Tipo de Artículo: *Artículo de Investigación Original*
Área: *Ciencias Exactas*

Análisis estadístico comparativo: Adecuación semántica CV-oferta entre ATS basados en TF-IDF+coseno y enfoque asistido por IA

Comparative Statistical Analysis: Semantic Adequacy between CV and Job Offer across TF-IDF+Cosine and AI-Assisted ATS Approaches

Cristian Sebastián Luna Pérez[1], **Juan Carlos Santillan Lima**[2]

Cita:
Luna Pérez, C. S., & Santillan Lima, J. C. (2026). Análisis estadístico comparativo: Adecuación semántica CV-oferta entre ATS basados en TF-IDF+coseno y enfoque asistido por IA. *TESLA Revista Científica, 6*(2), e667. https://doi.org/10.55204/trc.v6i2.e667

Recibido: 16/06/2026
Revisado: 20/06/2026–08/08/2026
Corregido: 07/09/2026
Aceptado: 11/09/2026
Publicado: 12/09/2026

[1] Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Dirección de Posgrados, EC060155, Panamericana Sur km 1½, Riobamba, Chimborazo, Ecuador.

[2] Escuela Superior Politécnica de Chimborazo, Facultad de Informática y Electrónica, EC060155, Panamericana Sur km 1½, Riobamba, Chimborazo, Ecuador.

[1]cristian.luna@espoch.edu.ec, [2]carlos.santillan01@espoch.edu.ec

[2]ORCID: 0000-0001-5812-7766

Resumen: Este estudio es un análisis estadístico comparativo de adecuación semántica para dos enfoques de procesamiento textual aplicados a procesos tipo ATS. La comparación es de un pipeline clásico basado en TF-IDF y similitud del coseno, y un pipeline asistido por inteligencia artificial mediante API de OpenAI. Se usó una muestra por conveniencia de currículums recibidos para una oferta laboral fija. Cada CV fue anonimizado y evaluado por ambos métodos bajo un diseño de investigación pareado intra-currículum; las métricas comparadas son: similarity_score, coverage, diversity, density, confidence y rationale_relevance. El análisis incluyó Shapiro-Wilk para evaluar la normalidad de las diferencias (deltas), prueba pareada t o Wilcoxon, tamaños de efecto, corrección Holm y para la concordancia se usó Bland-Altman e ICC. Los resultados evidencian diferencias sistemáticas entre ambos enfoques: el método clásico tiende a "inflar" la similitud y el puntaje global mientras que el asistido por inteligencia artificial tiende a ser más sensible al momento de captar conceptos inherentes no textuales.

Palabras clave: *ATS, TF-IDF, similitud semántica, análisis estadístico comparativo.*

Abstract: This study presents a comparative statistical analysis of semantic adequacy between two text-processing approaches applied to Applicant Tracking System (ATS) processes. The comparison involves a classical pipeline based on TF-IDF and cosine similarity and an artificial intelligence-assisted pipeline using the OpenAI API. A convenience sample of résumés submitted for a specific job vacancy was used. Each CV was anonymized and evaluated using both methods under a paired within-résumé research design. The metrics compared were similarity_score, coverage, diversity, density, confidence, and rationale_relevance. The statistical analysis included the Shapiro–Wilk test to assess the normality of the differences (deltas), paired t-test or Wilcoxon signed-rank test, effect sizes, Holm correction, and Bland–Altman analysis and the Intraclass Correlation Coefficient (ICC) to assess agreement. The results reveal systematic differences between the two approaches: the classical method tends to "inflate" similarity and the overall score, whereas the artificial intelligence-assisted method tends to be more sensitive in capturing inherent concepts that are not explicitly represented in the text.

Keywords: *ATS, TF-IDF, semantic similarity, comparative statistical analysis.*

INTRODUCCIÓN

La inteligencia artificial ha transformado la gestión organizacional para siempre, redujo la revisión humana casi por completo en los procesos de reclutamiento corporativo. En este contexto hoy, los Appliccant Tracking Systems (ATS) son indispensables en las infraestructuras digitales de las organizaciones, son softwares que sirven para administrar requerimientos laborales, filtrar candidatos y mejorar las decisiones de contratación. En la lista Fortune 500 de 2025, el fenómeno es marcado el 97,8 % de las empresas listadas utilizan un ATS detectables, según el reporte de uso de ATS de Jobscan (Purcell,2025). De manera consistente con los años siete pasados, Oracle ya referenciaba el mismo reporte en su documentación corporativa esto que sugiere un uso extendido de estos algoritmos (Oracle, 2019).

Esta tendencia es motivada por el aumento del volumen de postulaciones y con la necesidad operativa de mantener eficiencia, costos bajos, reproducibilidad y trazabilidad en los procesos de reclutamiento. (Suraj, Aruna y Binila, 2019). Para resumir, un ATS es un sistema que organiza y automatiza la fase inicial del reclutamiento: publicación de vacantes, recepción de currículos, estructuración de información y puntaje preliminar, su objetivo es reducir carga operativa y el número de entrevistas. (Suraj, Aruna y Binila, 2019).

Profundizando, los sistemas operan sobre currículos heterogéneos (.pdf/.doc/.docx, imágenes, etc.) por lo que es necesaria la normalización del texto. En la literatura, la preparación del CV incluye: extracción de

texto, eliminación de mayúsculas, limpieza de caracteres especiales, eliminación de elementos decorativos, estandarización y de ser necesario OCR cuando el documento no es seleccionable; además, se aplica una segmentación por secciones claves (experiencia, educación, etc.) para un procesamiento más consistente (Chafi, S., Kabil, M., & Kamouss, A. 2025). Este punto es importante, debido a que el poder del análisis semántico textual depende tanto del modelo aplicado como del texto comprable que se va a usar para cada candidato.

Muchos ATS son impulsados por enfoques basados en similitud textual y de recuperación informática como el TD-IDF y el coseno similitud, que funcionarían como una baseline eficiente para encontrar coincidencias del contenido de un currículo y el texto de una oferta fija de trabajo. (Sangwan, B., Sangeeta, & Diksha, 2025).

Sin embargo, los algoritmos clásicos pueden subestimar o sobrestimar la coincidencia de palabras claves; además, son carentes de contexto (más años de experiencia > menos años de experiencia), ignora jerarquías (director > técnico) u omitir ideas semejantes expresadas en palabras distintas, lo que puede generar dificultades para manejar currículos no estructurados o inducir a evaluaciones inestables. (Bevara, R. V. K., Dronavalli, S. C., Karedla, S. P., Rupeshkumar, S., & Xiao, T. 2025).

Para el presente documento investigativo, se ha propuesto dos enfoques basados en embeddings y modelos tipo transformer (encoders y decoders, incluyendo variantes de LLM) que ofrecen ventajas al capturar las relaciones semánticas y contexto, estos no se limitan al conteo de los n-términos coincidentes, lo que permite comparar currículo-oferta incluso cuando el contenido esta expresado de forma distinta (Bevara, R. V. K., Dronavalli, S. C., Karedla, S. P., Rupeshkumar, S., & Xiao, T. 2025). La necesitad de migrar de modelos TF-IDF hacia representaciones más robustas y modelos generativos está motivada por: la eficiencia de emparejamiento CV-vacante, y la reducción de sesgos en la toma de decisiones equivocas basadas en señales textuales superficiales.

Estos sesgos son recurrentemente listados en literatura sobre clasificación ocupacional y job screening, la evidencia sugiere que representaciones lingüísticas incorporan sesgos asociados, por ejemplo, al género, mismos que influyen en el resultado de clasificación y métricas de desempeño por subpoblaciones (Tyagi, S., Anuj, Qian, W., Xie, J., & Andrews, R. 2024). Por ello, las estrategias de mitigación que se proponen contemplan embeddings y re-ponderacion de datos para mejorar la imparcialidad de los modelos.

