ISSN 2395-8618

Postprocesamiento geométrico de espacios de embeddings para clasificación de textos con Voronoi

Johan Argenis Franco Rogel*

***Resumen*—Este artículo estudia si una simple corrección geométrica aplicada después de la extracción de incrustaciones mejora la separabilidad de clases en la clasificación de texto. Se evalúan seis familias de incrustaciones bajo dos condiciones, original y geométricamente mejorada, utilizando tres paradigmas: clasificación de Voronoi inducida por centroide, máquinas de vectores de soporte y agrupamiento no supervisado. El estudio abarca un corpus multiclase en español, un conjunto de datos de referencia de emociones de grano fino y un conjunto de datos de referencia de noticias. Demostramos que aumentar la isotropía puede destruir sistemáticamente la estructura discriminativa en ciertos regímenes. En el corpus en español, la corrección mejora consistentemente la clasificación de Voronoi y la clasificación supervisada. En el conjunto de datos de referencia de emociones, ayuda principalmente a Voronoi mientras que SVM permanece prácticamente sin cambios. En el conjunto de datos de referencia de noticias, degrada Voronoi, SVM y el agrupamiento en todos los modelos. El artículo también contrasta estos resultados con trabajos previos basados en métodos clásicos o débilmente supervisados y argumenta que la regularización geométrica debe tratarse como una intervención dependiente del conjunto de datos en lugar de un paso de preprocesamiento universalmente válido.**

***Palabras clave*— Embeddings textuales, isotropía, anisotropía, Voronoi, clasificación, AG News, GoEmotions.**

Geometric Post-processing of Embedding Spaces for Text Classification with Voronoi

***Abstract*—This paper studies whether a simple geometric correction applied after embedding extraction improves class separability in text classification. Six embedding families are evaluated under two conditions, original and geometrically improved, using three paradigms: centroid-induced Voronoi classification, support vector machines, and unsupervised clustering. The study covers a Spanish multi-class corpus, a fine-grained emotion benchmark, and a news topic benchmark. We show that increasing isotropy can systematically destroy discriminative structure in certain regimes. In the Spanish corpus, the correction improves Voronoi and supervised classification consistently. In the emotion benchmark, it mainly helps Voronoi while leaving SVM nearly unchanged. In the news benchmark, it degrades Voronoi, SVM, and clustering across all models. The paper also contrasts these results with prior work based on classical or weakly supervised methods and argues that geometric regularization should be treated as a dataset-dependent intervention rather than a universally valid preprocessing step.**

***Index Terms*—Text embeddings, isotropy, anisotropy, Voronoi, classification, AG News, GoEmotions.**

Manuscript received on 11/02/2026, accepted for publication on 09/04/2026. Corresponding author is Johan Argenis Franco Roge (m24ce052@cenidet.tecnm.mx).

The author is with Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos, México.

I. Introducción

Los embeddings textuales suelen evaluarse por similitud, recuperación o clasificación supervisada. Sin embargo, su utilidad práctica no depende solamente del codificador. También importa la forma en que los vectores ocupan el espacio. Cuando una parte importante de la varianza se concentra en pocas direcciones, las decisiones basadas en distancia pueden volverse inestables y textos semánticamente distintos pueden quedar artificialmente cercanos. En este trabajo, llamamos anisotropía a esa concentración dominante de la varianza en pocas direcciones, e isotropía a una ocupación más uniforme del espacio de representación [1–3]. Este problema ha motivado una línea de trabajo sobre anisotropía, isotropía y postprocesamiento geométrico [3–6].

La relevancia del problema va más allá de la tarea puntual estudiada aquí. La geometría del espacio de embeddings influye de manera directa en aplicaciones como búsqueda semántica, recuperación, clasificación y agrupamiento, donde la utilidad de una representación depende tanto del contenido semántico capturado como de la organización métrica que induce [6]. Además, benchmarks amplios recientes muestran que no existe una familia de embeddings universalmente dominante en todas las tareas, lo que vuelve especialmente relevante entender cuándo una corrección geométrica ayuda, cuándo es neutra y cuándo puede resultar perjudicial [7].

El interés de este trabajo no está en proponer un codificador nuevo, sino en estudiar si una intervención geométrica simple aplicada después de obtener los embeddings puede mejorar la separabilidad entre clases o, por el contrario, destruir estructura discriminativa ya presente en el espacio. Para ello se emplea un clasificador especialmente sensible a la estructura métrica: una partición tipo Voronoi inducida por centroides de clase. Si las clases forman regiones compactas y bien organizadas, la asignación al centroide más cercano debería funcionar razonablemente incluso sin aprender fronteras discriminativas. Para contextualizar esa lectura geométrica, también se evalúan máquinas de soporte vectorial (SVM) y varios algoritmos de agrupamiento.

El estudio integra tres colecciones con propiedades distintas: un corpus multiclase en español, una tarea fina de emociones

ISSN 2395-8618

y una tarea temática de noticias. El objetivo es observar no sólo si la corrección ayuda, sino en qué tipo de escenario lo hace y en cuáles deja de ser conveniente.

