

Generación Sintética de Datos Mediante la Estimación de la Matriz de Covarianza Regularizada: Análisis de Fidelidad OOD y Aplicación en la Detección de Emociones en Español

Synthetic Data Generation Through Regularized Covariance Matrix Estimation: Out-of-Distribution (OOD) Fidelity Analysis and Application to Emotion Detection in Spanish

Ireimis Leguen-de-Varona ^{1*} <https://orcid.org/0000-0002-1886-7644>

Julio Madera ¹ <https://orcid.org/0000-0001-5551-690X>

Alfredo Simon-Cuevas ² <https://orcid.org/0000-0002-6776-9434>

Leonardo Lastre Figueroa ¹ <https://orcid.org/0009-0009-1526-0108>

¹ Universidad de Camagüey “Ignacio Agramonte Loynaz”, Camagüey, Cuba

² Universidad Tecnológica de La Habana “José Antonio Echeverría”, CUJAE, La Habana, Cuba.

*Autor para la correspondencia. (ireimis.leguen@reduc.edu.cu)

RESUMEN

El desbalance de clases limita severamente la generalización de los modelos de aprendizaje automático, especialmente en espacios de alta dimensionalidad donde los métodos tradicionales pierden efectividad. Este trabajo propone y valida una técnica de sobremuestreo (*oversampling*) fundamentada en principios de la teoría estadística general, específicamente mediante un marco de balanceo probabilístico basado en la estimación robusta de la matriz de covarianza con el método de contracción de Ledoit-Wolf. Se extrajeron representaciones vectoriales (*embeddings*) de 768 dimensiones con RoBERTa aplicados al corpus TASS 2020 (detección de emociones en español) para generar instancias sintéticas a partir de la distribución multivariada aprendida. La novedad del enfoque radica en la validación sistemática de los datos generados mediante pruebas Fuera de Distribución (*OOD*), utilizando la distancia de Mahalanobis y un clasificador de dos muestras (*two-sample*) regularizado. Los resultados evidencian una mejora sustancial en la detección de clases minoritarias críticas: el F1-Score de "Sorpresa" subió del 10.9% al 26.4%, y el de "Miedo" del 0% al 7.1%. Asimismo, las pruebas OOD arrojaron un AUC cercano a 0.5, demostrando que los datos sintéticos son estadísticamente indistinguibles de la distribución real. El método propuesto alcanza un Macro F1 competitivo (35.51%) superando la línea base (28.08%), y ofrece la ventaja adicional de un bajo costo computacional al eliminar la necesidad de un ajuste fino (*fine-tuning*). Se concluye que el modelado probabilístico regularizado constituye un marco robusto y estadísticamente validado para mitigar el desbalance de clases en representaciones de alta dimensionalidad.

Palabras clave: aprendizaje desbalanceado; sobremuestreo probabilístico; análisis OOD; detección de emociones; embeddings contextuales

ABSTRACT

Class imbalance severely limits the generalization of machine learning models, especially in high-dimensional spaces where traditional methods lose effectiveness. This work proposes and validates an oversampling technique grounded in principles of general statistical theory, specifically through a probabilistic balancing framework based on the robust estimation of the covariance matrix using the Ledoit-Wolf shrinkage method. 768-dimensional vector representations (*embeddings*) were extracted using RoBERTa applied to the TASS

2020 corpus (emotion detection in Spanish) to generate synthetic instances from the learned multivariate distribution. The novelty of the approach lies in the systematic validation of the generated data through Out-of-Distribution (OOD) testing, using the Mahalanobis distance and a regularized two-sample classifier. The results show a substantial improvement in the detection of critical minority classes: the F1-Score for "Surprise" increased from 10.9% to 26.4%, and for "Fear" from 0% to 7.1%. Furthermore, the OOD tests yielded an AUC close to 0.5, demonstrating that the synthetic data are statistically indistinguishable from the real distribution. The proposed method achieves a competitive Macro F1 (35.51%), outperforming the baseline (28.08%), and offers the additional advantage of low computational cost by eliminating the need for fine-tuning. It is concluded that regularized probabilistic modeling constitutes a robust and statistically validated framework to mitigate class imbalance in high-dimensional representations.

Keywords: imbalanced learning; probabilistic oversampling; OOD analysis; emotion detection; contextual embeddings.

Recibido: 23/01/2026

Aceptado: 27/03/2026

Publicado: 01/04/2026

Introducción

El desbalance de clases constituye un desafío crítico en el aprendizaje automático que sesga los modelos hacia las clases mayoritarias, comprometiendo la detección de categorías minoritarias (Carvalho et al., 2025). Tradicionalmente, este problema se aborda mediante técnicas de sobremuestreo como SMOTE, que genera instancias por interpolación lineal. Sin embargo, en espacios de alta dimensionalidad como las representaciones vectoriales (*embeddings*) de 768 dimensiones de los modelos *Transformer*, SMOTE presenta limitaciones severas. Debido a la "maldición de la dimensionalidad", las distancias se concentran,

volviendo inestable la selección de vecinos cercanos y generando ruido semántico al ignorar la topología no lineal de los datos (Blagus y Lusa, 2013).