Por el contrario, la evidencia sugiere que sistemas basados en LLM pueden producir evaluaciones y rankings con razonamiento textual, aprovechando la contextualización del texto para propulsar el modelo más allá de las coincidencias literales (Deshmukh, A., Raut, A., & Deshmukh, V. 2026). Esto abre la pregunta bajo la cual se construye esta propuesta de investigación. ¿Qué tan comparables son las métricas obtenidas por los distintos enfoques en cuanto se evalúen a los mismos currículos contra la misma oferta?

Se propone un análisis estadístico comparativo de métricas de adecuación semántica CV-oferta entre dos pipelines: (i) enfoque clásico TF-IDF + coseno similitud y (ii) un enfoque asistido por un modelo de lenguaje basado en inteligencia artificial a través de la API de la arquitectura transformer, operado con un promt estandarizado para la normalización del texto, dotar de contexto, jerarquía y conciencia al momento de generar el mismo conjunto de métricas.

No se busca “remplazar” el modelo clásico, sino cuantificar las discrepancias, sesgos y convergencias de ambos enfoques, considerando que estas tecnologías están avanzando a soluciones hibridas que combinan modelos tradicionales y modelos de lenguajes automatizado para los procesos de reclutamiento (Chong, K. W., Ng, K. W., & Fu, Y. H. 2026). El artículo se enmarca como una comparación de métodos, donde la atención se centra en la evidencia estadística de diferencias pareadas, de acuerdo y concordancia, sobre las métricas comparables y no en proponer un nuevo algoritmo de reclutamiento.

METODOLOGÍA

Diseño de investigación: El estudio se definirá como una investigación de tipo no experimental, comparativa y transversal, bajo un enfoque cuantitativo que realiza una comparación pareada intra-currículum, donde cada CV constituye una unidad experimental y procesa bajo los procedimientos de medición semántica: (i)

un pipeline clásico basado en representaciones TF-IDF y similitud del coseno; y (ii) un pipeline asistido por IA implementado mediante la API de OpenAI.

Ambos enfoques trabajan con una única variable de control (Oferta de trabajo en texto plano) para garantizar la trazabilidad y comparabilidad, esto evita la confusión por cambios en el texto objetivo. Este diseño pareado permite estimar, por cada métrica X, pares $\left(X_i^{classic}, X_i^{OpenAI}\right)$ y analizar diferencias $d_i = X_i^{classic} - X_i^{OpenAI}$, siguiendo la lógica de comparación de métodos de medición (Bland & Altman, 1986). para cada CV donde $i = 1, \ldots, N$, esto controla la varibilidad de documentos y refuerza la rigurosidad estadística.

Muestreo y normalización de texto: Los inputs se recolectan en formatos heterogéneos, esto es un problema típico en este tipo de procesamiento de texto automático ya que los CVs difieren ampliamente en estructura, diseño y forma, esto dificulta el análisis y puede llegar derivar ciertas inconsistencias. Para mitigar esta variabilidad, de cada CV se extra el texto plano, es decir, se convierte a un archivo tipo .txt, con normalización básica (minúsculas, limpieza de separadores y caracteres decorativos) para homogenizar el contenido y facilitar replicabilidad.

Cuando el documento contiene texto incrustado como imagen o PDF escaneado, se activa un módulo de OCR. Esta decisión fundamentada en la literatura, para obtener una extracción más robusta se los documentos que requieren OCR se vuelven "machine-readable" cuando el parseo tradicional no es aplicable y evita ruido en el archivo (Nawalny et al., 2025/2026). Este problema ya fue descrito antes de cualquier emparejamiento semántico, el primer cuello de botella es extraer el cuerpo textual desde PDFs/imagenes no textuales (Alonso et al. 2025).

En nuestro caso se utilizó una muestra por conveniencia de 119 currículums recibidos para una única vacante laboral publicada por decanato de la Facultad de Informática y Electrónica de la ESPOCH (FIE). Tras aplicar criterios de elegibilidad, 35 estaban redactados en ingles por lo que fueron descartados bajo decisión del autor fundamentada más adelante, 15 por consistir en enlaces no ingesta-bles y 16 por otras no conformidades documentales (No cumplen con los requisitos mínimos); la muestra final fue n = 53 CVs. Esto garantiza que ambos pipelines procesen insumos idénticos y comparables condición indispensable del diseño pareado. Se aplicó anonimización de datos personales por tratarse de documentos sensibles, la unidad de análisis se centró en el par de mediciones obtenidas de un mismo grupo de objetos, donde la muestra final n= 53 superó el umbral de 30 unidades para una estimación estable del ICC (Koo & Li, 2016).

Ingesta y trazabilidad en base de datos: Con los CVs listos en texto plano, un script de ingesta ejecuta el modelo de la raíz y almacena los metadatos en una base de datos SQL. De forma simultánea la oferta laboral (tipo .txt) se ingesta como una variable de control para cada uno de los CV. Esta arquitectura es una base sólida por dos aspectos: primero la reproducibilidad al dejar los objetos de análisis en su forma natural y las pruebas pareadas limpias, al mantener fijo el texto objetivo. (Nawalny et al., 2025/2026).

Pipeline1: Enfoque clásico (TF-IDF + coseno similitud)

Este enfoque clásico calcula "similarity" (similitud semántica) a través de un vectorización numérica del texto propia del TF-IDF entre CV y la oferta propuesta. Estos modelos son reconocidos por su transparencia e interpretabilidad, el modelo pondera términos relevantes y el coseno permite comparar vectores de alta dimensionalidad en el texto no estructurado (Ajjam & Al-Raweshidy, 2026). Del mismo modo, de acuerdo con los autores Ajjam & Al-Raweshidy (2026), el objeto del presente no es presentar TF-IDF como "el mejor modelo", sino como un procedimiento simple, escalable y explicable para comparación de requisitos y postulaciones.

Pipeline 2: Enfoque asistido por IA (API)

El segundo enfoque usa un LLM como asistencia, este algoritmo es invocado mediante una API de OpenAI para robustecer un modelo clásico dotándolo de inteligencia artificial con el objetivo de tener capacidades que se asemejen a la de un reclutador de talento humano, mediante un requerimiento estandarizado (Temperatura, prompt, modelo, etc) y los"outputs" son los mismos del modelo clásico, pero esta vez fueron obtenidos con algo semejante a la consciencia, con contexto, razonamiento e interpretabilidad. Los outputs

son clasificados en variables pareadas entre modelos y almacenados en una base de datos estructurada tipo (JSON). En reclutamiento, emplear modelos basados en "keyword matching" pierden matices y relaciones implícitas; modelos tipo transformer capturan mejor las ideas del contenido de cada postulación (Liu, 2025). Esto es coherente con las recientes conclusiones donde las arquitecturas construidas con BERT y embeddings superan a TF-IDF en tareas de matching léxico a gran escala (Liu, 2025).

Sin embargo, el uso de modelos generativos requiere de controles que permitan la reproducibilidad. Estudios comparativos sobre pipelines de extracción con LLM refuerzan la idea de que la estandarización de los promts es un paso indiscutible para obtener resultados auditables (Nawalny et al., 2025/2026). Por ello, cada output del modelo, es sometido a validaciones de formato, coerción y recorte a un rango definido [-1,1; 0,1] si fuese necesario, un recalculo local de métricas derivadas para preservar trazabilidad.

Para comprobar si las muestras pareadas siguen una distribución normal, usaremos la prueba Shapiro-Wilk, esta prueba mide la diferencia de las métricas $d_i = X_i^{classic} - X_i^{OpenAI}$ (Shapiro & Wilk, 1965). Esta prueba además de proveernos los estadísticos W, p-valor y una interpretación gráfica, nos permitirá inferir sobre que prueba estadística utilizar en función de la normalidad de las diferencias. Usaremos la prueba t-pareada (paramétrica) en caso de que las diferencias sean aproximadamente normales para comprobar si la media es de las diferencias es cero $(H_0 : \mu_{d_i} = 0)$ (Student, 1908). O Wilcoxon (no paramétrica) si las diferencias no son normales y la mediana de las diferencias es cero $(H_0 : Me_{d_i} = 0)$ (Wilcoxon, 1945).