II. Trabajo relacionado

Los transformers consolidaron el uso de embeddings de oración y documento como una representación estándar para tareas de PLN [4, 8]. Más adelante, modelos orientados a nivel de oración como Sentence-BERT (SBERT) reforzaron la importancia de la geometría del espacio para búsqueda semántica, agrupamiento y clasificación [6]. En paralelo, varios trabajos mostraron que muchos espacios de representación son anisotrópicos y que correcciones simples pueden modificar de manera importante su comportamiento [3, 5].

En la tarea de emociones finas, Demszky et al. [9] introdujeron un banco de prueba de 28 etiquetas y reportaron una línea base con macro-$F_1 = 0.46$. Posteriormente, Alvarez-Gonzalez et al. [10] ampliaron la comparación de métodos y mostraron que representaciones estadísticas como frecuencia de término–frecuencia inversa de documento (TF–IDF) siguen siendo competitivas e incluso superan a algunas alternativas más complejas.

En la tarea de noticias, Zhang et al. [11] reportaron una referencia clásica basada en n-gramas, TF–IDF y regresión logística multinomial, con un error de prueba de 7.64%, equivalente a 92.36% de exactitud. En la dirección no supervisada, Stammbach y Ash [12] reportaron resultados con TF–IDF + k-means, SBERT + k-means, DocSCAN y SBERT + SVM, lo que ofrece un marco útil para comparar el rango de desempeño esperado en esa colección.

La diferencia principal con esos antecedentes es que aquí el foco no está en entrenar mejores clasificadores, sino en estudiar el efecto de una corrección geométrica post hoc sobre tres lecturas complementarias del espacio: Voronoi, SVM y clustering.

III. Metodología

A. Familias de embeddings y corrección geométrica

Se evaluaron seis familias de embeddings, resumidas en la Tabla I.

TABLE I
FAMILIAS DE EMBEDDINGS EVALUADAS.

Clave corta	Modelo
OpenAI 3-L	`text-embedding-3-large`
MiniLM	`all-MiniLM-L6-v2`
MPNet	`all-mpnet-base-v2`
mE5-base	`intfloat/multilingual-e5-base`
BGE-small	`BAAI/bge-small-en-v1.5`
mBERT	`google-bert/bert-base-multilingual-uncased`

Cada familia se evaluó en dos condiciones:

1. Embeddings originales, normalizados con norma L_2.
2. Embeddings corregidos geométricamente mediante una intervención post hoc de tipo all-but-the-top: centrado para remover la media global, análisis de componentes principales (PCA) para identificar componentes dominantes, anulación de las tres direcciones principales dominantes, reconstrucción y normalización final.

La transformación sigue la idea general de postprocesamiento propuesta por Mu y Viswanath [5]. En otras palabras, no se modifica ni se reentrena el codificador original; sólo se transforma geométricamente el espacio ya obtenido para atenuar sesgos globales asociados con la media y con unas pocas direcciones de varianza dominante. Dado un espacio de embeddings $X \in \mathbb{R}^{n\times d}$, con media empírica $\mu \in \mathbb{R}^d$ y matriz de componentes principales $V \in \mathbb{R}^{d\times d}$, la versión corregida se define como en la Ecuación 1:

$$X' = \mathrm{norm}\Big((X - \mathbf{1}\mu^\top)\, V\, (I - P_k)\, V^\top\Big), \tag{1}$$

donde P_k es el proyector diagonal que conserva las primeras k direcciones principales y $\mathrm{norm}(\cdot)$ indica normalización fila a fila con norma L_2.

En la configuración base se utilizó $k = 3$, siguiendo la heurística del experimento original. Para evitar que ese valor quedara como una decisión arbitraria, se añadió un análisis de sensibilidad para $k \in \{0, 1, 2, 3, 5, 10\}$ sobre tres familias representativas, con el fin de observar cómo cambia el rendimiento al remover distinto número de componentes.

Adicionalmente, se incorporaron métricas geométricas explícitas antes y después del postprocesamiento: $\lambda_1/\lambda_{\text{mean}}$, λ_1/λ_{10}, entropía espectral, rango efectivo, media y desviación estándar de cosenos, así como proporción de varianza explicada acumulada en las primeras componentes. De este modo, la interpretación geométrica ya no descansa solamente en resultados de clasificación. El análisis de sensibilidad se concentró deliberadamente en k, es decir, en el número de componentes principales anuladas, porque ese es el hiperparámetro central del operador all-but-the-top. Otras decisiones potenciales, como normalizaciones alternativas o variantes adicionales de distancia, quedan fuera del alcance de este estudio y se consideran extensiones naturales para trabajo futuro.

B. Protocolo de evaluación

Se utilizaron tres familias de evaluación.

B1. Clasificación Voronoi: Para cada colección y espacio de embeddings se calculó un centroide por clase real, definido en la Ecuación 2:

$$m_c = \frac{1}{|D_c|} \sum_{x_i \in D_c} x_i. \tag{2}$$

Posteriormente, cada instancia se asignó al centroide más cercano con distancia coseno, euclidiana y Manhattan. Este procedimiento es determinista una vez fijado el espacio. En consecuencia, Voronoi debe interpretarse como una prueba geométrica de organización de clases y no como un clasificador convencional de entrenamiento/prueba.