Este fenómeno afecta transversalmente al análisis de sentimientos y la detección de emociones, donde las distribuciones suelen ser marcadamente asimétricas. Investigaciones en diversos conjuntos de datos demuestran que incluso modelos robustos como BETO o TWilBERT difícilmente superan el 45% de Macro F1 debido a este sesgo intrínseco. Un caso representativo es el corpus TASS 2020 para español, donde categorías críticas como alegría o ira están significativamente infrarrepresentadas frente a clases más frecuentes, lo que limita la utilidad práctica de los clasificadores finales.

Para mitigar este problema, el estado del arte reciente ha recurrido a modelos generativos profundos como las Redes Generativas Antagónicas (GAN) o los Autocodificadores Variacionales (VAE). Si bien enfoques avanzados como el Aprendizaje de Representaciones de Emociones Balanceado por Clases (CBERL) (Shou et al., 2024) o las Redes Generativas Antagónicas acopladas a Autocodificadores Variacionales de Muchos a Muchos (M2M-VAEGAN) (Kang et al., 2025) logran generar representaciones sintéticas complejas, requieren un entrenamiento antagónico altamente costoso y arquitecturas densas que incrementan exponencialmente el costo computacional. Frente a esto, surge como alternativa el modelado probabilístico global mediante la matriz de covarianza (Leguen-de-Varona et al., 2024). Este enfoque evita las limitaciones de la interpolación local y, mediante el estimador de contracción de Ledoit-Wolf, garantiza estabilidad matemática en altas dimensiones, respetando la variabilidad real de los datos sin la complejidad técnica de las redes generativas.

Este artículo demuestra que el modelado probabilístico regularizado constituye un marco robusto para el balanceo en espacios vectoriales densos. Utilizando el corpus TASS 2020, se propone una metodología que no solo genera datos, sino que certifica su fidelidad mediante un análisis Fuera de Distribución (OOD), asegurando que las instancias sintéticas no degraden la estructura original del espacio latente.

Las contribuciones principales de este trabajo son:

1. El desarrollo de la técnica PMOTE-COV-LW, la cual emplea el modelado de la matriz de covarianza y la contracción de Ledoit-Wolf para el balanceo estadístico en escenarios de alta dimensionalidad.

2. Un protocolo de validación de fidelidad basado en el análisis OOD, que integra la distancia de *Mahalanobis* y un clasificador de dos muestras (*two-sample*) fuertemente regularizado para garantizar que los datos sintéticos sean estadísticamente indistinguibles de los reales.

Métodos o Metodología Computacional

Preparación del Dataset y Generación de Embeddings

Se utilizó el corpus TASS 2020 para la detección de emociones en español, compuesto por tweets etiquetados en siete categorías. El conjunto presenta un marcado desbalance: la clase mayoritaria ("Otras") representa aproximadamente el 30%, mientras que las minoritarias ("Miedo" y "Repulsión") apenas superan el 5% cada una. Para capturar el significado semántico en un espacio continuo, los textos se procesaron con el modelo preentrenado RoBERTa (RoBERTuito) (Pérez et al., 2022), extrayendo el embedding correspondiente al token de clasificación [CLS], lo que generó vectores densos de 768 dimensiones.

Método de Balanceo Propuesto: PMOTE-COV-LW

El método propuesto se fundamenta en el modelado probabilístico de la distribución de las clases minoritarias. A diferencia de SMOTE, que interpola entre puntos existentes, nuestro enfoque estima la distribución multivariada subyacente y genera nuevas instancias mediante muestreo de esta distribución.

Estimación de la Matriz de Covarianza: Para una clase minoritaria con n instancias representadas por vectores de dimensión p ($p=768$), se asume una distribución normal multivariante $N(\mu, C)$. La matriz de covarianza C cuantifica las relaciones lineales entre variables: una covarianza positiva indica una relación directa, negativa una relación inversa, y cero ausencia de relación lineal. Formalmente, se define como:

$$C = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \dots & Var(X_n) \end{pmatrix} \quad (1)$$

El desafío fundamental radica en estimar la matriz de covarianza C de manera robusta, especialmente cuando el número de atributos p es grande en relación con el número de instancias n , situación en la que la matriz de covarianza muestral S se vuelve singular o mal condicionada.

Para abordar esta limitación, se emplea el estimador de contracción (shrinkage) de Ledoit-Wolf (Ledoit y Wolf, 2022). Este estimador, C_{LW} se define como una combinación convexa de la matriz muestral S y una matriz objetivo estructurada F (típicamente la matriz identidad escalada por la varianza media):

$$C_{LW} = (1 - \delta)S + \delta F \quad (2)$$

El parámetro de contracción δ se calcula analíticamente para minimizar el error cuadrático medio esperado, proporcionando un equilibrio óptimo entre sesgo y varianza. El resultado es una matriz de covarianza siempre definida positiva y bien condicionada, incluso en escenarios de alta dimensionalidad donde $p > n$.