Con el objetivo de realizar un análisis comparativo adecuado, utilizaremos el ICC, métrica para medir el acuerdo absoluto entre métodos, cuantifica cuánta variabilidad total de las mediciones se debe a diferencias reales entre objetos (CVs) frente a la variabilidad propia del método utilizado. A diferencia de la correlación de Pearson la cual es ampliamente usada para medir la asociación, el coeficiente de correlación intraclase mide la concordancia, si se obtiene resultados altos, se puede considerar evidencia suficiente para suponer que los enfoques califican de forma similar a los mismo "inputs". (McGraw & Wong, 1996; Koo & Li, 2016)

2.1 HIPÓTESIS DE INVESTIGACIÓN

Hipótesis nula (H_0): No existen diferencias estadísticamente significativas entre las métricas semánticas obtenidas por el pipeline clásico (TF-IDF + similitud del coseno) y el pipeline asistido por OpenAI al evaluar los mismos CV frente a una misma oferta.

Hipótesis alternativa (H_1): Existen diferencias estadísticamente significativas entre las métricas semánticas obtenidas por ambos pipelines bajo el mismo escenario (mismos CV y la misma oferta).

RESULTADOS

El trabajo no compara algoritmos compara dos enfoques distintos, el clásico (Python – TF-IDF) mide similitud y el otro (TF-IDF + API OPEN AI) incluye contextualización semántica, permite a la vectorización numérica del texto interpretar: Contexto, Jerarquía, Posición, Equivalencias conceptuales entre habilidades y Requisitos.

Uno depende de coincidencias lexicales, el otro modela relaciones semánticas y entiende lo que significan. Los descriptivos tienen como objetivo resumir cada método y evaluar si estos cumplen los supuestos necesarios para aplicar cada prueba estadística. La primera Shapiro-Wilk: 7 de las 9 variables rechazan una distribución normal (Tabla 1), con valores de p menores a 0.05; una variable, "coverage", cumple con el supuesto de normalidad (p = 0.0693), mientras que "n_rationales" no es evaluable por presentar varianza nula."

Ya que las métricas no cumplieron el supuesto de normalidad del delta pareado, se aplicó la prueba no paramétrica de Wilcoxon para muestras relacionadas con el fin de comparar los resultados obtenidos por los pipelines Python y OpenAI. Esta prueba permite evaluar diferencias sistemáticas entre pares sin asumir normalidad en la distribución de los datos, así mismo se aplicó una prueba t-pareada para coverage.

Tabla 1

Prueba Shapiro-Wilk.

Variable / Métrica	W (Shapiro)	p-value (Shapiro)	Decisión
similarity_score	0.8698	0.0000	No normal -> No paramétrica
coverage	0.9594	0.0693	Normal -> Paramétrica
diversity	0.8929	0.0002	No normal -> No paramétrica
density	0.8484	0.0000	No normal -> No paramétrica
confidence	0.8858	0.0001	No normal -> No paramétrica
rationale_relevance	0.9532	0.0370	No normal -> No paramétrica
n_rationales	NaN	NaN	No evaluable (varianza nula)
latency_ms	0.9264	0.0029	No normal -> No paramétrica
db_score	0.8684	0.0000	No normal -> No paramétrica

Nota: *Para el análisis del supuesto de normalidad se usó la prueba de Shapiro-Wilk. La detección de distribuciones no normales pone en evidencia la necesidad de usar pruebas no paramétricas para el contraste de los datos.*

Tabla 2

Prueba de Wilcoxon y t pareada para coverage

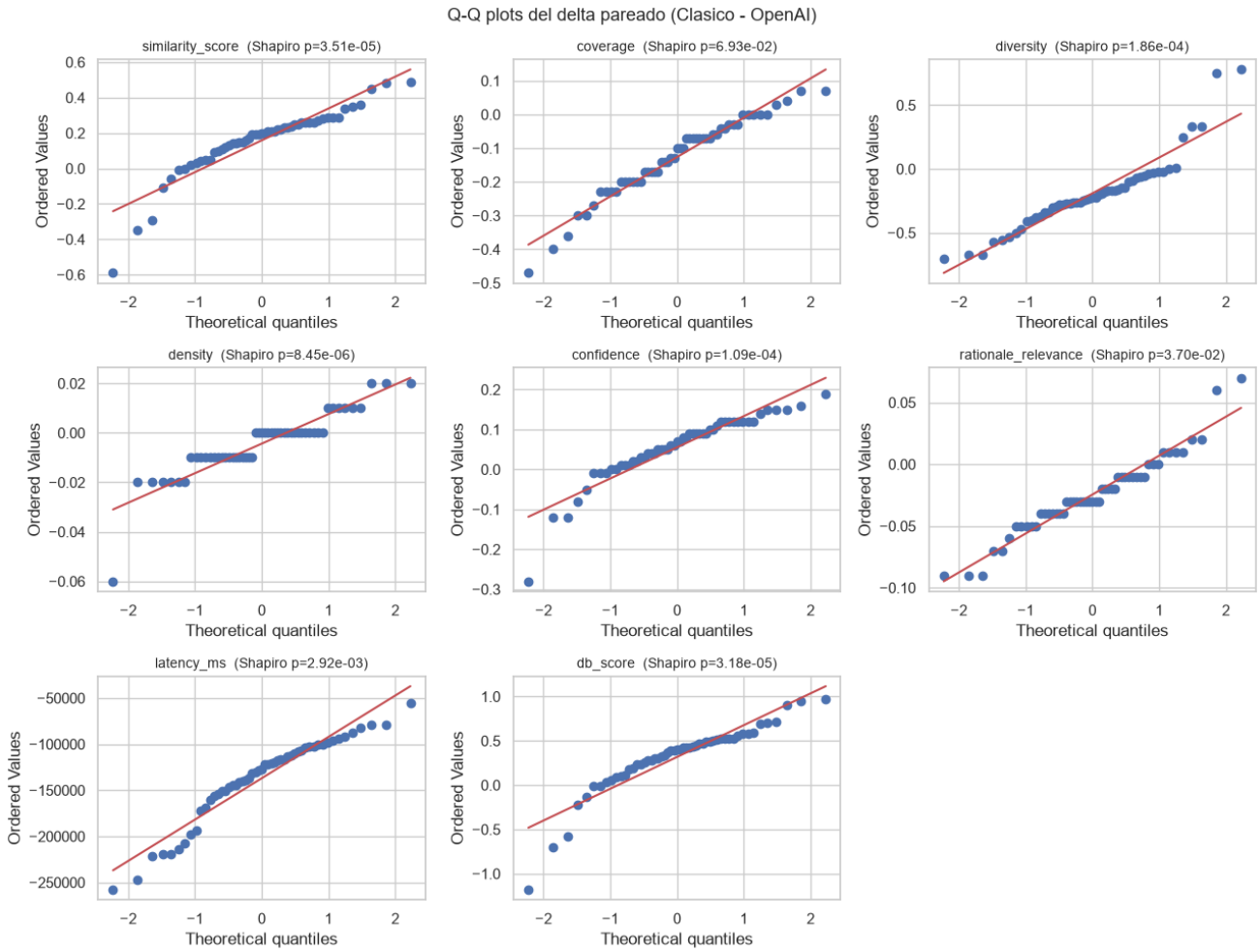
Variable / Métrica	Test	Estadístico	p-value	Tamaño del Efecto	Magnitud
similarity_score	Wilcoxon	159.0000	0.0000	0.6630	Grande
coverage	t-pareada	-7.9076	0.0000	-1.0862	Grande
diversity	Wilcoxon	199.5000	0.0000	0.6123	Grande
density	Wilcoxon	158.5000	0.0281	0.3016	Mediano
confidence	Wilcoxon	176.5000	0.0000	0.6268	Grande
rationale_relevance	Wilcoxon	152.0000	0.0000	0.6447	Grande
n_rationales	NaN	NaN	NaN	NaN	NaN
latency_ms	Wilcoxon	0.0000	0.0000	0.8701	Grande
db_score	Wilcoxon	166.0000	0.0000	0.6682	Grande