ISSN 2395-8618

TABLE II
CONJUNTOS DE DATOS UTILIZADOS EN EL ESTUDIO.

Colección	Clases	Muestras	Observaciones
Corpus multiclase en español	30	1200	Balanceado, 40 por clase
Subconjunto balanceado de emociones	28	1120	Sólo ejemplos de etiqueta única, 40 por clase
Subconjunto balanceado de noticias	4	1200	Balanceado, 300 por clase

Mide una propiedad más restringida que SVM o k vecinos más cercanos (k-NN): cuán compatible es el espacio con una descripción prototípica en la que cada clase queda representada por un solo núcleo y la asignación depende sólo de proximidad al centroide. Si esa hipótesis se cumple, Voronoi funciona como diagnóstico de separabilidad por prototipos; si no se cumple, su degradación indica que la clase requiere múltiples modos, fronteras no convexas o estructura local que un solo centroide no puede capturar.

Por ello, su función en este estudio es complementar la evaluación supervisada con una lectura geométrica del espacio, no sustituir una validación de entrenamiento/prueba ni competir directamente con clasificadores discriminativos.

B2. SVM: Se evaluaron las variantes lineal y RBF mediante validación cruzada estratificada de cinco particiones. Se reporta el mejor macro-F_1 entre ambas.

B3. Clustering: Se evaluaron k-means, agrupamiento aglomerativo con Ward, agrupamiento aglomerativo promedio con coseno, agrupamiento espectral y BIRCH. Como las etiquetas de grupo son arbitrarias, los grupos predichos se alinearon con las clases reales por mayoría antes de calcular macro-F_1. Esta familia debe leerse con un supuesto distinto al de Voronoi o SVM. Los algoritmos de agrupamiento optimizan compacidad, conectividad o cortes espectrales sin usar etiquetas; por ello, un espacio puede ser muy útil para clasificación supervisada y aun así ser poco compatible con particiones no supervisadas alineadas con las clases. En otras palabras, separabilidad supervisada y recuperación directa de grupos no son propiedades equivalentes del mismo espacio.

Se fijó una semilla global igual a 42 para el muestreo balanceado, el barajado, PCA, la validación cruzada y los procedimientos aleatorios de clustering. Para complementar la evaluación, también se calcularon intervalos de confianza bootstrap sobre las predicciones de Voronoi y SVM en los casos principales, así como comparaciones adicionales entre Voronoi, centroide más cercano y k-NN.

C. Conjuntos de datos

La Tabla II resume las colecciones utilizadas.

El corpus en español fue diseñado con clases cercanas semánticamente pero no triviales. La colección de emociones se convirtió a una variante balanceada de etiqueta única conservando exclusivamente ejemplos con una sola etiqueta agregada del recurso original [9]. La colección de noticias sigue el banco de prueba temático estándar introducido por Zhang et al. [11], pero aquí se evalúa sobre una muestra balanceada.

Con el fin de medir la dificultad del corpus en español mediante referencias simples, se añadieron tres líneas base con validación cruzada estratificada: TF–IDF + regresión logística, TF–IDF + SVM lineal y un clasificador aleatorio estratificado. Los resultados obtenidos fueron 0.6467 de exactitud y 0.6551 de macro-F_1 para TF–IDF + regresión logística, 0.9992 de exactitud y 0.9992 de macro-F_1 para TF–IDF + SVM lineal, y 0.0250 de exactitud con 0.0262 de macro-F_1 para la línea base aleatoria.

Esta cifra cercana a uno obliga a interpretar ese recurso con cautela: no representa un escenario abierto y ruidoso, sino un banco de prueba controlado donde una representación dispersa clásica ya separa casi linealmente las clases. Por ello, las ganancias observadas en ese corpus deben leerse como evidencia de que el postprocesamiento puede preservar o reforzar una geometría ya favorable, no como prueba suficiente de utilidad en escenarios difíciles.

IV. RESULTADOS

A. Efecto agregado de la corrección geométrica

La Tabla III muestra el cambio promedio inducido por la corrección geométrica a través de las seis familias de embeddings.

TABLE III
CAMBIO PROMEDIO INDUCIDO POR LA CORRECCIÓN GEOMÉTRICA A TRAVÉS DE LOS MODELOS. LOS VALORES POSITIVOS INDICAN MEJORA.

Colección	Voronoi	SVM	Clustering
Corpus en español	+0.0510	+0.1419	+0.0080
Emociones	+0.0543	-0.0055	-0.0006
Noticias	-0.3960	-0.0641	-0.4960

El patrón es claro. En el corpus en español la corrección ayudó de forma casi uniforme y especialmente fuerte en clasificación supervisada. En la tarea de emociones, la ganancia se concentra en Voronoi, lo que sugiere una mejor organización alrededor de centroides sin una mejora equivalente en la frontera discriminativa de SVM. En la tarea de noticias, la misma transformación resultó perjudicial en las tres familias de evaluación.

