Microagregación híbrida asistida por modelos de lenguaje

Sara Lanuza-Orna
Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili, Tarragona, España
sara.lanuza@urv.cat

Alberto Blanco-Justicia
Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili, Tarragona, España
alberto.blanco@urv.cat

***Resumen*—La microagregación es una técnica clásica del control estadístico de la divulgación (SDC) utilizada para garantizar el k-anonimato en microdatos. Sin embargo, los métodos tradicionales se centran en atributos numéricos y no capturan la semántica de los atributos textuales, reduciendo la utilidad de los datos anonimizados. Este trabajo presenta una microagregación híbrida que combina el algoritmo *Maximum Distance to Average Vector* (MDAV) con modelos de lenguaje preentrenados. En concreto, se emplea *Sentence-BERT*, una variante de BERT diseñada para generar representaciones semánticas de frases. Los atributos categóricos y numéricos se transforman conjuntamente en un espacio latente continuo, sobre el cual se aplica el procedimiento de agrupamiento de MDAV. La propuesta preserva las garantías de k-anonimato y mejora la coherencia semántica entre los registros agrupados. Los experimentos con datos mixtos muestran una menor pérdida de información y un coste computacional contenido respecto al MDAV clásico y a enfoques semánticos basados en ontologías.**

***Index Terms*—anonimización de datos, k-anonimato, microagregación, modelos de lenguaje, BERT, privacidad.**

I. Introducción

El creciente acceso a grandes volúmenes de datos de ámbitos sanitarios, gubernamentales y empresariales ha impulsado el desarrollo de métodos de anonimización que permitan compartir información sin comprometer la privacidad de las personas [1]. En este contexto, el *control de la confidencialidad estadística* (*Statistical Disclosure Control*, SDC) agrupa un conjunto de técnicas orientadas a evitar la reidentificación de individuos o la revelación de atributos sensibles mediante transformaciones controladas de los datos [2].

Las estrategias para alcanzar dicha protección suelen situarse entre dos extremos: los enfoques *utility-first* aplican métodos como la generalización, supresión o adición de ruido de forma heurística hasta alcanzar los niveles de protección suficientes para la publicación de los datos; por otro lado, los enfoques *privacy-first* se basan en modelos formales para guiar el proceso de anonimización. Entre dichos modelos, el k-anonimato se ha consolidado como uno de los enfoques más extendidos: garantiza que cada combinación de cuasi-identificadores aparezca en al menos k registros distintos [3], fijando así un riesgo máximo de reidentificación de $1/k$.

La microagregación permite asegurar el k-anonimato agrupando los registros en bloques de tamaño mínimo k y sustituyendo sus valores originales por un representante común, reduciendo el riesgo de reidentificación al tiempo que preserva parcialmente la estructura estadística del conjunto de datos.

Entre las técnicas de microagregación, el algoritmo *Maximum Distance to Average Vector* (MDAV) [4] se emplea como método para generar particiones que satisfacen el k-anonimato. Sin embargo, su aplicación en escenarios actuales presenta limitaciones relevantes. En particular, su formulación clásica opera exclusivamente con atributos numéricos y distancias euclídeas, lo que restringe su capacidad para tratar datos heterogéneos y, especialmente, variables categóricas cuyo contenido semántico no puede capturarse adecuadamente mediante métricas puramente geométricas.

En paralelo, los modelos de lenguaje preentrenados han transformado el procesamiento del lenguaje natural al generar representaciones vectoriales densas que capturan relaciones semánticas entre palabras, frases y conceptos. Entre ellos, BERT [5] destaca como un modelo basado en *transformers* que aprende representaciones contextuales profundas mediante tareas de enmascaramiento y predicción del contexto. Sobre esta arquitectura se construye *Sentence-BERT* [6], una variante entrenada con un esquema siamés para producir *embeddings* semánticos de alta calidad a nivel de frase o de registro completo. Estas representaciones permiten calcular similitudes en espacios vectoriales de forma consistente y eficiente, y resultan especialmente adecuadas en escenarios de anonimización en los que los atributos categóricos codifican información semántica que conviene preservar. Al incorporar esta información semántica en una codificación conjunta de los atributos categóricos y numéricos, es posible obtener datos anonimizados que mantienen relaciones conceptuales significativas sin recurrir a ontologías externas ni jerarquías específicas.

