Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

Trabajo previo a la obtención de título de Magister en Inteligencia Artificial Aplicada

AUTOR/ES:

Álvaro Xavier Trejo Rodríguez
Jennifer Mikaela Guerrero Duque
Cristhian Alberto Motoche Macas

TUTOR/ES:

Fernanda Paulina Vizcaíno Imacaña

Andrés Roberto Navas Castellanos

Tema:

Desarrollo de un sistema de procesamiento de lenguaje natural para extraer habilidades en hojas de vida y realizar coincidencia semántica con ofertas laborales.

Quito - Ecuador

Agosto – 2026

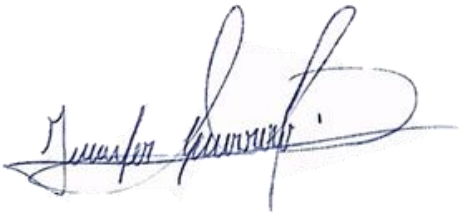
REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, **Álvaro Xavier Trejo Rodríguez, Jennifer Mikaela Guerrero Duque y Cristhian Alberto Motoche Macas,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma
Álvaro Xavier Trejo Rodríguez



Firma
Jennifer Mikaela Guerrero Duque

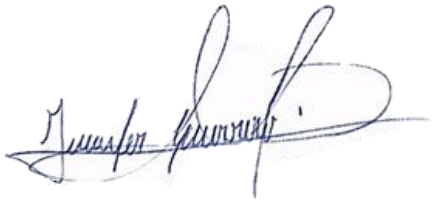
Firma
Cristhian Alberto Motoche Macas

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Álvaro Xavier Trejo Rodríguez, Jennifer Mikaela Guerrero Duque y Cristhian Alberto Motoche Macas**, en calidad de autores del trabajo de investigación titulado ***Desarrollo de un sistema de procesamiento de lenguaje natural para extraer habilidades en hojas de vida y realizar coincidencia semántica con ofertas laborales***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, agosto 2026

Firma
Álvaro Xavier Trejo Rodríguez

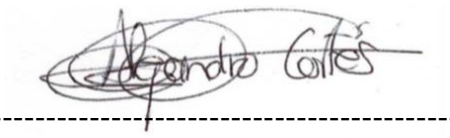


Firma
Jennifer Mikaela Guerrero Duque

Firma
Cristhian Alberto Motoche Macas

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López Director EIG y Karla Estefanía Mora Cajas Coordinadora UIDE,** declaramos que: **Álvaro Xavier Trejo Rodríguez, Jennifer Mikaela Guerrero Duque, y Cristhian Alberto Motoche Macas,** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López
Director de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

Dedicatoria

Con mucho agradecimiento, dedicamos este trabajo principalmente a Dios, quien nos ha guiado con paz y sabiduría, con fortaleza en los momentos difíciles y con su aliento cuando más lo necesitábamos.

A nuestras familias, por comprendernos y respetarnos al seguir este camino donde hubo sacrificios; su bondad fue clave durante todo este proceso. Finalmente, a nuestros amigos y parejas sentimentales, quienes nos han brindado apoyo incondicional en todo momento.

Agradecimiento

Agradecemos a todos quienes han formado parte de esta maestría: docentes, compañeros y personal auxiliar, quienes con esfuerzo, dedicación, apoyo y entrega fueron guía para nosotros a lo largo de esta formación académica.

Asimismo, agradecemos a Dios por su apoyo y guía incondicional a lo largo de esta maestría.

Finalmente, agradecemos a todos nuestros familiares, amigos y parejas sentimentales, quienes supieron brindarnos ánimos y aliento durante todo este proceso. Sin ellos, no se hubiese logrado este proyecto.

Resumen

Las pequeñas y medianas empresas suelen revisar las hojas de vida de manera manual y cuentan con recursos limitados para incorporar sistemas comerciales de reclutamiento. Esta situación hace que la preselección de candidatos demande tiempo, sea propensa a errores y dependa del criterio individual de quien realiza la revisión. En el presente trabajo se desarrolla el sistema de apoyo para reclutamiento en PYMEs denominado FreeATS, el cual es un sistema basado en procesamiento de lenguaje natural que analiza hojas de vida en español, compara la información de los candidatos con los requisitos de una vacante y genera un ranking de adecuación.

Para construir el sistema se consolidó un corpus de 2.713 hojas de vida, traducidas y adaptadas al español latinoamericano. El pipeline procesa documentos en formato PDF, DOCX y texto plano, y extrae habilidades, cargos y organizaciones mediante Reconocimiento de Entidades Nombradas y la taxonomía ESCO. Para el emparejamiento entre candidatos y vacantes se evaluaron enfoques léxicos y semánticos basados en TF-IDF, SBERT, TinyBERT y DistilBERT. Estos componentes se integraron en una aplicación web local que permite gestionar vacantes, cargar candidatos, consultar la información extraída y visualizar un ranking acompañado de un desglose de puntajes y requisitos no encontrados. La aplicación puede ejecutarse en equipos convencionales, sin una GPU dedicada ni una conexión permanente a servicios en la nube.

Para construir el sistema se consolidó un corpus de 2.713 hojas de vida, traducidas y adaptadas al español latinoamericano. El pipeline procesa documentos en formato PDF, DOCX y texto plano, y extrae habilidades, cargos y organizaciones mediante Reconocimiento de Entidades Nombradas y la taxonomía ESCO. Para el emparejamiento entre candidatos y vacantes se evaluaron enfoques léxicos y semánticos basados en TF-IDF,

SBERT, TinyBERT y DistilBERT. Estos componentes se integraron en una aplicación web local que permite gestionar vacantes, cargar candidatos, consultar la información extraída y visualizar un ranking acompañado de un desglose de puntajes y requisitos no encontrados. La aplicación puede ejecutarse en equipos convencionales, sin una GPU dedicada ni una conexión permanente a servicios en la nube.

El módulo de extracción alcanzó una puntuación F1 global de 0,9217, superando el umbral de 0,80 definido para el proyecto. En la comparación semántica, SBERT obtuvo un resultado de 0,7256 frente a 0,2735 de TF-IDF, mientras que DistilBERT ajustado para la tarea alcanzó un F1 de 0,9412 en clasificación binaria. La evaluación exploratoria del ranking mostró una correlación de Spearman de 0,232 con CareerCorpus, lo que evidencia una señal positiva, aunque todavía insuficiente para considerar cerrada su validación. En conjunto, los resultados demuestran que es viable implementar una solución de PLN útil y accesible para apoyar la preselección de candidatos en entornos con recursos limitados. No obstante, se requiere ampliar la validación humana, incorporar datos originalmente redactados en español y realizar pruebas con personal de recursos humanos de PYMEs ecuatorianas.

Palabras clave: procesamiento de lenguaje natural, hojas de vida, selección de personal, PYMEs, Reconocimiento de Entidades Nombradas, emparejamiento semántico.

Abstract

Small and medium-sized enterprises often review résumés manually and have limited resources to adopt commercial recruitment systems. As a result, candidate screening can be time-consuming, prone to errors, and highly dependent on the individual judgment of the person conducting the review. To address this problem, this study presents FreeATS, a natural language processing system that analyzes Spanish-language résumés, compares candidate information with job requirements, and produces a candidate–job suitability ranking. The system is intended to support recruitment activities in SMEs rather than replace human judgment in hiring decisions.

A corpus of 2,713 résumés, translated and adapted to Latin American Spanish, was consolidated to develop the system. The resulting pipeline processes PDF, DOCX, and plain-text documents and extracts skills, job titles, and organizations using Named Entity Recognition and the ESCO taxonomy. Lexical and semantic approaches based on TF-IDF, SBERT, TinyBERT, and DistilBERT were evaluated for candidate–job matching. These components were integrated into a local web application that allows users to manage job openings, upload résumés, review the extracted information, and examine a ranked list of candidates together with score breakdowns and missing requirements. The application can run on standard office computers without a dedicated GPU or a permanent connection to cloud services.

The extraction module achieved an overall F1 score of 0.9217, exceeding the 0.80 threshold established for the project. In the semantic comparison, SBERT achieved a score of 0.7256, compared with 0.2735 for TF-IDF, while the fine-tuned DistilBERT model reached an F1 score of 0.9412 in binary classification. The exploratory ranking evaluation produced a Spearman correlation of 0.232 with CareerCorpus, indicating a positive signal, although the

available evidence is not yet sufficient to consider the ranking fully validated. Overall, the findings show that a useful and accessible NLP solution can support candidate screening in resource-constrained environments. Further work should expand human validation, incorporate résumés originally written in Spanish, and evaluate the application with human resources professionals from Ecuadorian SMEs

Keywords: natural language processing, résumés, recruitment, SMEs, Named Entity Recognition, semantic matching.