B. Mejores resultados por colección

La Tabla IV resume el mejor resultado alcanzado en cada colección y familia de evaluación.

ISSN 2395-8618

TABLE IV
MEJOR MACRO-F_1 POR COLECCIÓN Y FAMILIA DE EVALUACIÓN.

Colección	Familia	Mejor modelo	Mejor macro-F_1
Corpus en español	Voronoi corregido	OpenAI 3-L	0.9800
	SVM corregido	OpenAI 3-L	0.9916
	Clustering corregido	OpenAI 3-L	0.1177
Emociones	Voronoi corregido	OpenAI 3-L	0.7938
	SVM original	OpenAI 3-L	0.4583
	Clustering original	OpenAI 3-L	0.2565
Noticias	Voronoi original	OpenAI 3-L	0.9123
	SVM original	OpenAI 3-L	0.9166
	Clustering original	mE5-base	0.8662

De aquí se desprenden dos observaciones. La primera es que la familia más fuerte se mantiene dominante en la mayor parte de los escenarios. La segunda es que el signo de la mejora depende más del tipo de colección que de la familia de embeddings. La misma intervención que ayudó tanto a modelos fuertes como débiles en el corpus en español es casi neutra en emociones y consistentemente negativa en noticias.

C. Comparación con trabajos previos

La Tabla V sitúa nuestros resultados junto a cifras reportadas previamente. Esta comparación debe leerse con cuidado. En noticias, la cercanía metodológica es mayor porque se trata de una tarea temática de etiqueta única. En emociones, la comparabilidad es parcial porque aquí se emplea una variante balanceada de etiqueta única, mientras que la tarea oficial es multietiqueta.

La comparación en noticias sugiere que los espacios originales ya eran competitivos frente a referencias clásicas y métodos débilmente supervisados, al menos en orden de magnitud. En particular, el mejor SVM original de este estudio queda numéricamente cerca de la referencia clásica de Zhang et al. [11] y de SBERT + SVM reportado por Stammbach y Ash [12]. El mejor clustering original también queda por encima de TF–IDF + k-means y muy cerca del rango de otros enfoques más fuertes, aunque aquí la comparación no es exacta porque cambian métrica y subconjunto.

En emociones la comparación es más delicada. El mejor SVM original de este estudio es numéricamente similar a la línea base temprana reportada por Demszky et al. [9], pero esa coincidencia no debe sobreinterpretarse porque la tarea oficial es multietiqueta y usa las particiones completas. El valor alto de Voronoi debe leerse, por tanto, como evidencia de organización geométrica dentro de la variante balanceada de etiqueta única, no como un reemplazo de las cifras oficiales.

Para atender este problema de comparabilidad, se implementaron además líneas base TF–IDF sobre los splits estándar. En noticias, TF–IDF + regresión logística alcanzó 0.9189 de exactitud y TF–IDF + SVM lineal alcanzó 0.9234, mientras que TF–IDF + k-means obtuvo 0.4336 de exactitud. En la tarea oficial multietiqueta de emociones, un esquema TF–IDF + regresión logística one-vs-rest alcanzó macro-$F_1 = 0.2014$ y micro-$F_1 = 0.3871$. Estas cifras no reemplazan la evaluación principal basada en embeddings, pero sí permiten anclar la discusión en métricas y configuraciones más cercanas a la literatura.

D. Robustez estadística y sensibilidad en k

Los intervalos bootstrap confirman que los signos principales del efecto no dependen de una sola partición muestral. En la tarea de emociones, la familia más fuerte pasó de 0.7598 a 0.7899 en Voronoi coseno, con intervalos al 95% de $[0.7374, 0.7818]$ y $[0.7707, 0.8121]$, respectivamente. En cambio, su mejor SVM pasó de 0.4550 a 0.4392, con intervalos de $[0.4295, 0.4779]$ y $[0.4126, 0.4628]$, lo que respalda la lectura de una mejora geométrica centrada en prototipos más que en fronteras supervisadas. En noticias, la caída es mucho más marcada: la misma familia pasa de 0.9119 a 0.4943 en Voronoi y de 0.9165 a 0.6019 en SVM, con intervalos claramente separados en ambos casos.

El análisis de sensibilidad sobre tres familias representativas también ayuda a interpretar por qué ocurre este cambio de signo. En emociones, al aumentar k disminuye la concentración espectral y aumenta el rango efectivo. Por ejemplo, en la familia más fuerte la razón $\lambda_1/\lambda_{\text{mean}}$ baja de 71.37 con $k = 0$ a 52.38 con $k = 3$ y a 36.07 con $k = 10$, mientras el rango efectivo sube de 411.01 a 443.07 y luego a 487.72. En paralelo, Voronoi mejora de 0.7640 a 0.7938 y alcanza 0.8319 con $k = 10$, pero SVM cae de 0.4616 a 0.4426 y después a 0.3890. En noticias ocurre lo contrario: para la misma familia, Voronoi pasa de 0.9096 con $k = 0$ a 0.4938 con $k = 3$ y 0.4832 con $k = 10$, mientras SVM cae de 0.9174 a 0.6012 y luego a 0.4169. Es decir, hacer el espacio más isotrópico no implica necesariamente conservar la señal discriminativa relevante.