Generación de Instancias Sintéticas: Para la generación eficiente y numéricamente estable de nuevas instancias, se realiza la descomposición de Cholesky de la matriz regularizada, $C_{LW} = LL^T$. Una nueva instancia sintética x se genera mediante:

$$x_{syn} = \mu + L \cdot z \quad (3)$$

donde μ es el vector de medias de la clase minoritaria y $z \sim N(0, I)$ es un vector de ruido blanco gaussiano. Este procedimiento asegura que x_{syn} constituye una muestra matemáticamente válida de la distribución multivariada aprendida, preservando tanto las varianzas individuales como las correlaciones entre atributos de la clase original.

Control de Outliers: Para evitar la generación de instancias extremas que puedan caer fuera de la distribución real, se aplica un recorte (*clipping*) basado en los límites observados en los datos originales de la clase minoritaria, restringiendo cada dimensión al rango $[\min c, \max c]$.

Pseudocódigo del Algoritmo de Balanceo Probabilístico

A continuación se presenta el pseudocódigo del algoritmo propuesto, que describe el flujo completo de trabajo: desde la extracción de embeddings hasta la generación de instancias sintéticas y el entrenamiento del

clasificador. Este algoritmo unifica la estimación robusta de la covarianza mediante el método de Ledoit-Wolf con un control explícito de outliers para garantizar la calidad de los datos sintéticos generados.

Algoritmo 1: Balanceo Probabilístico con Estimación de Covarianza de Ledoit-Wolf (PMOTE-COV-LW)

Entrada:

- D: Conjunto de datos original desbalanceado (textos y etiquetas)
- Memb: Modelo de lenguaje preentrenado (RoBERTa)
- C: Clasificador base seleccionado (MLP)
- B: Tamaño del lote (batch size) para la generación estocástica

Salida:

- D_bal: Conjunto de datos final balanceado
- D_syn: Conjunto exclusivo de instancias sintéticas (para análisis OOD)
- C_entrenado: Modelo clasificador optimizado

Procedimiento:

// Fase 1: Extracción y Preprocesamiento del Hiperespacio

1. Si existe columna de texto en D entonces
2. Extraer embeddings [CLS] usando Memb para cada documento.
3. Fin Si
4. Expandir embeddings obteniendo matriz X de tamaño $m \times n$ ($n=768$) y vector de clases Y.
5. Estandarizar X (centrado en media 0 y varianza 1) aplicando StandardScaler.
6. Inicializar D_syn como conjunto vacío.
7. N_maj = Contar instancias de la clase mayoritaria en Y.

// Fase 2: Aprendizaje de Distribución y Generación Estocástica

8. Para cada clase c en Y hacer:
9. N_c = Número de instancias de la clase c.
10. Si N_c es igual a N_maj entonces continuar a la siguiente clase.
11. K = N_maj - N_c // Número de muestras sintéticas requeridas.
12. X_c = Subconjunto de X donde Y = c.
13. Calcular vectores característicos de X_c: media μ_c , desviación estándar σ_c , mínimos min_c y máximos max_c.
14. // Estimar matriz de covarianza con Ledoit-Wolf
15. C_LW = CovarianzaShrinkage(X_c) // Estimador de Ledoit-Wolf.
16. // Descomposición matricial para inyección de ruido correlacionado
17. Intentar: L = Descomposición Cholesky(C_LW).
18. Si falla por matriz no definida positiva (singularidad):
19. L = Descomposición SVD(C_LW) // Fallback robusto.
20. // Generación de instancias por lotes
21. S_generados = conjunto vacío
22. Mientras $|S_generados| < K$ hacer:

```
23. batch_size = min(B, K - |S_generados|)
24. Z = Generar Ruido Gaussiano de tamaño (batch_size × n) con distribución N(0, I)
25. S_lote =  $\mu_c + (L \cdot Z^T)^T$  // Transformación afín al espacio original
26. // Recorte (clipping) para prevenir outliers extremos
27. S_lote = clip(S_lote, min_c, max_c)
28. Añadir S_lote a S_generados
29. Fin Mientras
30. Añadir S_generados a D_syn asignando la etiqueta de clase c.
31. Fin Para
// Fase 3: Consolidación y Entrenamiento
32. D_bal = X  $\cup$  D_syn // Unificar datos originales estandarizados y sintéticos
33. Guardar D_bal y D_syn independientemente en almacenamiento.
34. Entrenar el clasificador C (MLP) utilizando D_bal.
35. Retornar D_bal, D_syn, C_entrenado
```

Validación de datos sintéticos mediante análisis OOD

Para garantizar que los datos sintéticos generados son estadísticamente indistinguibles de los datos reales, se implementaron dos pruebas complementarias de OOD.

Distancia de Mahalanobis: Evalúa la similitud estructural considerando las correlaciones intrínsecas (Lee et al., 2018). Si las instancias sintéticas son de alta fidelidad, la distribución de sus distancias respecto al centroide real no debe presentar diferencias significativas frente a las distancias de los datos originales, validado mediante un test de Kolmogorov-Smirnov ($p > 0.05$).

