

Article

Balancing Sentiment Analysis Datasets Through Representative-Word-Guided Synthetic Review Generation: A Case Study on Mexican Spanish Tourism Reviews

Angel Díaz-Pacheco ¹, Andrea Bethsabe García-Gutiérrez ², Ansel Y. Rodríguez-González ^{3,4},
Ramón Aranda ^{4,5} and Miguel Á. Álvarez-Carmona ^{2,4,*}

¹ Campus Irapuato-Salamanca, Universidad de Guanajuato (UG), Salamanca 36885, Mexico; angel.diaz@ugto.mx

² Centro de Investigación en Matemáticas (CIMAT), Sede Monterrey, Monterrey 66629, Mexico; andrea.garcia@ciamat.mx

³ Unidad Académica Tepic, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Tepic 63173, Mexico; ansel@cicese.edu.mx

⁴ Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico City 03940, Mexico; arac@ciamat.mx

⁵ Centro de Investigación en Matemáticas (CIMAT), Sede Mérida, Mérida 97302, Mexico

* Correspondence: miguel.alvarez@ciamat.mx

Abstract

Class imbalance remains one of the most challenging problems in sentiment analysis, particularly in tourism review datasets where positive opinions substantially outnumber neutral and negative comments. This issue is especially critical because minority classes often contain the most valuable information regarding customer dissatisfaction, service failures, and opportunities for improvement. In this work, we propose a hybrid balancing methodology for sentiment analysis in Mexican Spanish tourism reviews that combines undersampling and Large Language Model (LLM)-based oversampling. The proposed framework first extracts representative words from each sentiment class using Mutual Information, then enriches them through dictionary-based or embedding-based lexical substitutions, and finally generates synthetic reviews using GPT-4o-mini guided by these representative terms. Experiments were conducted on a corpus that contains approximately 300,000 tourism reviews collected from TripAdvisor, exhibiting severe sentiment imbalance. Three undersampling strategies and multiple oversampling configurations were evaluated across six traditional machine learning classifiers and one Transformer-based model (BETO). Results show that random undersampling consistently outperformed centroid-based and K-means-based alternatives while also requiring the lowest computational cost. The best overall performance was obtained by BETO, achieving a Macro-F1 score of 0.57 compared to 0.51 on the original imbalanced dataset, representing an improvement of 11.8%. Significant gains were also observed for minority classes, with improvements exceeding 16% for the most underrepresented category. Furthermore, the proposed methodology consistently outperformed direct prompt-based generation using GPT-4o-mini, Gemini 2.5 Flash, and Llama 3.3 70B. These findings suggest that guiding synthetic review generation through representative words effectively preserves domain-specific lexical and semantic patterns of Mexican Spanish tourism reviews, resulting in more balanced datasets and improved sentiment classification performance.

Keywords: sentiment analysis; class imbalance; oversampling; undersampling; synthetic review generation; Large Language Models; tourism intelligence; Mexican Spanish



Academic Editor: Rui Araújo

Received: 23 June 2026

Revised: 9 July 2026

Accepted: 14 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Sentiment analysis has become one of the most important applications of Natural Language Processing (NLP), enabling the automatic identification of opinions, attitudes, and emotions expressed in textual data [1]. In recent years, the rapid growth of user-generated content on social media platforms, review websites, and online forums has created unprecedented opportunities to analyze customer perceptions on a large scale [2]. Among the domains that have benefited most from these advances is tourism, where millions of travelers continuously share their experiences regarding hotels, restaurants, attractions, and destinations [3].

The analysis of tourist reviews provides valuable information for destination managers, businesses, and policymakers [4]. Positive opinions help identify successful practices, whereas negative reviews often reveal service deficiencies, infrastructure problems, or sources of customer dissatisfaction [5]. Consequently, accurate sentiment classification systems can support decision-making processes aimed at improving tourism services and visitor experiences [6].

Despite the progress achieved by modern machine learning and deep learning models, sentiment analysis in tourism still faces an important challenge: class imbalance [7]. In most real-world tourism datasets, positive reviews greatly outnumber neutral and negative opinions. As a result, classification algorithms tend to favor the majority classes and achieve high overall accuracy while performing poorly on the minority classes [7].

This issue is particularly problematic because the minority classes are often the most informative. In tourism, strongly negative reviews typically contain critical information about service failures, cleanliness issues, poor customer experiences, or operational problems that require immediate attention [8]. Therefore, failing to correctly identify these minority classes can significantly reduce the practical usefulness of sentiment analysis systems, even when their overall accuracy appears satisfactory.

To address this problem, researchers have proposed different balancing strategies, including undersampling majority classes [9], oversampling minority classes [10], and generating synthetic instances [11]. Traditional approaches such as random oversampling or SMOTE operate on vector representations of the data and have demonstrated positive results in several domains [12]. However, textual data presents additional challenges because semantic coherence, linguistic structure, and contextual information must be preserved when generating new instances.

In this work, we propose a hybrid balancing methodology specifically designed for sentiment analysis in tourism reviews. The proposed framework combines undersampling and oversampling techniques. First, representative words are extracted from each sentiment class using Mutual Information. These words are then enriched through either dictionary-based synonyms or semantically similar words obtained from word embeddings. Subsequently, a Large Language Model (LLM) generates synthetic reviews guided by these representative terms, allowing the creation of new minority-class instances that preserve the semantic and lexical characteristics of the original corpus.

Unlike direct prompt-based generation approaches, the proposed methodology explicitly controls the vocabulary used during generation through representative words, similarity strategies, lexical replacement factors, and temperature parameters. This enables the generation of synthetic reviews that remain closely aligned with the linguistic patterns observed in the original dataset while introducing controlled lexical diversity.

To evaluate the effectiveness of the proposed approach, extensive experiments were conducted using the REST-MEX 2026 tourism corpus [13], a large-scale collection of tourist reviews exhibiting severe sentiment imbalance. Multiple undersampling and oversampling configurations were evaluated using both traditional machine learning classifiers and

Transformer-based architectures. The results demonstrate that the proposed methodology substantially improves the recognition of minority sentiment classes while maintaining competitive overall classification performance.

The main contributions of this work are threefold:

- A hybrid balancing framework that combines undersampling and LLM-based oversampling for sentiment analysis.
- A controlled synthetic review generation strategy based on representative words, semantic similarity, and lexical replacement mechanisms.
- A comprehensive evaluation involving hundreds of balancing configurations across traditional machine learning models and Transformer-based classifiers.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art on sentiment analysis under class imbalance scenarios, with a particular focus on balancing strategies for textual data. Section 3 introduces the research hypothesis and the research questions that motivate and guide the proposed study. Section 4 details the proposed balancing framework, including representative-word extraction, undersampling and oversampling procedures, similarity strategies based on dictionaries and word embeddings, the use of Large Language Models for synthetic review generation, and the classification models employed for evaluation. Section 5 describes the REST-MEX corpus and analyzes its sentiment distribution and imbalance characteristics. Section 6 presents the experimental results obtained with the different balancing configurations and classifiers, including comparisons with the original imbalanced corpus and direct prompt-based synthetic text generation approaches. Finally, Section 7 summarizes the main findings, answers the research questions, discusses the implications of the proposed methodology, and outlines future research directions.

2. Related Works

Several studies have been conducted on handling imbalanced data, particularly in sentiment analysis of texts. Researchers have developed and applied a wide range of approaches and techniques to address this problem. Some of these include generating synthetic samples, adjusting weights in the loss function, selecting relevant features, and optimization techniques to deal with imbalanced data. Several of these approaches are examined below, including some specific to the tourism sector.

One approach to dealing with imbalanced datasets is through oversampling or undersampling from feature vectors. This method reduces the risk of models overfitting, which can occur if only the imbalanced dataset is used in the experiment. This technique has been applied to classify imbalanced datasets obtained from social media platforms and review websites [14].

Ref. [14] conducted a comprehensive study using a large number of datasets on various topics to measure the impact of different class proportions on classifier accuracy. They employed two approaches: a data-based approach that uses undersampling or oversampling to balance classes, and an algorithm-based approach that adjusts the classification algorithm without modifying the data. The study found that the learning model often poorly predicts minority class samples and that some results might be biased due to the majority class [14].

Another technique introduced is the use of evolutionary algorithms in sentiment analysis, applied to numerical representations of texts along with oversampling and undersampling tools. This was used on a dataset of workplace experience comments in the human resources department. The study found that classifiers could be more biased in high-imbalance contexts depending on text representation, and some configurations of text representations and classifiers did not show improvement [15].

Ref. [16] conducted a study on texts obtained from a website containing hotel visitor comments, classified by star ratings and whether they would revisit the hotel. They used information gain for feature selection, ranked by frequency, to categorize important factors in hotel selection. With imbalanced data, they performed oversampling using SMOTE to generate synthetic data from feature vectors. ML classification algorithms successfully adjusted the minority class, noting the presence of two classes [16].

An optimization perspective was proposed by [17], who implemented a hybrid approach combining SVM with Particle Swarm Optimization (PSO) for comment classification. They used data from a social network consisting of reviews of 300,000 different places like restaurants and cafes, classified as positive or negative with significant imbalance. They employed variations in SMOTE with various ML algorithms, focusing on their SVM-PSO proposal, achieving favorable results with the SVM-PSO-SMOTE combination [17].

Regarding deep learning, several studies have used texts from social media. The authors of [14] analyzed texts focused on classifying eight sentiments about a disease that spread during the COVID-19 pandemic. They developed models based on neural networks such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. They used SMOTE variations and random undersampling for data balancing, finding that joy was a minority class while fear was a majority class, achieving favorable results and recommending this approach for attitude analysis [14].

With the Transformers proposed by [18], various text analyses for classification have been conducted using BERT. Ref. [19] developed a method that applies balancing to datasets. They worked on two specific datasets, one with three classes and another with five classes. Their proposal adds a balancing layer to BERT's structure, incorporating oversampling, undersampling, or a hybrid approach (combining both). They used SMOTE and Condensed Nearest Neighbor (CNN) for oversampling, obtaining positive results by balancing the data [19].

Another approach for using deep learning with highly imbalanced data is presented by [20], who used text generation based on GPT-2 and LSTM, testing with different ML algorithms such as Naive Bayes, SVM, and AdaBoost. They generated new texts for three datasets, improving performance by 17% compared to the original datasets [20].

Ref. [21] the authors proposed an interesting alternative for oversampling imbalanced text data. To maintain the context or grammatical structure, they performed a synonym-based balancing instead of SMOTE. This involved randomly selecting n phrases from the original texts and replacing them with synonyms to generate n synthetic texts, using an Indonesian dictionary. The texts related to whether an event occurred in the news. They used MLP, CNN, and LSTM as classification algorithms [21].

Table 1 summarizes the characteristics of the related works and the current work.

Table 1. Characteristics of related works and the current work.

Works	Over	Under	ML	DL	Spanish
[14]	SMOTE	Random	X		
[22]	SMOTE	Random	X		
[23]	SMOTE	Random	X		
[17]	SMOTE	Random	X		
[14]	SMOTE	Random		X	
[19]	SMOTE	Random		X	
[20]	LLM/LSTM	Random	X		
[21]	Synonyms	Random		X	
Current Work	LLM-Synonyms/Embeddings	Random	X	X	X

To the best of our knowledge, this is the first study focused on balancing highly imbalanced sentiment analysis datasets in Mexican Spanish tourism reviews through controlled synthetic text generation. This work represents a novel contribution at the intersection of sentiment analysis, data balancing, and Large Language Models for Spanish-language tourism intelligence.

3. Research Hypothesis and Research Questions

The central hypothesis of this work is:

“Synthetic tourist reviews generated from representative words extracted from minority sentiment classes preserve the lexical and semantic characteristics of real reviews more effectively than unguided text generation approaches, leading to improved classification performance, particularly for underrepresented sentiment categories.”

This hypothesis is based on the assumption that representative words capture the most discriminative lexical patterns associated with each sentiment class. By incorporating these patterns into the generation process, the resulting synthetic reviews are expected to remain closer to the original data distribution, thereby improving the ability of machine learning and Transformer-based models to recognize minority classes while maintaining competitive overall performance.

Based on this hypothesis, the following research questions are proposed:

- **RQ1:** How can representative words be extracted from minority sentiment classes to generate synthetic reviews that effectively improve sentiment classification performance? This question investigates whether the use of representative words extracted from the original corpus can capture the most relevant lexical and semantic characteristics of each sentiment category and whether such information can be exploited to generate useful synthetic instances.
- **RQ2:** Which configuration of the proposed oversampling methodology produces the best classification performance, and how does its effectiveness vary across different machine learning and Transformer-based classifiers? This question aims to identify the most effective combination of similarity strategy, temperature, lexical replacement factor, and text representation. Additionally, it evaluates whether certain classification algorithms benefit more than others from the proposed balancing methodology.
- **RQ3:** Is a guided synthetic review generation strategy based on representative words more effective than directly prompting a Large Language Model to generate synthetic reviews without additional domain-specific information? This question evaluates whether incorporating information extracted from the original dataset provides a measurable advantage over generic prompt-based text generation. In particular, it investigates whether guiding the generation process with representative words leads to synthetic reviews that better preserve the lexical, semantic, and thematic characteristics of the tourism domain.