E. Tablas detalladas por familia de evaluación

Las Tablas VI–XI condensan el comportamiento detallado de las seis familias de embeddings. En todos los casos, O indica el espacio original e I el espacio corregido. Para mejorar la lectura, Voronoi y SVM se integran en una misma tabla por colección, mientras que clustering se reporta por separado.

En este caso se aprecia una mejora generalizada tanto en Voronoi como en SVM. Las ganancias son particularmente visibles en mE5-base y mBERT, lo que sugiere que la corrección favorece tanto la organización alrededor de centroides como la separabilidad supervisada cuando el espacio original presenta anisotropía más marcada.

El clustering sigue siendo la familia más débil. Aunque hay mejoras puntuales, la transformación no produce un patrón tan estable como en Voronoi o SVM, lo que refuerza la idea de que separabilidad supervisada y estructura de clusters no son propiedades equivalentes.

En esta colección la mejora geométrica beneficia sobre todo a Voronoi. La ganancia es especialmente visible en mBERT, mientras que SVM se mantiene casi estable o incluso cae en algunos modelos fuertes. Esto indica que la corrección mejora

ISSN 2395-8618

TABLE V
COMPARACIÓN CON RESULTADOS PREVIOS DE LA LITERATURA BASADOS EN MÉTODOS NO NEURONALES O DÉBILMENTE SUPERVISADOS. EN NOTICIAS, LOS TRABAJOS PREVIOS REPORTAN EXACTITUD, MIENTRAS QUE AQUÍ SE REPORTA MACRO-F_1 SOBRE UN SUBCONJUNTO BALANCEADO; POR ELLO, LA COMPARACIÓN ES SÓLO APROXIMADA.

Colección	Método	Métrica	Resultado
Noticias	n-gramas + TF–IDF + regresión logística multinomial [11]	Acc.	92.36
	SBERT + SVM [12]	Acc.	92.10
	TF–IDF + k-means [12]	Acc.	49.50
	DocSCAN [12]	Acc.	84.10
	TF–IDF + regresión logística sobre split estándar	Acc.	91.89
	TF–IDF + SVM lineal sobre split estándar	Acc.	92.34
	TF–IDF + k-means sobre split estándar	Acc.	43.36
	Mejor SVM original de este estudio	Macro-F_1	0.9166
	Mejor clustering original de este estudio	Macro-F_1	0.8662
Emociones	Línea base BERT sobre la tarea oficial [9]	Macro-F_1	0.46
	Mejor modelo benchmarkeado por Alvarez-Gonzalez et al. [10]	Macro-F_1	0.47
	Mismo benchmark, mejor micro-F_1 [10]	Micro-F_1	0.57
	TF–IDF + regresión logística OvR en la tarea oficial	Macro-F_1	0.2014
	TF–IDF + regresión logística OvR en la tarea oficial	Micro-F_1	0.3871
	Mejor SVM original de este estudio	Macro-F_1	0.4583
	Mejor Voronoi corregido de este estudio	Macro-F_1	0.7938

TABLE VI
COMPORTAMIENTO DE VORONOI Y SVM EN EL CORPUS EN ESPAÑOL. SE REPORTA MACRO-F_1 PARA LAS TRES DISTANCIAS DE VORONOI Y PARA SVM LINEAL Y RBF, ANTES (O) Y DESPUÉS (I) DE LA CORRECCIÓN GEOMÉTRICA.

	Coseno		Euclidiana		Manhattan		SVM lineal		SVM RBF	
Modelo	O	I	O	I	O	I	O	I	O	I
OpenAI 3-L	0.9599	0.9758	0.9599	0.9758	0.9590	0.9800	0.9734	0.9916	0.9603	0.9701
MiniLM	0.6486	0.7174	0.6486	0.7174	0.6427	0.7118	0.6481	0.7497	0.6329	0.6611
MPNet	0.6488	0.7178	0.6488	0.7178	0.6391	0.7094	0.6988	0.7910	0.7069	0.7409
mE5-base	0.9242	0.9500	0.9242	0.9500	0.9211	0.9441	0.7360	0.9850	0.8066	0.9288
BGE-small	0.7462	0.7970	0.7462	0.7970	0.7526	0.7964	0.6561	0.8593	0.6772	0.7320
mBERT	0.6485	0.7263	0.6485	0.7263	0.6378	0.7214	0.3630	0.6660	0.3792	0.3805

TABLE VII
COMPORTAMIENTO DE CLUSTERING EN EL CORPUS EN ESPAÑOL. SE REPORTA MACRO-F_1 POR ALGORITMO, ANTES (O) Y DESPUÉS (I) DE LA CORRECCIÓN GEOMÉTRICA.