Clasificador Two-Sample Regularizado: Siguiendo a Lopez-Paz y Oquab (2017), se entrenó una Regresión Logística (con penalización L2 estricta y validación cruzada) para intentar distinguir entre el conjunto S y R. Un rendimiento cercano al azar ($AUC \approx 0.5$) certifica que no existen diferencias sistemáticas explotables entre ambas distribuciones.

Clasificador MLP

El conjunto de datos balanceado se utilizó para entrenar un clasificador ligero basado en un Perceptrón Multicapa (MLP). La arquitectura de la red consistió en una capa de entrada de 768 dimensiones, una capa oculta con 256 neuronas empleando la función de activación ReLU, una capa de Dropout del 20%, y una capa de salida (*softmax*) para la clasificación multiclase. El entrenamiento se realizó asignando el 80% de los datos

para el aprendizaje y reservando el 20% restante para la prueba. Se empleó el optimizador Adam con una tasa de aprendizaje de $2e-4$ y una función de pérdida de entropía cruzada categórica. Para garantizar la estabilidad en la predicción de las clases minoritarias y prevenir el sobreajuste, se implementaron técnicas de regularización adicionales al Dropout: se aplicó una penalización por decaimiento de pesos (L2) con un valor de $2e-4$ y se habilitó un criterio de parada temprana (*early stopping*) para interrumpir el entrenamiento cuando no se observaran mejoras en el conjunto de validación. Finalmente, para mitigar la sensibilidad del MLP a la inicialización aleatoria de los pesos, todos los experimentos se ejecutaron cinco veces utilizando semillas aleatorias distintas (*random seeds*).

Resultados y discusión

Análisis del Balanceo y Rendimiento por Clase

Las Figuras 1, 2 y 3 evidencian el impacto directo de la técnica PMOTE-Cov-LW. Antes del balanceo, el clasificador MLP presenta un fuerte sesgo hacia la clase mayoritaria "Otras", ignorando sistemáticamente las clases minoritarias, lo que se refleja en una exactitud nula para estas últimas (Figura 2) y en altas tasas de error al intentar separarlas de las dominantes (Figura 3). Tras aplicar el balanceo, se observa una distribución predictiva mucho más equitativa, permitiendo al modelo reconocer emociones previamente invisibilizadas y reduciendo significativamente los márgenes de error entre clases como "Sorpresa" y "Alegría".

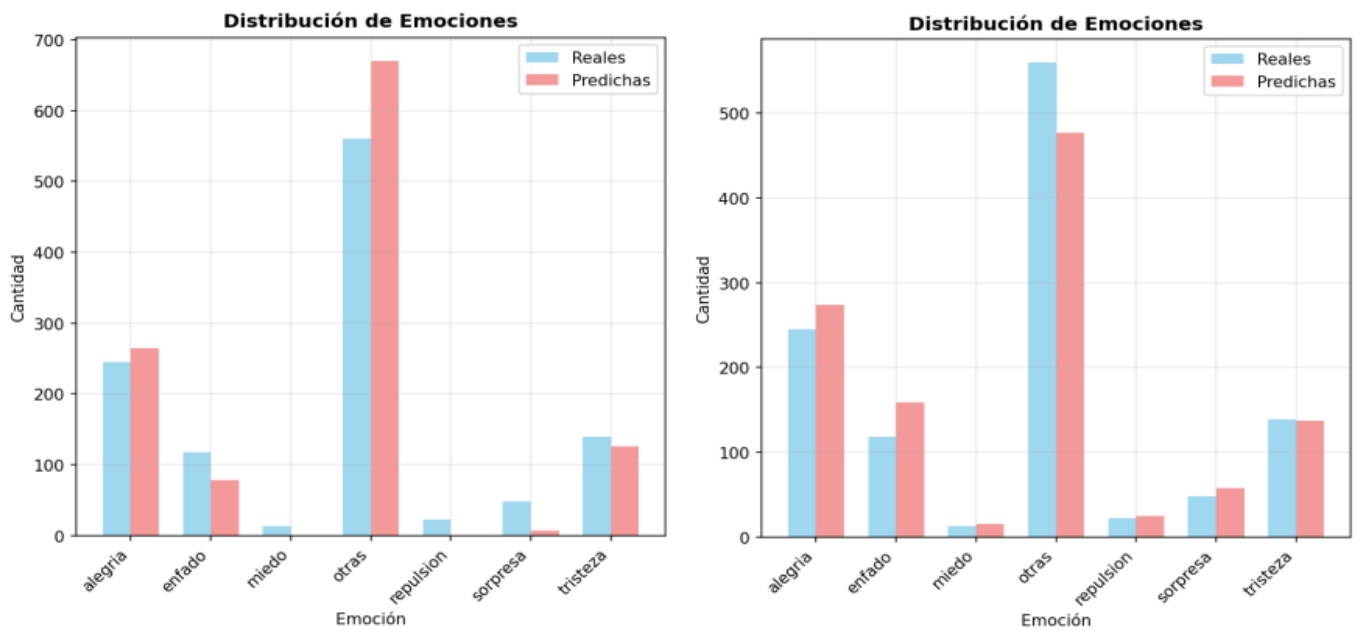


Fig. 1– Cantidades de clase predichas por el modelo MLP. Izquierda: entrenado con datos sin balancear. Derecha: entrenado con datos balanceados mediante PMOTE-Cov-LW.