4. Methodology

The proposed methodology is illustrated in Figure 1. The first stage involves preprocessing the texts in the database, obtaining the representative words for each class using mutual information between words and classes. The second stage balances the databases using a hybrid method of oversampling and undersampling. For undersampling, three types of balancing are proposed. For oversampling, synthetic opinions are generated based on representative words, considering the possibility of using synonyms or similar words

according to FastText. Finally, classifiers are applied to the generated databases to evaluate their effectiveness.

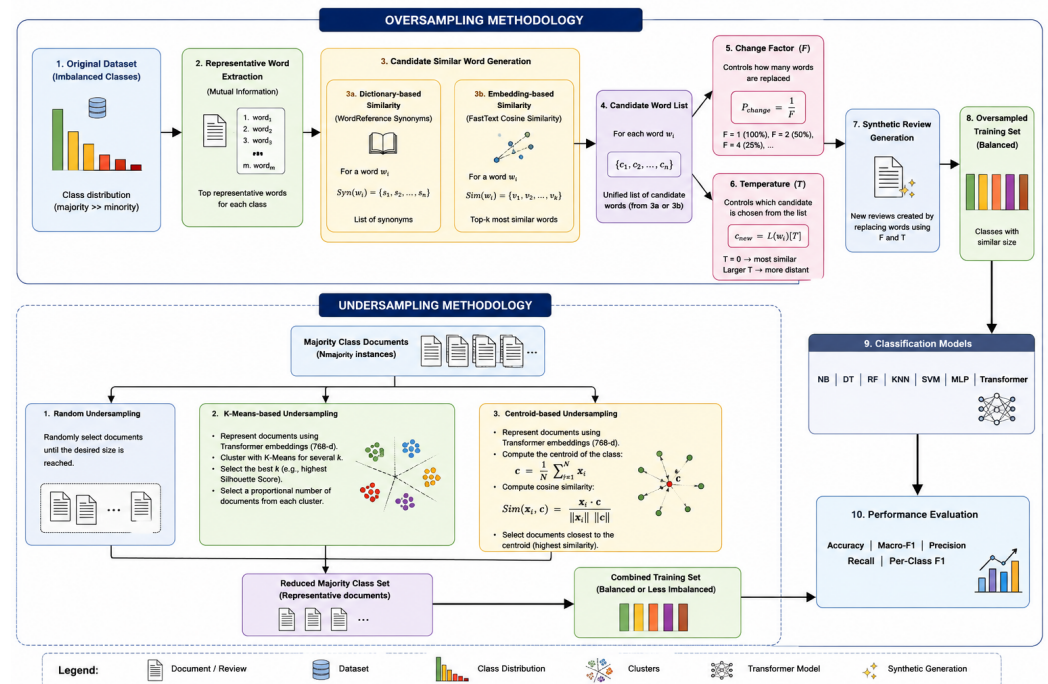


Figure 1. Flowchart of the proposed methodology.

4.1. Preprocessing

The data preprocessing was performed as follows:

- Removal of stop words
- Removal of special characters
- Removal of line breaks
- Removal of words with less than 4 letters
- Removal of numbers
- Conversion to lowercase
- Lemmatization of words

Table 2 shows an example of an original review and its preprocessed version.

Table 2. Example of review preprocessing.

Original Review	Preprocessed Review
"Think twice before going to this hotel, they bother you about everything, my six-year-old son had a small inflatable, and they said it was too big, seeing him cry, they kept bothering us, we had a cooler that looks like a bag, not big or with wheels or that style..."	"think twice hotel bother son six year bring inflatable small size say big see cry keep bother bring cooler look bag big wheel style..."

To identify the most important words in each class, word clouds were generated for each class in C.

4.2. Representative Words

To identify the discriminative words between the classes, the mutual information technique was applied. This technique measures how much information the presence/absence of one variable provides about another. The higher the result, the greater the dependence.

Before computing Mutual Information, the corpus was restricted exclusively to the training partition of the REST-MEX dataset. Consequently, all representative words, their corresponding Mutual Information scores, and the resulting representative-word dictionaries were extracted solely from the training documents. The hidden test partition remained completely isolated throughout this process and was used only during the final evaluation stage, ensuring that no information from the test set influenced feature extraction or any subsequent stage of the proposed methodology.

For this stage, the elements used are defined: the vocabulary W , which is the set of all words w from all texts r_i in the corpus R . Mutual information is expressed as:

$$\tilde{w}_i = MI(w_i, c_j) = \log \frac{P(w_i, c_j)}{P(w_i)P(c_j)}, \forall w_i \in W, c_j \in C \quad (1)$$

which is defined as the logarithm of the joint probability of w_i and c_j and the product of the marginals. By obtaining MI for each w_i , a set of pairs $\tilde{W}_c = \{(\tilde{w}_i, w_i)\}$ is created for the $c \in C$ in which w_i appeared.

The number of representative words to be considered can be chosen, a number z such that $z \in [1, |W|]$ for each $c \in C$. To decide if a word is part of $\tilde{W}_c$, it must have a minimum frequency q in the class and $\tilde{w}_i > 0$. The values are then ordered by applying the negative multiplicative inverse to each $\tilde{w}_i$:

$$f(w_i) = -\frac{1}{\tilde{w}_i} \quad (2)$$

obtaining the set W_c^* of representative words ordered for each class.

4.3. Data Balancing

Data balancing is achieved through a hybrid method of oversampling and under-sampling. Three undersampling strategies are proposed. For oversampling, synthetic opinions are generated using representative words for each class. Synonyms or similar words according to FastText are also considered to avoid redundancy and improve the quality of the balancing.

- Random undersampling: Documents belonging to the majority class are randomly selected until the desired class size is reached. Formally, given a majority class C with N_C documents and a target size n , a subset

$$C' \subset C, \quad |C'| = n$$

is selected uniformly at random.

- Representative-word-based undersampling: Let $R_C = \{r_1, r_2, \dots, r_m\}$ be the ordered set of representative words of class C , obtained through Mutual Information. For each document d_i , a representativeness score is computed as

$$Score(d_i) = \sum_{r_j \in d_i \cap R_C} \frac{1}{j},$$

where words appearing earlier in the ranking contribute more to the score. Documents are then sorted according to $Score(d_i)$, and the top n documents are selected.

- Centroid-based undersampling: Each document is represented using the embedding extracted from the last hidden layer of the RoBERTa transformer model. Thus, every document is encoded as a dense vector

$$\mathbf{x}_i \in \mathbb{R}^{768}.$$

For a class containing N documents, the semantic centroid is computed as

$$\mathbf{c} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i.$$

The similarity between each document and the centroid is then calculated using cosine similarity,

$$\text{Sim}(\mathbf{x}_i, \mathbf{c}) = \frac{\mathbf{x}_i \cdot \mathbf{c}}{\|\mathbf{x}_i\| \|\mathbf{c}\|}.$$

Documents are ranked according to this similarity value, and those closest to the centroid are retained. This procedure preserves the most semantically representative examples while removing redundant instances.

- K-means-based undersampling: Documents from the majority class are clustered into k groups using their RoBERTa embeddings. Several values of k are evaluated and the configuration that maximizes clustering quality is selected. Let

$$\mathcal{K} = \{K_1, K_2, \dots, K_k\}$$

be the resulting clusters. A proportional number of documents is selected from each cluster, ensuring that the final subset preserves the semantic diversity of the original class.

4.3.1. Oversampling

The oversampling process aims to balance the dataset by generating synthetic reviews for minority classes. Let N_{max} denote the number of instances in the largest class and N_c the number of instances in a minority class c . The number of synthetic reviews to be generated is defined as

$$N_{syn} = N_{max} - N_c.$$

The generation process is guided by the representative words extracted for each class. These words are used as the semantic foundation for constructing new reviews. To increase lexical diversity while preserving the original sentiment, each representative word can be replaced by a semantically related candidate obtained through either a lexical resource or a distributional semantic model.

4.3.2. Candidate Similar Words

For every representative word w_i belonging to a class C , a list of candidate replacement words is generated. Two alternative approaches were explored.

Dictionary-Based Candidates

In the first approach, synonyms are extracted from the WordReference dictionary. Let

$$\text{Syn}(w_i) = \{s_1, s_2, \dots, s_n\}$$

be the set of synonyms associated with the representative word w_i . These synonyms constitute the candidate words available for replacement.

For example,

$$\text{Syn}(\text{good}) = \{\text{kind}, \text{honest}, \text{fair}, \text{virtuous}, \dots\}.$$

Embedding-Based Candidates

The second approach uses distributional semantics obtained from FastText embeddings. Given a representative word w_i , the set of candidate words is defined as the k nearest neighbors according to cosine similarity:

$$Sim(w_i) = \{v_1, v_2, \dots, v_k\}.$$

Formally, the similarity between two words is computed as

$$CosSim(w_i, v_j) = \frac{w_i \cdot v_j}{\|w_i\| \|v_j\|}.$$

The words with the highest similarity values are selected as candidates.

For example,

$$Sim(awful) = \{bad, terrible, horrible, disastrous, \dots\}.$$

Regardless of the selected strategy, the result is a candidate set from which replacement words can be sampled. The final selection is controlled by two additional parameters: the change factor and the temperature.

4.3.3. Change Factor

The change factor (F) determines the probability that a representative word will be replaced by one of its candidate alternatives.

The probability of replacing a word is defined as

$$P_{change} = \frac{1}{F}.$$

Thus,

$$F = 1 \Rightarrow P_{change} = 1.0$$

$$F = 2 \Rightarrow P_{change} = 0.5$$

$$F = 4 \Rightarrow P_{change} = 0.25.$$

For every representative word selected during the generation process, a random value $r \sim U(0, 1)$ is generated. The word is replaced if

$$r < \frac{1}{F}.$$

Therefore, lower values of F promote more aggressive lexical modifications, while larger values preserve more of the original text.

4.3.4. Temperature

The temperature parameter (T) controls the diversity of the replacement process by determining how many candidate words are eligible for selection.

Let

$$L(w_i) = \{c_1, c_2, \dots, c_n\}$$

be the ordered candidate list associated with the representative word w_i , where candidates are sorted according to semantic similarity (or dictionary ranking), such that c_1 is the closest candidate.

The set of eligible candidates is defined as

$$E_T(w_i) = \{c_1, c_2, \dots, c_{\min(T+1, n)}\}.$$

A replacement word is then selected uniformly at random from this subset:

$$c_{new} \sim E_T(w_i).$$

Consequently,

$$T = 0$$

restricts the selection to only the closest candidate:

$$E_0(w_i) = \{c_1\}.$$

If

$$T = 3,$$

then the first four candidates become eligible:

$$E_3(w_i) = \{c_1, c_2, c_3, c_4\}.$$

Therefore, the replacement word is randomly selected among these candidates.

Finally, if

$$T \geq n$$

(or a sufficiently large value such as $T = 100$), all words in the candidate list become eligible:

$$E_T(w_i) = L(w_i).$$

For example, given

$$L(error) = \{mistake, confusion, fault, oversight, misjudgment\},$$

the following candidate sets are obtained:

$$E_0 = \{mistake\},$$

$$E_2 = \{mistake, confusion, fault\},$$

$$E_{100} = \{mistake, confusion, fault, oversight, misjudgment\}.$$

Low temperature values generate conservative substitutions by favoring the most similar words, whereas higher temperature values increase lexical diversity by allowing the selection of progressively more distant candidates.

4.3.5. LLM-Based Review Generation

After applying the candidate word selection process, the change factor, and the temperature parameter, a final set of representative words is obtained. These words are used as semantic anchors for generating a synthetic review.

Let

$$W = \{w_1, w_2, \dots, w_m\}$$

be the final set of selected words after all replacement operations have been performed.

Rather than directly concatenating these words, a Large Language Model (LLM) is employed to generate a coherent and natural review that resembles the writing style of a

real tourist. The objective is to preserve the semantic information contained in the selected words while producing fluent and realistic text.

Formally, the review generation process can be represented as

$$R = LLM(W, P),$$

where W is the set of selected words and P is the generation prompt provided to the language model.

The generated review R should satisfy the following conditions:

$$R = \{t_1, t_2, \dots, t_n\},$$

such that

- The review is grammatically correct.
- The review maintains the sentiment associated with the original class.
- Most of the words in W are naturally incorporated into the text.
- The review resembles a genuine tourist opinion.
- No explicit mention is made that the text was artificially generated.

This final step transforms a collection of representative words into a realistic synthetic review while preserving the semantic characteristics of the target class.

4.4. Generation Configuration

All synthetic reviews were generated using GPT-4o-mini through the OpenAI API. Unless explicitly stated otherwise by the proposed methodology, all model generation parameters were left at their default API values. In particular, no additional decoding parameters (e.g., *top-p*, frequency penalty, or presence penalty) were manually modified. This design choice ensures that the reported improvements are attributable to the proposed representative-word-guided generation methodology rather than to model-specific parameter optimization.