En este trabajo proponemos integrar representaciones semánticas generadas mediante modelos de lenguaje en el proceso de microagregación. Nuestro objetivo es extender MDAV a escenarios con datos mixtos que incluyen tanto atributos numéricos como categóricos, evitando la dependencia de recursos externos como ontologías o jerarquías. Para ello, cada registro completo se codifica de manera conjunta en un espacio vectorial mediante un modelo *Sentence-BERT* preentrenado, de modo que la información semántica y la información numérica quedan integradas en una única representación latente. Sobre estas representaciones se aplica el algoritmo MDAV, obteniendo una microagregación híbrida que mantiene las garantías de k-anonimato, mejora la coherencia interna de los grupos formados y reduce la pérdida de información en comparación con la microagregación tradicional y con métodos semánticos basados en ontologías [7].

II. Conceptos preliminares

II-A. Microagregación y el algoritmo MDAV

La microagregación transforma un conjunto de datos en grupos de al menos k registros, sustituyendo los valores originales por un representante común. Esta estrategia garantiza el k-anonimato, ya que cada registro queda indistinguible de al menos $k-1$ elementos adicionales. Entre los distintos métodos existentes, el algoritmo MDAV [4] destaca por su simplicidad y buena relación entre privacidad y preservación de utilidad.

MDAV opera de forma iterativa siguiendo una estrategia de selección de elementos extremos. En cada paso, se calcula el centroide de los registros aún no agrupados y se identifica el registro que se encuentra a mayor distancia de dicho centro. A continuación, se localiza un segundo registro que maximiza la distancia respecto al primero. Con cada uno de estos dos puntos de referencia se forman grupos tomando sus $k-1$ vecinos más próximos, lo que da lugar a dos bloques k-anónimos por iteración. El procedimiento continúa mientras queden al menos $3k$ registros por procesar. Cuando el número restante se sitúa entre $2k$ y $3k$, los datos se dividen en dos grupos finales; si es inferior a $2k$, se agrupan todos en un único bloque.

El procedimiento completo se resume en el Algoritmo 1.

Algorithm 1 MDAV

Entrada: Conjunto de datos X con n registros y m cuasi-identificadores; parámetro k.
Salida: Una partición de X que satisfaga el k-anonimato.

while $|X| \geq 3k$ **do**
  a. Selección del centroide
  Calcular el centroide $\bar{x}$ de los registros restantes en X.
  Sea x_r el registro más alejado de $\bar{x}$ (distancia euclídea).
  Sea x_s el registro más alejado de x_r.
  b. Formación de dos grupos
  Formar el grupo C_r con x_r y sus $k-1$ vecinos más cercanos.
  Eliminar de X los registros pertenecientes a C_r.
  Formar el grupo C_s con x_s y sus $k-1$ vecinos más cercanos.
  Eliminar de X los registros pertenecientes a C_s.
end while
if $2k \leq |X| < 3k$ **then**
  Formar un grupo con el registro más alejado del centroide actual y sus $k-1$ vecinos más cercanos.
  Asignar los registros restantes a un segundo grupo.
end if
if $0 < |X| < 2k$ **then**
  Formar un único grupo con todos los registros restantes.
end if

II-B. Medida de distorsión para datos mixtos: NCP

La evaluación de la pérdida de información introducida por la microagregación requiere métricas capaces de reflejar el grado de generalización aplicado a los datos. En escenarios con atributos numéricos y categóricos, una medida ampliamente utilizada es el *Normalized Certainty Penalty* (NCP) [8], que permite cuantificar la distorsión de forma homogénea en dominios heterogéneos.

Sea A_j un atributo y sea $A_j(C_z)$ el conjunto de valores que toma dentro del grupo C_z. La penalización asociada a dicho atributo en el grupo se define como:

$$\text{NCP}_{A_j}(C_z) = \begin{cases} \dfrac{\text{range}\,(A_j(C_z))}{\text{range}\,(A_j(T))}, & \text{si } A_j \text{ es numérico,} \\ \dfrac{|\text{dom}(A_j(C_z))|}{|\text{dom}(A_j)|}, & \text{si } A_j \text{ es categórico,} \end{cases} \tag{1}$$

donde el rango de un atributo en un conjunto C_z se define como:

$$\text{range}(A_j(C_z)) = \text{máx}(A_j(C_z)) - \text{mín}(A_j(C_z)). \tag{2}$$

En el caso de atributos numéricos, esta penalización mide la fracción del rango global ocupado por los valores presentes en el grupo. Para atributos categóricos, cuantifica el número de categorías distintas dentro del grupo con respecto al dominio total del atributo.

Una vez calculadas las penalizaciones por atributo y grupo, el NCP total del conjunto anonimizado T' se obtiene mediante el promedio ponderado:

$$\text{NCP}(T') = \sum_{j=1}^{m} w_j \frac{1}{n} \sum_{C_z \in \mathcal{C}} \sum_{\mathbf{x}_i \in C_z} \text{NCP}_{A_j}(C_z), \tag{3}$$

donde w_j representa la importancia relativa del atributo A_j. Valores de NCP próximos a cero indican baja distorsión y, por tanto, mayor utilidad de los datos anonimizados; valores próximos a uno reflejan una pérdida sustancial de información.

