PMOTE-UC-CUJAE at Rest-Mex 2025: Evaluation of Probabilistic Data Augmentation Models for Mining Mexican Tourist Reviews

Ireimis Leguen-de-Varona[1,*,†], Julio Madera[1,†], Alfredo Simon-Cuevas[2,†], Leonardo Lastre Figueroa[1,†] and Yoan Martínez-López [3,4,†]

[1]University of Camagüey Ignacio Agramonte Loynaz, Camagüey, Cuba
[4]Technological University of Havana José Antonio Echeverría, CUJAE, Cuba
[3]University of Cordoba, Cordoba, Spain
[4]Plénitas, C/ Le Corbusier, 14005 Córdoba, Spain

Abstract

This paper presents the contribution of the PMOTE-UC-CUJAE team to the 2025 edition of the Rest-Mex shared task, focused on sentiment polarity classification in Spanish-language tourist reviews related to Mexico's Magical Towns. We propose a strategy based on probabilistic data augmentation techniques, employing covariance matrix estimation via the Ledoit–Wolf method, alongside Lasso regression and Elastic Net, applied to CLS embeddings generated using the RoBERTa-base-bne model.The resulting balanced datasets were used to train lightweight multilayer perceptron (MLP) classifiers, avoiding the need for computationally intensive transformer fine-tuning. Despite technical limitations that restricted our participation to a single subtask, the results demonstrate significant improvements over the baseline in key metrics such as Macro F1 for polarity. Notably, our models achieved balanced performance across all sentiment classes, including those with fewer examples, confirming the effectiveness of probabilistic oversampling in imbalanced contexts.These findings highlight the potential of probabilistic data augmentation methods for multilingual sentiment analysis tasks and reinforce the feasibility of efficient, transformer-free solutions in resource-constrained environments.

Keywords

Sentiment Analysis, Class Imbalance, Probabilistic Data Augmentation, RoBERTa, Ledoit-Wolf, Elastic Net, Lasso Regression, Mexican Magical Towns

1. Introduction

One of the most significant challenges in supervised classification problems is class imbalance, which occurs when one or more classes are represented by considerably fewer examples compared to others. This situation is common in real-world applications such as fraud detection, medical diagnosis, fault monitoring, risk analysis, and sentiment analysis—where minority classes are often the most critical. When models are trained on imbalanced data, they tend to optimize overall accuracy at the expense of properly detecting minority classes, thus compromising the practical usefulness of the system.

The field of sentiment analysis is not exempt from this phenomenon. Despite recent advances in transformer-based techniques, there remains a need for methods capable of generating semantically coherent synthetic data in high-dimensional spaces to effectively address imbalance. Classical oversampling methods, such as SMOTE [1] and its variants (SMOTE-Tomek Links [2], Borderline-SMOTE, [3],SPIDER [4], SMOTE–RSB* [5] ADASYN [6], among others), exhibit limitations when applied to dense representations like language embeddings, as they were originally designed for low-dimensional feature spaces.

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
ireimis.leguen@reduc.edu.cu (I. Leguen-de-Varona); julio.madera@reduc.edu.cu (J. Madera); asimon@ceis.cujae.edu.cu (A. Simon-Cuevas); leonardo.lastre@reduc.edu.cu (L. L. Figueroa); yoan.martinez@plenitas.com (Y. M.)
0000-0002-1886-7644 (I. Leguen-de-Varona); 0000-0001-5551-690X (J. Madera); 0000-0002-6776-9434 (A. Simon-Cuevas); 0009-0009-1526-0108 (L. L. Figueroa); 0000-0002-1950-567X (Y. M.)

Most of these approaches rely on interpolation between k-nearest neighbors to generate new samples. While this is generally effective in low-dimensional domains, their performance tends to degrade when dealing with high-dimensional embeddings. To overcome this limitation, a new oversampling strategy was proposed in 2024 based on shrinkage estimation of the covariance matrix using the Ledoit–Wolf method—known as SMOTE-COV HD [7]. This technique demonstrated promising results in sentiment classification under high-dimensional imbalance.

Probabilistic approaches have thus emerged as competitive alternatives, modeling the statistical distribution of minority classes based on their embeddings. This work is situated within Subtask 1 of the Rest-Mex 2025 competition [8], [9], unlike to others editions [10, 11, 12], aims to automatically classify the sentiment polarity of Spanish-language tourist reviews, specifically those related to Mexico's Magical Towns.

In this context, we propose a strategy based on probabilistic data augmentation using covariance matrix estimation via the Ledoit–Wolf method, as well as Lasso regression and Elastic Net. Our approach extracts CLS vector representations from the RoBERTa-base-bne model and generates new synthetic instances for the minority classes from multivariate distributions. These balanced representations are then classified using a lightweight multilayer perceptron (MLP), which avoids the need to fine-tune large transformer models, thereby reducing computational costs and offering a viable solution for resource-constrained environments.

