4B-Instruct-2507",
"town": "Veracruz",
"type": "Package",
"stars": "3 stars",
"language": "Andaluz",
"review": "Pues no me ha salido como esperaba, pero tampoco fue tan malo. El guía se
    adaptó un poco, pero el transporte era un poco lento, o sea, no fue lo mejor. 3
    estrellas por el medio.",
"user_profile": {"person": {"gender": "male", "family_status": "divorced", "company": "
    trip to reflect on life and set new goals", "work": "accountant", "age": 51, "sons":
     2}, "review": {"stars": "3 stars", "emotion": "mixed feelings", "length": "between
    15 and 50 words", "tone": "professional", "expressivity": "detailed but not too much
    ", "topics": ["Guide's ability to adapt to the group", "Quality of transportation
    contracted"]}, "language": {"variant": "Andaluz"}}
}
{
"model": "Qwen3-4B-Instruct-2507",
"town": "QuintanaRoo",
"type": "Hotel",
"stars": "3 stars",
"language": "Gallego",
"review": "O hotel está ben situado, xa que está a poucos pasos da praza e todo o mundo
    pasa por ela. A cama foi xente de baña, pero a verdade é que o barullo da catedral
    ás 6 da mañá xa me pegou como unha toca de xogo. Non me atopei a durmir, e xa non
```

```
    podo dicir que foi unha experiencia perfecta. A xente con mobilidade reducida ten
    dito que non é ideal, e o baixo de cama non é nada confortable. A visibilidade en
    días de choiva é máis ou menos normal, pero non me gustou tanto o aspecto de todo o
    lugar. Xa non vou volver a durmir así.",
"user_profile": {"person": {"gender": "female", "family_status": "married", "company": "
    university trip", "work": "nurse", "age": 61, "sons": 1}, "review": {"stars": "3
    stars", "emotion": "uncertain", "length": "more than 50 words but less than 100
    words", "tone": "humorous", "expressivity": "explicit", "topics": ["Whether the
    activity lasts as long as indicated", "Access for people with reduced mobility", "
    Comfort of the bed", "How the experience is affected (rain, heat, visibility)", "
    Adaptation for older people, children, or people with reduced mobility", "Places
    with good views"]}, "language": {"variant": "Gallego"}}
}
```


Below are some examples with a polarity of 2 generated for the HRP_UDGEN dataset.


```
{
"model": "Qwen3-4B-Instruct-2507",
"town": "Álamos",
"type": "Package",
"stars": "2 stars",
"language": "Paisa",
"review": "Pues, mijo, esperábamos más, el paquete era bonito pero la verdad, los
    talleres no pasaban, los equipos estaban en mal estado y no hubo participación real
    en los mercados. El staff fue frio, no hubo atención personal. Solo un 2 estrellas
    por eso.", "user_profile": {"person": {"gender": "male", "family_status": "divorced"
    ,"company": "with the new partner", "work": "architect","age": 61, "sons": 2}, "
    review": {"stars": "2 stars", "emotion": "disappointed", "length": "between 0 and 50
     words", "tone": "emotional", "expressivity": "very detailed", "topics": [ "Depth
    and relevance of information", "Common areas", "Participation in markets, workshops,
     demonstrations","Professionalism of the staff", "Condition of equipment (boats,
    bicycles, kayaks, etc.)", "Local or low-impact options","Available amenities (
    hairdryer, minibar, safe, etc.)"] }, "language": {"variant": "Paisa"}}
}
{
"model": "Qwen3-4B-Instruct-2507",
"town": "Tlalpujahua",
"type": "Hotel",
"stars": "2 stars",
"language": "Castellano",
"review": "El lugar era mágico, pero no cumplía con lo descrito. Sin calefacción, sin
    servicio útil, y las habitaciones eran inadecuadas para un voluntario.",
"user_profile": {"person": {"gender": "no-binary", "family_status": "dating", "company":
     "trip to volunteer and give back to the community", "work": "doctor", "age": 30, "
    sons": 0}, "review": {"stars": "2 stars", "emotion": "dissatisfied", "length": "
    between 0 and 15 words", "tone": "formal", "expressivity": "neutral", "topics": ["
    Accuracy of descriptions", "Organization", "Problem-solving ability of the staff", "
    Beauty of the landscape or visual appeal", "Safety in activities (snorkeling,
    trekking, boat trips, etc.)"]}, "language": {"variant": "Castellano"}}
}
{
"model": "Qwen3-4B-Instruct-2507",
"town": "SanMiguel",
```