III. Trabajo relacionado

La microagregación se ha consolidado como una técnica eficaz para lograr k-anonimato gracias al equilibrio que ofrece entre simplicidad del procedimiento y utilidad de los datos resultantes. El algoritmo MDAV original [4] agrupa los datos usando distancias euclídeas y reemplaza cada grupo por su centroide, minimizando la suma de errores cuadráticos dentro de los grupos. Para mejorar la flexibilidad de la agrupación, V-MDAV permite tamaños de grupo variables entre k y $2k-1$ mediante un factor de ganancia que controla el crecimiento de los grupos y aumenta la homogeneidad intragrupo [9]. Posteriormente, F-MDAV introdujo técnicas de reutilización de normas, precomputación y ordenaciones parciales para reducir la complejidad computacional en entornos de *big data* sin incrementar la distorsión [10], mientras que estrategias de particionado recursivo dividen el conjunto de datos en macroceldas y aplican MDAV de manera local, reduciendo la complejidad de cuadrática a casi lineal para tamaños de datos muy grandes [11].

El tratamiento de datos heterogéneos ha dado lugar a variantes que combinan diferentes tipos de atributos en un marco unificado. MDAV-generic extiende el algoritmo a atributos continuos, ordinales y nominales mediante funciones de agregación específicas para cada tipo (como la media, la mediana o la moda) y procedimientos de normalización adecuados, de forma que se preserva en la medida de lo posible la semántica de cada dominio [12]. Por su parte, las versiones con ponderaciones y sensibilidades adaptativas introducen pesos distintos por atributo para reflejar su importancia relativa y su contribución al riesgo o a la pérdida de utilidad, ajustando la métrica de distancia al contexto de aplicación [13]. En cuanto a atributos categóricos, Martínez et al. propusieron *Semantic-MDAV* [7], que sustituye la distancia euclídea por una medida de similitud semántica basada en ontologías como WordNet y define centroides semánticos que resumen el contenido de cada grupo. Aunque esta aproximación mejora la coherencia de los grupos cuando existen jerarquías semánticas

bien definidas, su dependencia de ontologías externas y la necesidad de tratar manualmente valores no cubiertos por ellas limita su aplicabilidad y conlleva un coste computacional elevado.

IV. Microagregación híbrida propuesta

Nuestra propuesta, denominada *MDAV-SBERT*, integra modelos de lenguaje en el proceso de microagregación para gestionar de manera conjunta atributos numéricos y categóricos. La idea central consiste en codificar cada registro completo en un espacio vectorial común mediante un modelo de lenguaje preentrenado, obteniendo así una representación uniforme apta para el cálculo de similitudes. Sobre este espacio latente se aplica posteriormente el algoritmo MDAV clásico, manteniendo las garantías de k-anonimato y la lógica de agrupamiento basada en la selección de puntos extremos respecto al centroide.

Sea $\mathcal{D}$ un conjunto de n registros con atributos numéricos $\mathbf{x}_u m \in \mathbb{R}^p$ y atributos categóricos $x_t ext \in \mathcal{T}$. El método recibe como entrada el parámetro k y un modelo de lenguaje preentrenado LM (*Sentence-BERT*). A continuación se describen sus fases principales:

1. *Preprocesamiento:* Los valores numéricos y los atributos categóricos con valores vacíos o nulos se sustituyen por un marcador neutro. A continuación, los atributos de cada registro se concatenan para formar una cadena mixta $v_i \in V$. Por ejemplo, para un registro con valores (x_0, x_1, x_2) para los atributos *attr0*, *attr1*, y *attr2*, la cadena mixta resultante sería "'`attr0:` x_0`, attr1:` x_1`, attr2:` x_2'". Esta cadena se transforma posteriormente en una representación vectorial empleando el modelo de lenguaje, obteniéndose $S = \mathrm{LM}(V)$ mediante codificación por lotes para reducir el tiempo de ejecución.
   Reducción opcional de la dimensionalidad: El método contempla tres configuraciones para ajustar la dimensión latente empleada por MDAV: (i) *full*, que mantiene íntegramente la dimensión original del embedding; (ii) *sqrt*, que proyecta los vectores a $d = \lceil p^{1/2} \rceil$; (iii) *cube*, que proyecta a $d = \lceil p^{1/3} \rceil$. En todos los casos, p denota el número de atributos de los cuasi-identificadores en el esquema original (antes del preprocesamiento), y d se restringe a no superar la dimensión del embedding generado por LM.
2. *Agrupamiento de los registros:* Mientras existan suficientes elementos por agrupar, se calcula el centroide global en el espacio latente S y se identifica el registro más alejado de dicho centroide. A continuación se selecciona el registro más distante de este primero, y ambos actúan como puntos de referencia para formar grupos de tamaño k añadiendo sus $k-1$ vecinos más próximos. El procedimiento se repite de manera iterativa hasta asignar todos los registros.
3. *Generalización:* Para cada grupo C_j se calcula la media de los atributos numéricos y la moda de los atributos categóricos. Todos los registros del grupo se reemplazan por estos valores representativos, garantizando que los k elementos sean indistinguibles en los cuasi-identificadores.