Índice

Contenido

Índice de tablas

Índice de figuras

CAPÍTULO 1

1. Introducción

1.1.Definición del proyecto

La digitalización se ha convertido en un factor relevante para la competitividad y sostenibilidad de pequeñas y medianas empresas ecuatorianas (PYMEs). Sin embargo, las PYMEs ecuatorianas enfrentan problemas durante este proceso debido a la poca formación en digitalización, la escasez de recursos económicos y la falta de la infraestructura necesaria (Sáenz de Viteri et al., 2026). Todos estos factores marcan una distancia entre la aspiración y lo que ocurre realmente en cuanto a la digitalización en ámbitos muy concretos como es el caso de la gestión del talento humano.

En lo que respecta al reclutamiento, la tecnología adquirió un papel diferenciador, ya que ha permitido que las PYMEs puedan encontrar profesionales con perfiles más adecuados para satisfacer las necesidades de la empresa, y simplifica los procesos de selección (Biea et al., 2024). Sin embargo, la mayoría de las PYMEs todavía no cuentan con herramientas propias del procesamiento automatizado de hojas de vida y dependen de plataformas de terceros que les generan costos que no parecieran eficientes para empresas con ciclos de contratación esporádicos (Nikolaou, 2021).

En el caso de las PYMEs, el proceso de selección de personal se basa en métodos que se desarrollan en su mayoría de forma manual o informal (dependiendo generalmente de referidos o del uso de anuncios con escasa cobertura). La ausencia de herramientas adecuadas lleva a que la selección de personal sea ineficaz, muy diferente de lo que hacen grandes empresas las cuales poseen medios o tecnologías con las cuales procesar un gran volumen de información y así acelerar el proceso de selección de personal (Nikolaou, 2021). La selección

para el área de talento humano representa un gran esfuerzo que consiste en recolectar; analizar, y procesar hojas de vida hasta poder llegar a la mejor elección. Este proceso manual es propenso a errores y a retrasos tanto para las propias empresas como para los candidatos.

El núcleo del problema se encuentra en la complejidad de poder analizar de manera eficiente el contenido textual no estructurado de las hojas de vida y de poderlo cruzar con las exigencias marcadas en las descripciones de puestos de trabajo (Sridevi & Suganthi, 2022). Algunas investigaciones recientes han intentado encontrar respuesta a este problema dividiéndolo en diferentes vías: la aplicación de sistemas basados en el PLN para poder extraer las competencias relevantes de las hojas de vida y compararlas con competencias de habilidades aplicando reglas de similitud, como Jaccard o Coseno (Saatçı et al., 2024) o la aplicación de modelos de aprendizaje profundo que puedan predecir habilidades de alto nivel de las hojas de vida, incluso cuando las hojas de vida no mencionan ni proponen explícitamente tales habilidades (Jiechieu et al., 2021). Sin embargo, estas soluciones han sido desarrolladas prioritariamente en inglés y en entornos de alta disponibilidad computacional, lo cual limita su poder de aplicación en PYMEs latinoamericanas.

Falta una línea de investigación que se enfoque particularmente en el idioma español y que pueda funcionar en un entorno de recursos limitados. Existen trabajos en la línea de la extracción de habilidades, donde las técnicas han avanzado sobre conjuntos de datos anotados en inglés (Otani et al., 2025) pero persiste un vacío metodológico para el español. Por otra parte, la aplicación de métodos del estado del arte, como los grandes modelos de lenguaje, a situaciones con recursos limitados, como ocurre por ejemplo en países en desarrollo, es desafiante en la práctica, por lo que es necesario revisar enfoques más livianos y eficientes (Wang et al., 2023). De la misma manera, los sistemas Applicant Tracking Systems (ATS) que se encuentran disponibles actualmente con soluciones en español no son adaptables a las condiciones de las PYMEs, ya que son soluciones de pago basadas en la nube.

Para resolver la problemática, el presente proyecto propone FreeATS, el cual es un sistema que implementa técnicas de procesamiento de lenguaje natural (PLN) que permita analizar de forma automática el contenido textual de las hojas de vida y descripciones de puestos con énfasis en su funcionamiento en sistemas con recursos limitados. El sistema extraerá habilidades a partir de técnicas como el Reconocimiento de Entidades Nombradas (NER), calculará métricas de similitud semántica para el cálculo de la compatibilidad entre candidatos y vacantes, y presentará un ranking de los candidatos más aptos para la posición. Con respecto a los modelos, se plantean alternativas ligeras como DistilBERT y TinyBERT, debido a que estos modelos son versiones pequeñas y rápidas de BERT, que buscan mantener el 97% del rendimiento con una fracción de los parámetros (Sanh et al., 2019), lo cual los hace adecuados para entornos sin infraestructura de GPU dedicada.

El sistema se evaluará y validará como resultado de la preparación y selección de conjuntos de datos abiertos, cada uno de los cuales se organizará en función de su función en el pipeline: el Resume Dataset PDF y el Resume Dataset DOCX, que se encuentran en Kaggle, forman el corpus principal del sistema una vez traducido y adaptado al español latinoamericano; el CareerCorpus proporciona perfiles estructurados para realizar las pruebas de emparejamiento; y para la validación en condiciones reales se analizó el dataset de emparejamiento vacante–hoja de vida, el dataset Data Voto Informado 2026, además de una colección propia de hojas de vida del contexto ecuatoriano. De la misma manera, se usará la taxonomía European Skills, Competences, Qualifications and Occupations (ESCO) como referencia taxonómica conceptual y estructural desde la cual se realizará la extracción de habilidades y ocupaciones a partir de texto no estructurado, esto es, una taxonomía abierta y multilingüe que permite alinear las habilidades extraídas con estándares de mercado laboral (Zhang et al., 2023).

En este sentido, el sistema extraerá habilidades relevantes de las hojas de vida, realizará una coincidencia semántica con una descripción laboral, calculará el índice de compatibilidad entre candidatos y vacantes, y generará un ranking según su adecuación al perfil que hay que cubrir, de esta manera, se convierte en un primer filtro automatizado para ayudar en el proceso de selección de reclutamiento de PYMEs. Su interfaz web local e intuitiva permitirá que el personal de recursos humanos sin conocimientos técnicos pueda operar el sistema desde sus propios equipos de oficina.

1.2.Justificación e importancia del proyecto

El presente proyecto es relevante porque plantea la democratización del acceso a la tecnología de Inteligencia Artificial (IA) y herramientas de PLN en beneficio de las PYMEs. El libre acceso a la IA es cada vez más común, por lo que los empleados no técnicos pueden desarrollar soluciones de IA personalizadas para resolver diversas tareas operativas correspondientes a sus áreas de trabajo (Sundberg & Holmström, 2023); sin embargo, este proceso de democratización no ha llegado equitativamente a toda la gestión del talento humano de las PYMEs en ciertos países, como Ecuador (Biea et al., 2024; Sáenz de Viteri et al., 2026).

Los sesgos algorítmicos, la calidad de los datos y el cumplimiento normativo plantean desafíos relevantes en las PYMEs, que operan con escasez de recursos (Koman et al., 2024), por lo tanto se vuelve necesario desarrollar soluciones que al mismo tiempo sean eficientes tecnológicamente y accesibles. Por tal motivo, el proyecto plantea utilizar técnicas de PLN eficaces y ejecutables en entornos con recursos limitados, sin depender de infraestructura en la nube. De esta manera, la transformación digital en el área de recursos humanos será accesible para las PYMEs.

La viabilidad del proyecto se basa en dos premisas: la existencia de datos abiertos de hojas de vida y de descripciones de puestos, y la disponibilidad de herramientas de código abierto para PLN como spaCy. Esta biblioteca ha demostrado ser eficaz en tareas como la extracción y categorización automática de información de hojas de vida mediante NER. La conjunción de estos recursos permite la implementación de un sistema funcional sin necesidad de inversiones adicionales en infraestructura especializada.