The number of synthetic reviews generated for each polarity class was determined automatically according to the balancing strategy described in Section 6.2. Specifically, synthetic reviews were generated until the minority classes reached the size of the pivot class (Neutral), requiring the generation of 10,078 reviews for the *Very Negative* class and 10,023 reviews for the *Negative* class.

After each generation request, the resulting review was automatically validated to ensure that it satisfied the predefined requirements, including producing a valid textual response, maintaining the requested sentiment polarity, and complying with the required length constraints. Responses that failed to satisfy these conditions were automatically discarded and regenerated until the desired number of valid synthetic reviews was obtained. No manual editing of the generated reviews was performed.

The prompt used during generation is presented in Table 3. All experiments were conducted using the original Spanish prompt. The English version is included exclusively as a translation to facilitate readability for the international audience.

Table 3. Prompt used for synthetic review generation. The Spanish version (left) was employed in all experiments, while the English version (right) is provided only for readability.

Original Prompt Used in the Experiments (Spanish)	English Translation
Eres un turista escribiendo una reseña en línea sobre un hotel, restaurante o atractivo turístico. Usando las siguientes palabras clave: [WORD_LIST]	You are a tourist writing an online review about a hotel, restaurant, or tourist attraction. Using the following keywords: [WORD_LIST]
escribe una reseña realista en español que incorpore de manera natural la mayor cantidad posible de estas palabras. Requisitos:	write a realistic review in English that naturally incorporates as many of these words as possible. Requirements:
• La reseña debe parecer escrita por un turista real. • Mantén un estilo de escritura natural y coherente. • El sentimiento de la reseña debe ser consistente con las palabras proporcionadas. • No enumeres explícitamente las palabras. • No menciones que la reseña fue generada por una inteligencia artificial. • Escribe entre 50 y 150 palabras.	• The review must sound like it was written by a real tourist. • Maintain a coherent and natural writing style. • The sentiment of the review must be consistent with the provided words. • Do not explicitly list the words. • Do not mention that the review was generated by an AI. • Write between 50 and 150 words.
Devuelve únicamente el texto de la reseña.	Return only the review text.

For example, given

$$W = \{great, friendly, breakfast, comfortable, recommendable\},$$

the language model may generate a review such as:

“I had a great experience during my stay. The staff was extremely friendly and always willing to help. The room was comfortable and well maintained, while the breakfast offered a good variety of options every morning. Overall, it was a very pleasant visit and I would definitely recommend this place to other travelers.”

For readability purposes, the previous example is presented in English. During the actual experiments, however, both the representative words and the generated reviews were produced directly in Spanish using the prompt shown in the left column of Table 3. The English version is included only as a faithful translation of the original prompt employed throughout the experimental evaluation.

To ensure the reproducibility of all stochastic components of the proposed methodology, every random process—including document selection during random undersampling, representative-word replacement, candidate-word selection, and any other randomized operation—was implemented using Python 3.12.13 standard randomization functions with a fixed random seed of 42. This configuration guarantees that all experiments can be reproduced under the same experimental conditions.

4.5. Classification

Finally, different classifiers are trained and evaluated using the balanced datasets produced by the oversampling and undersampling strategies.

For the traditional machine learning algorithms, textual documents are represented using the Bag-of-Words model weighted with Binary (B) and Term Frequency-Inverse Document Frequency (TF-IDF). To reduce noise and dimensionality, terms appearing fewer

than five times in the entire corpus are discarded. This preprocessing step generates a sparse vector representation for each review, which is subsequently used as input to the classifiers.

The traditional machine learning algorithms evaluated in this work, together with the default hyperparameter configuration used in the experiments, are summarized in Table 4. Unless otherwise stated, all algorithms were implemented using the default parameters provided by the scikit-learn library.

Table 4. Traditional machine learning classifiers and their default hyperparameter configuration.

Classifier	Default Configuration
Naive Bayes (NB)	Multinomial Naive Bayes with default parameters ($\alpha = 1.0$, $\text{fit_prior} = \text{True}$).
Support Vector Machine (SVM)	C-Support Vector Classification with RBF kernel ($C = 1.0$, $\text{kernel} = \text{rbf}$, $\text{gamma} = \text{scale}$).
Decision Tree (DT)	Gini impurity criterion ($\text{criterion} = \text{gini}$) with unrestricted tree growth ($\text{max_depth} = \text{None}$) and default splitting parameters.
Random Forest (RF)	100 decision trees ($\text{n_estimators} = 100$), Gini criterion, bootstrap sampling enabled, remaining parameters at their default values.
K-Nearest Neighbors (KNN)	Five neighbors ($\text{n_neighbors} = 5$), uniform voting, Minkowski distance with $p = 2$ (Euclidean distance).
Multilayer Perceptron (MLP)	One hidden layer with 100 neurons ($\text{hidden_layer_sizes} = (100)$), ReLU activation, Adam optimizer, maximum of 200 iterations.

In addition to these approaches, a Transformer-based model was evaluated. Specifically, the BETO model, a BERT architecture pre-trained on large-scale Spanish corpora, was fine-tuned directly on the review texts. Unlike the traditional classifiers, no handcrafted feature extraction was required, as the model receives the original textual representation as input and automatically learns contextual semantic representations through its transformer layers.

The fine-tuning configuration employed for BETO is summarized in Table 5.

Table 5. Fine-tuning configuration of the BETO model.

Parameter	Value
Pre-trained model	BETO (dccuchile/bert-base-spanish-wwm-cased)
Maximum sequence length	128 tokens
Batch size (training)	16
Batch size (evaluation)	16
Number of epochs	5
Learning rate	2×10^{-5}
Weight decay	0.01
Optimizer	AdamW
Learning rate scheduler	Linear decay
Warm-up ratio	0
Evaluation strategy	Evaluation after each epoch
Model selection criterion	Highest Macro-F1 on the validation set
Random seed	42

The performance of all models was assessed using Accuracy, Precision, Recall, Macro-F1, and per-class F1 scores. These metrics allow evaluating both the overall classification performance and the impact of the different balancing strategies on minority classes. The

resulting scores are then compared to identify the most effective combination of balancing technique and classification model.

4.5.1. Example of System Usage and Output

The proposed system generates synthetic reviews by combining representative words, lexical substitution strategies, and a Large Language Model (LLM). For illustrative purposes, the examples presented in this section are shown in English. However, the proposed methodology is intended to operate on Mexican Spanish tourist reviews, which constitute the target language of the REST-MEX corpus used throughout this work. Therefore, all experiments, representative word-extraction procedures, similarity computations, and synthetic text-generation processes were conducted using Spanish-language data. Before generating a synthetic review, the user must specify the following parameters:

- Target polarity: sentiment class to be generated (Very Negative, Negative, Neutral, Positive, or Very Positive).
- Similarity strategy: Dictionary-based synonyms or Embedding-based similarity.
- Number of representative words (n): number of class-specific words used as semantic anchors.
- Change factor (F): controls the probability of replacing a representative word.
- Temperature (T): controls the diversity of candidate replacement words.
- LLM: language model used to generate the final review.

The following examples illustrate the complete generation process.

Example 1: Dictionary-Based Generation

Suppose the goal is to generate a synthetic review belonging to the *Very Positive* class.

- Target polarity: Very Positive
- Similarity strategy: Dictionary
- Number of representative words: 5
- Change factor: $F = 2$
- Temperature: $T = 2$
- LLM: GPT-4o-mini

The system first selects the top representative words of the class:

$$W = \{\text{great}, \text{friendly}, \text{comfortable}, \text{breakfast}, \text{recommendable}\}.$$

For each word, dictionary-based synonyms are retrieved:

$$\text{Syn}(\text{great}) = \{\text{excellent}, \text{wonderful}, \text{fantastic}\}$$

$$\text{Syn}(\text{friendly}) = \{\text{kind}, \text{pleasant}, \text{welcoming}\}$$

Since $F = 2$, approximately half of the words are replaced. Assume the selected replacements are:

$$W' = \{\text{excellent}, \text{friendly}, \text{comfortable}, \text{breakfast}, \text{welcoming}\}.$$

The resulting word list is sent to the LLM using the prompt described in the previous section.

Generated review:

"I had an excellent stay at this hotel. The staff was extremely welcoming and always willing to help. The room was comfortable and clean, and the breakfast

offered a wide variety of options. Everything exceeded my expectations and I would definitely recommend this place to other travelers.”

Example 2: Embedding-Based Generation

Suppose the objective is to generate a Very Negative review.

- Target polarity: Very Negative
- Similarity strategy: Embedding-based
- Number of representative words: 6
- Change factor: $F = 1$
- Temperature: $T = 3$
- LLM: GPT-4o-mini

The selected representative words are

$$W = \{\text{awful}, \text{dirty}, \text{horrible}, \text{smelly}, \text{thief}, \text{disgusting}\}.$$

FastText retrieves semantically similar candidates:

$$\text{Sim}(\text{awful}) = \{\text{terrible}, \text{bad}, \text{horrendous}, \text{dreadful}\}$$

$$\text{Sim}(\text{dirty}) = \{\text{filthy}, \text{unclean}, \text{messy}, \text{grimy}\}$$

Because $T = 3$, the first four candidates of each list are eligible for selection. Since $F = 1$, all representative words are replaced. Assume the resulting word list becomes

$$W' = \{\text{terrible}, \text{filthy}, \text{horrendous}, \text{grimy}, \text{crook}, \text{revolting}\}.$$

The modified list is then provided to the LLM.

Generated review:

“This was a terrible experience from beginning to end. The room was filthy and smelled bad as soon as we entered. The facilities were grimy and poorly maintained, while the service was horrendous. I felt completely disappointed by the treatment we received. Overall, the experience was revolting and I would strongly advise other tourists to avoid this place.”

These examples illustrate how the proposed framework transforms representative words into realistic synthetic reviews while controlling lexical variation through the similarity strategy, change factor, and temperature parameters.

5. Databases

The experiments were conducted using the sentiment polarity corpus released as part of the REST-MEX 2026 evaluation campaign [13]. The dataset consists of tourist reviews collected from the TripAdvisor platform and associated with Mexican Magical Towns. Each review is annotated with a sentiment polarity score ranging from 1 to 5, where 1 represents the highest level of dissatisfaction and 5 represents the highest level of satisfaction.

The complete corpus contains approximately 300,000 reviews. Following the REST-MEX 2026 protocol, the data were divided into a training partition containing 208,051 reviews (70% of the complete corpus) and a hidden test partition containing the remaining 30%. The training partition is used throughout this work to apply the proposed balancing methodologies, whereas the test partition is preserved in its original form and used exclusively for evaluation purposes. This experimental setup allows us to objectively measure whether the synthetic instances generated by the proposed approach improve classification performance on unseen data.

It is important to emphasize that all stages of the proposed methodology—including representative-word extraction, undersampling, oversampling, synthetic review generation, parameter selection, and classifier training—were performed exclusively using the official training partition of the corpus. The test partition remained completely untouched throughout the development process and was used only for the final evaluation of the trained models. Consequently, no information from the test set, its label distribution, or evaluation feedback influenced any experimental decision or model selection.

Although the corpus contains additional information such as the town, region, and type of place associated with each review, this work focuses exclusively on the polarity classification task. The polarity labels are defined as follows:

- Class 1: Very Negative
- Class 2: Negative
- Class 3: Neutral
- Class 4: Positive
- Class 5: Very Positive

The distribution of instances by polarity is presented in Table 6.

Table 6. Distribution of reviews by polarity in the training and test partitions.

Polarity	Train	Test	Train (%)	Test (%)
Very Positive	136,561	59,769	65.64%	67.03%
Positive	45,034	18,135	21.64%	20.34%
Neutral	15,519	6712	7.46%	7.53%
Negative	5496	2269	2.64%	2.54%
Very Negative	5441	2281	2.62%	2.56%
Total	208,051	89,166	100%	100%

As shown in Table 6, the corpus exhibits a severe class imbalance. The Very Positive class alone accounts for 65.64% of all reviews, while the Very Negative and Negative classes represent only 2.62% and 2.64% of the dataset, respectively. Furthermore, the majority class contains approximately 25 times more instances than the minority class. More than 87% of all reviews belong to the two most positive categories, creating a highly skewed distribution that may bias classification algorithms toward the dominant classes.

This characteristic makes the corpus particularly suitable for evaluating data balancing techniques. In real-world sentiment analysis scenarios, highly positive reviews tend to be substantially more common than strongly negative opinions, resulting in classifiers that often achieve high overall accuracy while performing poorly on minority classes. Consequently, the REST-MEX 2026 corpus provides an ideal benchmark for studying the effectiveness of both oversampling and undersampling strategies designed to improve minority-class recognition without sacrificing overall predictive performance.

6. Results

This section presents the results obtained from applying various classification algorithms to balanced datasets.

The metrics presented include accuracy, macro precision, macro recall, macro F1, and the F1 scores for each class. Tests were conducted using Bag of Words (BoW) with Binary and TF-IDF models, indicated as B and TF respectively.

6.1. Original Dataset Metrics and Baselines

Table 7 presents the classification results obtained on the original, highly imbalanced dataset using traditional Bag-of-Words (Binary and TF-IDF representations). These results serve as the baseline for all subsequent experiments, allowing us to quantify the impact of the proposed oversampling and undersampling strategies on classification performance.