Algorithm 2 MDAV-SBERT

Entrada: datos $\mathcal{D}$, parámetro k y modelo de lenguaje LM.
Salida: partición k-anónima.
a. Preprocesamiento
Sustituir valores vacíos por un marcador neutro y concatenar los atributos numéricos y categóricos de cada registro en una cadena mixta v_i.
Transformar las cadenas $v_i \in V$ en vectores latentes $S = \mathrm{LM}(V)$ mediante *Sentence-BERT*.
Inicializar el conjunto X a partir de S con todos los registros e índices y precomputar y almacenar normas en el conjunto X para la reutilización en el cálculo de las distancias entre registros.
b. Agrupamiento de los registros
while $|X| \geq 3k$ **do**
    Calcular el centroide global $\bar{z}$ de $\{z_i : i \in X\}$.
    Seleccionar x_r como el registro más alejado de $\bar{s}$.
    Seleccionar x_s como el registro más alejado de x_r.
    Formar el grupo C_r con x_r y sus $k-1$ vecinos más cercanos a x_r; eliminar C_r de X.
    Formar el grupo C_s con x_s y sus $k-1$ vecinos más cercanos a x_s; eliminar C_s de X.
end while

if $2k \leq |X| < 3k$ **then**
    Calcular el centroide $\bar{z}$ de los registros restantes.
    Formar un grupo C con el registro más alejado de $\bar{z}$ y sus $k-1$ vecinos más cercanos; eliminar C de X
else if $0 < |X| < 2k$ **then**
    Formar un único grupo C con todos los registros restantes.
end if

c. Generalización
for cada grupo C **do**
    Sustituir los atributos numéricos por su media y los categóricos por el valor más frecuente.
end for
return Conjunto anonimizado con grupos generalizados.

Además, se incorporan pequeñas variaciones al esquema clásico de MDAV con el fin de adaptarlo mejor al espacio latente generado por el modelo de lenguaje. En primer lugar, la proximidad entre registros y centroides se evalúa mediante la distancia del coseno,

$$d_{Sc}(\mathbf{c}, \mathbf{Y}) = 1 - \frac{\langle \mathbf{Y}, \mathbf{c} \rangle}{\|\mathbf{Y}\| \, \|\mathbf{c}\|}, \tag{4}$$

donde $\langle \mathbf{Y}, \mathbf{c} \rangle$ denota el producto escalar entre el centroide y los vectores restantes del dataset $Y = \{y_1, ..., y_n\}$, mientras que $\|\mathbf{c}\|$ y $\|\mathbf{Y}\|$ corresponden a sus normas euclídeas. Esta medida resulta especialmente adecuada en espacios de alta dimensión, donde el ángulo entre vectores captura relaciones semánticas con mayor fidelidad que la distancia euclídea. Asimismo, para reducir el coste computacional, se reutilizan cálculos intermedios, en particular la norma $\|\mathbf{Y}\|$ y los productos escalares, que permanecen invariantes mientras no cambian los centroides, evitando recalcular desde cero todas las distancias en cada iteración. Finalmente, la selección de los $k-1$ vecinos más próximos se acelera mediante técnicas de ordenación parcial, que permiten identificar los elementos requeridos sin ordenar completamente el conjunto restante, manteniendo el comportamiento del algoritmo original pero con una mejora apreciable en eficiencia.

V. Evaluación experimental

V-A. Metodología experimental

El conjunto de datos *Adult* de la UCI contiene 48 842 registros y 14 atributos demográficos y socioeconómicos. Para la evaluación se definieron dos configuraciones de cuasi-identificadores (QI):

- *QI7*: conjunto compuesto por siete atributos categóricos de tipo demográfico y socioeconómico (*workclass*,

marital-status, *occupation*, *relationship*, *race*, *sex* y *native-country*). Este escenario reproduce un conjunto de QI estrictamente categóricos.