Desde un punto de vista económico, la viabilidad del proyecto se refuerza si se consideran los costes que representan las actuales soluciones comerciales en el mercado. Las PYMEs que adoptan sistemas ATS comerciales enfrentan modelos de licencias por suscripción que son muy difíciles de justificar en el caso de PYMEs que tienen ciclos de contratación temporales (Nikolaou, 2021). Mientras que las PYMEs que adoptan el proceso de selección manual, el personal de reclutamiento designado emplea un tiempo considerable (Biea et al., 2024; Nikolaou, 2021). De esta manera, FreeATS elimina estos obstáculos al no requerir licencias, suscripciones ni infraestructura especializada, siendo ejecutable en ordenadores de oficina convencionales con un mínimo de 8 GB de RAM y sin necesidad de conexión permanente a internet.

De este modo, los profesionales del área de recursos humanos utilizarán un sistema ejecutable en sus computadores de uso cotidiano y podrán realizar una evaluación inicial de los candidatos.

1.3.Alcance

El alcance del presente proyecto es la construcción de una herramienta que, a partir de una hoja de vida en formato PDF, DOCX o texto plano, permite extraer información de interés como habilidades, educación o historia laboral, realice una coincidencia semántica

con una descripción laboral, y ofrezca una lista de los candidatos más aptos como la primera fase del proceso de selección automatizado. La herramienta se presentará como una interfaz web que funcionará localmente, podrá ejecutarse en ordenadores de recursos limitados, se desarrollará mediante componentes de código libre y mediante datos abiertos disponibles en la comunidad académica y de investigación.

El resultado final del presente proyecto será una interfaz web local, ejecutable desde cualquier navegador mediante un servidor ligero y diseñada para operar utilizando los recursos de la propia computadora del usuario, evitando la dependencia de infraestructura especializada o soluciones de alto costo. La interfaz le permitirá al usuario subir hojas de vida en diferentes formatos (PDF, Word o texto plano), visualizar el resultado del análisis de habilidades y categorías extraídas por medio del modelo de PLN de una forma amigable, e ingresar los requisitos de una vacante específica; con ello, se consultará el ranking de adecuación candidato-vacante generado por el modelo. Una interfaz local facilita el uso de la herramienta para profesionales de recursos humanos de PYMEs que no tengan conocimientos técnicos especializados y evita la necesidad de ejecutar comandos, instalar programas, realizar configuraciones avanzadas, más que una simple configuración inicial. El prototipo será funcional y de validación, demostrando la viabilidad técnica del pipeline de PLN en condiciones reales con recursos limitados.

Este proyecto no prevé integrar el sistema con plataformas externas para reclutar, con los sistemas de HRMS (Human-Resource Management Systems) o con las bases de datos de la propia empresa. Del mismo modo, no se prevé la automatización total de los procesos de reclutamiento, o de la toma de decisiones de contratación: el sistema es una herramienta que apoya la decisión humana; no es un sustituto. Este mismo proyecto queda excluido de la evaluación de competencias conductuales o blandas que requieren procedimientos de toma de decisiones cuya evaluación no puede realizarse a partir de una hoja de vida.

1.4.Hipótesis

La implementación de un sistema basado en técnicas de procesamiento de lenguaje natural facilita la extracción de información de hojas de vida, ayuda en la comparación respecto a los requisitos de vacantes laborales, y presenta a los candidatos más aptos de forma más eficiente que el filtrado manual típico de PYMEs y que los enfoques clásicos basados en similitud léxica como TF-IDF, en entornos de computación con recursos limitados. Se utilizarán métricas de precisión, exhaustividad y puntuación F1, esperando obtener valores superiores a 0,80 en la identificación de habilidades, y métricas de correlación con rankings humanos (e.g. Spearman) para evaluar la calidad del ranking de candidatos generado.

1.5.Objetivos

1.5.1. Objetivo general

Desarrollar un sistema basado en técnicas de procesamiento de lenguaje natural que permita extraer información de hojas de vida en español, compararlas con los requisitos de vacantes laborales, y generar un ranking de adecuación candidato-vacante presentado a través de una interfaz web local, funcional en entornos de recursos computacionales limitados.

1.5.2. Objetivos específicos

- Definir las categorías de información relevante a extraer de las hojas de vida, incluyendo habilidades técnicas, formación académica y experiencia laboral. Con el fin de estructurar el esquema de extracción del sistema, mediante una revisión de la literatura y análisis de taxonomías abiertas como ESCO.
- Implementar un pipeline de procesamiento de texto que incluya las etapas de limpieza, normalización y estructuración de los datos extraídos de hojas de vida en formato PDF, DOCX y texto plano, con soporte para documentos en español.

- Diseñar e implementar un módulo de comparación semántica entre la información extraída de las hojas de vida y los requisitos especificados en las descripciones de puestos de trabajo, evaluando su desempeño frente a tres baselines: un enfoque léxico basado en TF-IDF con similitud coseno, un enfoque semántico intermedio basado en sentence-transformers (SBERT), y un modelo eficiente contextual como TinyBERT sin fine-tuning.
- Desarrollar un modelo de clasificación y ranking que categorice a los candidatos según su nivel de adecuación al perfil requerido, considerando dimensiones como habilidades, experiencia y formación académica, evaluando la calidad del ranking generado mediante métricas de correlación con rankings humanos como Spearman.
- Evaluar el desempeño del sistema mediante métricas estándar de precisión, exhaustividad (recall) y puntuación F1 para el módulo de extracción de habilidades, y métricas de correlación con rankings humanos (Spearman) para el módulo de emparejamiento, con el fin de determinar la efectividad del primer filtro automatizado generado frente a los baselines definidos.
- Implementar una interfaz web local, ejecutable desde cualquier navegador sin conexión a internet ni infraestructura en la nube, que permite la interacción con el sistema en condiciones de bajos recursos computacionales, sin requerir conocimientos técnicos especializados por parte del usuario final, permitiéndole cargar hojas de vida, consultar habilidades clave y visualizar el ranking de adecuación candidato–perfil generado por el sistema.

CAPÍTULO 2

2. Revisión de literatura

2.1.Estado del arte

En los últimos años, se incrementó la implementación de modelos basados en técnicas de PLN como herramienta complementaria para el proceso de contratación en el área de talento humano, agilizando tareas como la extracción y clasificación de información relevante en hojas de vida (Otani et al., 2025). Diversos estudios han demostrado la utilidad de este enfoque al permitir reducir el esfuerzo manual que implica revisar las hojas de vida y mejorar la eficiencia en la selección de candidatos (Trovão et al., 2025).

De acuerdo con la literatura revisada, el preprocesamiento, la extracción de información y la comparación entre hojas de vida y vacantes suelen abordarse mediante técnicas de PLN como TF-IDF y los embeddings, junto al uso de métricas de similitud de Jaccard y de coseno para determinar el grado de emparejamiento entre los candidatos y las vacantes laborales (Sridevi & Suganthi, 2022; Saatçı et al., 2024).

En el desarrollo del estado del arte se distinguen tres grandes oleadas en la automatización del reclutamiento. La primera oleada corresponde a los sistemas basados en reglas y en estadística (Saatçı et al., 2024; Sridevi & Suganthi, 2022), que utilizan el método TF-IDF, con unos sistemas eficaces pero rígidos frente a variaciones semánticas. La segunda oleada corresponde a los modelos de representación profunda (Deshmukh & Raut, 2024; Li et al., 2025), que giran en torno a la semántica mediante arquitecturas Transformer y que alcanzan altas precisiones con un alto costo, lo que implica una barrera para implementarlos en entornos con recursos limitados. La tercera oleada, corresponde a los modelos destilados y

eficientes (Kiruthiga Devi et al., 2025; Marliyas et al., 2025), que intentan conciliar ambas posturas mediante arquitecturas ligeras como la DistilBERT, y es en esta corriente donde se sitúa el interés del presente proyecto.

Primera ola: Enfoques basados en métodos estadísticos y machine learning.

Saatçı et al. (2024) proponen un sistema de screening de hojas de vida basado en PLN, donde se aplican técnicas de tokenización (segmentación del texto en unidades mínimas), eliminación de palabras vacías (stopwords), lematización (reducción de cada palabra a su forma base, por ejemplo “trabajando” a “trabajar”) y cálculo de emparejamiento semántico para clasificar y filtrar candidatos. Según Saatçı et al. (2024), el sistema propuesto permitió automatizar el filtrado y la clasificación de candidatos.