Nota: *La magnitud del tamaño del efecto se estimó mediante Cohen para las variables normales (paramétricas) y a través del coeficiente r (basado en Z) para aquellas variables con distribución no normal o no paramétricas.*

El resultado Cohen's $|dz| = |-1{,}09|$ (Tabla 2) de la prueba t-pareada dz $\approx$ 1.09 indica: diferencia grande y no solo estadística, sino prácticamente relevante. Además, el p_value la probabilidad de observar la diferencia en caso de que no existiera dicha diferencia es casi cero. En conclusión, el modelo asistido por IA produce racionales significativamente más relevantes y de mejor calidad que el modelo clásico. Un análisis basado únicamente en coincidencias léxicas puede fallar cuando las habilidades de un curriculum no están expresadas con las mismas palabras, pero sí con el mismo significado.

Figura 1

Shapiro-Wilk de deltas pareados



Nota: *Gráfico para interpretar visualmente el supuesto de normalidad de las diferencias entre ambos enfoques. Las desviaciones que se observan respecto a la diagonal y los p-valores de la prueba sirven como justificación de los modelos de contraste.*

Con respecto a los efectos r de Wilcoxon de las restantes variables (Tabla 2) todos son significativos, existen efectos grandes (≈ 0.6 - 0.67) y muy grandes (latency_ms, r ≈ 0.87). Esto significa que para la mayoría de los inputs los resultados de los enfoques son sistemáticamente mayores y diferentes uno del otro. Es decir, no es ruido y los resultados no dependen de normalidad ni de outliers, sin importar que prueba se use, la conclusión es la misma, los resultados de ambos pipelines son diferentes de forma consistente. Para reforzar la metodología, aplicamos la corrección estricta HOLM para cada métrica (Tabla 3), los p_holm demuestra que no hay resultados "frágiles" y descartan falsos positivos (error tipo I). Así mismo, la concordancia de la dirección de los resultados refuerza la robustez del análisis comparativo.

Tabla 3

Pruebas estadísticas y corrección Holm

Variable / Métrica	Test	p-value	p-value (Holm)	Rechaza H0 (Holm)
similarity_score	Wilcoxon	0.0000	0.0000	True (Sí)
coverage	t-pareada	0.0000	0.0000	True (Sí)
diversity	Wilcoxon	0.0000	0.0000	True (Sí)
density	Wilcoxon	0.0281	0.0281	True (Sí)
confidence	Wilcoxon	0.0000	0.0000	True (Sí)
rationale_relevance	Wilcoxon	0.0000	0.0000	True (Sí)
latency_ms	Wilcoxon	0.0000	0.0000	True (Sí)
db_score	Wilcoxon	0.0000	0.0000	True (Sí)

Nota: *Resultados de la prueba no paramétrica de Wilcoxon con corrección secuencial de Holm para las diferencias pareadas (Python - OpenAI).*

Tenemos evidencia suficiente en todas las métricas evaluadas, estas presentan diferencias estadísticamente significativas ($p < 0.05$ y $p\_holm < 0.05$). Además, para evitar errores tipo I y poder descartar todos los falsos positivos y confirmar que las discrepancias entre "pipelines" no se deben a las comparaciones como tal, sino diferencias sistemáticas y reproducibles bajo el diseño pareado intra currículum.

Usamos las pruebas Pearson/Spearman, Bland -Altman e ICC con el objetivo de medir asociación y concordancia. Se descubrió que los enfoques no son equivalentes y aunque ordenan los inputs (CVs) en forma similar, no concuerdan en magnitud (Tabla 4). El pipeline clásico calcula valores sistemáticamente más altos para la variable similarity que el asistido por IA. No hay evidencias de concentraciones cerca a cero por lo que el error es sistemático y no aleatorio.

Tabla 4

Asociación, acuerdo y concordancia

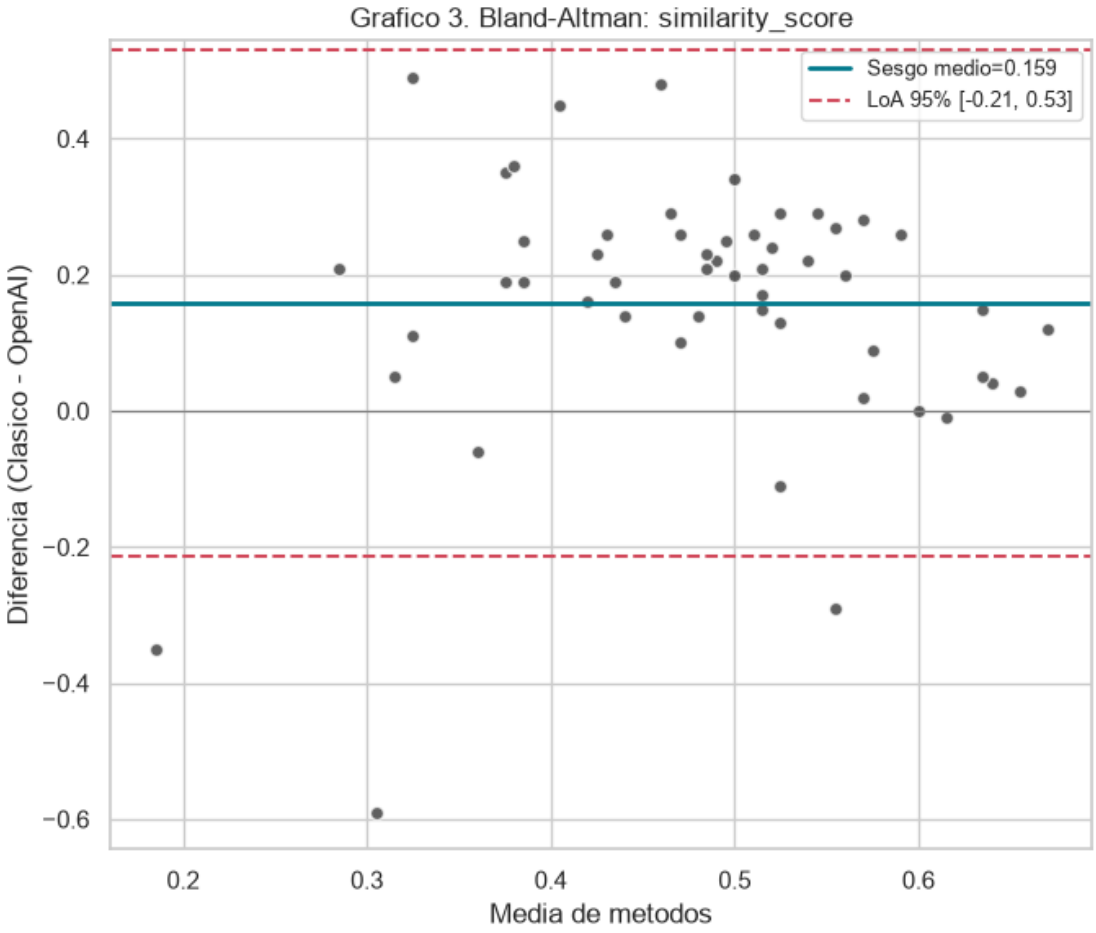
Variable / Métrica	Pearson (r)	Spearman (ρ)	ICC (A,1)	Sesgo (Bias)	LOA Inferior (95 %)	LOA Superior (95 %)
similarity_score	0.0950	0.3451	0.0586	0.1587	-0.2125	0.5298
coverage	0.5139	0.5314	0.2635	-0.1262	-0.3540	0.1015
diversity	0.0369	0.1808	0.0257	-0.1908	-0.7602	0.3787
density	0.5994	0.6186	0.5763	-0.0043	-0.0294	0.0207
confidence	0.2500	0.3242	0.1863	0.0555	-0.1043	0.2152
rationale_relevance	0.2241	0.2313	0.1521	-0.0245	-0.0865	0.0374
n_rationales	NaN	NaN	NaN	0.0000	NaN	NaN
latency_ms	0.1687	0.1509	0.0001	-136962.5472	-226013.7529	-47911.3414
db_score	0.0961	0.3385	0.0591	0.3181	-0.4237	1.0600