- *QI14*: conjunto formado por la totalidad de los atributos disponibles. Este escenario representa un caso mixto que combina atributos numéricos y categóricos. Solo se evaluó *MDAV-SBERT*, dado que *Semantic-MDAV* no puede operar con atributos numéricos.

V-A1. Variaciones experimentales: En ambos escenarios se analizó el comportamiento de *MDAV-SBERT* variando el tamaño del lote (*batch size*) empleado durante la codificación del modelo preentrenado, con valores de 128, 256, 512 y 1024, así como la dimensionalidad de la proyección latente. Para ello, se evaluaron las siguientes variantes:

- *Semantic-MDAV*: implementación de referencia basada en ontologías léxicas, donde la distancia euclídea se sustituye por una medida semántica (por ejemplo, Wu–Palmer) y los centroides se construyen a partir de WordNet. Este método solo puede aplicarse en escenarios puramente categóricos.
- *MDAV-SBERT (full)*: variante que emplea los embeddings completos generados por *Sentence-BERT*, sin aplicar reducción dimensional.
- *MDAV-SBERT (sqrt)*: variante que proyecta los embeddings a una dimensión latente $d_{\text{sqrt}} = \lceil p^{1/2} \rceil$, donde p es el número de atributos de los cuasi-identificadores en su definición original (antes del preprocesamiento). Si d_{sqrt} supera la dimensión del embedding generado por el modelo de lenguaje, se utiliza esta última como cota superior.
- *MDAV-SBERT (cube)*: variante análoga que utiliza $d_{\text{cube}} = \lceil p^{1/3} \rceil$, lo que induce una compresión más agresiva de la representación latente sin exceder la dimensión original del embedding.

V-A2. Diseño experimental: El parámetro de privacidad k toma valores entre 10 y 100 en incrementos regulares, con el fin de analizar la evolución del compromiso entre privacidad y utilidad. Para cada combinación de método, configuración de cuasi-identificadores, tamaño de lote, esquema de truncamiento y valor de k, se realizaron tres ejecuciones independientes y se reportó la media de las métricas obtenidas.

Se llevaron a cabo tres experimentos complementarios:

1. *Evaluación de utilidad*: mediante el *Normalized Certainty Penalty* (NCP), calculado sobre el conjunto completo de cuasi-identificadores.
2. *Análisis de tiempos por fase*: evaluando el coste computacional asociado a cada una de las tres etapas del proceso:
   - *codificación*: obtención de los embeddings generados por *Sentence-BERT* para los atributos categóricos y numéricos, así como la posible proyección a una dimensión latente reducida;
   - *microagregación*: ejecución del algoritmo MDAV sobre las representaciones resultantes;
   - *generalización*: sustitución de los valores originales por los correspondientes centroides numéricos o valores modales en atributos categóricos.

   Cada gráfica de tiempos por fase muestra el tiempo medio de ejecución de cada etapa, obtenido promediando los valores registrados para todos los niveles de anonimización comprendidos entre $k = 10$ y $k = 100$.
3. *Tiempo medio por k*: análisis detallado de la evolución del coste computacional a medida que aumenta el nivel de anonimización impuesto por el parámetro k.

Todos los experimentos se ejecutaron en un entorno WSL con Ubuntu, equipado con 32 GiB de memoria RAM y ejecutándose sobre un sistema dotado de un procesador AMD Ryzen 9 9900X (12 núcleos, 24 hilos) y una GPU NVIDIA GeForce RTX 2060.

V-B. Resultados y discusión

V-B1. Resultados para QI7: La Fig. 1 muestra la evolución del NCP en función de k para la configuración *QI7*. Se observa que todas las variantes de *MDAV-SBERT* presentan niveles de distorsión notablemente inferiores a los de *Semantic-MDAV*. Mientras que este último supera el 70 % de NCP incluso para valores moderados de k, las variantes basadas en modelos de lenguaje se sitúan en el intervalo aproximado del 20–47.5 %. En conjunto, los resultados indican que las representaciones latentes generadas por *MDAV-SBERT* permiten obtener agrupamientos más consistentes que los derivados de la distancia ontológica empleada en *Semantic-MDAV*.

En cuanto a las variantes de truncamiento, las tres configuraciones (*full*, *sqrt* y *cube*) exhiben una evolución del NCP con forma similar, aunque con desplazamientos sistemáticos en su nivel absoluto. La versión *cube* presenta consistentemente los valores más elevados dentro del grupo de métodos basados en BERT, seguida por *sqrt*, mientras que la configuración *full* obtiene la distorsión más reducida. No obstante, incluso las variantes con mayor reducción dimensional se mantienen claramente por debajo de los valores obtenidos mediante *Semantic-MDAV*.