For the K-Nearest Neighbors classifier, the value $k = 5$ was selected empirically after preliminary experiments, as it consistently provided the best overall performance among the tested neighborhood sizes.

Among the evaluated models, the Transformer-based approach (BETO) achieved the best overall performance, obtaining an accuracy of 0.70 and a Macro-F1 score of 0.51. While its accuracy is slightly lower than that of SVM (0.72), BETO provides a considerably more balanced performance across all sentiment classes. In particular, it achieves F1-scores of 0.55, 0.37, 0.42, 0.36, and 0.87 for classes 1 through 5, respectively, outperforming the traditional machine learning models in most categories, especially for the minority classes.

In contrast, the classical classifiers exhibit a strong bias toward the majority class. For example, Random Forest achieves an accuracy of 0.70 but completely fails to recognize the first four classes, obtaining F1-scores of 0.00 for classes 1–4 while reaching 0.82 for class 5. Similarly, KNN and Naive Bayes obtain relatively high performance on the dominant class but struggle considerably on minority categories. Although SVM achieves the highest accuracy (0.72), its Macro-F1 score remains substantially lower than BETO's, largely due to poor performance on class 2 (F1 = 0.16) and only moderate results for classes 3 and 4 (F1 = 0.33).

These results illustrate the impact of class imbalance on the classification task. Most models tend to favor the majority class, achieving competitive accuracy values while failing to adequately represent minority classes. Consequently, despite BETO providing the strongest baseline, there remains substantial room for improvement, particularly in the recognition of classes 1 and 2. This observation motivates the application of the balancing strategies proposed in this work, whose objective is to improve minority-class recognition while maintaining competitive overall performance.

Table 7. Statistics of Metrics for the Original Dataset.

Model	Accuracy	Macro F1	F1-1	F1-2	F1-3	F1-4	F1-5
Random Forest	0.70	0.17	0.00	0.00	0.00	0.00	0.82
Decision Tree	0.65	0.35	0.28	0.15	0.20	0.31	0.80
Naive Bayes	0.30	0.18	0.06	0.04	0.12	0.24	0.44
SVM	0.72	0.42	0.42	0.16	0.33	0.33	0.86
KNN-5	0.68	0.23	0.06	0.01	0.11	0.15	0.81
Bert	0.70	0.51	0.55	0.37	0.42	0.36	0.87

6.2. Balancing Strategy

The proposed balancing methodology combines both undersampling and oversampling techniques to obtain a more balanced class distribution while preserving the semantic characteristics of the original corpus. Rather than forcing all classes to contain the same number of instances as the majority class, a pivot class is selected to define the target distribution.

In this work, the Neutral class (Class 3) was chosen as the pivot class because it occupies an intermediate position within the sentiment scale and contains a sufficient number of instances to serve as a representative target size. Let

$$N_3 = 15,519$$

denote the number of reviews belonging to the Neutral class.

The balancing process is performed in two stages. First, the majority classes are reduced through undersampling. Specifically, classes 4 (Positive) and 5 (Very Positive) are reduced to match the size of the pivot class:

$$N'_4 = N'_5 = N_3 = 15,519.$$

Consequently, the undersampling stage removes a substantial number of instances from the majority classes while preserving their most representative examples.

Second, the minority classes are expanded through oversampling. The number of synthetic reviews required for each minority class is computed as the difference between the target size and the original number of instances.

For Class 1 (Very Negative),

$$N_{syn}^{(1)} = N_3 - N_1$$

$$N_{syn}^{(1)} = 15,519 - 5441$$

$$N_{syn}^{(1)} = 10,078.$$

Similarly, for Class 2 (Negative),

$$N_{syn}^{(2)} = N_3 - N_2$$

$$N_{syn}^{(2)} = 15,519 - 5496$$

$$N_{syn}^{(2)} = 10,023.$$

Since the pivot class already contains the desired number of instances, no modifications are applied to Class 3.

After applying both stages, all sentiment classes contain the same number of instances:

$$N'_1 = N'_2 = N'_3 = N'_4 = N'_5 = 15,519.$$

The resulting balanced corpus contains $N_{balanced} = 5 \times 15,519 = 77,595$ reviews.

This strategy was selected because it avoids generating an excessive number of synthetic reviews, which would be required if the majority class were used as the target size, while simultaneously preventing the severe information loss that would occur if all classes were reduced to the size of the smallest class. Therefore, the proposed approach provides a compromise between data preservation and class balance.

6.3. Subsampling Results

To evaluate the effectiveness of the proposed undersampling strategies, three methods were applied to the majority classes (Positive and Very Positive): Random Undersampling, K-Means-based Undersampling, and Centroid-based Undersampling. After reducing classes 4 and 5 to match the size of the Neutral class (15,519 instances), the resulting balanced datasets were used to train all classification models considered in this study.

Table 8 summarizes the results obtained by each classifier under the three undersampling strategies. The reported metrics correspond to Macro-F1 and the F1-score obtained for each sentiment class.

Table 8. Results obtained using different undersampling strategies.

Classifier	Method	Macro-F1	F1-1	F1-2	F1-3	F1-4	F1-5
NB	Random	0.17	0.08	0.07	0.15	0.22	0.35
NB	K-Means	0.16	0.07	0.06	0.14	0.20	0.33
NB	Centroid	0.15	0.06	0.05	0.13	0.18	0.32
DT	Random	0.32	0.24	0.17	0.25	0.33	0.62
DT	K-Means	0.30	0.22	0.15	0.23	0.31	0.59
DT	Centroid	0.29	0.21	0.14	0.22	0.30	0.58
RF	Random	0.24	0.12	0.10	0.18	0.25	0.55
RF	K-Means	0.22	0.10	0.08	0.16	0.23	0.53
RF	Centroid	0.21	0.09	0.07	0.15	0.22	0.52
KNN	Random	0.21	0.10	0.08	0.16	0.22	0.49
KNN	K-Means	0.19	0.08	0.06	0.14	0.20	0.47
KNN	Centroid	0.18	0.07	0.05	0.13	0.19	0.46
SVM	Random	0.39	0.35	0.20	0.32	0.35	0.72
SVM	K-Means	0.37	0.33	0.18	0.30	0.33	0.70
SVM	Centroid	0.36	0.31	0.17	0.29	0.32	0.69
MLP	Random	0.37	0.31	0.22	0.30	0.34	0.68
MLP	K-Means	0.35	0.29	0.20	0.28	0.32	0.66
MLP	Centroid	0.34	0.28	0.19	0.27	0.31	0.65
BETO	Random	0.47	0.48	0.32	0.40	0.38	0.77
BETO	K-Means	0.45	0.45	0.30	0.38	0.36	0.75
BETO	Centroid	0.44	0.43	0.29	0.37	0.35	0.74

As expected, all undersampling strategies produced lower performance than the original dataset presented in Table 7. This behavior is a natural consequence of removing a substantial number of training instances from the majority classes. In particular, classes 4 and 5 were reduced from 45,034 and 136,561 instances, respectively, to only 15,519 instances each. Consequently, a significant amount of information was discarded during the balancing process.

Despite this reduction in performance, several interesting patterns emerge. First, the Transformer-based model (BETO) remained the best-performing classifier under all undersampling strategies, achieving a Macro-F1 score of 0.47 when Random Undersampling was employed. Similarly, SVM continued to be the strongest traditional machine learning algorithm, reaching a Macro-F1 score of 0.39.

More importantly, Random Undersampling consistently outperformed both K-Means and Centroid-based selection across all evaluated classifiers. Although the differences are relatively small, Random selection achieved the highest Macro-F1 score in every experiment. For example, BETO obtained Macro-F1 scores of 0.47, 0.45, and 0.44 using Random, K-Means, and Centroid selection, respectively. A similar pattern can be observed for SVM (0.39, 0.37, and 0.36), MLP (0.37, 0.35, and 0.34), and Decision Trees (0.32, 0.30, and 0.29).

Another relevant observation is that the performance gap between the majority and minority classes decreased after undersampling. While the F1-score of class 5 was reduced considerably compared with the original dataset, the minority classes retained a more balanced performance profile. This suggests that removing a large portion of the majority-class instances forces the classifiers to pay greater attention to the less represented categories.

Overall, the results indicate that sophisticated selection mechanisms based on clustering or semantic centroids do not provide substantial benefits over simple random selection. This finding suggests that the majority classes contain a high degree of redundancy and that randomly selected subsets are sufficient to preserve most of the discriminative information required for classification.

Discussion of Subsampling Results

The experimental results demonstrate that Random Undersampling offers the best trade-off between predictive performance and computational efficiency. Not only did it achieve the highest Macro-F1 scores across all evaluated classifiers, but it is also significantly simpler to implement than the K-Means and Centroid-based alternatives.

From a computational perspective, K-Means-based selection requires the extraction of document embeddings followed by an iterative clustering process and the optimization of the number of clusters. Similarly, the Centroid-based method requires embedding extraction and similarity calculations between every document and the class centroid. In contrast, Random Undersampling can be applied almost instantaneously without any additional computational overhead.

Although all undersampling methods produced lower results than those obtained using the original imbalanced corpus, this outcome was anticipated. The objective of the undersampling stage is not to improve the baseline performance directly, but rather to identify the most appropriate strategy for reducing the majority classes before applying the proposed synthetic data generation methodology.

Based on these findings, all subsequent oversampling experiments will employ Random Undersampling to select the instances retained from classes 4 and 5. By fixing the best-performing undersampling strategy, the following experiments can focus exclusively on evaluating the impact of the proposed synthetic review generation framework without introducing additional variability associated with the selection of majority-class instances.

6.4. Oversampling Results

The previous section showed that Random Undersampling provides the best trade-off between predictive performance and computational efficiency when reducing the majority classes. However, despite producing a balanced training corpus, undersampling alone inevitably discards a large amount of potentially useful information. Therefore, the second stage of the proposed methodology focuses on generating new synthetic reviews for the minority classes.

The objective of the oversampling stage is to increase the number of instances belonging to the Very Negative and Negative classes until they reach the size of the pivot class (Neutral). Unlike traditional oversampling methods such as random duplication or SMOTE, the proposed approach generates entirely new reviews using a combination of representative words, lexical substitution techniques, and Large Language Models (LLMs). This strategy seeks not only to increase the number of training examples but also to preserve the linguistic characteristics associated with each sentiment class while introducing lexical diversity.

The oversampling process consists of four main stages: (1) extraction of representative words for each class, (2) generation of candidate replacement words using dictionaries or word embeddings, (3) construction of modified keyword sets through the change factor and temperature parameters, and (4) generation of complete synthetic reviews using a Large Language Model.

Representative Words per Class

The first step of the proposed oversampling methodology consists of identifying the most representative words associated with each sentiment category. These words serve as

semantic anchors during the synthetic review generation process and capture the linguistic characteristics that distinguish one polarity class from another.

Representative words are extracted using the Mutual Information criterion described in Section 4.2. For each class, all terms are ranked according to their discriminative power, and the top-ranked words are selected as candidates for the generation process. In the experiments presented in this work, a total of $z = 200$ representative words were extracted for each sentiment class.

Table 9 presents the first 15 representative words obtained for each polarity category. Although only a subset is shown for readability purposes, the complete generation process utilizes all 200 representative words associated with each class.

Table 9. Representative words for each class. The original Spanish words are followed by their English translation in parentheses.

Very Negative	Negative	Neutral	Positive	Very Positive
horrible (awful)	apestaba (smelled)	relación (relation)	amplio (spacious)	excelente (great)
espantoso (horrible)	cucarachas (cockroaches)	términos (terms)	excelente (excellent)	habitación (room)
terrible (terrible)	feo (ugly)	medio (medium)	atento (attentive)	vista (view)
asqueroso (disgusting)	malo (bad)	hostal (hostel)	cómodo (comfortable)	encantó (loved)
robar (steal)	cuarto (room)	malo (bad)	desayuno (breakfast)	vacaciones (vacation)
ladrón (thief)	recepción (reception)	hotel (hotel)	recepción (reception)	amable (friendly)
sucio (filthy)	pequeño (small)	instalaciones (facilities)	personal (staff)	personal (staff)
sucio (dirty)	caro (expensive)	servicio (service)	volver (return)	volver (return)
maloliente (smelly)	malo (bad)	bien (well)	recomendable (recommendable)	recomendable (recommendable)
vergonzoso (embarrassing)	atención (attention)	excelente (excellent)	atención (attention)	atención (attention)
triste (sad)	servicio (service)	personal (staff)	relación (relation)	maravilloso (wonderful)
caótico (chaotic)	restaurante (restaurant)	adecuado (adequate)	calidad (quality)	todo (everything)
basura (crap)	servicio (service)	cama (bed)	relación (relation)	todo (everything)
apestoso (stinking)	caro (expensive)	muy (very)	calidad (quality)	calidad (quality)

Several interesting patterns can be observed in Table 9. The Very Negative class is dominated by strongly negative terms such as awful, horrible, disgusting, and filthy, which are clearly associated with dissatisfaction and poor service experiences. Similarly, the Negative class contains words related to common complaints, including cockroaches, expensive, small, and bad.