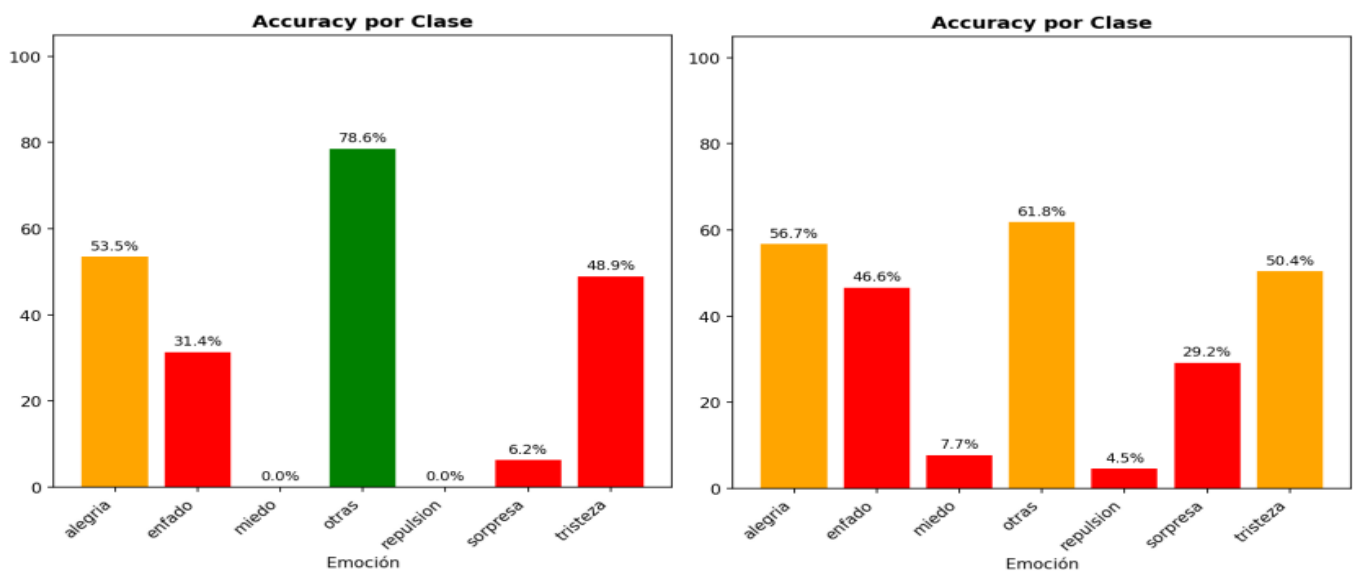


Fig. 2– Accuracy por clase antes y después del balanceo.

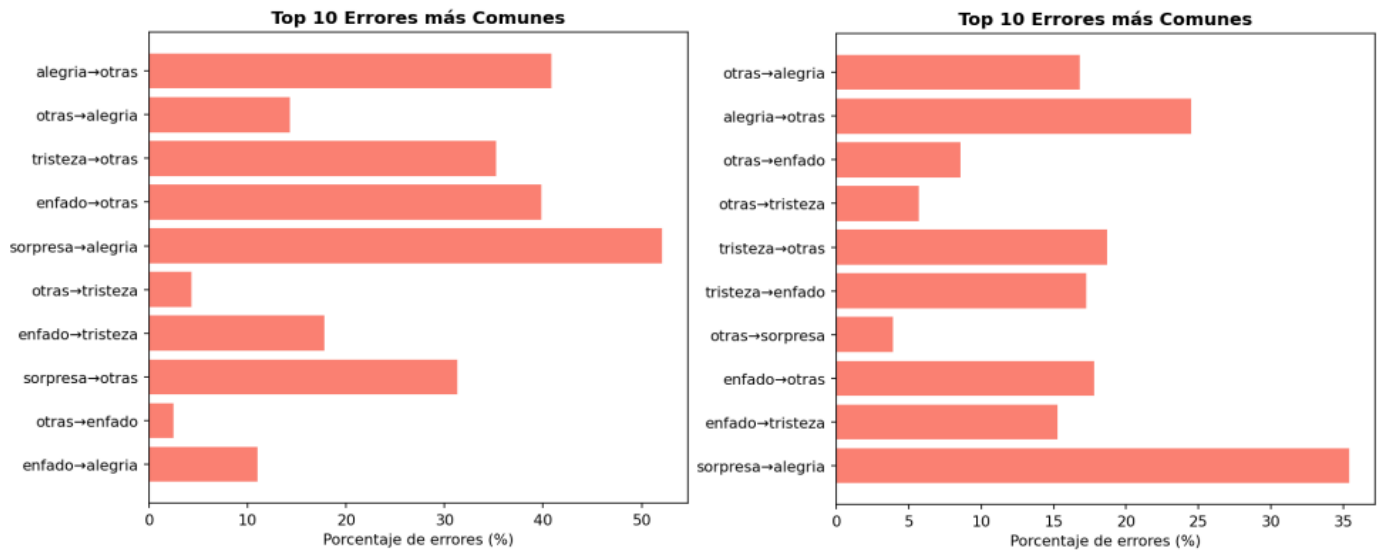


Fig. 3– Matrices de confusión normalizadas (errores de clasificación).