Nota: *La asociación (Pearson/Spearman) mide la covariación y el mantenimiento del orden; por su parte, el ICC utilizado para medir concordancia absoluta cuantifica si los métodos asignan valores idénticos, y no meramente correlacionados entre sí.*

El análisis de Bland-Altman describe el sesgo medio y los límites de concordancia (LoA 95 %), distinguiendo entre errores sistemáticos y aleatorios en los diferentes enfoques. Los enfoques se clasifican de forma similar, pero difieren en magnitud: el coeficiente de correlación intraclase (ICC) de concordancia absoluta es más alto para density (0.58) y cero para la diversity (003). Similarityscore y dbscore combinan una correlación de Spearman moderada (aprox. 0.34) con un ICC bajo (0.06) es decir existe concordancia ordinal sin concordancia absoluta. Además, el gráfico Bland-Altman muestra evidencia de un sesgo positivo en las variables “similarity_score” (0.16) y en la métrica “db_score” (0.32).

Figura 2

Bland–Altman de la métrica de similitud



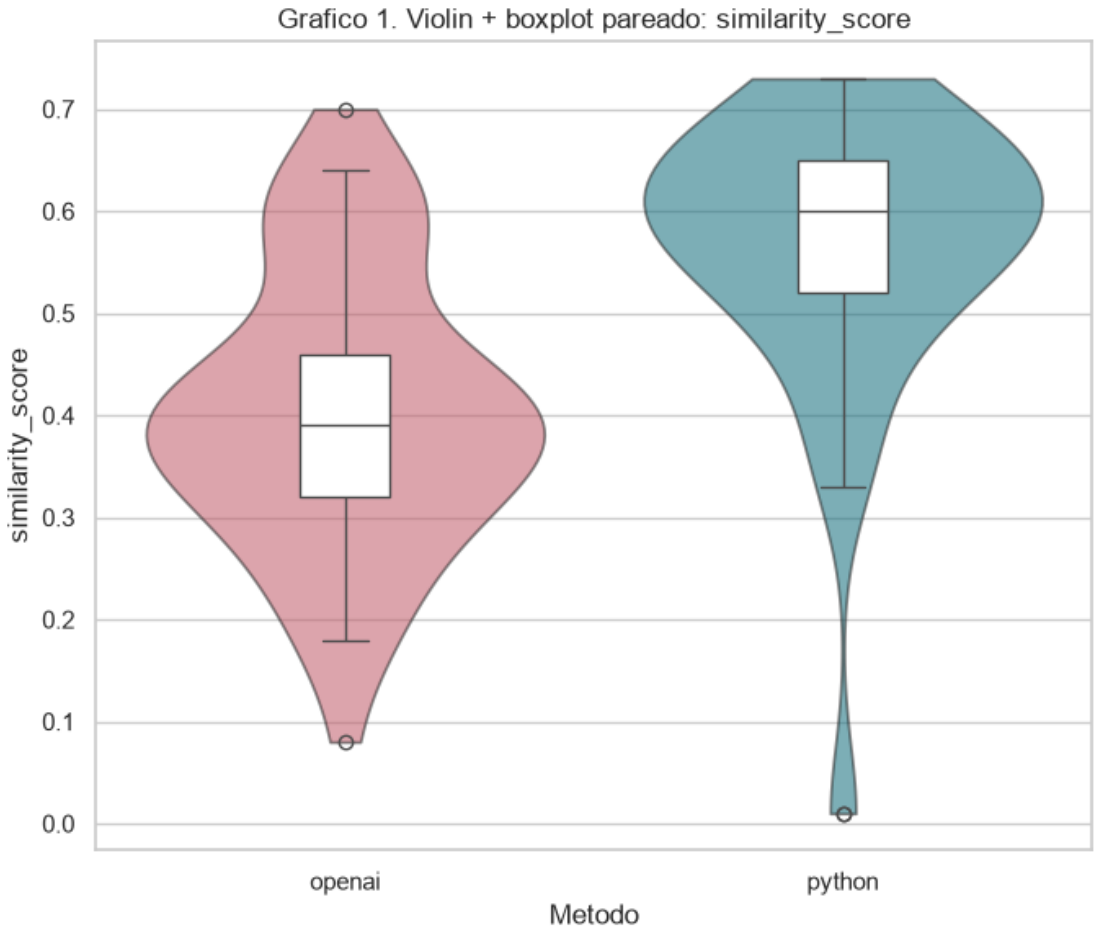
Nota: *Gráfico de Bland–Altman para la concordancia de* similarity_score*. Existe un sesgo medio positivo entre los deltas pareados (Python - OpenAI), con dispersión heterocedástica que se estabiliza en rangos medios y altos.*

Para la variable coverage existe coherencia y concordancia moderada como muestra. Pearson = 0.513 y Spearman = 0.531 (Tabla 4). Si un enfoque aumenta, el otro también, por lo tanto, tenemos evidencia para afirmar que es probable que el enfoque asistido por IA reconoce equivalencias o sinónimos con mayor facilidad, mientras que TF-IDF depende más de coincidencias léxicas. No obstante, el Coeficiente de Correlación Intraclase muestra una concordancia absoluta débil (ICC= 0.2635).

Por otro lado, diversity es la métrica con mayor dependencia del enfoque metodológico. La presencia de una asociación lineal y un acuerdo absoluto prácticamente nulo(p = 0.1808 y ICC = 0.0257) confirma que cada procesamiento computa la variabilidad léxica como criterios ortogonales. Finalmente, el comportamiento de db_score replica al patrón de similarity_score. Ambas métricas registran asociaciones muy similares ($r < 0.10$), concordancia entre modelos casi nula (ICC < 0.06) y un ordenamiento que se puede considerar “débil”. El sesgo positivo del BIAS = 0.318 significa que uno de los enfoques (modelo clásico) tiene tendencia a sobre puntuar más que el otro de forma recurrente. Es así, que se demuestra que nuestros resultados particularmente, la nota final “db_score” y el parecido “similarity_score” entre los textos dependen por completo de su enfoque.