2. Methodology

The proposed approach follows a systematic sequence of steps to perform sentiment polarity classification on tourist reviews using probabilistic data augmentation and lightweight models. The process is detailed as follows:

1. **Dataset Preparation:** The dataset is examined to identify the relevant columns for each task, particularly the text column and the label column indicating the sentiment polarity (ranging from 1 to 5).
2. **Text Vectorization:** Each review is tokenized using the pre-trained model `roberta-base-uncased`. The `[CLS]` token embedding (a 768-dimensional vector) is extracted for each review. The label column is mapped to numerical indices for classification purposes.
3. **Data Augmentation:** If class imbalance is detected, one of three probabilistic oversampling techniques is applied to generate synthetic examples for the minority classes. The three variants are:
    - **Ledoit–Wolf (LW) shrinkage:** The covariance matrix is estimated with shrinkage to reduce estimation error [13]:
    $$\Sigma_{\text{LW}} = (1 - \delta)S + \delta F \tag{1}$$
    where S is the sample covariance matrix, F is a structured target matrix (such as a scaled identity matrix), and δ is a shrinkage parameter.
    - **Lasso Regression:** The synthetic instances are generated based on regression coefficients obtained by minimizing [14]:
    $$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2}$$
    Where:
    y_i observed value of the dependent variable
    x_{ij} value of predictor j for observation i
    β_j model coefficients
    λ regularization parameter that controls the penalty

- **Elastic Net:** Combines Lasso and Ridge penalties to handle correlated features [15]:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3}$$

Where:

y_i observed value of the dependent variable
x_{ij} value of predictor j for observation i
β_j model coefficients
λ_1 Lasso (L1) regularization parameter
λ_2 Ridge (L2) regularization parameter

4. **Classifier Training:** A lightweight Multilayer Perceptron (MLP) is trained on the balanced dataset. The data is split into 80% for training and 20% for testing. During training, the loss and accuracy are monitored, and Macro F1-scores are calculated on the test set.
5. **Model Persistence:** After training, the MLP weight dictionary and a JSON file mapping the class labels to indices are saved for later use.
6. **Deployment:** These files are then used to classify new review data, replicating the tokenization and embedding pipeline and predicting sentiment classes.

This pipeline allows efficient classification without transformer fine-tuning while leveraging semantically coherent synthetic samples to improve the performance of the model under severe class imbalance.

3. Results

A detailed comparison between our three submitted systems (labeled as HM) and the reference baseline system (labeled as BL) for Subtask 1 of Rest-Mex 2025 is shown in Table 1. This subtask focuses on sentiment polarity classification of tourist reviews. The table reports the overall performance in terms of Macro F1 and Accuracy, along with F1 scores for each sentiment class (from 1 to 5), providing a clearer picture of how each method handles class imbalance.

Run	Place	Method	Macro F1	Accuracy	F1_C1	F1_C2	F1_C3	F1_C4	F1_C5
PMOTE-UC-CUJAE_1	HM	Ledoit-Wolf	**0.4087**	61.41%	**0.3987**	**0.3782**	0.3228	0.3551	0.5801
PMOTE-UC-CUJAE_2	HM	Lasso	0.3783	61.57%	0.3413	0.3464	**0.3275**	0.3379	0.5778
PMOTE-UC-CUJAE_3	HM	ElasticNet	0.3708	**61.73%**	0.2836	0.3261	0.3195	**0.3852**	**0.5951**
Baseline	BL	Baseline	0.1584	65.54%	0.0000	0.0000	0.0000	0.0000	0.7968

Table 1
Performance comparison between our proposals and the baseline system.

Overall, the model **PMOTE-UC-CUJAE_1**, based on Ledoit-Wolf covariance shrinkage, achieved the best global performance with a **Macro F1 of 0.4087** and an **Accuracy of 61.41%**, outperforming the baseline system, which only reached a Macro F1 of 0.1584 and Accuracy of 65.54%. Although the baseline scored higher in accuracy, this value is misleading in imbalanced scenarios, as its performance focused almost entirely on the majority class (F1 = 0.7968 for class 5), with zero scores for the remaining classes (F1 = 0 for classes 1–4).

In contrast, all three of our models exhibited **non-zero performance across all classes**, demonstrating a **superior ability to generalize** and handle class imbalance. Specifically:

- **Ledoit-Wolf** showed the highest F1 scores for classes 1 (0.3987) and 2 (0.3782), achieving the most balanced performance.

- **Elastic Net** performed well in classes 2 (0.3464) and 3 (0.3275), though with slightly lower F1 for class 1.
- **Lasso regression** also maintained consistent F1s above 0.28 for minority classes, highlighting its reliability in similar tasks.