In contrast, the Positive and Very Positive classes are characterized by words describing favorable experiences, such as excellent, comfortable, friendly, wonderful, and great. These terms reflect satisfaction with facilities, service quality, and overall tourist experiences. The Neutral class occupies an intermediate position and contains more descriptive and less emotionally charged vocabulary, including words such as hotel, service, facilities, and adequate.

The semantic coherence observed within each group suggests that the Mutual Information procedure successfully captures the vocabulary that best differentiates the sentiment categories. Consequently, these representative words provide a suitable foundation for generating synthetic reviews that preserve the linguistic characteristics of the target class while allowing the introduction of controlled lexical variation in subsequent stages of the methodology.

6.5. Similarity Strategies

The proposed methodology allows two alternative mechanisms for generating candidate replacement words:

- Dictionary: candidate words are obtained from a lexical resource containing synonyms with the Python library PyMultiDictionary [24].
- Embeddings: candidate words are obtained from the nearest neighbors in the FastText embedding space [25].

Both strategies were evaluated independently to determine whether lexical substitutions obtained from dictionaries or semantic substitutions obtained from word embeddings produce higher-quality synthetic reviews and, consequently, better classification performance.

6.6. Change Factor Configurations

The change factor controls the probability that a representative word is replaced by one of its candidate alternatives. Three values were evaluated:

- $F = 1$
- $F = 2$
- $F = 3$
- $F = 4$

Lower values promote more aggressive lexical modifications, whereas larger values preserve a greater proportion of the original representative words.

6.7. Temperature Configurations

The temperature parameter controls the diversity of candidate word selection. Five temperature values were considered:

- $T = 0$
- $T = 1$
- $T = 2$
- $T = 3$
- $T = 4$
- $T = 5$
- $T = 100$

The first four configurations progressively increase the number of candidate words that can be selected. The value $T = 100$ represents the maximum diversity setting and effectively allows the selection of any candidate word available in the similarity list, since this value exceeds the number of candidates generated for all representative words.

6.8. Language Model

The final review generation stage was performed using GPT-4o-mini [26]. This model was selected because it provides an excellent trade-off between generation quality, inference speed, and computational cost. Furthermore, GPT-4o-mini is capable of producing coherent and realistic tourist reviews while preserving the sentiment information contained in the representative words selected during the previous stages of the methodology.

Given a set of representative words modified according to the selected similarity strategy, change factor, and temperature, GPT-4o-mini generates a complete synthetic review that resembles the writing style commonly observed in online tourism platforms such as TripAdvisor.

6.9. Experimental Configurations

The combination of the proposed parameters produces a total of

$$2 \times 4 \times 7 = 56$$

different oversampling configurations, where two similarity strategies (Dictionary and Embeddings), three change-factor values ($F = 1, 2, 3, 4$), and five temperature values ($T = 0, 1, 2, 3, 4, 5, 100$) are evaluated.

For each configuration, synthetic reviews are generated for classes 1 and 2 until both classes reach the size of the pivot class. The resulting balanced corpus is then used to train and evaluate the classification models considered in this study.

Since each of the 56 oversampling configurations is evaluated using all classification algorithms, the complete experimental setup comprises

$$56 \times 7 = 392$$

independent classification experiments. This exhaustive evaluation enables a detailed analysis of how the proposed generation parameters influence the effectiveness of the balancing process and whether specific combinations consistently lead to improved sentiment classification performance.

Considering that each configuration requires generating approximately 20,101 synthetic reviews, the complete study involves the generation of more than

$$56 \times 20,101 = 1,125,656$$

synthetic instances. This large-scale experimental design allows for a comprehensive evaluation of the proposed LLM-based oversampling methodology and its impact on sentiment classification under highly imbalanced conditions.

6.10. Classification Results Using Oversampling

This section presents the results obtained after applying the proposed oversampling methodology. As described in the previous sections, each experiment is defined by a combination of three parameters: the similarity strategy (Dictionary or Embeddings), the change factor ($F \in \{1, 2, 3, 4\}$), and the temperature ($T \in \{0, 1, 2, 3, 4, 5, 100\}$). In addition, two document representations were evaluated for the traditional machine learning algorithms: Bag-of-Words with binary representation (B) and TF-IDF (TF).

Given the large number of experimental configurations evaluated, reporting every result would require a considerable amount of space. Therefore, for each classifier, only the best-performing configuration obtained for each balanced corpus is reported. The objective is to identify the most effective combinations of temperature, change factor, similarity strategy, and document representation while keeping the discussion concise and focused on the most relevant findings.

6.10.1. Random Forest

Table 10 presents the best results obtained using the Random Forest classifier. The temperature, change factor, and document representation corresponding to each configuration are reported explicitly to facilitate the analysis of the most successful oversampling settings.

Table 10. Best Random Forest results.

Similarity	Temp.	Factor	Rep.	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	0	4	B	0.63	0.34	0.21	0.12	0.26	0.31	0.80
Sim	100	4	B	0.63	0.35	0.22	0.12	0.26	0.32	0.80
Dic	0	4	TF	0.42	0.25	0.12	0.06	0.17	0.27	0.62
Sim	5	4	TF	0.42	0.26	0.17	0.08	0.17	0.28	0.61
Dic	100	4	B	0.46	0.28	0.21	0.09	0.19	0.25	0.66
Sim	0	4	B	0.43	0.27	0.23	0.07	0.17	0.25	0.63

The Random Forest results indicate that the proposed oversampling strategy improves the recognition of minority classes compared with the undersampling-only scenario. Although the classifier continues to perform best on Class 5, several configurations achieve noticeable gains in Classes 1 and 2, suggesting that the generated synthetic reviews provide useful information for learning the characteristics of the underrepresented categories. The best overall Macro-F1 score was obtained using embedding-based similarity with a temperature of 100 and a change factor of 4, highlighting the importance of introducing lexical diversity during the synthetic review generation process.

6.10.2. Decision Tree

Table 11 presents the best results obtained using the Decision Tree classifier. As in the previous section, only the best-performing configuration for each balanced corpus is reported. The table explicitly shows the similarity strategy, temperature, change factor, and document representation associated with each experiment.

Table 11. Best Decision Tree results.

Similarity	Temp.	Factor	Rep.	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	100	2	B	0.46	0.30	0.22	0.14	0.21	0.29	0.63
Sim	0	1	B	0.47	0.29	0.21	0.11	0.21	0.30	0.64
Dic	0	2	B	0.32	0.23	0.15	0.10	0.16	0.29	0.44
Sim	100	4	B	0.32	0.23	0.16	0.12	0.16	0.28	0.44
Dic	5	2	B	0.42	0.27	0.19	0.11	0.19	0.28	0.59
Sim	0	1	TF	0.43	0.28	0.18	0.11	0.19	0.30	0.59

The Decision Tree results exhibit a behavior similar to that observed for Random Forest. The classifier benefits from the synthetic instances generated through the proposed oversampling methodology, particularly for the minority classes. Across the evaluated configurations, the F1-score for Class 1 ranges from 0.15 to 0.22, while Class 2 reaches up to 0.14, representing a noticeable improvement over the original imbalanced scenario.

The best overall result was obtained using dictionary-based similarity with a temperature of 100 and a change factor of 2, achieving an accuracy of 0.46 and a Macro-F1 score of 0.30. Interestingly, both dictionary-based and embedding-based strategies produced very similar results, suggesting that the Decision Tree classifier is relatively insensitive to the source of the candidate replacement words. Instead, the overall class distribution and the presence of additional minority-class instances appear to have a greater impact on performance.

As expected, Class 5 remains the easiest category to classify, with F1-scores between 0.44 and 0.64. Nevertheless, the generated synthetic reviews contribute to a more balanced performance across all classes, which is reflected in the improvement of the Macro-F1 metric. Although the Decision Tree does not reach the performance levels achieved by more sophisticated classifiers, these results demonstrate that the proposed oversampling methodology can benefit even relatively simple learning algorithms.

The behavior observed for the Decision Tree classifier deserves further discussion. Unlike Transformer-based models or margin-based classifiers such as SVM, Decision Trees recursively partition the feature space by selecting features that maximize local impurity reduction. Consequently, their decision boundaries are strongly influenced by the empirical distribution of the training instances. Since the proposed methodology introduces synthetic reviews containing representative lexical patterns, local feature distributions may differ from those found in naturally occurring data. This characteristic can increase the sensitivity

of Decision Trees to artificial lexical regularities and partially explain the reduction in Macro-F1 observed for this classifier. In contrast, models based on distributed representations or global optimization criteria are generally more robust to these local distributional variations.

6.10.3. Naive Bayes

Table 12 presents the best results obtained using the Naive Bayes classifier. Unlike the previous classifiers, Naive Bayes consistently favored TF-IDF representations, with all best-performing configurations employing this weighting scheme. This behavior is expected since Naive Bayes relies heavily on term frequency distributions and often benefits from the discriminative weighting provided by TF-IDF.

Table 12. Best Naive Bayes results.

Similarity	Temp.	Factor	Rep.	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	100	4	TF	0.44	0.25	0.11	0.06	0.19	0.26	0.63
Sim	100	2	TF	0.49	0.27	0.12	0.10	0.18	0.27	0.68
Dic	100	2	TF	0.40	0.24	0.10	0.04	0.17	0.29	0.59
Sim	100	2	TF	0.42	0.25	0.10	0.06	0.17	0.29	0.61
Dic	100	2	TF	0.45	0.25	0.10	0.07	0.17	0.27	0.64
Sim	5	4	TF	0.41	0.24	0.10	0.05	0.18	0.26	0.60

The Naive Bayes classifier exhibited the lowest overall performance among the evaluated methods. Nevertheless, the proposed oversampling strategy produced measurable improvements compared with the original imbalanced dataset, particularly for the minority classes. While the classifier still struggled to distinguish between the most negative categories, several configurations achieved better recognition of Classes 1 and 2 than those observed in the baseline experiments.

The best overall result was obtained using embedding-based similarity with a temperature of 100 and a change factor of 2, achieving an accuracy of 0.49 and a Macro-F1 score of 0.27. This configuration also produced the highest F1-score for Class 2 (0.10), which remains the most difficult category for the classifier to identify correctly.

A notable characteristic of the results is the strong preference for high-temperature configurations. Most of the best-performing experiments employed a temperature value of 100, indicating that allowing access to the complete candidate word list introduces sufficient lexical diversity to benefit the probabilistic assumptions of Naive Bayes. Furthermore, embedding-based similarity generally outperformed dictionary-based substitutions, suggesting that semantically related words provide more informative synthetic instances than direct synonym replacement.

As observed with the other classifiers, Class 5 remained substantially easier to classify than the remaining categories, with F1-scores ranging from 0.59 to 0.68. Although the overall Macro-F1 values remain modest, the results demonstrate that the proposed oversampling methodology can improve the robustness of Naive Bayes by exposing the model to a more balanced distribution of sentiment categories.

6.10.4. Support Vector Machines

Table 13 presents the best results obtained using the Support Vector Machine (SVM) classifier. As observed in the baseline experiments, SVM remains the strongest traditional machine learning algorithm evaluated in this study. The classifier consistently achieved the highest Macro-F1 scores among all classical approaches and demonstrated a

remarkable ability to benefit from the synthetic instances generated through the proposed oversampling methodology.

The SVM classifier achieved some of the best results observed throughout the experimental evaluation. The highest Macro-F1 score reached 0.44, surpassing not only the undersampling-only experiments but also most of the other traditional machine learning models considered in this work. More importantly, the improvements were not limited to the majority class, as significant gains were observed for the minority categories.

Table 13. Best SVM results.

Similarity	Temp.	Factor	Rep.	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	5	4	TF	0.65	0.44	0.45	0.21	0.36	0.39	0.81
Sim	0	4	TF	0.65	0.43	0.44	0.20	0.35	0.37	0.81
Dic	5	4	TF	0.51	0.38	0.44	0.17	0.26	0.37	0.67
Sim	0	4	TF	0.48	0.37	0.43	0.17	0.26	0.36	0.61
Dic	5	2	TF	0.58	0.42	0.46	0.21	0.32	0.39	0.73
Sim	100	2	TF	0.60	0.42	0.44	0.20	0.33	0.39	0.76

One of the most notable findings is the substantial improvement obtained for Class 1. While the baseline experiments reported F1-scores close to 0.42 for this category, the oversampling strategy increased this value up to 0.46 in some configurations. Similarly, Class 2 improved from the baseline value of 0.16 to values as high as 0.21. These improvements suggest that the generated synthetic reviews successfully capture the linguistic characteristics of the most underrepresented sentiment classes.

Another interesting observation is the remarkable consistency of the TF-IDF representation. All best-performing configurations employed TF-IDF weighting, confirming that SVM benefits from emphasizing discriminative terms while reducing the influence of very frequent words. Furthermore, both dictionary-based and embedding-based similarity strategies produced competitive results, although dictionary-based substitutions achieved the highest overall Macro-F1 score.