Otras investigaciones se han centrado en explorar el uso de algoritmos de machine learning, como Support Vector Machines (SVM), K-Nearest Neighbors (KNN) y regresión logística, en combinación con técnicas de extracción de texto, como bag of words (representación de un texto como un conjunto de palabras sin considerar su orden) y TF-IDF (medida que pondera la relevancia de una palabra en un documento en relación con su frecuencia del total de documentos). Estas metodologías han demostrado su eficacia para analizar y categorizar hojas de vida, y medir la adecuación entre perfiles y vacantes (Sridevi & Suganthi, 2022). Asimismo, el uso de Named Entity Recognition (NER) ha sido útil para extraer información clave, como educación, trayectoria profesional y habilidades destacadas de los candidatos, reorganizando hojas de vida en datos y formatos comparables (Senger et al., 2024). Este tipo de modelos tiene como ventaja el bajo costo computacional, pero su principal barrera es la baja precisión ante un vocabulario técnico variado o el diferente tipo de estructuras que pueden tener las hojas de vida.

Segunda ola: Modelos avanzados en Transformers y aprendizaje profundo. Los trabajos más recientes apuntan hacia modelos más avanzados basados en técnicas de deep learning y arquitecturas tipo Transformer. Diversos autores coinciden en que la precisión mejora bastante al usar estas arquitecturas (Kinger et al., 2024; Deshmukh & Raut, 2024); sin embargo, el coste computacional continúa constituyendo el principal obstáculo para su uso en entornos con restringidos recursos (Liu et al., 2023) .

En otras investigaciones se han propuesto modelos que combina la visión por computadora usando YOLOv5 y DistilBERT (Kinger et al., 2024) para realizar parsing y ranking automático de hojas de vida, permitiendo extraer información estructurada desde distintos formatos y mejorar la estandarización del análisis, alcanzando una precisión cercana al 96%. De manera equivalente, Sridevi y Suganthi (2022) plantean un sistema basado en técnicas de PLN y machine learning para el análisis, categorización y ranking de candidatos, enfatizando la importancia del contexto semántico para mejorar la precisión del emparejamiento. A diferencia de los métodos TF-IDF y similitud del coseno, los modelos híbridos son más robustos en contextos donde las hojas de vida no tienen una estructura homogénea y su vocabulario técnico es variado, sin embargo, requieren más recursos computacionales (Deshmukh & Raut, 2024).

Bajo la misma premisa, Deshmukh y Raut (2024) implementaron una solución híbrida que combina múltiples enfoques de aprendizaje automático y PLN para el procesamiento de hojas de vida, consiguiendo una extracción estructurada de datos y campos importantes. Este tipo de investigaciones supera con creces a métodos basados en reglas textuales o el uso de palabras clave (Deshmukh & Raut, 2024). A diferencia del enfoque estadístico de la primera ola, estos modelos muestran una mejor comprensión contextual, a pesar de estar desplegados localmente.

La literatura también abarca propuestas enfocadas en el equilibrio entre eficiencia y precisión. Otani et al. (2025) presentan una revisión de técnicas, desafíos y tendencias en el parsing automatizado de hojas de vida, destacando el uso combinado de métodos estadísticos y semánticos. Este enfoque permitió identificar estrategias para estructurar textos en representaciones organizadas, optimizando el proceso de buscar y filtrar candidatos sin condicionar la calidad del proceso en sí.

Análisis de emparejamiento candidato - vacantes. Sobre el emparejamiento entre candidatos y vacantes laborales, cabe señalar que las investigaciones son favorables a la elaboración del análisis semántico en este campo. Li et al. (2025) presentan un modelo de deep learning e implementación de grafos, que permite mayor compatibilidad entre los requerimientos de las vacantes y las hojas de vida a través de representaciones de tipo estructuradas, que obtiene un F1-score de 0,91; mientras que Jiechieu y Tsopze (2021) son más proclives a utilizar el machine learning para hacer parsing y predecir el dominio laboral así como la categoría ocupacional del candidato. En este punto, cabe señalar que el modelo de Li et al. (2025) es preciso en su semántica, superando a los de Saatçı et al. (2024) y Sridevi y Suganthi (2022); no obstante, su elevado costo computacional hace que no sea adecuado para entornos con pocos recursos, como es el caso de las PYMEs.

Las investigaciones empíricas en el campo refuerzan la tesis de la importancia de estos modelos y sistemas: se ha concluido que la automatización y el ranking de candidatos pueden acelerar la selección en comparación con los métodos tradicionales; además, se señala que los modelos híbridos que combinan PLN y machine learning mejoran tanto la eficiencia como la calidad de la selección (Abhishek et al., 2025).

Tercera ola: Aplicaciones en PYMEs y automatización integral del reclutamiento. De un lado Kiruthiga Devi et al. (2025) dan a conocer el ATS con IA para las

PYMEs después de conseguir un 45% menos de tiempo de filtrado y una mejora del 30% en el emparejamiento, lo que demuestra que se pueden aplicar herramientas de PLN en entornos de recursos que son limitados; del otro lado Marliyas et al. (2025) corroboran la revisión anterior al expresar a través de una revisión sistemática que se puede conseguir un recorte de entre el 55% y el 60% del tiempo de reclutamiento y un recorte de los costes de operación en el rango del 20% y el 30%. Estas revisiones corroboran que se mantiene la idea de la importancia de una correcta automatización del proceso de reclutamiento para poder gestar modelos de reclutamiento eficaces en un contexto de presupuestos limitados.

A pesar de que Kiruthiga Devi et al. (2025) validan la capacidad de la IA aplicada a PYMEs desde su perspectiva, a menudo recurren a plataformas cerradas o servicios en la nube de terceros, lo que limita su personalización con datos abiertos, además de su uso en países que cuentan con una infraestructura tecnológica restringida.

Limitaciones y vacíos de investigación. Los modelos investigados, en especial aquellos basados en arquitecturas de BERT y redes neuronales profundas, son computacionalmente exigentes, lo que dificulta su aplicación en escenarios reales de PYMEs. A su vez, la dependencia de conjuntos de datos etiquetados incrementa aún más los costos y restringe la posibilidad de escalar las soluciones. También persisten problemas de generalización, dado que gran parte de las investigaciones se concentran en sectores específicos, como el tecnológico. Finalmente, se enfrentan retos vinculados con los sesgos presentes en los datos y con la transparencia de los modelos, factores que condicionan la confianza y la aceptación de estas herramientas en procesos sensibles como la gestión de talento y la toma de decisiones en recursos humanos.

A pesar de los avances reportados por Li et al. (2025) y Deshmukh y Raut (2024) en cuanto a la precisión semántica de la clasificación de habilidades por desambiguación

automática, el dilema permanece en la literatura: a mayor capacidad de clasificación de habilidades, mayor dependencia de hardware especializado o servicios en la nube caros. Autores como Kiruthiga Devi et al. (2025) validan el uso de la IA en las PYMEs, pero los modelos suelen basarse en interfaces cerradas, y la mayoría de los artículos revisados investigan solamente corpus del idioma inglés, lo cual genera un vacío metodológico para contextos hispanohablantes como Ecuador (Butt et al., 2026).

De esta manera, se identifica la brecha de investigación: en la literatura reciente no se ha detectado un sistema que alimente el potencial de los modelos con la ligereza necesaria para ejecutarse en la infraestructura habitual de las PYMEs en los países en desarrollo, usando únicamente estándares de open data. Por lo tanto, el presente proyecto se enfocará en investigar y desarrollar esta limitación mediante el diseño de un sistema de PLN, combinando técnicas de procesamiento de texto ligeras y modelos de aprendizaje automático, todo sobre entornos controlados que equilibrarán el rendimiento, el consumo de recursos y la facilidad de aplicación, simulando el contexto de empresas PYMEs.

En la Tabla 1, se presenta la comparación en literatura de procesamiento de hojas de vida.

Tabla 1

Comparativa de métodos de procesamiento de hojas de vida

Referencia	Enfoque Principal	Técnicas Clave	Rendimiento	Costo	Limitación para PYMEs
Saatçı et al. (2024)	Estadístico / NLP Tradicional	Tokenización, TF-IDF, Coseno	Eficiencia alta en filtrado	Bajo	Precisión limitada ante lenguaje semántico variado.

Referencia	Enfoque Principal	Técnicas Clave	Rendimiento	Costo	Limitación para PYMEs
Kinger et al. (2024)	Híbrido (CV + NLP)	YOLOv5 + DistilBERT	96% Precisión	Medio a Alto	Requiere GPUs para la parte de visión y parsing complejo.
Li et al. (2025)	Deep Learning / Grafos	GNN + GPT	F1-score: 0.91	Muy Alto	Inviable en servidores locales modestos o sin suscripciones a API.
Kiruthiga Devi et al. (2025)	ATS basado en IA	Automatización de flujo	45% de ahorro de tiempo	Medio	Enfoque comercial/cerrado, difícil de personalizar con datos abiertos.
Propuesta de Investigación	Híbrido Ligero	NLP + ML + Datos Abiertos	Por determinar	Bajo (Local)	Ninguna (Diseñado para recursos limitados).