Figura 3

Violín y boxplot pareado de similitud

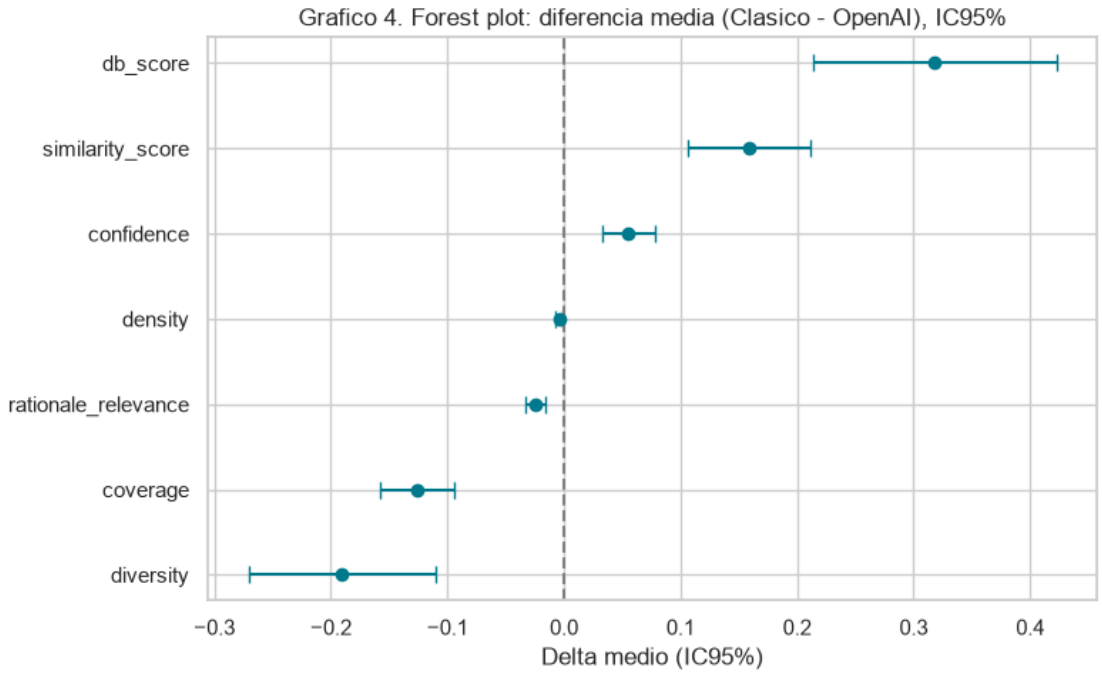


Nota: *Se observa una mayor concentración de puntajes altos en el método clásico, con mayor presencia de valores inferiores atípicos.*

OpenAI suele procesar valores inferiores con respecto a el método clásico y tiende a mostrar una cobertura más alta. La única métrica donde realmente hay una concordancia razonable entre ambos es en la métrica "density". Es decir, los enfoques coinciden cuando dependen de la estructura del documento y difieren en cuando la métrica necesariamente requiere interpretación. La Figura 4 nos muestra la diferencia de medias (Delta) entre métodos, junto con su IC95 %, lo que nos permite identificar sesgos, las métricas cuyo IC95 % no cruza el cero, la diferencia promedio puede considerarse consistente y no atribuible a un componente estocástico.

Figura 4

Representación de deltas mediante un forest plot”



Nota: *Gráfico de bosque de diferencias de media con intervalos de confianza del 95 % que presenta sesgos estadísticamente significativos.*

Existe un sesgo positivo en las tres primeras métricas (de arriba hacia abajo), lo que sugiere que el pipeline clásico tiende a asignar valores más altos a los obtenidos por el enfoque asistido por OpenAI. Por el contrario, el sesgo negativo esta prensente en las métricas “coverage” y “diversity”, sugiriendo que OpenAI genera puntuaciones más altas cuando se trata de cobertura y léxica. La métrica “density” es la única que se está próxima a cero, posicionándose como la métrica con mayor compatibilidad.

DISCUSIÓN

Los resultados de la investigación sugieren que la comparación entre los pipelines utilizados no busca comparar puntajes, sino buscar evidencias suficientes sobre si ambos métodos son equivalentes entre sí. Bajo el diseño pareado intra-currículum con una variable de control fija, el estudio demostró diferencias sistemáticas en la mayoría de las métricas, lo que sugiere que el enfoque es capaz de incorporar sesgos a cada una de las magnitudes evaluadas.

En particular el pipeline basado en TF-IDF tiene valores superiores en sus métricas de similitud se puede decir que existe una sobre ponderación en los términos que realmente coinciden entre la oferta y el CV analizado, incluso cuando el CV y la oferta no tengan relaciones conceptuales fuertes.

A diferencia del otro enfoque planteado (Open AI) el cual muestra mayor tendencia a puntuar más alto en métricas que dependen de diversidad y cobertura, es decir, tiene mayor sensibilidad a captar acuerdos conceptuales semánticos, sinónimos e incluso interpretaciones de alta relevancia, esto se puede explicar a que no necesita de solapamiento literal de caracteres. De ninguna manera esto quiere decir que estos enfoques puedan intercambiarse entre sí; las pruebas realizadas para medir acuerdo y concordancia nos sugieren que aunque los CVs tengan un ordenamiento parecido la mayoría de las veces sus magnitudes no coinciden. Indiscutiblemente esto representa un problema que puede llegar a afectar las decisiones sen el ámbito de reclutamiento de personal en el ámbito organizacional, o peor aún en el funcionamiento del ATS, donde las variaciones, por insignificantes que parezcan pueden ser decisivas entre si excluye o no un candidato conocido como sesgo de priorización.

Esta investigación debe interpretarse como un caso aislado y aplicado a una realidad especifica que cuenta con: una oferta de trabajo particular, idioma especifico, modelos con parámetros controlados, ubicación geográfica y área de conocimientos específicas, es correcto tratarla como una aproximación. Sin embargo, es arriesgado generalizar los resultados obtenidos a otros dominios, esta generalización requiere de validación y

evidencia adicional. No obstante, el enfoque estadístico planteado usa una estructura trazable y replicable para evaluar estas tecnologías útiles para evidenciar diferencias de algoritmos, bajo ningún contexto se debe usar como una herramienta digital de reclutamiento corporativo, especialmente cuando las instituciones buscan habilidades ajenas al conocimiento y propias del ser humano.

APLICACIONES PRÁCTICAS O FUTURAS LÍNEAS DE INVESTIGACIÓN

Para las organizaciones esta investigación brinda un soporte auditable y útil para su transición de sistemas de reclutamiento basados en coincidencias léxicas hacia sistemas asistidos o potenciados con LLMs. Después de demostrar que los modelos clásicos especialmente los basados en TF-IDF con coseno similitud, sobrestima la similitud, los departamentos de reclutamiento y recursos humanos pueden agregar estas tecnologías como un filtro adicional para evitar la exclusión errónea de los candidatos calificados que han empleado sinónimos en sus CVs.

Este protocolo funciona como un filtro de control y no como una herramienta que toma decisiones por si sola. Además, es responsabilidad de las direcciones ajustar los modelos, la temperatura, directrices, prompts y semillas según su realidad especifica, no controlar dichos parámetros conduce a la no replicabilidad y no auditabilidad, por tanto, las decisiones se pueden considerar instables. Practicar estas recomendaciones significa generar proceso transparentes, útiles, auditables, estables y controlados.

Vale la pena destacar que aquella falta de concordancia en magnitudes númericas abre un nuevo horizonte de investigación centrada en funciones propias de transferencia matemática que nos permitan homologar los puntajes obtenidos en diseños de investigación inter-modelo. La investigación y control de la migración de los algoritmos tradicionales hacia nuevos algoritmos basados en tecnología Transformers es necesaria, probar que estos algoritmos son “mejores” es imperante al momento de tomar decisiones y no generar conclusiones muy generalizadas.

Investigaciones futuras podrían centrar sus esfuerzos en proponer índices dinámicos construidos sobre la densidad de cada documento, con miras a reducir la sobreestimación propias de solapamientos textuales, se busca mejorar la eficiencia computacional del conteo de los n-gramas en términos de razonamiento contextual. Para finalizar, existen dos limitaciones importantes que podrían abrir nuevos caminos i) trabajar con más de un idioma para evitar que el descarte arbitrario de postulantes, ya que es una restricción propia de la forma de evaluación y no de fondo, ii) diversificar la variable de control mediante la ingesta de varias ofertas de trabajo para comprar sus resultados con el fin de probar la capacidad de inferencia de resultados del modelo cuando se traba con otros perfiles, áreas del conocimiento y dominios.