	K-means		Ward		Promedio coseno		Espectral		BIRCH	
Modelo	O	I	O	I	O	I	O	I	O	I
OpenAI 3-L	0.0022	0.0022	0.0219	0.0151	0.0646	0.1177	0.0022	0.0057	0.0471	0.0236
MiniLM	0.0452	0.0380	0.0909	0.0517	0.1751	0.1725	0.0328	0.0420	0.1431	0.1615
MPNet	0.0994	0.0711	0.1578	0.1200	0.2189	0.2036	0.0704	0.0667	0.2129	0.2139
mE5-base	0.0022	0.0022	0.0072	0.0089	0.0222	0.0286	0.0022	0.0022	0.0022	0.0115
BGE-small	0.0133	0.0082	0.0124	0.0119	0.0933	0.0865	0.0039	0.0084	0.0738	0.0192
mBERT	0.0040	0.0022	0.0022	0.0054	0.0022	0.0047	0.0080	0.0105	0.0022	0.0080

TABLE VIII
COMPORTAMIENTO DE VORONOI Y SVM EN LA COLECCIÓN DE EMOCIONES. SE REPORTA MACRO-F_1 PARA LAS TRES DISTANCIAS DE VORONOI Y PARA SVM LINEAL Y RBF, ANTES (O) Y DESPUÉS (I) DE LA CORRECCIÓN GEOMÉTRICA.

	Coseno		Euclidiana		Manhattan		SVM lineal		SVM RBF	
Modelo	O	I	O	I	O	I	O	I	O	I
OpenAI 3-L	0.7626	0.7938	0.7626	0.7938	0.7747	0.7883	0.4541	0.4233	0.4583	0.4426
MiniLM	0.6068	0.6506	0.6068	0.6506	0.5945	0.6373	0.3247	0.2961	0.3183	0.3181
MPNet	0.6010	0.6440	0.6010	0.6440	0.6001	0.6188	0.3018	0.2753	0.2971	0.2934
mE5-base	0.5762	0.6221	0.5762	0.6221	0.5639	0.6040	0.2855	0.3207	0.3150	0.3216
BGE-small	0.5793	0.5965	0.5793	0.5965	0.5861	0.5925	0.3703	0.3204	0.3766	0.3424
mBERT	0.3505	0.5140	0.3505	0.5140	0.3451	0.4998	0.1852	0.2243	0.2015	0.2267

la geometría de prototipos más de lo que mejora la frontera supervisada.

TABLE IX
COMPORTAMIENTO DE CLUSTERING EN LA COLECCIÓN DE EMOCIONES. SE REPORTA MACRO-F_1 POR ALGORITMO, ANTES (O) Y DESPUÉS (I) DE LA CORRECCIÓN GEOMÉTRICA.

	K-means		Ward		Promedio coseno		Espectral		BIRCH	
Modelo	O	I	O	I	O	I	O	I	O	I
OpenAI 3-L	0.2183	0.2070	0.1795	0.2021	0.1128	0.1655	0.2565	0.2187	0.2082	0.2014
MiniLM	0.1486	0.1621	0.1188	0.1258	0.0877	0.1390	0.1412	0.1465	0.1180	0.1249
MPNet	0.1628	0.1454	0.1362	0.1348	0.1010	0.1366	0.1310	0.1657	0.1360	0.1284
mE5-base	0.1360	0.1366	0.1442	0.1058	0.0697	0.1141	0.1392	0.1410	0.0025	0.1345
BGE-small	0.1656	0.1640	0.1501	0.1581	0.0719	0.1398	0.1837	0.1696	0.1438	0.1669
mBERT	0.1014	0.0992	0.1002	0.1413	0.0582	0.1043	0.1035	0.1404	0.1060	0.1271

TABLE X
COMPORTAMIENTO DE VORONOI Y SVM EN LA COLECCIÓN DE NOTICIAS. SE REPORTA MACRO-F_1 PARA LAS TRES DISTANCIAS DE VORONOI Y PARA SVM LINEAL Y RBF, ANTES (O) Y DESPUÉS (I) DE LA CORRECCIÓN GEOMÉTRICA.

	Coseno		Euclidiana		Manhattan		SVM lineal		SVM RBF	
Modelo	O	I	O	I	O	I	O	I	O	I
OpenAI 3-L	0.9123	0.4938	0.9123	0.4938	0.9096	0.4883	0.9140	0.2891	0.9166	0.6012
MiniLM	0.8796	0.4529	0.8796	0.4529	0.8681	0.4480	0.8694	0.3090	0.8817	0.6360
MPNet	0.8720	0.5012	0.8720	0.5012	0.8762	0.4881	0.8833	0.4000	0.8867	0.6973
mE5-base	0.8964	0.5076	0.8964	0.5076	0.8883	0.4915	0.9032	0.3734	0.9015	0.6047
BGE-small	0.8815	0.4598	0.8815	0.4598	0.8830	0.4523	0.8807	0.2883	0.8833	0.5482
mBERT	0.8491	0.5115	0.8491	0.5115	0.8549	0.5098	0.8682	0.4270	0.8747	0.7307

TABLE XI
CLUSTERING EN NOTICIAS. MACRO-F_1 ANTES (O) Y DESPUÉS (I).