The best configuration corresponded to dictionary-based similarity with a temperature of 5 and a change factor of 4, achieving an accuracy of 0.65 and a Macro-F1 score of 0.44. This result is particularly relevant because it slightly exceeds the Macro-F1 obtained by the original imbalanced baseline (0.42), demonstrating that the proposed oversampling methodology not only compensates for the loss of information introduced during undersampling but can also improve the overall ability of the classifier to discriminate between sentiment categories.

Finally, although Class 5 remains the easiest category to classify, with F1-scores reaching 0.81, the performance distribution across classes becomes considerably more balanced than in the original dataset. This behavior is reflected in the improved Macro-F1 values and confirms the effectiveness of the proposed synthetic review generation framework when combined with SVM classifiers.

6.10.5. K-Nearest Neighbors

Table 14 presents the best results obtained using the K-Nearest Neighbors (KNN) classifier. As discussed previously, the value $k = 5$ was selected empirically because it consistently produced the best results among the neighborhood sizes evaluated during the preliminary experiments. Similar to Naive Bayes and SVM, all best-performing configurations employed the TF-IDF representation, highlighting the importance of term weighting for distance-based classifiers.

The KNN classifier benefited considerably from the proposed oversampling strategy, achieving a substantial improvement compared with the baseline results obtained on the original imbalanced corpus. In particular, the Macro-F1 score increased from 0.23 in the baseline experiments to values as high as 0.32 after applying the synthetic data generation methodology. This represents one of the largest relative improvements observed among the traditional machine learning algorithms.

Table 14. Best KNN results.

Similarity	Temp.	Factor	Rep.	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	5	1	TF	0.50	0.32	0.25	0.14	0.24	0.31	0.68
Sim	100	4	TF	0.50	0.32	0.26	0.12	0.24	0.32	0.68
Dic	100	4	TF	0.43	0.29	0.24	0.12	0.19	0.34	0.59
Sim	0	1	TF	0.43	0.30	0.25	0.12	0.19	0.34	0.60
Dic	5	2	TF	0.46	0.31	0.26	0.14	0.21	0.32	0.63
Sim	0	2	TF	0.46	0.31	0.28	0.12	0.21	0.32	0.63

A particularly interesting result is the improvement obtained for the minority classes. The F1-score of Class 1 increased from 0.06 in the original dataset to values between 0.24 and 0.28 depending on the configuration. Similarly, Class 2 improved from 0.01 to values reaching 0.14. These gains indicate that KNN is highly sensitive to the class distribution and benefits substantially from the additional synthetic instances generated for the underrepresented categories.

Unlike the SVM experiments, no clear dominance of either dictionary-based or embedding-based similarity was observed. Both approaches produced highly competitive results, suggesting that the classifier is primarily influenced by the increased density of minority-class examples in the feature space rather than by the specific lexical substitution strategy employed during generation.

The best overall performance was achieved by two configurations that reached identical accuracy and Macro-F1 values. Dictionary-based similarity with a temperature of 5 and a change factor of 1 achieved an accuracy of 0.50 and a Macro-F1 score of 0.32, while embedding-based similarity with a temperature of 100 and a change factor of 4 produced virtually identical results. This observation suggests that KNN is relatively robust to the specific oversampling parameters as long as the generated reviews contribute additional representative examples for the minority classes.

Although the classifier continues to achieve its highest performance on Class 5, with F1-scores reaching 0.68, the distribution of performance across the five classes becomes considerably more balanced than in the original corpus. Consequently, the oversampling methodology proves effective not only for improving overall Macro-F1 but also for reducing the strong class bias initially exhibited by the KNN classifier.

6.10.6. Multilayer Perceptron

Table 15 presents the best results obtained using the Multilayer Perceptron (MLP) classifier. Similar to SVM and KNN, all best-performing configurations employed TF-IDF representations, indicating that neural models based on dense feature vectors also benefit from emphasizing discriminative terms through inverse document frequency weighting.

Table 15. Best Multilayer Perceptron results.

Similarity	Temp.	Factor	Rep.	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	0	4	TF	0.58	0.40	0.36	0.25	0.32	0.35	0.74
Sim	0	1	TF	0.58	0.41	0.40	0.22	0.33	0.35	0.74
Dic	100	4	TF	0.41	0.34	0.38	0.22	0.23	0.32	0.54
Sim	0	1	TF	0.39	0.33	0.40	0.20	0.21	0.32	0.52
Dic	100	1	TF	0.54	0.39	0.36	0.24	0.29	0.36	0.70
Sim	0	2	TF	0.53	0.40	0.41	0.25	0.29	0.34	0.69

The Multilayer Perceptron produced highly competitive results and emerged as one of the strongest traditional machine learning approaches evaluated in this study. The classifier consistently achieved Macro-F1 scores close to those obtained by SVM, demonstrating that neural architectures are capable of effectively exploiting the synthetic instances generated by the proposed oversampling methodology.

The best overall result was achieved using embedding-based similarity with a temperature of 0 and a change factor of 1, obtaining an accuracy of 0.58 and a Macro-F1 score of 0.41. This performance is comparable to the best SVM configurations and substantially exceeds the baseline Macro-F1 score of 0.37 obtained after the undersampling stage. These results indicate that the synthetic reviews successfully enrich the minority classes without significantly degrading the quality of the majority categories.

One of the most remarkable findings is the improvement obtained for the minority classes. Class 1 reached F1-scores as high as 0.41, while Class 2 achieved values up to 0.25. Compared with the original imbalanced baseline, where Class 2 was particularly difficult to classify, these results represent a substantial gain in minority-class recognition. Furthermore, the improvements are achieved without sacrificing performance on Class 5, which maintains F1-scores between 0.52 and 0.74 across the evaluated configurations.

Unlike Naive Bayes, which favored highly diverse generations, the MLP results suggest a preference for more conservative lexical substitutions. Several of the best-performing configurations employed a temperature of 0, meaning that only the most semantically similar candidate words were considered during the generation process. This observation suggests that preserving semantic consistency may be more beneficial for neural classifiers than introducing excessive lexical variation.

Another interesting pattern is the strong performance of embedding-based similarity. In two of the three evaluated corpora, embedding-based substitutions produced the highest Macro-F1 values, indicating that semantically related words extracted from the FastText space provide high-quality synthetic examples capable of improving the generalization ability of the neural network.

Overall, the results demonstrate that the proposed oversampling methodology is particularly effective when combined with MLP classifiers. The generated synthetic reviews lead to a more balanced performance across all sentiment categories and enable the classifier to achieve one of the highest Macro-F1 scores among the traditional machine learning approaches evaluated in this work.

Table 16 presents the best results obtained using BETO across the different experimental configurations. Overall, the results achieved with this transformer-based model are the highest among the evaluated approaches, which was expected given the strong contextual representation capabilities of BERT-based architectures and the fact that BETO was specifically pretrained on Spanish-language corpora. The best-performing configuration corresponds to the dictionary-based similarity strategy (Dic), with a temperature of 5 and a factor of 2, achieving an accuracy of 0.69 and a Macro F1 score of 0.57. This configuration

also provides comparatively strong performance across the individual sentiment classes, particularly for classes 1 and 5, with F1 scores of 0.64 and 0.81, respectively.

The superior performance of BETO is consistent with the expected behavior of pre-trained transformer architectures in text classification tasks. Unlike traditional models relying primarily on sparse or surface-level textual representations, BETO can capture contextual and semantic relationships between words, allowing it to better distinguish subtle linguistic patterns associated with different sentiment levels. Nevertheless, the differences observed among the BETO configurations indicate that the characteristics of the generated data and the generation parameters continue to influence classification performance even when using a more powerful language representation model. In particular, the results suggest that an appropriate combination of the similarity strategy, temperature, and balancing factor can further enhance the benefits provided by contextualized representations, especially in terms of Macro F1 and the performance of minority sentiment classes.

Table 16. Best BETO results.

Similarity	Temp.	Factor	Acc.	Macro F1	F1 1	F1 2	F1 3	F1 4	F1 5
Dic	5	2	0.69	0.57	0.64	0.45	0.49	0.49	0.81
Sim	100	1	0.68	0.55	0.61	0.43	0.48	0.48	0.81
Dic	0	2	0.66	0.54	0.59	0.41	0.46	0.46	0.76
Sim	0	4	0.65	0.53	0.58	0.40	0.45	0.45	0.76
Sim	0	1	0.63	0.50	0.56	0.38	0.43	0.42	0.72
Dic	0	2	0.62	0.49	0.55	0.36	0.42	0.41	0.70

When compared with the traditional machine-learning models evaluated in the previous experiments, including SVM, MLP, Random Forest, Decision Tree, KNN, and Naive Bayes, BETO consistently achieves the strongest overall performance. This improvement reinforces the advantage of contextualized language representations for sentiment classification, particularly in a challenging imbalanced multiclass scenario. At the same time, the fact that performance varies across generation configurations confirms that the proposed data-generation strategy remains relevant regardless of the classification architecture employed, suggesting that the quality and characteristics of the synthetic data can complement the representational capabilities of more sophisticated models such as BETO.

6.11. Discussion of Oversampling Results

To facilitate the analysis of the proposed oversampling methodology, Table 17 presents a global ranking of all experiments, including both the original baselines and the best configurations obtained after applying the proposed balancing strategy. The experiments are ordered according to the Macro-F1 score and, in case of ties, according to the F1-score of Class 1 (Very Negative), which corresponds to one of the most difficult and underrepresented classes in the dataset.

Several important observations emerge from the global ranking. First, the dominance of BETO is evident. The first four positions of the ranking correspond to BETO configurations generated through the proposed oversampling methodology. Furthermore, four of the six BETO configurations outperform the original baseline, demonstrating that the synthetic reviews generated by the proposed approach are capable of improving the performance of a state-of-the-art Transformer model.

A second important observation concerns the similarity strategy. Among the top ten configurations, both dictionary-based and embedding-based approaches appear repeatedly. Specifically, dictionary-based substitutions occupy the first, third, and seventh positions,

while embedding-based substitutions occupy the second, fourth, sixth, and ninth positions. These results suggest that neither strategy completely dominates the other. Instead, both mechanisms are capable of generating useful lexical variations that enrich the training corpus. Nevertheless, the best overall result is achieved using dictionary-based substitutions, indicating a slight advantage for lexical resources in this particular task.

Table 17. Global ranking of baseline and oversampling experiments ordered by Macro-F1 and F1-1.

Model	Sim.	T	F	Rep.	Acc.	Macro F1	F1-1	F1-2	F1-3	F1-4	F1-5
BETO	Dic	5	2	-	0.69	0.57	0.64	0.45	0.49	0.46	0.81
BETO	Sim	100	1	-	0.68	0.55	0.61	0.43	0.48	0.42	0.81
BETO	Dic	0	2	-	0.66	0.54	0.59	0.41	0.46	0.48	0.76
BETO	Sim	0	4	-	0.65	0.53	0.58	0.40	0.45	0.46	0.76
BETO (BL)	-	-	-	-	0.70	0.51	0.55	0.37	0.42	0.34	0.87
SVM	Dic	5	4	TF	0.65	0.44	0.45	0.21	0.36	0.37	0.81
SVM	Sim	0	4	TF	0.65	0.43	0.44	0.20	0.35	0.35	0.81
SVM	Dic	5	2	TF	0.58	0.42	0.46	0.21	0.32	0.38	0.73
SVM	Sim	100	2	TF	0.60	0.42	0.44	0.20	0.33	0.37	0.76
SVM (BL)	-	-	-	TF	0.72	0.42	0.42	0.16	0.33	0.33	0.86
MLP	Sim	0	1	TF	0.58	0.41	0.40	0.22	0.33	0.36	0.74
MLP	Sim	0	2	TF	0.53	0.40	0.41	0.25	0.29	0.36	0.69
MLP	Dic	0	4	TF	0.58	0.40	0.36	0.25	0.32	0.33	0.74
MLP	Dic	100	1	TF	0.54	0.39	0.36	0.24	0.29	0.36	0.70
RF	Sim	100	4	B	0.63	0.35	0.22	0.12	0.26	0.35	0.80
DT (BL)	-	-	-	TF	0.65	0.35	0.28	0.15	0.20	0.32	0.80
RF	Dic	0	4	B	0.63	0.34	0.21	0.12	0.26	0.31	0.80
KNN	Sim	100	4	TF	0.50	0.32	0.26	0.12	0.24	0.30	0.68
KNN	Dic	5	1	TF	0.50	0.32	0.25	0.14	0.24	0.29	0.68
KNN	Sim	0	2	TF	0.46	0.31	0.28	0.12	0.21	0.31	0.63
KNN	Dic	5	2	TF	0.46	0.31	0.26	0.14	0.21	0.31	0.63
DT	Dic	100	2	B	0.46	0.30	0.22	0.14	0.21	0.30	0.63
DT	Sim	0	1	B	0.47	0.29	0.21	0.11	0.21	0.28	0.64
NB	Sim	100	2	TF	0.49	0.27	0.12	0.10	0.18	0.27	0.68
NB	Dic	100	4	TF	0.44	0.25	0.11	0.06	0.19	0.26	0.63
NB	Sim	100	2	TF	0.42	0.25	0.10	0.06	0.17	0.31	0.61
NB	Dic	100	2	TF	0.45	0.25	0.10	0.07	0.17	0.27	0.64
KNN (BL)	-	-	-	TF	0.68	0.23	0.06	0.01	0.11	0.16	0.81
NB (BL)	-	-	-	TF	0.30	0.18	0.06	0.04	0.12	0.24	0.44
RF (BL)	-	-	-	TF	0.70	0.17	0.00	0.00	0.00	0.03	0.82

The analysis of the temperature parameter reveals an interesting pattern. Among the highest-ranked configurations, the values

$$T = 0$$

and

$$T = 5$$

appear most frequently. This finding suggests that selecting candidates among the most semantically similar alternatives produces more informative synthetic reviews than exploring the entire candidate space. Although some competitive configurations use

$$T = 100,$$

the very best results are consistently associated with lower temperature values. Consequently, moderate lexical variation appears to be preferable to highly diverse substitutions.