```
"type": "Restaurant",
"stars": "2 stars",
"language": "Nicaraguense",
"review": "Vos dijiste que era un lugar mágico, pero al final fue un fraude. El precio era loco, el servicio como si nos hubieran olvidado, y el atardecer no valió el dinero. Pues no, no fue una experiencia, fue una pérdida de tiempo.",
"user_profile": {"person": {"gender": "female", "family_status": "divorced", "company": "trip to escape the routine and find oneself", "work": "actor", "age": 52, "sons": 3}, "review": {"stars": "2 stars", "emotion": "disappointed", "length": "between 15 and 50 words", "tone": "sarcastic", "expressivity": "explicit", "topics": ["Comparison between cost and services offered", "Condition of elevators and shared areas", "Condition of equipment (boats, bicycles, kayaks, etc.)", "Feeling of safety", "Guide's ability to adapt to the group", "Punctuality", "Time to take photos", "Gentler alternatives"]}, "language": {"variant": "Nicaraguense"}}
}
```


Below are some examples with a polarity of 2 generated for the R1R_UDGEN dataset.


```
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Álamos",
"type": "Hotel",
"stars": "1 star",
"language": "Paisa",
"review": "Bien, pues, vino todo mal. Cuando reservé, me dijeron que era un hotel super cómodo, bacano. Pero, no. No tuvimos wifi. ¿Qué es eso? ¿En el siglo XXI? ¿Qué hacemos sin internet? Pues nada, ni siquiera podía buscar recetas para cocinar. Y no solo eso, las habitaciones eran pequeñas y húmedas. La gente decía que era lindo, pero yo no. No sé qué ven en este lugar. Por lo menos la piscina hubiera sido una salvación. Pero no, tampoco tenían esa. En serio, ¿dónde están los puntos positivos? Todo lo demás era una mierda. La comida, la atención, todo. No recomendarída este sitio a nadie, ni a mis mismos primos.", "user_profile": {"person": {"gender": "female", "family_status": "divorced", "company": "with the family dog", "work": "chef", "age": 52, "sons": 3}, "review": {"stars": "1 star", "emotion": "miserable", "length": "between 51 and 100 words", "tone": "constructive", "expressivity": "very detailed", "topics": ["Clarity of prior instructions", "WiFi performance"]}, "language": {"variant": "Paisa"}}
}


{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Guerrero",
"type": "Restaurant",
"stars": "1 star",
"language": "Mexicano",
"review": "Qué mala experiencia tuvimos en este lugar, güey. La mesa estaba sucia y el ambiente en la naturaleza era un desastre. Ni hablar del equipo, los botes estaban viejos y los kayak, con agujeros. El wifi funcionaba mal y no había ningún lugar donde pudiese desconectarme de la tecnología. La organización fue un caos, ni modo que las actividades duras como indicado, no manches. El mesero, no te cuento, no prestaba atención al cliente. En resumen, una decepción total.",
"user_profile": {"person": {"gender": "male", "family_status": "married", "company": "
```

```
    trip to disconnect from technology and reconnect with nature", "work": "athlete", "
    age": 30, "sons": 2}, "review": {"stars": "1 star", "emotion": "angry", "length": "
    more than 50 words but less than 100 words", "tone": "professional", "expressivity":
     "detailed but not too much", "topics": ["Overall cleanliness of the accommodation",
     "Condition of the natural environment", "Interactivity: demonstrations, anecdotes,
    participation", "Condition of equipment (boats, bicycles, kayaks, etc.)", "WiFi
    performance", "Soundproofing", "Whether the activity lasts as long as indicated", "
    Organization"]}, "language": {"variant": "Mexicano"}}
}
{
"model": "Qwen2.5-7B-Instruct-Uncensored",
"town": "Nayarit",
"type": "Package",
"stars": "1 star",
"language": "Castellano",
"review": "Estuve tan decepcionado con este viaje, pensé que era todo lo que prometían
    pero resultó ser una trampa. El lugar estaba lleno de casas y edificios en muy mala
    forma, totalmente diferente a lo que se les vendía. La artesanía no era nada
    especial, era lo mismo que veías en un mercado o en una tienda de aeropuerto. No
    merecía ni un poco de mi tiempo y dinero. Si buscas un destino real para un festival
     o evento, busca algo más.",
"user_profile": {"person": {"gender": "no-binary", "family_status": "divorced", "company
    ": "trip to attend a festival or event", "work": "accountant", "age": 51, "sons": 3}
    , "review": {"stars": "1 star", "emotion": "hopeless", "length": "more than 50 words
     but less than 100 words", "tone": "critical", "expressivity": "very detailed", "
    topics": ["Accuracy of descriptions"]}, "language": {"variant": "Castellano"}}
}
```