These results confirm that the three probabilistic data augmentation methods allowed the MLP classifiers to learn more representative patterns across all sentiment categories, particularly for under-represented ones. Unlike the reference system, our models achieved consistent results without relying solely on the dominant class.

4. Conclusions and Future Work

This work presents a probabilistic data augmentation framework designed to address polarity classification in highly imbalanced contexts, such as the Rest-Mex 2025 task. By using regularized covariance estimators (Ledoit–Wolf), Lasso regression, and Elastic Net, we generated synthetic examples for minority classes based on the statistical structure of RoBERTa CLS embeddings, improving distribution without introducing semantic noise.

The experimental results validate the effectiveness of this strategy: the Ledoit–Wolf-based variant achieved the highest Macro F1 score (0.4087), significantly outperforming the reference system (0.1584) and showing balanced performance across all polarity classes. Although the Lasso and Elastic Net models also yielded competitive results, the superior performance of Ledoit–Wolf highlights the benefits of directly modeling global covariance structure in high-dimensional spaces.

Beyond classification accuracy, this approach greatly reduces computational costs. Unlike transformer fine-tuning, which requires extensive computational resources, our system uses a lightweight MLP-based classifier, making it a practical solution for real-world sentiment analysis scenarios in tourism and related domains.

Future work will focus on:

- Extending the approach to other multilingual or multimodal datasets where class imbalance is also critical.
- Exploring block-structured or hierarchical covariance estimation to improve the scalability of the Ledoit–Wolf method.
- Incorporating semantic validation or diversity constraints during synthetic generation to enhance coverage and reduce redundancy.
- Evaluating the approach in few-shot or zero-shot learning contexts, where synthetic generation may be crucial in the absence of labeled data.

This study reinforces the role of probabilistic data augmentation as a robust, scalable, and interpretable strategy to tackle imbalanced sentiment classification in complex domains.

Acknowledgments

We would like to express our sincere gratitude to the organizers of Rest-Mex 2025 for designing and coordinating this valuable shared task. Competitions of this nature not only promote research in low-resource languages and real-world applications, but also foster academic collaboration, methodological exchange, and the development of practical solutions in the field of Natural Language Processing. We especially appreciate the opportunity to participate and contribute to the scientific community through this platform.

We also thank the National Program of Science and Technology PN223LH004: Automation, Robotics and Artificial Intelligence, of the Ministry of Science, Technology and Environment of Cuba, for supporting this work under project PN223LH004-038: Theoretical contributions to AI in the management of complex data problems.

Declaration on Generative AI

We declare that the present manuscript has been written entirely by the authors and that no generative artificial intelligence tools were used in its preparation, drafting, or editing.

References

[1] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[2] G. Batista, R. Prati, M. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explorations Newsletter 6 (2004) 20–29.

[3] H. Han, W. Wang, B. Mao, Borderline-smote: A new over-sampling method in imbalanced datasets learning, in: Proceedings of the International Conference on Intelligent Computing (ICIC05), 2005, pp. 878–887.

[4] J. Stefanowski, S. Wilk, Selective pre-processing of imbalanced data for improving classification performance, in: DaWaK 2008, Turin, 2008, pp. 283–292.

[5] E. Ramentol, F. Herrera, R. Bello, Y. Caballero, Y. Sánchez, Edición de conjuntos de entrenamiento no balanceados usando operadores genéticos y conjuntos aproximados, in: Universidad de Camagüey, 2009.

[6] H. He, Y. Bai, E. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: IEEE International Joint Conference on Neural Networks (IJCNN), 2008, pp. 1322–1328.

[7] I. Leguen-de Varona, J. Madera, H. Gonzalez, L. Tubex, T. Verdonck, Oversampling method based on covariance matrix estimation in high-dimensional imbalanced classification, in: Progress in Artificial Intelligence and Pattern Recognition, Springer, 2024.

[8] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of rest-mex at iberlef 2025: Researching sentiment evaluation in text for mexican magical towns, volume 75, 2025.

[9] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.

[10] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[11] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[12] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

[13] O. Ledoit, M. Wolf, The power (non) linear shrinking: A review and guide to covariance matrix estimation, Journal of Multivariate Analysis (2022).

[14] Y. Lu, Y. Yin, Applying logistic lasso regression for the diagnosis of atypical crohn's disease, Computers in Biology and Medicine (2022).

[15] T. Kovács, A. Ruckstuhl, H. Obrist, P. Bühlmann, Graphical elastic net and target matrices: Fast algorithms and software for sparse precision matrix estimation, Journal of Computational and Graphical Statistics (2021).

Online Resources

The results and official rankings of the shared task can be accessed through the following link:

- Rest-Mex 2025 Results