Reporte de métricas

Para evaluar cuantitativamente el impacto del método de balanceo propuesto, se compararon las métricas de F1-Score y Exactitud (Accuracy) obtenidas por el clasificador MLP entrenado con el conjunto de datos original desbalanceado frente al entrenamiento con el conjunto balanceado mediante PMOTE-Cov-LW. Los resultados de clasificación reportados en las Tablas 1 y 2 corresponden al promedio de 5 ejecuciones independientes del modelo, presentando tanto los valores desagregados por clase como las métricas globales.

Tabla 1 - F1-Score (%).

Clase	Clasificador MLP entrenado sin balanceo	Clasificador MLP entrenado con dataset balanceado usando Matriz de covarianza de Ledoit-Wolf (PMOTE-Cov-LW)
General	31.86	35.51
Alegría	51.5	53.6

Clase	Clasificador MLP entrenado sin balanceo	Clasificador MLP entrenado con dataset balanceado usando Matriz de covarianza de Ledoit-Wolf (PMOTE-Cov-LW)
Sorpresa	10.9	26.4
Tristeza	51.3	50.7
Otras	78.5	66.7
Enfado	37.8	39.7
Miedo	0	7.1
Repulsión	0	4.3

Tabla 2 - Accuracy (%).

Clase	Clasificador MLP entrenado sin balanceo	Clasificador MLP entrenado con dataset balanceado usando Matriz de covarianza de Ledoit-Wolf (PMOTE-Cov-LW)
General	59.30	54.67
Alegría	53.5	56.7
Sorpresa	6.2	29.2
Tristeza	48.9	50.4
Otras	78.6	61.8
Enfado	31.4	46.6
Miedo	0	7.7
Repulsión	0	4.5

Resulta destacable el incremento en el F1-Score de la clase "Sorpresa" (del 10.9% al 26.4%). Asimismo, emociones extremas como "Miedo" y "Repulsión", originalmente ignoradas (0%), alcanzaron un F1-Score del 7.1% y 4.3% respectivamente.

Análisis de errores críticos: El modesto rescate de la clase "Miedo" en comparación con "Sorpresa" exige un análisis detallado. Este fenómeno se explica por un fuerte solapamiento semántico en el espacio latente de RoBERTa. Investigaciones previas (Bostan y Klinger, 2018; Acheampong et al., 2020) indican que emociones de alta intensidad y valencia negativa (miedo, tristeza, ira) suelen compartir estructuras léxicas y contextos sintácticos análogos, agrupándose de manera densa según el modelo circunflejo del afecto (Russell, 1980). En consecuencia, las fronteras de decisión en estas regiones del hiperespacio son mucho más difusas, lo que provoca confusión cruzada y dificulta la separabilidad de las instancias sintéticas del "Miedo" frente a otras emociones negativas con mayor densidad de datos reales.

Como contraparte a la ganancia en clases minoritarias, se observa una reducción en el rendimiento de la clase dominante "Otras", lo que explica el descenso de la exactitud global (del 59.30% al 54.67%). Este trade-off es un comportamiento intrínseco de los métodos de balanceo. No obstante, en aplicaciones del mundo real (e.g., monitorización psicológica), el costo de ignorar una emoción crítica es inaceptable, haciendo que este sacrificio sea deseable. La robustez global del modelo queda confirmada por el F1-Score General (Macro F1), que aumentó del 31.86% al 35.51%.

Resultados de las pruebas de OOD

Tabla 3 -Resultados de las pruebas OOD.

Prueba	Métrica	Resultado	Criterio de evaluación
Distancia de Mahalanobis	Mediana (sintéticos vs reales)	Dentro del rango intercuartílico	Deben ser similares
	Rangos percentiles (10-90)	Coincidentes	Deben coincidir
	Test Kolmogorov-Smirnov (KS)	Estadístico D = 0.3984, p-valor > 0.05	p>0.05 indica misma distribución
	Histogramas	Superposición significativa	Distribuciones similares
Clasificador Two-Sample	AUC	0.4566 ≈ 0.5	AUC ≈ 0.5 indica indistinguibilidad

Prueba	Métrica	Resultado	Criterio de evaluación
	Configuración	LogisticRegression (C pequeño), bagging, cross-val	Regularización para evitar sobreajuste

Como detalla la Tabla 3, la evaluación OOD arroja resultados concluyentes sobre la fidelidad topológica de los datos. La distancia de Mahalanobis evidenció una superposición estructural, con una mediana para instancias sintéticas ubicada dentro del rango intercuartílico real, confirmando (mediante un test KS con $p > 0.05$) la ausencia de diferencias significativas. De manera complementaria, el clasificador two-sample (con regularización estricta para evitar memorización) obtuvo un AUC de 0.4566, cercano al umbral aleatorio de 0.5 (Lopez-Paz y Oquab, 2017). Estas métricas validan empíricamente que, a diferencia de la interpolación local que inyecta ruido semántico, PMOTE-Cov-LW produce instancias semánticamente válidas y estadísticamente indistinguibles de los datos originales.