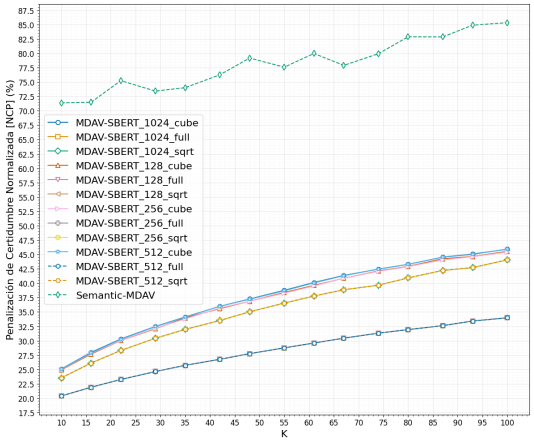


Figura 1. Evolución del NCP en función de k para la configuración *QI7*.

El desglose temporal por fases, representado en la Fig. 2, muestra que la codificación BERT constituye la parte dominante del coste total en todas las configuraciones. La fase de MDAV incrementa su tiempo en las variantes sin reducción dimensional (*full*), consecuencia directa de operar en un espacio latente de mayor dimensión; aun así, su contribución permanece claramente por debajo de la fase de codificación. El tamaño del lote presenta un impacto limitado: los tiempos de codificación son casi constantes en los cuatro valores

considerados y, en la fase de MDAV, únicamente se observa una ligera mejora al aumentar el tamaño de lote en las variantes *full*. El lote de tamaño 128 presenta un tiempo de codificación ligeramente superior.

Este comportamiento sugiere que, para la configuración *QI7*, el coste computacional en la fase de codificación está dominado por componentes que no se benefician de manera significativa del incremento del paralelismo en GPU al aumentar el *batch size*. La penalización observada para el lote de tamaño 128 es coherente con una menor eficiencia computacional asociada a lotes pequeños, mientras que para tamaños de 256 o superiores no se aprecian mejoras significativas, lo que indica un comportamiento estable en esta fase.

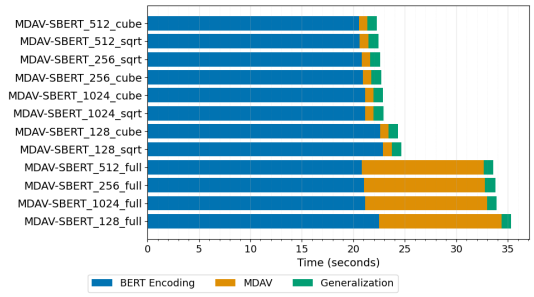


Figura 2. Desglose temporal por fases (codificación, MDAV y generalización) para *QI7*.

La Fig. 3 muestra el tiempo medio total en función de k. *Semantic-MDAV* resulta significativamente menos eficiente que cualquier variante de *MDAV-SBERT*; por ejemplo, para $k = 10$ la relación aproximada es de $\frac{2000\,\text{s}}{25\text{–}30\,\text{s}} \approx 67\text{–}80\times$. Entre las variantes basadas en BERT, las representaciones sin reducción (*full*) presentan tiempos superiores debido a la mayor dimensionalidad del espacio latente, en consonancia con lo observado en la Fig. 2. No obstante, todas las variantes se mantienen en el rango de decenas de segundos y muestran una tendencia decreciente conforme aumenta k, consecuencia del menor número de grupos a formar para niveles de anonimización más elevados.

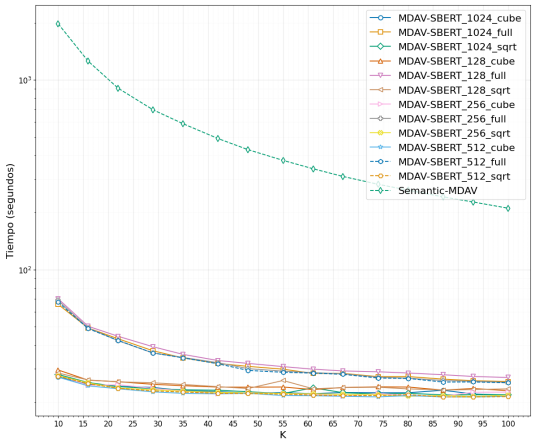


Figura 3. Tiempo medio de ejecución en función de k para *QI7*, con el eje vertical representado en escala logarítmica.