	K-means		Ward		Promedio coseno		Espectral		BIRCH	
Modelo	O	I	O	I	O	I	O	I	O	I
OpenAI 3-L	0.8501	0.2515	0.8291	0.2876	0.3756	0.3621	0.5916	0.3104	0.8244	0.3528
MiniLM	0.8474	0.2550	0.7939	0.2617	0.3674	0.3214	0.8255	0.2479	0.5055	0.3270
MPNet	0.8089	0.3249	0.8059	0.3024	0.3628	0.2894	0.5615	0.3246	0.7819	0.3041
mE5-base	0.8662	0.2801	0.8014	0.2896	0.1052	0.3057	0.6191	0.2582	0.1000	0.3267
BGE-small	0.8530	0.2624	0.8040	0.2919	0.1067	0.2765	0.8352	0.2532	0.7994	0.2353
mBERT	0.8075	0.3921	0.8065	0.2986	0.1085	0.4245	0.5929	0.2740	0.1351	0.3894

El clustering muestra aquí un comportamiento mixto y mucho menos estable que Voronoi. Hay mejoras por algoritmo y por modelo, pero no aparece una tendencia dominante comparable a la observada en la organización por centroides.

ISSN 2395-8618

En noticias aparece el patrón más contundente: la corrección geométrica reduce de forma drástica el rendimiento en Voronoi y SVM, lo que indica que las direcciones removidas contenían información temática relevante.

El efecto negativo también se mantiene en clustering, aunque con variación entre algoritmos.

V. Discusión

La evidencia conjunta descarta una interpretación universal del postprocesamiento geométrico. Si sólo se observara el corpus en español, la conclusión natural sería que eliminar direcciones principales dominantes casi siempre ayuda. En ese escenario hubo ganancias consistentes en Voronoi y, sobre todo, en SVM, con un aumento promedio de $+0.1419$ en macro-F_1 para clasificación supervisada. Sin embargo, ese mismo corpus es demasiado fácil para una representación sparse clásica, por lo que no debe tomarse como evidencia principal de generalización sino como un caso controlado de alta separabilidad.

Sin embargo, al ampliar la observación a las otras dos colecciones, esa lectura deja de sostenerse. La tarea de emociones muestra que la corrección puede mejorar la organización alrededor de prototipos sin traducirse en una mejora clara para clasificación supervisada. Esto sugiere que Voronoi y SVM responden a propiedades distintas: Voronoi mejora cuando las clases se vuelven más compatibles con un esquema de centroides, mientras que SVM sólo mejora si la transformación preserva o refuerza fronteras discriminativas útiles.

La tarea de noticias revela el modo de falla opuesto. Aquí ya no basta con decir que la corrección "probablemente" eliminó información relevante. El operador aplicado puede escribirse como una proyección lineal $R_k = V(I - P_k)V^\top$ sobre el subespacio ortogonal a las primeras componentes principales. Si dos clases c y c' tienen medias μ_c y $\mu_{c'}$, entonces su separación entre centroides pasa de $\|\mu_c - \mu_{c'}\|$ a $\|R_k(\mu_c - \mu_{c'})\|$. Por tanto, si una parte sustancial de la diferencia entre clases está alineada con las direcciones removidas, la distancia entre centroides y el margen supervisado necesariamente se reducen. Eso es exactamente lo que sugieren los resultados: al aumentar k, la razón $\lambda_1/\lambda_{\text{mean}}$ cae y el rango efectivo sube, pero Voronoi y SVM empeoran de forma simultánea. En la familia más fuerte de noticias, por ejemplo, Voronoi cae de 0.9096 a 0.4938 y SVM de 0.9174 a 0.6012 entre $k = 0$ y $k = 3$. El punto no es sólo que el espacio se vuelva más isotrópico, sino que la proyección está suprimiendo direcciones que también contienen separación entre clases.

Esta lectura es consistente con una parte importante de la literatura reciente. Por un lado, varios trabajos advierten que calibrar isotropía no produce mejoras uniformes entre modelos y tareas [1], y que la anisotropía no debe tratarse como causa única del desempeño semántico [2]. Por otro, Mickus et al. [3] muestran que imponer isotropía puede entrar en tensión con la presencia de clusters y con objetivos de clasificación lineal. Nuestros resultados encajan con esa interpretación: la corrección geométrica puede ayudar cuando atenúa sesgos globales dominantes, pero puede perjudicar cuando las direcciones removidas también contienen señal discriminativa relevante para separar clases.

Estas observaciones también aclaran el papel metodológico de Voronoi. No pretende sustituir a una evaluación train/test, sino responder otra pregunta: ¿cuánto de la organización de clases ya está presente en la geometría del espacio si cada clase se representa con un solo prototipo? En ese sentido, su aportación principal no es competir con SVM, sino diagnosticar si la estructura del embedding es compatible con una partición prototípica estable. En el corpus en español y en la tarea de emociones, Voronoi siguió siendo informativo e incluso mejoró con la corrección. En noticias, en cambio, su caída después de la transformación señaló que se había dañado precisamente la estructura métrica de centroides que hacía útil al espacio original.