Comparativa con el Estado del Arte

Para situar el rendimiento del método propuesto en el contexto de las investigaciones previas sobre el corpus TASS 2020, se realizó una comparación con los principales enfoques reportados en la literatura. Estos incluyen técnicas basadas en fine-tuning de modelos Transformer (TWilBERT, BERT en español), enfoques híbridos léxico-estadísticos con ensambles, y combinaciones de características lingüísticas con embeddings clasificados mediante SVM. Adicionalmente, se incluye como línea base el clasificador MLP entrenado sin balanceo para evidenciar el impacto directo de nuestra metodología. La Tabla 4 resume los valores de Macro F1 reportados por cada enfoque, junto con el resultado obtenido por la propuesta (PMOTE-COV-LW + MLP).

Tabla 4 - Comparativa con otros enfoques en el corpus TASS 2020.

Enfoque / Trabajo	Modelo Clasificador	Macro F1 (%)
Fine-tuning TWilBERT (Civit-Masot y otros, 2020)	Transformer	44.7
Fine-tuning BERT en español (López y Azzopardi, 2020)	Transformer	44.7
Lexicón SEL + Stacking	Ensamble	43
Características lingüísticas + embeddings (Franco-Salvador y otros, 2020)	SVM	37.9

Enfoque / Trabajo	Modelo Clasificador	Macro F1 (%)
Sin balanceo previo	MLP	28.08
PMOTE-COV-LW (nuestra propuesta)	MLP	35.51

La Tabla 4 sitúa el método propuesto (35.51% Macro F1) por encima de la línea base no balanceada (28.08%). Es innegable que los modelos Transformer con fine-tuning completo (TWilBERT, BERT) alcanzan mayores valores (44.7%); sin embargo, esto implica actualizar cientos de millones de parámetros, exigiendo un costo computacional masivo y hardware especializado.

Discusión sobre eficiencia computacional: A nivel de manipulación de datos, la literatura reciente emplea asiduamente modelos generativos profundos (GANs o VAEs) como CBERL o M2M-VAEGAN para combatir el desbalance. Aunque efectivos, estos enfoques requieren arquitecturas densas y un entrenamiento antagónico altamente inestable y costoso (Kang et al., 2025; Shou et al., 2024). En contraste directo, PMOTE-Cov-LW fundamenta su funcionamiento en principios estadísticos cerrados (estimador de Ledoit-Wolf). Al prescindir del entrenamiento de redes adversarias, logra una eficiencia computacional notablemente superior. Acoplando representaciones estáticas precalculadas con un clasificador MLP ligero, el método optimiza una fracción minúscula de parámetros, alineándose con los principios de la Inteligencia Artificial Verde (Green AI).

Finalmente, los enfoques actuales del estado del arte omiten pruebas OOD, careciendo de garantías de que sus instancias sintéticas respeten la distribución original. El marco metodológico propuesto destaca por certificar formalmente que la mitigación del desbalance se logra con datos topológicamente seguros y estadísticamente fiables.

Conclusiones

Este trabajo presentó y validó PMOTE-Cov-LW, un marco de balanceo probabilístico diseñado para mitigar el desbalance de clases en escenarios de alta dimensionalidad. Mediante el uso del estimador de contracción de Ledoit-Wolf, el método superó las deficiencias geométricas inherentes a los métodos de interpolación

local, los cuales tienden a inyectar ruido semántico en espacios latentes profundos. La robustez teórica de la propuesta fue validada mediante un riguroso análisis Fuera de Distribución (OOD): las pruebas de Mahalanobis y un clasificador de dos muestras fuertemente regularizado $AUC \approx 0.5$ certificaron que los datos sintéticos preservan la correlación multivariada original y son estadísticamente indistinguibles de los reales. Desde una perspectiva empírica, la aplicación de este marco en el corpus TASS 2020 mejoró sustancialmente la sensibilidad del modelo ante emociones críticas. El clasificador elevó el F1-Score de la clase "Sorpresa" (del 10.9% al 26.4%) y rescató categorías extremas como el "Miedo" y la "Repulsión", que pasaron del 0% a alcanzar un 7.1% y 4.3%, respectivamente, incrementando el Macro F1 global al 35.51%. Si bien estas métricas predictivas se sitúan por debajo de arquitecturas masivas basadas en fine-tuning, la metodología propuesta destaca por su eficiencia. Al emplear clasificadores ligeros sobre embeddings estáticos, elimina los costos computacionales exponenciales y se alinea con los principios de la Inteligencia Artificial Verde (Green AI).