CONCLUSIONES

La presente investigación provee suficiente evidencia para para rechazar la hipótesis nula, formulado bajo este diseño pareado intra-curriculum con una única variable de control, los resultados son contundentes y sistemáticos. En su totalidad las métricas son estadísticamente significativas, tal es el caso de: la corrección Holm ($p_{holm} < 0.05$) tienen también efectos r grandes que van desde 0.6 a 0.87, también es útil rescatar el valor cohes´s $|dz| \approx 1.09$ para coverage. Esto sustenta nuestra afirmación formulada sobra las discrepancias, las diferencias son propias de los algoritmos y no del azar así mismo son trazables, no cuestionables y reproducibles.

Con respecto a los sesgos se tiene suficiente evidencia para afirmar que el pipeline basado únicamente en el TF-IDF tiende a sobrecalificar las métricas globales con esto genera sesgos positivos en variables como “similarity_score” y “db_score” con resultados de +0.16 y +0.32 respectivamente. Caso contrario con el algoritmo asistido por IA el cual presenta mayor sensibilidad a captar ideas no expresadas con constructos textuales, es decir, es capaz de identificar sinónimos e interpretar texto, en el diagrama de bosque donde se trató la diferencia de medias, la variable con concordancia casi absoluta es “density” ICC = 0.58, sesgo ≈ 0. Esto nos sugiere que las variables que evalúan estructuras de documentos convergen y no lo hacen cuando se trata de variables que requieren interpretación.

Con respecto a las discrepancias evidenciadas en al análisis de asociación y concordancia, se concluye

que a pesar de que el ordenamiento sigue un patrón muy similar en ambos métodos, existe una concordancia "débil" en la mayoría de los resultados ICC <0.27, por lo tanto, no es apresurado advertir que ambos enfoques no pueden ser considerados intercambiables, es importante entender que estos umbrales son inherentes al modelo de análisis y que no se pueden transferir de uno a otro, un postulante puede ser excluido con un modelo y aceptado en otro (sesgo de priorización).

Finalmente, las conclusiones son propias de esta realidad especifica la cual sigue una configuración fija de algoritmos, modelos y temperatura, no se debe interpretar el modelo como una generalización a otros ámbitos o dominios. Es decir, la contribución de este texto investigativo es el protocolo estadístico de comparación y no los resultados de este, debe ser entendido como un marco investigativo auditable para la migración de sistemas ATS basados en solapamiento léxico a un modelo con asistencia o que funcione en su totalidad por LLMs.

FINANCIACIÓN

La presente investigación no recibió financiación externa específica de organismos públicos, comerciales o entidades sin fines de lucro. Los recursos necesarios para el desarrollo del estudio fueron asumidos por los autores.

CONFLICTO DE INTERESES

Los autores declaran que no existen conflictos de intereses financieros, profesionales, institucionales o personales que pudieran haber influido en el desarrollo, análisis, interpretación o presentación de los resultados de esta investigación.

CONTRIBUCIÓN DE AUTORÍA

Tabla 5

Contribución de autoría según taxonomía CRediT

Participar activamente en:	Autor 1	Autor 2
Conceptualización	X	X
Análisis formal	X	X
Adquisición de fondos	X	
Investigación	X	X
Metodología	X	X
Administración del proyecto		X
Recursos	X	
Redacción – borrador original	X	
Redacción – revisión y edición	X	X
La discusión de los resultados	X	X
Revisión y aprobación de la versión final del trabajo	X	X

Nota: *Contribuciones declaradas de acuerdo con la taxonomía CRediT.*

REFERENCIAS

Ajjam, M.-H., & Al-Raweshidy, H. S. (2026). AI-driven semantic similarity-based job matching framework for recruitment systems. *Information Sciences, 724*, 122728. https://doi.org/10.1016/j.ins.2025.122728

Alonso, R., Dessí, D., Meloni, A., & Reforgiato Recupero, D. (2025). A novel approach for job matching and skill recommendation using transformers and the O*NET database. *Big Data Research, 39*, 100509. https://doi.org/10.1016/j.bdr.2025.100509

Bestgen, Y. (2014). Improving the validity of the automated essay scoring system under conditions of text complexity variation. Journal of Writing Research, 6(2), 123–143. https://doi.org/10.17239/jowr-2014.06.02.2

Bevara, R. V. K., Mannuru, N. R., Karedla, S. P., Lund, B., Xiao, T., Pasem, H., Dronavalli, S. C., &

Rupeshkumar, S. (2025). Resume2Vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching. *Electronics, 14*(4), 794. https://doi.org/10.3390/electronics14040794

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Chafi, S., Kabil, M., & Kamouss, A. (2025). Optimizing automatic CV classification with contrastive and generative learning. *Procedia Computer Science, 265*, 342–349. https://doi.org/10.1016/j.procs.2025.07.190

Chong, K. W., Ng, K. W., & Fu, Y. H. (2026). Resume data extract and job recruitment chatbot features for AI-based resume screening & analytics. MethodsX, 16, 103775. https://doi.org/10.1016/j.mex.2025.103775

Deshmukh, A., Raut, A., & Dahake, S. (2025). *Intelligent resume screening system using BERT and cosine similarity*. *International Journal of Artificial Intelligence, 14*(4), 3404–3411. https://doi.org/10.11591/ijai.v14.i4.pp3404-3411

Deshmukh, A., Raut, A., & Deshmukh, V. (2026). Scalable resume screening using large language model Meta AI version 3. IAES International Journal of Artificial Intelligence (IJ-AI), 15(1), 953–961. https://doi.org/10.11591/ijai.v15.i1.pp953-961

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Liu, X. (2025). Deep learning-based intelligent resume-position matching system: Semantic understanding and recommendation of BERT model in massive recruitment data. In *Proceedings of MLSC 2025*. https://doi.org/10.1145/3778450.3778452

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46

Mendoza, L. E., & Ramírez, J. P. (2020). Inteligencia artificial aplicada a la gestión del talento humano: Retos y oportunidades. Revista Iberoamericana de Tecnología y Sociedad, 12(3), 45–60.

Mkrtchyan, A. (2025). Text-based approaches for exploring job matching techniques (BS in Data Science thesis). American University of Armenia, Zaven & Sonia Akian College of Science and Engineering.

Murel, J., & Kavlakoglu, E. (2024). ¿Qué es el modelado de temas?. IBM Research.

Myneni, M. B., & Quadhari, F. S. (2025). *[Artículo sobre resume parsing/matching y uso de cosine similarity y embeddings]*. *IJECE, 12*(9), 63–71.

Nawalny, M., Łepicki, M., Latkowski, T., et al. (2025/2026). Comparative evaluation of GPT-4o, GPT-OSS-120B and Llama-3.1-8B-Instruct language models in a reproducible CV-to-JSON extraction pipeline. *Applied Sciences, 16*, 217. https://doi.org/10.3390/app16010217

Oracle. (s. f.). What is an applicant tracking system? https://www.oracle.com/human-capital-management/recruiting/what-is-applicant-tracking-system/Oracle

Psico Smart. (2024). Incorporating artificial intelligence in candidate experience management. Retrieved from https://psicosmart.com/en/blogs/blog-incorporatingartificial-intelligence-in-candidate-experiencemanagement10305#:~:text=According%20to%20a%20study%20by,per%2Dhire%20by%2025%25.

Purcell, K. (2025, 14 de julio). 2025 Applicant Tracking System (ATS) Usage Report: Key Shifts and Strategies for Job Seekers. Jobscan. https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/

Salazar, E. (2023). Análisis del impacto de Inteligencias Artificiales en la experiencia de los usuarios universitarios: un enfoque en la minería de datos y el análisis de sentimiento. Caso de estudio: ChatGPT. Pontificia Universidad Católica del Ecuador.

Saliu, S. (2025). The algorithmic turn in talent acquisition: A critical analysis of AI-mediated recruitment

technologies. Iconic Research and Engineering Journals, 8(9), 755–767.

Sangwan, B., Sangeeta, & Diksha. (2025). Intelligent Resume Analyzer. Documento presentado en la conferencia IRDO 2025.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591–611.