V-B2. Resultados para QI14: La Fig. 4 presenta los valores de NCP para la configuración *QI14*, que incluye todos los atributos categóricos y numéricos. En comparación con *QI7*, los niveles de distorsión abarcan un rango más amplio (aprox. 21–64 %), reflejando el incremento de complejidad al considerar un conjunto más extenso y heterogéneo de cuasi-identificadores. Las diferencias entre las variantes de truncamiento reproducen el patrón previamente observado: *cube* induce mayor pérdida de utilidad, *full* obtiene los valores más bajos de NCP y *sqrt* se sitúa en una posición intermedia.

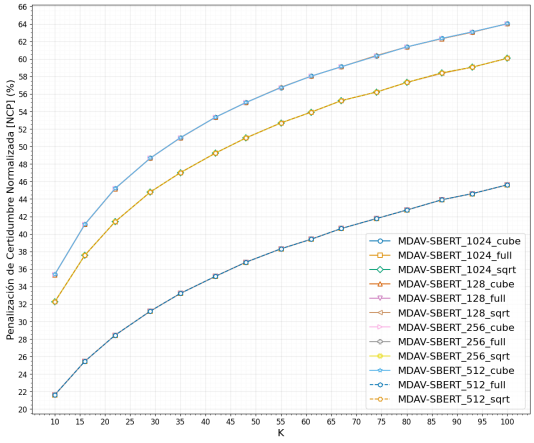


Figura 4. Evolución del NCP en función de k para la configuración *QI14*.

La Fig. 5 muestra el tiempo desglosado por fases. El comportamiento general coincide con el observado en *QI7*: la codificación BERT domina el tiempo total, mientras que la fase de MDAV incrementa su coste en la variante *full* debido a la mayor dimensionalidad. En comparación con *QI7*, los tiempos totales son ligeramente superiores, lo cual es esperable por el mayor número de atributos considerados en *QI14*. El tamaño de lote tiene un impacto reducido y las diferencias apreciables se concentran en las variantes sin reducción. A diferencia de *QI7*, el lote de tamaño 128 no presenta un desfase significativo respecto al resto.

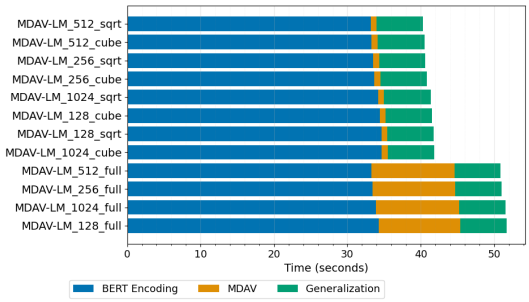


Figura 5. Desglose temporal por fases para *QI14*.

El tiempo medio total en función de k, representado en la Fig. 6, muestra también una tendencia decreciente a medida que aumenta el nivel de anonimización. Las variantes *full* presentan tiempos mayores, mientras que *sqrt* y *cube* reducen de forma apreciable el coste total con un impacto moderado en la utilidad. En todos los casos los tiempos se mantienen en el rango de decenas de segundos, con un comportamiento análogo al observado en *QI7*.

V-B3. Discusión conjunta: Los resultados evidencian que *MDAV-SBERT* ofrece un compromiso favorable entre utilidad y coste computacional en ambas configuraciones analizadas.

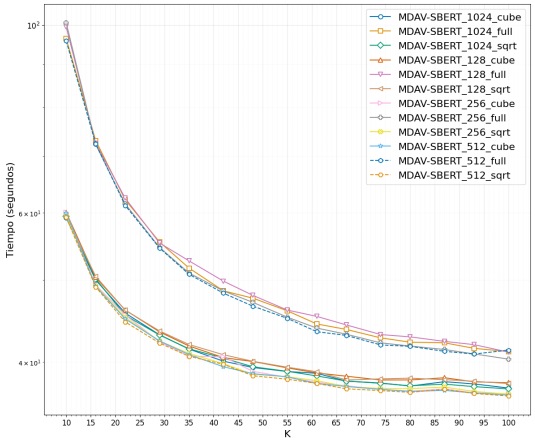


Figura 6. Tiempo medio de ejecución en función de k para *QI14*, con el eje vertical representado en escala logarítmica.

Como cabía esperar, la distorsión medida mediante NCP es mayor en *QI14* que en *QI7*, reflejando la mayor complejidad y heterogeneidad del conjunto de cuasi-identificadores. Sin embargo, en ambos casos las variantes basadas en modelos de lenguaje superan de manera consistente a *Semantic-MDAV*, tanto en términos de utilidad como de eficiencia computacional.