Como líneas de trabajo futuro, se proyecta la comparación empírica de PMOTE-Cov-LW con métodos algorítmicos como Focal Loss y Weighted Loss para cuantificar rigurosamente las ventajas relativas entre el balanceo a nivel de datos y a nivel de función de pérdida. Adicionalmente, se explorará la extensión de este marco probabilístico a dominios multimodales (texto, imagen y audio simultáneos), la integración de estimadores de covarianza adaptativos para flujos de datos en tiempo real, y técnicas de regularización por bloques para escalar eficientemente a dimensiones ultra-altas.

Agradecimientos

Los autores agradecen el apoyo brindado por el Ministerio de Ciencia, Tecnología y Medio Ambiente de Cuba a través del Programa Nacional de Ciencia y Tecnología (PN223LH004), en el marco del proyecto PN223LH004-038: Contribuciones teóricas a la IA en la gestión de problemas de datos complejos.

Referencias

- Acheampong, F. A., Wenyu, C. y Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2020, 2(7), e12189.
- Blagus, R. y Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 2013, 14, 1-16.
- Bostan, L. A. M. y Klinger, R. An analysis of annotated corpora for emotion classification in text. En: *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, p. 2104-2119.
- Carvalho, M., Pinho, A. J. y Bras, S. Resampling approaches to handle class imbalance: A review from a data perspective. *Journal of Big Data*, 2025, 12(1), 71.
- Civit-Masot, J., et al. ELiRF-UPC at TASS 2020: TwilBERT for sentiment analysis and emotion detection in Spanish tweets. En: *CEUR Workshop Proceedings*. Aachen: CEUR-WS, 2020, vol. 2664.
- Franco-Salvador, M., et al. UMUTeam at TASS 2020: Combining linguistic features and machine-learning models for sentiment classification. En: *CEUR Workshop Proceedings*. Aachen: CEUR-WS, 2020, vol. 2664.
- García-Vega, M., et al. Overview of TASS 2020: Introducing emotion detection. En: *CEUR Workshop Proceedings*. Aachen: CEUR-WS, 2020, vol. 2664.
- Kang, F., Feng, T. y Lin, J. VAE-GAN-Guided Cross-Class Generation: A Class Imbalance Data Augmentation Method for Network Intrusion Detection. *Electronics*, 2025, 14(11), 2103.
- Ledoit, O. y Wolf, M. The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 2022, 20(1), 187-218.
- Lee, K., Lee, K., Lee, H. y Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 2018, 31, 7167-7177.
- Leguen-de-Varona, I., Madera, J., Gonzalez, H., Tubex, L. y Verdonck, T. Oversampling Method Based Covariance Matrix Estimation in High-Dimensional Imbalanced Classification. En: Hernández Heredia, Y., Milián Núñez, V. y Ruiz Shulcloper, J. (Eds.). *Progress in Artificial Intelligence and Pattern Recognition. IWAIPR 2023*. Lecture Notes in Computer Science. Cham: Springer, 2024, vol. 14335, p. 16-23.
- López, M. y Azzopardi, J. Fine-tuning BERT for Spanish sentiment analysis at TASS 2020. En: *CEUR Workshop Proceedings*. Aachen: CEUR-WS, 2020, vol. 2664.

- Lopez-Paz, D. y Oquab, M. Revisiting classifier two-sample tests. En: *International Conference on Learning Representations (ICLR)*, 2017.
- Mahalanobis, P. C. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 1936, 2(1), 49-55.
- Pérez, J. M., Furman, D. A., Alonso Alemany, L. y Luque, F. M. RoBERTuito: a pre-trained language model for social media text in Spanish. En: *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*. Marseille: European Language Resources Association, 2022, p. 7235-7243.
- Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980, 39(6), 1161-1178.
- Shou, Y., Meng, T., Ai, W., Yin, C. y Li, K. Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Artificial Intelligence*, 2024, 5(12), 6472-6487.

Conflicto de interés

Los autores declaran que no existen conflictos de interés relacionados con la realización de esta investigación. Los resultados presentados son producto del trabajo científico de los autores y no han sido influenciados por intereses comerciales, financieros o de otra índole. Los autores autorizan la distribución y uso del artículo para fines académicos y científicos.

Contribuciones de los autores

Conceptualización: Ireimis Leguen-de-Varona, Julio Madera

Curación de datos: Ireimis Leguen-de-Varona, Leonardo Lastre Figueroa

Análisis formal: Ireimis Leguen-de-Varona, Julio Madera, Alfredo Simon-Cuevas

Investigación: ireimis Leguen-de-Varona, Leonardo Lastre Figueroa

Metodología: Ireimis Leguen-de-Varona, Julio Madera, Alfredo Simon-Cuevas

Administración del proyecto: Julio Madera, Alfredo Simon-Cuevas

Software: Ireimis Leguen-de-Varona, Leonardo Lastre Figueroa

Supervisión: Julio Madera, Alfredo Simon-Cuevas

Validación: Ireimis Leguen-de-Varona, Julio Madera

Visualización: Ireimis Leguen-de-Varona, Leonardo Lastre Figueroa

Redacción – borrador original: Ireimis Leguen-de-Varona

Redacción – revisión y edición: Julio Madera, Alfredo Simon-Cuevas