Student. (1908). The probable error of a mean. Biometrika, 6(1), 1–25. https://doi.org/10.2307/2331554

Sun, H., Lin, H., Yan, H., Song, Y., Gao, X., & Yan, R. (2025). MockLLM: A multi-agent behavior collaboration framework for online job seeking and recruiting. In *Proceedings of KDD '25*. https://doi.org/10.1145/3711896.3737051

Suraj, M., Aruna Kumari, K., & Binila B. Chandran. (2019). A descriptive study on applicant tracking system: Automation software for recruitment and selection. International Journal of Research and Analytical Reviews, 6(1), 223–230.

Tyagi, S., Anuj, Qian, W., Xie, J., & Andrews, R. (2024). Enhancing gender equity in resume job matching via debiasing-assisted deep generative model and gender-weighted sampling. International Journal of Information Management Data Insights, 4, 100283. https://doi.org/10.1016/j.jjimei.2024.100283

Vyshotravk, D. (2025). Smart application for recruiting based on natural language processing methods, transformer models and siamese neural network architecture. *Volume 17*(Issue 5), 156.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83.

ANEXOS

Tabla 6

Operacionalización computacional de las variables utilizadas

Variable en el código	Denominación en el artículo	Tipo y rango	Cálculo en el pipeline clásico — Python/TF-IDF	Cálculo en el pipeline asistido por OpenAI	Observación metodológica
method	Método de análisis semántico	Categórica nominal: python / openai	Toma el valor python cuando la observación procede del pipeline TF-IDF.	Toma el valor openai cuando la observación procede de la API.	Es la variable independiente del estudio.
similarity_score	Similitud semántica global	Continua ([0,1])	Se vectorizan la oferta y el CV mediante TF-IDF con n-gramas de palabras de 1 a 2 términos. Se calcula la similitud del coseno. Cuando el resultado es prácticamente cero, se aplica un respaldo con n-gramas de caracteres char_wb de 3 a 5.	El modelo estima la similitud semántica global entre CV y oferta conforme al prompt estandarizado. La respuesta es convertida a número y limitada al rango ([0,1]).	Es la métrica principal de adecuación semántica.
coverage	Cobertura de términos o conceptos relevantes	Continua ([0,1])	Se extraen hasta 30 términos y bigramas relevantes de la oferta. Después se calcula: coverage = present.mean(), equivalente a términos relevantes presentes en el CV divididos para el total de términos seleccionados.	El modelo identifica aproximadamente 30 términos o conceptos clave de la oferta, considera equivalencias semánticas y estima la proporción encontrada en el CV.	El clásico mide principalmente presencia léxica; OpenAI puede considerar sinónimos y equivalencias.
diversity	Diversidad de términos relevantes	Continua ([0,1])	Se calcula como unique_rel / total_rel, donde unique_rel es el número de términos relevantes distintos presentes y total_rel es el total de ocurrencias relevantes en el CV. Solo se calcula cuando total_rel > 0.	El modelo estima la razón entre términos relevantes distintos y el total de ocurrencias relevantes identificadas en el CV.	Aunque la definición solicitada es semejante, en OpenAI es una estimación contextual y en Python es un conteo determinista.

Variable en el código	Denominación en el artículo	Tipo y rango	Cálculo en el pipeline clásico — Python/TF-IDF	Cálculo en el pipeline asistido por OpenAI	Observación metodológica
density	Densidad relativa de contenido relevante	Continua ([0,1])	Se calcula como total_rel / total_tokens, donde total_rel representa las ocurrencias de términos relevantes y total_tokens el total de tokens o palabras procesables del CV.	El modelo estima el total de ocurrencias relevantes y lo divide para la longitud aproximada del CV en palabras.	Es la métrica con mayor concordancia observada entre métodos.
confidence	Confianza de la evaluación	Continua ([0,1])	Se utiliza una variable proxy: 0.5 × (similarity_score + (1 - sims_std)). sims_std es la desviación estándar de las similitudes entre las oraciones del CV y la oferta. El resultado se limita a ([0,1]).	El modelo declara su confianza considerando suficiencia, claridad y consistencia de la evidencia textual. El valor se valida y limita a ([0,1]).	No posee exactamente la misma operacionalización en ambos métodos; debe interpretarse como una métrica exploratoria.
rationales	Evidencias textuales de la coincidencia	Lista de fragmentos de texto	El CV se divide en oraciones. Se calculan sus similitudes frente a la oferta y se seleccionan las tres oraciones con mayor valor.	El modelo devuelve hasta tres citas textuales del CV, con una longitud máxima de 240 caracteres, que respaldan la evaluación.	No es una variable numérica directa; constituye la fuente para n_rationales y rationale_relevance.
rationale_relevance	Relevancia de las evidencias textuales	Continua ([0,1])	Se calcula localmente como la media de las similitudes del coseno entre cada rationale y la oferta, usando TF-IDF con n-gramas de 1 a 2 y stopwords en español.	Se aplica exactamente el mismo cálculo local a los fragmentos generados por OpenAI; el modelo no determina directamente el valor final.	Es una de las variables más comparables porque el procedimiento final de cálculo es común para ambos pipelines.
n_rationales	Número de evidencias textuales	Discreta, número entero	Se calcula mediante len(rationales), normalmente con un máximo de tres fragmentos.	Se cuenta el número de fragmentos válidos devueltos por el modelo, también con un máximo de tres.	Al presentar varianza nula no fue evaluable mediante contrastes o medidas de concordancia.
latency_ms	Tiempo de procesamiento	Continua, milisegundos	Se mide el tiempo transcurrido desde el inicio hasta el final del procesamiento de cada CV mediante time.perf_counter() y se convierte a milisegundos.	Incluye el tiempo de envío a la API, procesamiento remoto, recepción y validación de la respuesta.	Es una variable operativa, no una métrica de adecuación semántica.
db_score	Puntaje transformado para la base de datos	Continua ([-1,1])	Se calcula mediante db_score = 2 × similarity_score - 1.	Se aplica la misma transformación al similarity_score generado por OpenAI.	Es una transformación lineal de similarity_score; no constituye una métrica semántica independiente.
score	Campo de puntaje almacenado en SQL	Continua ([-1,1])	En la tabla prediction, el campo score recibe el valor de db_score.	En la tabla prediction, también recibe el db_score transformado.	score es el nombre del campo en SQL; conceptualmente equivale a db_score.
delta o (d_i)	Diferencia pareada Python - OpenAI	Continua	Se calcula después de unir los resultados: (d_i=X_i^{classic}-X_i^{OpenAI}).	No se calcula dentro del pipeline OpenAI; se obtiene durante el análisis estadístico.	Es la variable analizada por Shapiro-Wilk, t pareada, Wilcoxon y el forest plot.
bias	Sesgo medio entre métodos	Continua	Corresponde a la media de los deltas: (\bar d=\frac{1}{n}\sum d_i).	Se calcula conjuntamente utilizando los pares de ambos métodos.	Un valor positivo indica valores superiores en Python; uno negativo indica valores superiores en OpenAI.
loa_lower / loa_upper	Límites de acuerdo de Bland–Altman	Continua	Se calculan como bias - 1.96 × SD(delta) y bias + 1.96 × SD(delta).	Se obtienen de las diferencias entre ambos métodos.	Muestran el intervalo en el que se espera aproximadamente el 95 % de las diferencias individuales.
p_holm	Valor p ajustado por Holm	Continua ([0,1])	Se obtiene ajustando conjuntamente los valores p de las pruebas realizadas para cada métrica.	No pertenece a ningún pipeline; se calcula en la etapa estadística comparativa.	Controla el error familiar por comparaciones múltiples, pero no garantiza la eliminación absoluta de falsos positivos.

Nota: *Las variables de tiempo se expresan en milisegundos y n_rationales como conteo entero.*