2.2.Marco Teórico

2.2.1. Extracción de información y reconocimiento de entidades nombradas

NER es una de tarea de PLN que consiste en localizar e identificar en determinados documentos no estructurados los fragmentos de texto que pertenecen a categorías predeterminadas (habilidades, cargos o entidades, entre otros) e identificarlos con una única etiqueta semántica (Seow et al., 2025). En FreeATS, el modelo NER reconoce tres tipos de entidades: SKILL (habilidades técnicas y habilidades blandas), JOB_TITLE (cargos o puestos de trabajo) y ORG (empresas, instituciones educativas y otras organizaciones).

Un gold standard es un conjunto de referencia anotado y revisado manualmente por personas expertas; se utiliza para estimar el desempeño del modelo frente a criterio humano (Wissler et al., 2014). Un silver standard es un conjunto de referencias cuyas etiquetas se generan automática o semiautomáticamente; reduce el costo de anotación, pero puede heredar

errores y sesgos del procedimiento que produjo las etiquetas (Rebholz-Schuhmann et al., 2011).

Un elemento metodológico central presente en este proyecto es la diferenciación entre la anotación gold-standard y la silver-standard. La gold-standard corresponde a la verificada por evaluadores humanos, mientras que la silver-standard se deriva de un proceso automático a partir de otro proceso que se manifiesta como fiable. En FreeATS se utilizó un silver standard porque no se dispuso de un corpus público en español anotado manualmente con las categorías SKILL, JOB_TITLE y ORG. Las etiquetas SKILL se generaron mediante emparejamiento léxico-semántico con ESCO (Comisión Europea, 2024), mientras que los campos estructurados se obtuvieron con GPT-4o-mini de OpenAI. El proceso no incluyó anotadores humanos. Por ello, la F1 resultante mide fidelidad al etiquetado automático y no desempeño frente a un gold standard independiente (Rebholz-Schuhmann et al., 2011; Wissler et al., 2014)

Similitud semántica y modelos de embeddings. La comparación entre la información de una hoja de vida y las condiciones de una oferta de trabajo requiere que ambas representaciones de texto tengan un espacio común en el que se puedan comparar. La forma más sencilla es Term Frequency-Inverse Document Frequency (TF-IDF), una representación dispersa en forma de vector que pondera cada término mediante su cantidad de repeticiones en el documento y su rareza en el corpus (Lan, 2022). TF-IDF es computacionalmente barato, pero solo es capaz de capturar coincidencias léxicas literales: dos textos con un significado similar, pero que emplean sinónimos o fórmulas diferentes, obtendrán un índice de similitud bajo en el marco de este esquema.

Los modelos de embeddings de oraciones basados en Sentence-BERT (SBERT) proyectan oraciones completas en un espacio vectorial denso, donde la similitud del coseno refleja cercanía semántica (Reimers & Gurevych, 2019). Estos modelos combinan la red BERT con un objetivo de entrenamiento determinado para producir embeddings de oración comparables mediante similitud coseno (Reimers & Gurevych, 2019), a partir de la arquitectura Transformer bidireccional original (Devlin et al., 2019). FreeATS utiliza el modelo multilingüe sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, que produce embeddings de 384 dimensiones y puede ejecutarse en CPU (Reimers & Gurevych, 2019).

Un tercer enfoque que se evalúa es el uso de transformers comprimidos sin entrenamiento específico para la tarea, como TinyBERT (Jiao et al., 2020), el cual es un modelo destilado que reduce el tamaño de BERT, pero preserva la mayoría de sus capacidades en tareas generales. Un hallazgo metodológico importante de este proyecto es que los embeddings de un modelo Transformer sin entrenamiento específico para la tarea presentan anisotropía: los vectores se concentran en una región estrecha del espacio vectorial, de tal manera que casi cualquier par de textos obtiene una similitud coseno artificialmente elevada. Este fenómeno implica que la puntuación en bruto de un modelo zero-shot no es comparable a la de un modelo entrenado para la tarea, sino que debe interpretarse de manera conjunta a la métrica de ranking relativo, como Precision@K, y no de manera aislada (Ethayarajh, 2019).

Finalmente, FreeATS introduce un cuarto enfoque: el fine-tuning (ajuste fino) de un modelo de DistilBERT multilingüe (Sanh et al., 2019) como clasificador binario matching / no-matching (coincide / no coincide) en pares vacante-candidato. A diferencia de TF-IDF, SBERT o TinyBERT, que devuelven un score de similitud sin nunca haber visto ejemplos

etiquetados del dominio, el modelo fine-tuning aprende directamente en el conjunto de pares de entrenamiento etiquetados, lo cual le permite capturar patrones propios del problema de candidato vacante.

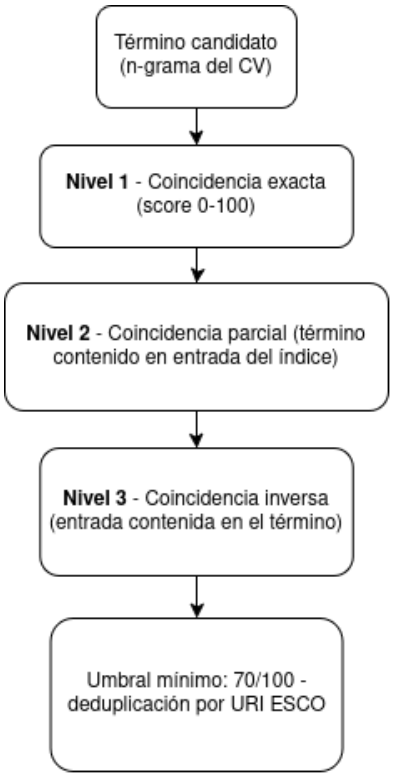
2.2.2. Taxonomías ocupacionales: ESCO

Los European Skills, Competences, Qualifications and Occupations (ESCO), son una taxonomía multilingüe creada por la Comisión Europea, es el documento que estandariza la descripción de ocupaciones, competencias y cualificaciones ocupacionales a través de identificadores únicos, etiquetas preferentes y alternativas por idioma. En los últimos años su uso en la investigación sobre PLN aplicada al mercado de trabajo ha aumentado como recurso para normalizar la enorme variabilidad con la que los candidatos describen sus propias habilidades (Zhang et al., 2023).

En FreeATS, ESCO ejerce una doble funcionalidad: por un lado, funciona como un vocabulario controlado contra el que se normalizan las habilidades detectadas en cada hoja de vida (para evitar que, por ejemplo, "Python", "programación en Python" y "desarrollo con Python" se traten como tres entidades separadas); por otro lado, ofrece la estructura jerárquica que puede ser extendida al emparejamiento de cargos u ocupaciones, yendo más allá de las habilidades individuales. El mecanismo concreto mediante el cual se realiza esta normalización sigue una búsqueda jerárquica de tres niveles de coincidencia, como se observa en la Figura 1.

Figura 1

Búsqueda jerárquica de tres niveles contra el índice de habilidades ESCO.



2.2.3. Validación de sistemas de ranking: correlación de Spearman

El proceso de evaluación del nivel de calidad de un sistema de ordenamiento de candidatos consiste en comparar el ordenamiento que produce el sistema seleccionado con respecto a un orden de referencia normalmente un ranking de los evaluadores humanos expertos y el coeficiente de correlación de rangos de Spearman (ρ) es la métrica estándar utilizada para esta comparación. A diferencia del coeficiente de Pearson, que establece que la relación entre las distintas variables continuas es lineal, el coeficiente de correlación de rangos de Spearman se limita a comprobar si el orden relativo de los distintos elementos del ranking coincide con respecto al de otro ranking sin suponer linealidad ni normalidad, esto es, sin establecer ningún supuesto sobre los rankings (Spearman, 1904). Las correlaciones de

Spearman van desde -1, que indica que existe un orden inverso, hasta 1, que indica que se obtiene el mismo orden de la referencia y que 0 significa que no se encuentra correlación alguna.