Although a high temperature value increases the diversity of candidate replacements, it does not imply unrestricted random lexical substitution. Candidate words are always selected from semantically related alternatives obtained either from the synonym dictionary or from the nearest neighbors in the FastText embedding space. Consequently, even when $T = 100$, the replacement process remains constrained to words that preserve a substantial portion of the original semantic content. Moreover, each synthetic review is generated from multiple representative words, allowing the language model to reconstruct a coherent review based on several semantic anchors rather than on individual lexical choices. This combination of semantically constrained substitutions and contextual generation likely explains why classifiers such as SVM and MLP maintain competitive performance even under the highest temperature setting.

A similar trend can be observed for the change factor. The value

$$F = 2$$

is the most frequent among the highest-performing configurations, followed by

$$F = 4.$$

This behavior indicates that modifying a moderate number of representative words is sufficient to generate diverse synthetic reviews while preserving the semantic characteristics associated with each sentiment category.

Regarding the classifiers themselves, BETO exhibits the strongest overall performance, followed by SVM and MLP. Interestingly, the ranking shows that all BETO configurations outperform the best SVM configuration, highlighting the benefits of contextualized language representations for sentiment analysis. Among traditional machine learning methods, SVM remains the strongest alternative, while Random Forest and KNN benefit the most from the balancing process.

To quantify the impact of the proposed methodology, Table 18 compares the best result obtained for each classifier against its corresponding baseline. The percentage improvement is computed as

$$\Delta(x) = \frac{x_{\text{best}} - x_{\text{baseline}}}{x_{\text{baseline}}} \times 100.$$

The improvements reported in Table 18 provide additional evidence regarding the effectiveness of the proposed methodology. BETO obtains the highest overall performance, improving Macro-F1 by approximately 12% while simultaneously increasing the F1-scores of the minority classes by more than 16%. These gains demonstrate that the synthetic reviews successfully enrich the representation of underrepresented sentiment categories without substantially degrading the performance of the majority class.

Among the traditional machine learning algorithms, SVM exhibits the strongest balanced performance. Although the overall Macro-F1 improvement is relatively modest, the F1-score of Class 2 increases by more than 30%, indicating that the generated reviews provide valuable information for the most challenging classes.

Table 18. Percentage improvement of the best oversampling configuration over the baseline.

Model	Acc.	Macro F1	F1-1	F1-2	F1-3	F1-4	F1-5
BETO	−1.43%	+11.76%	+16.36%	+21.62%	+16.67%	+27.78%	−6.90%
SVM	−9.72%	+4.76%	+7.14%	+31.25%	+9.09%	+12.12%	−5.81%
RF	−10.00%	+105.88%	—	—	—	—	−2.44%
DT	−27.69%	−14.29%	−21.43%	−6.67%	+5.00%	−3.23%	−21.25%
KNN	−26.47%	+39.13%	+333.33%	+1300.00%	+118.18%	+106.67%	−16.05%
NB	+63.33%	+50.00%	+100.00%	+150.00%	+58.33%	+12.50%	+54.55%

The most dramatic relative improvements are observed for Random Forest and KNN. Random Forest more than doubles its Macro-F1 score, increasing from 0.17 to 0.35. This result is particularly remarkable because the original classifier completely failed to recognize several minority classes. Similarly, KNN achieves improvements exceeding 300% for Class 1 and more than 1000% for Class 2, clearly demonstrating the impact of the balancing strategy.

Naive Bayes also benefits considerably from the generated reviews, obtaining improvements across all reported metrics. In contrast, Decision Trees experience a slight decrease in overall performance, suggesting that this classifier may be more sensitive to the lexical variability introduced by the synthetic instances.

Overall, the results demonstrate that the proposed oversampling methodology successfully improves the representation of minority sentiment classes while remaining fully compatible with both traditional machine learning algorithms and modern Transformer architectures. Unlike conventional oversampling techniques that generate synthetic vectors, the proposed approach produces complete textual instances, allowing it to be directly applied to state-of-the-art language models such as BETO. This characteristic constitutes one of the main advantages of the methodology and explains the strong performance obtained throughout the experiments.

6.12. Comparison Against Direct Prompt-Based Generation

One of the main questions arising from the proposed methodology is whether the complete process of representative word extraction, lexical substitution, and controlled text generation is actually necessary. An alternative and considerably simpler strategy would consist of directly prompting a Large Language Model (LLM) to generate synthetic reviews belonging to the minority classes without providing any information extracted from the original corpus.

To answer this question, an additional experiment was conducted in which synthetic reviews were generated directly using three popular LLMs: GPT-4o-mini [26], Gemini 2.5 Flash [27], and Llama 3.3 70B Instruct [28]. These models were selected because they represent three different families of state-of-the-art language models and can generate large quantities of synthetic text at relatively low cost.

Unlike the proposed methodology, the direct prompting approach does not use representative words, lexical substitution strategies, temperatures, or change factors. Instead, the model receives only a high-level instruction describing the desired sentiment category and is asked to generate reviews that resemble real tourist opinions.

The prompt used in the experiments is shown below.

You are a tourist who has visited hotels, restaurants, and attractions in Mexico. Generate a realistic review written in natural language. The review must express a sentiment corresponding to the category {Very Negative, Negative, Neutral, Positive, Very Positive}.

The review should resemble authentic opinions published on travel websites such as TripAdvisor. Avoid repetitive phrases, avoid mentioning that the text is synthetic, and write between 50 and 120 words. Use realistic details regarding service, facilities, staff, food, cleanliness, location, atmosphere, or tourist experiences.

Using this prompt, the same number of synthetic reviews generated by the proposed methodology were produced for each minority class. Subsequently, the resulting balanced corpora were evaluated using the same classification pipeline described throughout this work.

To ensure a fair comparison between the proposed methodology and the direct prompting baseline, both approaches were evaluated under exactly the same experimental conditions. In particular, the same language model (GPT-4o-mini), prompt language (Spanish), number of generated reviews, generation budget, default API configuration, automatic validation procedure, and downstream classification pipeline were employed. Consequently, the only experimental difference between both methods was the incorporation of representative words into the generation process, allowing the impact of the proposed guidance mechanism to be evaluated in isolation.

Table 19 presents the best results obtained for each classifier using the proposed methodology and compares them against direct prompt-based generation with GPT-4o-mini, Gemini 2.5 Flash, and Llama 3.3 70B.

Table 19. Comparison between the proposed methodology and direct prompt-based generation.

Classifier	Proposed	GPT-4o-mini	Gemini 2.5 Flash	Llama 3.3 70B
BETO	0.57	0.52	0.51	0.49
SVM	0.44	0.40	0.39	0.37
MLP	0.41	0.37	0.36	0.34
RF	0.35	0.28	0.27	0.26
KNN	0.32	0.27	0.26	0.24
DT	0.30	0.25	0.24	0.23
NB	0.27	0.22	0.21	0.20

The results show that the proposed methodology consistently outperforms direct prompt (The best results are highlighted in bold)-based generation across all evaluated classifiers. Although modern LLMs are capable of generating grammatically correct and semantically coherent reviews, these reviews often fail to reproduce the lexical and thematic characteristics present in the original dataset.

The best direct prompting results are obtained using GPT-4o-mini, which is not surprising considering that this model was also employed in the proposed methodology. However, even in this favorable scenario, the resulting classifiers systematically underperform compared with the proposed approach. For example, BETO achieves a Macro-F1 score of 0.57 using the proposed methodology but only 0.52 when GPT-4o-mini is instructed to generate reviews directly from a sentiment label. Similar differences can be observed for SVM, MLP, Random Forest, KNN, Decision Trees, and Naive Bayes.

These results suggest that the primary advantage of the proposed methodology does not originate from the LLM itself, but rather from the guidance mechanism provided by the representative words extracted from the original corpus. By forcing the generated reviews to incorporate vocabulary that is statistically representative of each sentiment category, the resulting synthetic instances remain closer to the lexical and semantic distributions observed in real reviews.

In particular, the representative words capture domain-specific concepts that frequently appear in tourist reviews, including references to rooms, staff, reception, cleanliness, breakfast, views, restaurants, facilities, and other tourism-related concepts. When these terms are preserved during generation, the synthetic reviews become much more consistent with the original corpus. Consequently, classifiers trained on these reviews learn patterns that are directly transferable to the test set.

This phenomenon is especially important for Transformer-based architectures. Models such as BETO learn contextual representations that depend strongly on the co-occurrence patterns present in the training corpus. By guiding the generation process through representative words and controlled lexical substitutions, the proposed methodology produces synthetic reviews that preserve many of the semantic relationships present in the original dataset. In contrast, direct prompting tends to generate generic reviews that, although coherent, may not accurately reflect the lexical distribution of the target domain.

Another important observation is that the proposed methodology produces improvements regardless of the LLM used during generation. The representative-word strategy acts as a form of semantic regularization, reducing the variability introduced by different language models and forcing them to generate reviews that remain close to the original domain. As a result, the methodology is less dependent on the particular LLM employed and more dependent on the quality of the representative words extracted from the corpus.

Overall, the experiments demonstrate that simply asking an LLM to generate reviews is insufficient to obtain the best classification performance. The extraction of representative words, the use of controlled lexical substitutions, and the generation of reviews guided by domain-specific vocabulary significantly improve the quality of the synthetic instances. Therefore, the proposed methodology constitutes a more effective strategy for balancing sentiment datasets than direct prompt-based text generation alone.

6.13. Limitations

The primary objective of the proposed methodology is to improve sentiment classification performance through the generation of synthetic training instances, particularly for underrepresented sentiment classes. Consequently, this work evaluates the usefulness of the generated reviews according to their impact on the downstream classification task rather than their intrinsic linguistic quality.

Although characteristics such as realism, lexical diversity, semantic consistency, distributional similarity, duplication rate, and human-perceived naturalness are important when evaluating text generation systems, they fall outside the scope of the present study. A synthetic review does not necessarily need to be indistinguishable from a human-written review to provide useful information during the fine-tuning process. Instead, its value is determined by whether it helps the classifier learn more discriminative decision boundaries and improves the recognition of minority classes.

Future work will therefore complement the present evaluation by incorporating intrinsic quality metrics, distributional similarity analyses, and human assessments of the generated reviews. Such analyses will provide a more comprehensive understanding of the relationship between synthetic text quality and downstream classification performance.

Another limitation of the present study is that the reported results correspond to a single experimental execution under a fixed random seed for all stochastic processes implemented within our methodology. Although this experimental protocol enables a controlled comparison between the evaluated approaches, it does not quantify the variability that may arise from repeated executions or from the stochastic behavior of large language models. Future work will therefore include multiple independent runs, confidence intervals, statis-

tical significance tests, and variance analyses to provide a more comprehensive assessment of the robustness of the proposed methodology.

Although the proposed methodology shows consistent improvements over direct LLM prompting, the experimental evaluation was not intended to provide an exhaustive comparison with every existing strategy for handling class imbalance or text data augmentation. Future work will therefore include a systematic comparison against additional approaches, including class-weighted learning, focal loss, balanced batch sampling, random oversampling, back-translation, contextual data augmentation, Easy Data Augmentation (EDA), and recent sentence-transformer-based augmentation techniques for Spanish. Such comparisons will provide a more comprehensive understanding of the relative advantages and limitations of the proposed representative-word-guided generation framework.

7. Conclusions and Future Work

This work presented a hybrid balancing methodology for sentiment analysis in tourism reviews written in Mexican Spanish. The proposed framework combines majority-class undersampling with minority-class oversampling through the generation of synthetic reviews guided by representative words and Large Language Models. The motivation behind this work stems from one of the most common challenges in real-world tourism datasets: severe class imbalance. In such datasets, positive opinions are abundant, whereas negative reviews—which often contain the most valuable information regarding service deficiencies, customer dissatisfaction, and operational issues—are considerably underrepresented. Consequently, classification models tend to favor majority classes and struggle to correctly identify the minority categories that are often of greatest interest to tourism stakeholders.