Las diferencias entre las variantes de *MDAV-SBERT* son estables en las dos configuraciones: la versión *full* logra la menor distorsión, seguida por *sqrt*, mientras que *cube* concentra la mayor pérdida de utilidad. No obstante, incluso las configuraciones más agresivas en la reducción dimensional mantienen niveles de NCP muy inferiores a los de *Semantic-MDAV*, lo que indica que las proyecciones latentes compactas preservan suficiente estructura semántica para producir agrupamientos coherentes.

Desde la perspectiva computacional, las configuraciones *sqrt* y *cube* reducen de manera significativa el tiempo total frente a la variante *full*, especialmente en *QI14*, donde la dimensionalidad del embedding original es mayor. En cualquier caso, todas las variantes de *MDAV-SBERT* se mantienen en un rango de ejecución de decenas de segundos, muy inferior al requerido por *Semantic-MDAV*. En conjunto, los resultados demuestran que la integración de modelos de lenguaje en microagregación constituye una alternativa eficaz, flexible y computacionalmente viable para el tratamiento de datos categóricos y mixtos.

VI. Conclusiones

Se ha presentado una técnica de microagregación híbrida basada en modelos de lenguaje que extiende el algoritmo MDAV a conjuntos de datos mixtos. La propuesta codifica conjuntamente los atributos categóricos y numéricos mediante *Sentence-BERT*, utilizando las representaciones obtenidas como espacio latente sobre el que se ejecuta el esquema de microagregación de MDAV con reutilización de distancias. Asimismo, se han incorporado configuraciones de reducción de dimensionalidad basadas en truncamientos dependientes del número de cuasi-identificadores, lo que permite ajustar automáticamente la dimensión latente para equilibrar coste computacional y utilidad.

La evaluación experimental en el conjunto *Adult* para las configuraciones *QI7* y *QI14* muestra que las variantes de *MDAV-SBERT* obtienen valores de *Normalized Certainty Penalty* inferiores y tiempos de ejecución considerablemente más reducidos que *Semantic-MDAV*. En particular, las configuraciones con reducción dimensional mantienen una utilidad muy próxima a la versión sin truncamiento, al tiempo que disminuyen de forma significativa el tiempo de codificación. Estos resultados ponen de manifiesto que la incorporación de modelos de lenguaje preentrenados en microagregación ofrece un compromiso adecuado entre privacidad, utilidad y eficiencia, sin depender de recursos ontológicos externos.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el MCIN/AEI/10.13039/501100011033 y por "ERDF A way of making Europe", en el marco de los proyectos PID2021-123637NB-I00 ("CURLING") y PID2024-156914OB-C42 ("CLEARING-IT"), así como por los fondos NextGenerationEU/PRTR a través de INCIBE (proyecto "HERMES" y Cátedra INCIBE-URV de ciberseguridad). La primera autora cuenta con el apoyo de una beca Martí i Franquès (2025PMF-PIPF-13) de la Universitat Rovira i Virgili.

Referencias

[1] A. Badshah, A. Daud, R. Alharbey, A. Banjar, A. Bukhari, and B. Alshemaimri, "Big data applications: overview, challenges and future," *Artificial Intelligence Review*, vol. 57, no. 11, p. 290, 2024.

[2] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf, *Statistical disclosure control*. John Wiley & Sons, 2012.

[3] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[4] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Transactions on Knowledge and data Engineering*, vol. 14, no. 1, pp. 189–201, 2002.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[7] S. Martínez, D. Sánchez, and A. Valls, "Semantic adaptive microaggregation of categorical microdata," *Computers & Security*, vol. 31, no. 5, pp. 653–672, 2012.

[8] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 785–790.

[9] A. Solanas, A. Martínez-Ballesté, and J. Domingo-Ferrer, "V-mdav: A multivariate microaggregation with variable group size," pp. 917–925, 2006.

[10] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A. M. Mezher, J. Parra-Arnau, and J. Forné, "The fast maximum distance to average vector (f-mdav): An algorithm for k-anonymous microaggregation in big data," *Engineering Applications of Artificial Intelligence*, vol. 90, p. 103531, 2020.

[11] E. Pallarès, D. Rebollo-Monedero, A. Rodríguez-Hoyos, J. Estrada-Jiménez, A. M. Mezher, and J. Forné, "Mathematically optimized, recursive prepartitioning strategies for k-anonymous microaggregation of large-scale datasets," *Expert Systems with Applications*, vol. 144, p. 113086, 2020.

[12] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 195–212, 2005.

[13] Z. Zhang, B. Han, H.-C. Chao, F. Sun, L. Uden, and D. Tang, "A new weight and sensitivity based variable maximum distance to average vector algorithm for wearable sensor data privacy protection," *IEEE Access*, vol. 7, pp. 104 045–104 056, 2019.