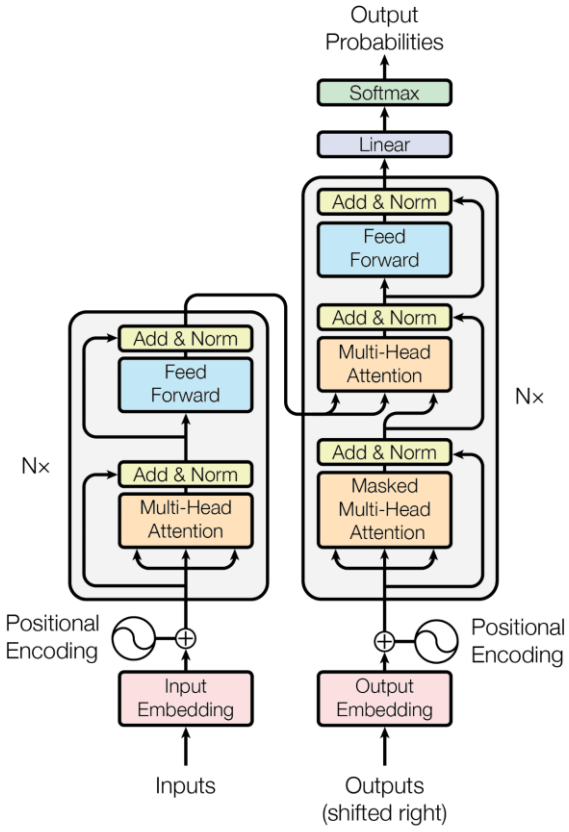
Un aspecto metodológico clave es la distinción entre ranking por candidato y ranking por vacante. Ambos problemas no son equivalentes, por lo que validar uno a partir del otro sin una transformación adecuada puede generar correlaciones espurias. Esta limitación, identificada durante la validación con el dataset de Vanetik y Kogan (2023), constituye un aporte metodológico del presente trabajo.

2.2.4. Arquitectura Transformer y mecanismos de atención

Los modelos BERT, DistilBERT, TinyBERT y Sentence-BERT que se evaluó en FreeATS se basan en bloques Transformer (Devlin et al., 2019). De este modo, a diferencia de las arquitecturas que procesan la secuencia de texto token a token, el Transformer (diagrama en la siguiente figura) procesa la secuencia completa en paralelo mediante un mecanismo de autoatención (self-attention), que calcula, para cada token, una ponderación de cuán importante es cada uno de los demás tokens de la secuencia para decidir su representación contextual, como se observa en la Figura 2.

Figura 2

Arquitectura Transformer



Para cada token, el modelo genera tres proyecciones lineales: una consulta (query), que representa la información buscada; una clave (key), que describe contra qué se compara; y un valor (value), que contiene la información que se combinará. La autoatención se calcula como el producto punto escalar entre consultas y claves, que se normaliza mediante una función softmax, para ponderar la combinación lineal de los valores (Vaswani et al., 2017).

Esta arquitectura es adecuada para el dominio de FreeATS, dado que permite capturar dependencias de largo alcance en el interior de una hoja de vida. Esta capacidad permite relacionar una habilidad descrita en la sección de experiencia con un puesto mencionado

varios párrafos antes. Los modelos recurrentes pueden perder información al procesar secuencias largas debido a problemas de desvanecimiento o explosión del gradiente (Pascanu et al., 2013). La autoatención reduce la longitud de la ruta entre posiciones y facilita el modelado de dependencias distantes (Vaswani et al., 2017). BERT incluye también el preentrenamiento bidireccional, introduciendo dos tareas auxiliares en el preentrenamiento: el modelado de lenguaje enmascarado por el que se enmascaran aleatoriamente tokens de la entrada y el modelo debe predecirse a partir de un contexto completo (izquierdo y derecho a la vez), y la predicción de la siguiente oración que permite al modelo aprender las relaciones que existen entre un par de oraciones. Este preentrenamiento bidireccional es lo que permite diferenciar BERT con respecto a arquitecturas autorregresivas unidireccionales, y es la base que sustenta a DistilBERT o los modelos SBERT utilizados en este proyecto.

Los modelos de aprendizaje profundo que han sido destilados como DistilBERT y TinyBERT reducen el número de capas Transformer y, en algunas ocasiones, la dimensión oculta de las representaciones, a través de un procedimiento de destilación de conocimiento en el que un modelo más pequeño es entrenado para tratar de replicar las salidas de un modelo más grande ya preparado (Sanh et al., 2019; Jiao et al., 2020). Esta técnica es la que permite el funcionamiento de FreeATS en computadores sin GPU dedicadas sin perder la capacidad de comprensión contextual de las arquitecturas Transformer completas.

2.2.5. Métricas de evaluación en traducción automática

Teniendo en cuenta que la traducción del corpus principal al español latinoamericano es una de las etapas delicadas del pipeline de FreeATS, resulta necesario explicar el marco teórico de las métricas que se utilizarán para comparar las alternativas de traducción: Bilingual Evaluation Understudy (BLEU) es la métrica clásica de evaluación de traducción

automática: mide la precisión de los n-gramas (secuencias contiguas de entre una y cuatro palabras) coincidentes entre la traducción candidata y una o varias traducciones de referencia, aplicando a la vez una penalización por brevedad que desincentiva la traducción artificialmente corta (Papineni et al., 2002). Su principal desventaja es que trabaja a nivel de coincidencia léxica exacta, penaliza por tanto las traducciones correctas que ofrecen sinónimos o reformulaciones aceptables.

Por otro lado, Translation Edit Rate (TER) parte de postulados diferentes: mide la cantidad mínima de operaciones de edición como inserciones, eliminaciones, sustituciones y reordenamientos de bloques necesarias para convertir la traducción candidata en la traducción de referencia y luego la normaliza a la longitud de esta última (Snover et al., 2006). A diferencia de BLEU, un valor de TER más bajo es mejor, ya que implica menor esfuerzo de corrección.

Las métricas chrF y chrF++ aparecen como respuesta a la limitación común de los modelos BLEU y TER con respecto a lenguas morfológicamente ricas, como es el caso del español. Dado que se calculan sobre n-gramas de palabras completas, son muy severas penalizando variaciones morfológicas menores (por ejemplo, género y número) que no influyen sobre la corrección semántica de la traducción. ChrF calcula un F-score sobre n-gramas de caracteres, lo que le confiere una cierta tolerancia a estas variaciones de carácter superficial; mientras que chrF++ extiende chrF con la inclusión adicional de n-gramas de palabras, intentando encontrar un equilibrio entre sensibilidad morfológica y relativo alcance sintáctico (Popović, 2015).

Por otro lado, COMET es una métrica neuronal entrenada para predecir juicios humanos de calidad de traducción a partir del texto fuente, la traducción candidata y la

referencia (Rei et al., 2020). Puede reconocer equivalencia semántica aun cuando exista poco solapamiento léxico. En contraposición a BLEU, TER, chrF y chrF++, COMET captura equivalencia semántica aun cuando la superposición léxica es baja, por lo que se convierte en la métrica individual que más se alinea con la evaluación de calidad que realizaría un traductor humano.

2.2.6. Sistemas de recuperación de información y búsqueda vectorial

El módulo de comparación semántica de FreeATS requiere, dada una vacante, poder recuperar de forma eficiente a los candidatos más similares a partir de un corpus que puede llegar a tener miles de hojas de vida. Este problema corresponde, en esencia, al problema de la búsqueda del vecino más cercano (nearest neighbor search) en las dimensiones de un espacio vectorial de alta dimensionalidad. La búsqueda exacta, que consiste en calcular la similitud entre la consulta y cada uno de los vectores del corpus, tiene una complejidad lineal respecto a la dimensión del corpus, lo que es computacionalmente viable para los volúmenes de corpus que se manejan en este proyecto pero no para corpus de millones de documentos, que suelen operar con índices aproximados (Approximate Nearest Neighbor, ANN) que intercambian exactitud por velocidad.

Facebook AI Similarity Search (FAISS) es una librería de código abierto que implementa búsqueda exacta y una variedad de familias de índices aproximados. FAISS ofrece índices aproximados basados en tres familias principales: cuantización de producto, grafos de proximidad como Hierarchical Navigable Small World (HNSW), un método de búsqueda aproximada que organiza los vectores en un grafo de proximidad multicapa para reducir el número de comparaciones durante la consulta (Malkov & Yashunin, 2020), y particionamiento del espacio mediante agrupamiento aptos para ejecutarse de forma eficiente

tanto en CPU como en GPU (Johnson et al., 2019). FreeATS utiliza IndexFlatIP, un índice exacto que calcula el producto punto entre la consulta y cada vector almacenado. Antes de indexarlos, los embeddings de SBERT se normalizan a norma unitaria. En estas condiciones, el producto punto equivale a la similitud del coseno. Esta elección evita errores de aproximación y mantiene un costo aceptable para un corpus de miles de hojas de vida. Para corpus de millones de documentos sería necesario evaluar índices aproximados. La elección de un índice exacto en lugar de uno aproximado refleja el objetivo del proyecto de priorizar la reproducibilidad de los resultados a partir de la precisión de los resultados en vez de la velocidad, en un contexto en el que el volumen de los datos de una PYME no compensa la mayor complejidad de un índice aproximado.