El comportamiento de clustering encaja con esta lectura y ayuda a explicar por qué sus resultados son más bajos e inestables. Un clasificador supervisado sólo necesita una frontera que separe etiquetas; un algoritmo no supervisado necesita, además, que la densidad inducida por el embedding forme grupos compactos y bien desconectados sin conocer las clases. Esa exigencia es bastante más fuerte. Por eso puede haber espacios donde Voronoi y SVM sean razonables, pero k-means, Ward o BIRCH no recuperen bien las clases. La conclusión no es que clustering sea irrelevante, sino que está midiendo otra propiedad: concordancia entre estructura no supervisada y etiquetas, no sólo separabilidad discriminativa.

De esta evidencia se desprende un criterio práctico preliminar. La corrección geométrica parece más conveniente cuando el espacio presenta alta concentración espectral, clases potencialmente multimodales y una mejora inicial de Voronoi al remover pocas componentes, porque en ese régimen la anisotropía dominante actúa más como sesgo global que como señal de clase. En cambio, cuando hay pocas clases temáticas bien separadas y las primeras componentes ya sostienen gran parte de la estructura global discriminativa, volver el espacio más isotrópico puede ser perjudicial. En términos operativos, una razón $\lambda_1/\lambda_{\text{mean}}$ alta no basta por sóla para justificar la corrección; debe leerse junto con la respuesta temprana de Voronoi y SVM para k pequeños.

También conviene delimitar el alcance empírico del estudio. Aunque las seis familias evaluadas cubren estilos de embedding heterogéneos, incluyendo modelos propietarios, sentence-transformers y encoders multilingües, no constituyen un inventario exhaustivo del estado del arte. Esta decisión es coherente con observaciones de benchmarks amplios, donde no emerge un embedding universalmente dominante para todas las tareas [7]. Como extensión futura, sería razonable incorporar familias multilingües más recientes y más generales, como las representadas por M3-Embedding [13], para evaluar si el patrón observado se mantiene bajo encoders de mayor cobertura funcional.

ISSN 2395-8618

VI. Conclusiones

Este trabajo deja una lección simple pero robusta: el postprocesamiento geométrico de embeddings depende del tipo de colección. En el corpus en español mejoró de manera consistente Voronoi y SVM, aunque ese resultado debe relativizarse por la facilidad del recurso. En la tarea de emociones ayudó sobre todo a Voronoi y apenas modificó SVM. En la tarea de noticias deterioró las tres familias de evaluación en todos los modelos.

La implicación práctica es directa. La regularización geométrica debe tratarse como una hipótesis a validar y no como un paso por defecto. Si la tarea depende de geometría de prototipos, asignación al centroide más cercano o particiones basadas en centroides, la intervención puede ayudar. Si el espacio original ya codifica separación temática fuerte, la misma corrección puede ser perjudicial.

Desde el punto de vista computacional, el postprocesamiento también introduce un costo adicional respecto del uso directo del encoder, ya que requiere al menos centrado, estimación de componentes principales y reconstrucción del espacio. En este trabajo esa etapa se ejecutó offline y no se evaluó como restricción principal. Sin embargo, en escenarios de gran escala o de baja latencia convendría explorar variantes truncadas o aleatorizadas de PCA, cuya literatura muestra ventajas claras para descomposiciones aproximadas sobre matrices grandes [14].

Como trabajo futuro, resulta natural estudiar si el signo del efecto puede anticiparse a partir de diagnósticos geométricos previos a la clasificación. En particular, conviene analizar de manera explícita la proyección de los centroides y de la dispersión intra-clase sobre las primeras componentes principales, para cuantificar cuánta separación entre clases está siendo removida por P_k. Si eso fuera posible, el análisis de isotropía podría pasar de ser una explicación posterior a convertirse en una señal útil para selección de modelo o de postprocesamiento.

References

[1] Y. Ding, K. Martinkus, D. Pascual, S. Clematide, and R. Wattenhofer, "On isotropy calibration of transformer models," in *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pp. 1–9, 2022.

[2] A. Fuster Baggetto and V. Fresno, "Is anisotropy really the cause of bert embeddings not being semantic?," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4271–4281, 2022.

[3] T. Mickus, S.-A. Gronroos, and J. Attieh, "Isotropy, clusters, and classifiers," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 75–84, 2024.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[5] J. Mu and P. Viswanath, "All-but-the-top: Simple and effective postprocessing for word representations," in *International Conference on Learning Representations*, 2018.

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of EMNLP-IJCNLP*, 2019.

[7] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[9] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020.

[10] N. Alvarez-Gonzalez, A. Kaltenbrunner, and V. Gomez, "Uncovering the limits of text-based emotion detection," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2560–2583, 2021.

[11] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems*, 2015.

[12] D. Stammbach and E. Ash, "Docscan: Unsupervised text classification via learning from neighbors," in *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, 2022.

[13] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024.

[14] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.