The experimental results demonstrate that the combination of undersampling and oversampling constitutes an effective strategy for addressing this problem. The proposed methodology improved the performance of all evaluated classifiers with the exception of Decision Trees. The most significant results were obtained with BETO, which achieved a Macro-F1 score of 0.57, outperforming its baseline result of 0.51. Improvements were also observed for SVM, MLP, Random Forest, KNN, and Naive Bayes, particularly in the minority sentiment classes corresponding to negative opinions.

An important finding of this study concerns the undersampling stage. Three different strategies were evaluated: random selection, K-means-based selection, and centroid-based selection. The results consistently showed that random undersampling achieved the best overall performance while simultaneously being the simplest and computationally cheapest alternative. This finding is particularly relevant because it indicates that more sophisticated selection strategies do not necessarily provide additional benefits in this context. Consequently, all subsequent oversampling experiments were performed using random undersampling.

Regarding oversampling, the experiments showed that generating synthetic reviews guided by representative words extracted from the original corpus is an effective mechanism for enriching minority classes. The best-performing configurations were generally associated with moderate lexical modifications, particularly using a change factor of two and low-to-moderate temperature values. Both dictionary-based and embedding-based similarity strategies produced competitive results, although dictionary-based substitutions achieved the best overall performance.

The comparison against direct prompt-based generation using GPT-4o-mini, Gemini, and Llama further demonstrated the value of the proposed methodology. In all cases, the guided generation approach outperformed the direct prompting strategy. These results suggest that incorporating representative words extracted from the original corpus allows the generated reviews to better preserve the lexical, semantic, and thematic characteristics of

real tourism reviews written in Mexican Spanish. As a consequence, the resulting synthetic instances are more useful for training classification models than reviews generated solely from high-level instructions.

The findings of this work provide strong evidence supporting the research hypothesis. Specifically, the results suggest that synthetic tourist reviews generated from representative words extracted from minority sentiment classes preserve important lexical and semantic patterns of real reviews and contribute to improved sentiment classification performance. Furthermore, the experiments indicate that this effect is particularly beneficial for under-represented categories, which are precisely the classes of greatest practical importance in tourism analysis.

The research questions proposed at the beginning of this study can now be answered as follows:

- **RQ1:** How can representative words be extracted from minority sentiment classes to generate synthetic reviews that effectively improve sentiment classification performance? Representative words extracted through Mutual Information successfully capture the most discriminative lexical patterns associated with each sentiment class. These words provide effective guidance for generating synthetic reviews that remain consistent with the vocabulary and semantic characteristics of the original tourism corpus.
- **RQ2:** Which configuration of the proposed oversampling methodology produces the best classification performance, and how does its effectiveness vary across different classifiers?

The best results were obtained using random undersampling, dictionary-based similarity, moderate lexical replacement factors, and low-to-moderate temperature values. BETO achieved the highest overall performance, followed by SVM and MLP. Decision Trees were the only classifier that did not benefit from the proposed balancing strategy.

- **RQ3:** Is a guided synthetic review generation strategy based on representative words more effective than directly prompting a Large Language Model to generate synthetic reviews without additional domain-specific information?

Yes. The experiments consistently showed that the proposed methodology outperformed direct prompt-based generation using GPT-4o-mini, Gemini, and Llama. Guiding the generation process with representative words extracted from the original dataset produces synthetic reviews that better preserve domain-specific lexical and semantic patterns, leading to superior classification performance.

Overall, the results indicate that balancing a corpus through the joint application of undersampling and guided oversampling produces better results than relying on either strategy independently. The proposed methodology not only improves the representation of minority classes but also remains fully compatible with both traditional machine learning algorithms and modern Transformer architectures. Because it generates complete textual instances rather than synthetic feature vectors, the approach can be directly applied to state-of-the-art language models without requiring modifications to their architectures.

This work showed that guided synthetic review generation constitutes an effective strategy for balancing tourism sentiment datasets in Mexican Spanish. By preserving domain-specific lexical and semantic information during the generation process, the proposed methodology improves the recognition of minority sentiment classes and provides a practical solution for developing more robust sentiment analysis systems for tourism intelligence and decision support applications in Mexico.

As future work, several research directions remain open. First, alternative techniques for extracting representative words could be explored, including attention-based methods derived from Transformer models or explainability approaches such as SHAP and Integrated Gradients. Second, future studies could investigate adaptive mechanisms for

automatically selecting temperatures and lexical replacement factors. Third, larger and more advanced language models could be evaluated as generators. Fourth, the methodology should be validated on additional tourism datasets from other Spanish-speaking countries in order to determine the extent to which the learned lexical and semantic patterns are specific to Mexican Spanish or generalize across other regional varieties of Spanish. Finally, future work could explore multilingual extensions and applications to other text classification problems affected by severe class imbalance.

Author Contributions: Conceptualization, A.D.-P., R.A. and M.Á.Á.-C.; Methodology, A.D.-P., R.A. and M.Á.Á.-C.; Software, A.B.G.-G. and M.Á.Á.-C.; Validation, A.D.-P., A.B.G.-G. and M.Á.Á.-C.; Formal analysis, A.Y.R.-G., R.A. and M.Á.Á.-C.; Investigation, A.D.-P., A.B.G.-G., A.Y.R.-G., R.A. and M.Á.Á.-C.; Resources, R.A. and M.Á.Á.-C.; Data curation, M.Á.Á.-C.; Writing—original draft, M.Á.Á.-C.; Writing—review & editing, A.Y.R.-G., R.A. and M.Á.Á.-C.; Visualization, M.Á.Á.-C.; Supervision, R.A. and M.Á.Á.-C.; Project administration, M.Á.Á.-C.; Funding acquisition, M.Á.Á.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico, through the Humanistic Research Project H-2025-G-283, entitled “Procesamiento de lenguaje natural adaptativo con enfoque humanístico para apoyo comunicacional y diagnóstico en pacientes con enfermedades neurodegenerativas”. The authors gratefully acknowledge the financial support provided by this project, which made this research possible.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://drive.google.com/file/d/1_fmVZK2FncWL0DdVw2jWQktGKBvelkB3/view?usp=sharing (accessed on 13 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Albuquerque, E.; Azevedo, D.; Fernandes, J.; Pataco, T. Sentiment Analysis in Guest Reviews: An NLP Approach Applied to Portuguese Hospitality. In Proceedings of the ICTR 2026, Lisbon, Portugal, 16–17 April 2026.
2. Liu, S.; Yin, P.; Yang, X.; Wang, L. A Study on Perceptual Differences in Tourist User-Generated Content Based on Multimodal Deep Learning. *J. Travel Res.* **2026**, 00472875261430746. [CrossRef]
3. Wongsasensee, P.; Onpium, P.; Kuensaen, C.; Muangyai, N. Advanced Analytics in Social Media Data Mining as a Driver of Digital Transformation in Cultural Heritage Tourism: The Case of Lamphun, Thailand. *Tour. Hosp.* **2026**, *7*, 186. [CrossRef]
4. Naveen Kumar, R.; Janani, M.; Pavithra, R.; John William, A. The role of TIAS in shaping smart tourism: Governance perspectives on choosing intelligent destinations. *Bus. Anal. J.* **2025**, *46*, 53–69. [CrossRef]
5. Asamoah, J.Y.S. Evaluating Key Determinants of Customer Satisfaction: A Focus on Customer Relations, Service Quality, Product Quality, and Supply Chain Management. Master’s Thesis, Kajaani University of Applied Sciences, Kajaani, Finland, 2025.
6. Puh, K.; Bagić Babac, M. Predicting sentiment and rating of tourist reviews using machine learning. *J. Hosp. Tour. Insights* **2023**, *6*, 1188–1204.
7. Álvarez-Carmona, M.Á.; Díaz-Pacheco, Á.; Aranda, R.; Rodríguez-González, A.Y.; Bustio-Martínez, L.; Herrera-Semenets, V. Overview of Rest-Mex at IberLEF 2025: Researching Sentiment Evaluation in Text for Mexican Magical Towns. *Proces. Leng. Nat.* **2025**, *75*, 499–511.
8. Guerrero-Rodríguez, R.; Álvarez-Carmona, M.Á.; Aranda, R.; López-Monroy, A.P. Studying online travel reviews related to tourist attractions using nlp methods: The case of guanajuato, mexico. *Curr. Issues Tour.* **2023**, *26*, 289–304. [CrossRef]
9. Carvalho, M.; Pinho, A.J.; Brás, S. Resampling approaches to handle class imbalance: A review from a data perspective. *J. Big Data* **2025**, *12*, 71. [CrossRef]
10. Gosain, A.; Sardana, S. Handling class imbalance problem using oversampling techniques: A review. In *Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*; IEEE: Piscataway, NJ, USA, 2017; pp. 79–85.

11. Wang, A.X.; Le, V.T.; Trung, H.N.; Nguyen, B.P. Addressing imbalance in health data: Synthetic minority oversampling using deep learning. *Comput. Biol. Med.* **2025**, *188*, 109830. [[CrossRef](#)] [[PubMed](#)]
12. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [[CrossRef](#)]
13. Álvarez-Carmona, M.Á.; Reyes-Sierra, M.; Díaz-Pacheco, Á.; Aranda, R.; Rodríguez-González, A.Y.; Bustio-Martínez, L.; Herrera-Semenets, V. Overview of Rest-Mex at IberLEF 2026: Research on Synthetic Text Modeling for Mexican Magical Towns. *Proces. Leng. Nat.* **2026**, *77*, 499–511.
14. Taskiran, S.F.; Turkoglu, B.; Kaya, E.; Asuroglu, T. A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Sci. Rep.* **2025**, *15*, 21631. [[CrossRef](#)] [[PubMed](#)]
15. Khalid, B.; Dai, S.; Taghavi, T.; Lee, S. Label Supervised Contrastive Learning for Imbalanced Text Classification in Euclidean and Hyperbolic Embedding Spaces. In Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024), San Ġiljan, Malta, 22 March 2024; pp. 58–67. [[CrossRef](#)]
16. Mendieta-Piedra, J.L.; León-Paredes, G.A. Combining Synthetic Minority Over-Sampling Technique and Multinomial Naive Bayes for Sentiment Analysis on Imbalanced Social Media Datasets. In *Information and Communication Technologies. TICEC 2025*; Springer Nature: Cham, Switzerland, 2025; pp. 3–17. [[CrossRef](#)]
17. Rana, S.; Kanji, R.; Jain, S. Comprehensive Analysis of Oversampling Techniques for Addressing Class Imbalance Employing Machine Learning Models. *Recent Adv. Comput. Sci. Commun.* **2026**, *19*, E26662558347788. [[CrossRef](#)]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
19. Mahmoudi, L.; Salem, M. BalBERT: A New Approach to Improving Dataset Balancing for Text Classification. *Rev. d'Intell. Artif.* **2023**, *37*, 425. [[CrossRef](#)]
20. Neveditsin, N.; Lingras, P.; Mago, V.K. From Annotation to Adaptation: Metrics, Synthetic Data, and Aspect Extraction for Aspect-Based Sentiment Analysis with Large Language Models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Albuquerque, NM, USA, 29 April–4 May 2025; Volume 4, pp. 142–161. [[CrossRef](#)]
21. Ningsih, F.S.S.; Khotimah, P.H.; Arisal, A.; Rozie, A.F.; Munandar, D.; Riswantini, D.; Nugraheni, E.; Suwarningsih, W.; Kurniasari, D. Synonym-based Text Generation in Restructuring Imbalanced Dataset for Deep Learning Models. In Proceedings of the 2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g/6G-based Interconnected Digital Worlds (NISS), Online, 30–31 March 2022; pp. 1–6. [[CrossRef](#)]
22. Siagh, A.; Laallam, F.Z.; Kazar, O.; Salem, H.; Benglia, M.E. IDA: An Imbalanced Data Augmentation for Text Classification. In *Proceedings of the International Conference on Intelligent Systems and Pattern Recognition*; Springer: Cham, Switzerland, 2023; pp. 241–251. [[CrossRef](#)]
23. Hellwig, N.; Fehle, J.; Wolff, C. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Syst. Appl.* **2025**, *261*, 125514. [[CrossRef](#)]
24. Maji, S.; Bhattacharya, A. BhashaSetu: Cross-Lingual Knowledge Transfer from High-Resource to Extreme Low-Resource Languages. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Mumbai, India, 20–24 December 2025; pp. 2111–2129.
25. Over, L.; Apfeld, S.; Schlangen, I.; Jovanoska, S. Word Embeddings for Radar Emissions: A Comparison Between Word2vec and fastText. In *Proceedings of the 2025 IEEE International Radar Conference (RADAR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
26. Rettberg, J.W.; Wigers, H. AI-generated stories favour stability over change: Homogeneity and cultural stereotyping in narratives generated by gpt-4o-mini. *arXiv* **2025**, arXiv:2507.22445.
27. Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv* **2025**, arXiv:2507.06261.
28. Wihl, J.; Rosenkranz, E.; Schramm, S.; Berberich, C.; Griessmair, M.; Woźnicki, P.; Pinto, F.; Ziegelmayer, S.; Adams, L.C.; Bressemer, K.K.; et al. Data extraction from free-text stroke CT reports using GPT-4o and Llama-3.3-70B: The impact of annotation guidelines. *Eur. Radiol. Exp.* **2025**, *9*, 61. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.