2.2.7. Aprendizaje por transferencia y fine-tuning de modelos pre-entrenados

En los modelos de lenguaje, una tarea general de preentrenamiento consiste en predecir tokens enmascarados a partir de grandes corpus no etiquetados (Devlin et al., 2019)

El paradigma de aprendizaje por transferencia (transfer learning) consiste en reutilizar el conocimiento adquirido por un modelo a raíz de un preentrenamiento muy largo sobre una tarea general. Por ejemplo, en el ámbito de los modelos de lenguaje, la predicción de tokens enmascarados sobre grandes volúmenes de texto no etiquetado para resolver una tarea específica bien distinta, para la cual se dispone de una cantidad mucho menor de datos etiquetados (Devlin et al., 2019). En FreeATS no es viable entrenar desde cero un modelo de lenguaje de gran escala debido al volumen de datos y al cómputo que requiere.

El fine-tuning es el procedimiento por medio del cual se entrena para una tarea determinada, continuando el entrenamiento de sus parámetros sobre un conjunto de datos específico de dicha tarea, para lo cual se suele fijar una tasa de aprendizaje menor que la de

dicho preentrenamiento, de modo que el objetivo de ajustar el modelo no implique destruir el conocimiento lingüístico general ya adquirido. En el caso del clasificador de emparejamiento vacante-candidato, el fine-tuning hace referencia a la totalidad de los parámetros del modelo distilbert-base-multilingual-cased, añadiendo sólo la capa de clasificación densa sobre la representación del token especial [CLS], el cual es un token especial colocado al inicio de la secuencia (Devlin et al., 2019), que en el caso de la arquitectura BERT, focaliza una representación agregada de la secuencia completa tras haberla procesado por las capas de auto-atención. La tasa de aprendizaje de 2×10^{-5} para controlar la magnitud de cada actualización de los pesos y el planificador con calentamiento gradual son procedimientos estándar en el fine-tuning de modelos Transformer teniendo como objetivo mitigar el riesgo de que las primeras actualizaciones de gradiente distorsionen los pesos ya aprendidos.

2.2.8. Consideraciones éticas y de equidad en sistemas de selección automatizada

La inserción de inteligencia artificial en los procesos de selección de personal también conlleva cuestiones éticas que trascienden el rendimiento técnico del sistema (Koman et al., 2024; Otani et al., 2025). La literatura sobre equidad algorítmica explica que los modelos de aprendizaje automático entrenados en datos históricos pueden replicar o amplificar sesgos contenidos en estos datos. Un ejemplo es la infrarrepresentación de determinados perfiles demográficos en los corpus de entrenamiento aunque no haya ninguna intención de discriminar por parte de quienes diseñan el sistema del sistema (Koman et al., 2024). En el caso de FreeATS, esta posibilidad de sesgos se da a través de dos vías concretas, y en particular en esta aplicación: en el diseño del corpus de entrenamiento de Resume Dataset PDF y DOCX que no documentan la composición demográfica del corpus original y los heurísticos curados manualmente en el módulo de ranking híbrido que, al no haberse

optimizado estadísticamente, pueden introducir preferencias no intencionadas hacia determinados perfiles de redacción de la hoja de vida.

FreeATS se concibe como un primer filtro que reduce el número de hojas de vida que deben revisarse manualmente, no como un sistema autónomo de contratación

Por tal motivo, el diseño de FreeATS se formula como una herramienta para ser un primer filtro que reduce el número de hojas de vida que deben revisarse manualmente y no como un sistema autónomo de contratación (Otani et al., 2025).

2.2.9. Procesamiento de documentos semiestructurados

Las hojas de vida no son texto abierto como puede ser una información, o una publicación en las redes sociales, pero no son tampoco un esquema rígido y predecible como el que se puede localizar en un formulario con campos predeterminados. En la literatura se las denomina documentos semiestructurados, ya que las hojas de vida suelen incluir secciones de datos personales, formación, experiencia y habilidades, pero su orden y presentación varían entre candidatos que no se halla marcada en el documento, ni de forma explícita ni de forma general en el formato del archivo (Zhang & Wang, 2018). Es esta misma premisa la que determina la razón de ser de la combinación de las dos estrategias complementarias que presenta FreeATS para la extracción de la información en el pipeline: extracción asistida por modelo de lenguaje, que utiliza la capacidad de comprensión del contexto de un modelo que ha sido entrenado sobre grandes cantidades de textos para la inferencia de la sección a la que pertenece cada fragmento de texto, incluso cuando el archivo no tiene cabecera, y la parte del NER, entrenada para el dominio, que localiza menciones de interés dentro del flujo de texto continuo, sin depender de que el documento original preserve una estructura de secciones reconocible.

En FreeATS se combinó dos estrategias complementarias: la extracción mediante un modelo de lenguaje general y la extracción mediante un modelo NER especializado. La primera ofrece mayor flexibilidad ante formatos heterogéneos, pero depende de una API externa. La segunda se ejecuta localmente, aunque su desempeño depende de la calidad del silver standard utilizado para entrenarla. La extracción asistida por un modelo de lenguaje general se beneficia mucho de la existencia de una comprensión general del contexto del documento extraído, lo que la hace más robusta ante aplicaciones con formatos de documento heterogéneos, aunque ello no sea gratuito, en el sentido de que incurre en el coste que suponga la dependencia de una API externa; la extracción mediante NER se puede ejecutar localmente, pero funciona acotada por la calidad de los materiales utilizados para el entrenamiento del modelo, en la medida en que se establecen como silver-standard. La decisión de que el pipeline en modo de extracción debe mantener ambas estrategias y no desechar una o la otra tiene que ver con la complementariedad de las ventajas y desventajas que ofrecen ambas opciones.

2.2.10. Reproducibilidad y trazabilidad en pipelines de procesamiento de datos

Un flujo de procesamiento de datos como el que se implementó en FreeATS conlleva un tipo de riesgos para la reproducibilidad, si no se tienen en cuenta las relaciones de dependencia entre etapas. La producción científica dentro de la ingeniería de datos y la ciencia de los datos reproducibles, indica o recomienda que de manera general se tomen en cuenta determinadas prácticas. Entre ellas se encuentran la explicitación del nombre de los scripts que muestran secuencialmente el orden en que deben ejecutarse estos scripts, persistiendo los artefactos intermedios en formatos abiertos y autodescriptivos, y la persistencia mediante logs de ejecución que permitan auditar que, dadas unas configuraciones, se ha llegado a un determinado resultado.

Como se presenta en la Tabla 2, el flujo de trabajo de FreeATS sigue estas tres reglas: enumerar de forma continua los scripts, utilizar siempre JavaScript Object Notation (JSON) como formato de intercambio entre etapas (corpus_esco_anotado.json, corpus_final.json, vacantes_procesadas.json, etc.) y, finalmente, crear el log estructurado en cada etapa (ner_evaluation.json o comparacion_metodos_final.json, por ejemplo), que es el recurso de trazabilidad.

Tabla 2

Artefactos de procesamiento y estructuración de datos

Archivo	Script productor	Elementos principales
corpus_esco_anotado.json	06_esco_integration.py	Metadatos: total_cvs, fecha_generado, esco_version y umbral_similitud. La lista cvs contiene identificación y procedencia del CV; resumen, experiencia, educación y certificaciones; habilidades detectadas; texto limpio; y habilidades_esco. Cada coincidencia ESCO registra termino_cv, label_esco, uri, tipo, score y metodo.
corpus_final.json	07_merge_corpus.py	Metadatos: version, fase, fecha_generado, total_cvs, estadisticas y schema_campos. La lista cvs integra identificación; nombre y ubicación; experiencia, empresas y cargos; sector; educación, certificaciones e idiomas; resumen y logros; habilidades_detectadas y habilidades_esco; textos normalizados; y estadísticas de procesamiento.
vacantes_procesadas.json	08_process_vacancies.py	Metadatos: total_vacantes, fecha_generado, modelo, traducidas y costo_total_usd. La lista vacantes contiene cargo, empresa, sector, ubicación, descripción original/traducida, experiencia mínima y deseada, educación, títulos aceptados, habilidades requeridas/deseadas, certificaciones, idiomas, resumen y _meta de trazabilidad.

Archivo	Script productor	Elementos principales
ner_evaluation.json	14_evaluate_ner.py	fecha_evaluacion, modelo, total_ejemplos_test, umbral_objetivo y cumple_umbral; metricas_globales con precision, recall y F1; metricas_por_tipo con precision, recall y F1 para cada entidad, como SKILL, JOB_TITLE y ORG.
comparacion_metodos_final.json	18_evaluate_distilbert.py	fecha_generado; metodos_evaluados, con metodo, precision, recall, F1 o score y descripcion; detalle de DistilBERT, incluida matriz de confusión y desglose por fuente de etiqueta; detalle NER; resumen TF-IDF/SBERT; y resumen TinyBERT.

2.2.11. Métricas de evaluación de sistemas de recuperación: Precision@K

La métrica P@K mide la proporción de elementos relevantes dentro de los primeros K resultados devueltos por un sistema de recuperación de información. En FreeATS, cada consulta corresponde a una vacante y los resultados son los candidatos ordenados según su nivel de adecuación. Su fórmula es:

$$P@K = \frac{número\ de\ candidatos\ relevantes\ entre\ los\ K\ primeros}{K}$$

Por ejemplo, si cuatro de los cinco primeros candidatos son relevantes, entonces P@5 = 4/5 = 0,80. En este proyecto se usaron P@5 y P@10, porque permiten evaluar la calidad de las primeras posiciones del ranking. Estas posiciones representan el grupo reducido de candidatos que un reclutador revisará en primer lugar.

En la evaluación exploratoria, la relevancia se definió de forma binaria mediante la coincidencia de la variable sector_industria: un candidato se consideró relevante cuando su sector profesional coincidía con el sector de la vacante. Este criterio funciona como un

ground truth débil y no sustituye una valoración humana sobre la adecuación entre candidato y vacante.

Otras métricas de ranking son Normalized Discounted Cumulative Gain (NDCG) y Mean Reciprocal Rank (MRR). NDCG evalúa la calidad del ordenamiento y permite distintos niveles de relevancia, por ejemplo, una escala de 0 a 3 (donde 0 es nula y 3 es alta) (Järvelin & Kekäläinen, 2002). No se aplicó porque el conjunto de evaluación no incluye relevancia graduada. MRR, en cambio, calcula el inverso de la posición del primer resultado relevante (Voorhees, 1999). Aunque admite relevancia binaria, no considera cuántos candidatos adecuados aparecen en las primeras posiciones. Por estas razones, P@K fue seleccionada como la métrica principal de la evaluación exploratoria.

2.2.12. Consideraciones específicas del procesamiento de lenguaje natural en español

El desarrollo de FreeATS en español requiere considerar características lingüísticas que influyen en varias decisiones de diseño. El español tiene una morfología rica en flexiones de género y número, además de una conjugación verbal compleja. Esto aumenta la variabilidad superficial entre expresiones semánticamente equivalentes, como «desarrollador», «desarrolladora» y «desarrolladores».

En la normalización previa al índice invertido de habilidades ESCO, se convirtió el texto a minúsculas y se eliminaron acentos. Estas operaciones reducen algunas variaciones ortográficas, aunque no eliminan las diferencias morfológicas de género y número. Por ello se dio mayor peso a métricas basadas en n-gramas de caracteres, como chrF y chrF++, frente a BLEU, que se basa en n-gramas de palabras. Las métricas de caracteres presentan mayor

tolerancia ante variaciones morfológicas menores que no alteran el significado de la traducción (Papineni et al., 2002; Popović, 2015).

La disponibilidad de recursos de PLN en español es menor que en inglés, lo que constituye una limitación metodológica. El corpus principal (Resume Dataset PDF y Resume Dataset DOCX) se obtuvo originalmente en inglés y debió traducirse automáticamente, ya que no se identificó un corpus abierto en español de volumen comparable. De igual manera, el modelo SBERT utilizado (paraphrase-multilingual-MiniLM-L12-v2) es multilingüe y no exclusivo para español (Reimers & Gurevych, 2019). Esto puede limitar parcialmente la especialización de sus representaciones semánticas frente a un modelo monolingüe entrenado con un corpus amplio en español. Su elección respondió a su disponibilidad, soporte multilingüe, eficiencia computacional y respaldo dentro de la comunidad de código abierto.

2.2.13. Extracción de información basada en instrucciones frente a modelos entrenados específicamente

La extracción de información en FreeATS combina dos paradigmas que conviene fundamentar teóricamente. El primero es el prompting, detallado en la extracción de campos estructurados. Consiste en dar instrucciones a un modelo de lenguaje de propósito general para que presente la información requerida siguiendo un esquema predefinido en lenguaje natural, sin haber sido entrenado para esa tarea (Liu et al., 2023). En este caso, GPT-4o-mini fue presentado como un modelo compacto orientado a equilibrar capacidad y costo de inferencia.

El segundo paradigma es el entrenamiento específico para la tarea (task-specific). Bajo este enfoque, un modelo compacto de NER, implementado con spaCy, se entrena desde cero con datos etiquetados del propio dominio. Su objetivo exclusivo es identificar las tres

categorías de entidades definidas en FreeATS. spaCy permite incorporar y entrenar componentes estadísticos específicos para el reconocimiento y etiquetado de entidades dentro de un pipeline de PLN (Honnibal et al., 2020).

Ambos paradigmas implican un balance entre flexibilidad y especialización, discutido en la literatura reciente (Zaratiana et al., 2024). La extracción basada en instrucciones no requiere datos etiquetados ni entrenamiento adicional, lo que reduce costos y permite adaptar el esquema de extracción modificando el texto de la instrucción. Sin embargo, cada extracción depende de una llamada a la API (interfaz de programación de aplicaciones), lo que implica tiempo y costo de inferencia por documento, además de la dependencia de un proveedor externo.

La extracción mediante un modelo entrenado, en cambio, requiere una etapa de preparación de datos y entrenamiento. En FreeATS, esta necesidad se resolvió mediante anotaciones silver-standard generadas automáticamente para evitar el costo de una anotación manual. Una vez finalizado el entrenamiento, la inferencia puede ejecutarse localmente y no depende de servicios externos. No obstante, el modelo presenta menor flexibilidad para incorporar nuevas categorías de entidades, ya que esto requeriría un nuevo ciclo de preparación de datos y entrenamiento.

La decisión de mantener ambos paradigmas de manera paralela, en lugar de optar por uno de ellos, responde a la intención de que FreeATS pueda operar bajo distintas prioridades de despliegue, con o sin acceso a servicios en la nube.

2.2.14. Privacidad y protección de datos personales en el procesamiento de hojas de vida

La hoja de vida es un documento que contiene información personal (nombre, datos de contacto y trayectoria laboral o académica). Al someterse a un tratamiento automatizado, esta información plantea cuestiones de privacidad que, aunque exceden el alcance técnico de este trabajo, son importantes y deben explicitar en el marco teórico que fundamenta el diseño del sistema.

En la literatura sobre protección de datos en sistemas de información se distingue entre los datos que permiten identificar directamente a una persona y aquellos que podrían facilitar su identificación al combinarse con otra información de referencia (Pilán et al., 2022).

Tabla 3

Tipos de datos personales presentes en las hojas de vida

Tipo de dato	Ejemplos
Datos de identificación directa	Nombre, correo electrónico y número de teléfono
Datos que podrían facilitar una identificación indirecta	Trayectoria laboral específica, institución educativa y procedencia geográfica

En el diseño de FreeATS, como se observa en la Tabla 3, esta consideración se refleja en dos decisiones específicas. Por un lado, la recopilación propia de hojas de vida en español se realizó sobre la base de la participación voluntaria (con fines exclusivamente académicos y de investigación, y sin vinculación con una oferta laboral existente). Esto permitió evitar la

ambigüedad ética que supondría recopilar datos personales bajo la apariencia de un proceso de selección real.

Por otro lado, el conjunto de datos de emparejamiento vacante–hoja de vida desarrollado por Vanetik y Kogan (2023) se utilizó en la forma anonimizada proporcionada por sus autores. Asimismo, el pipeline de FreeATS no realizó ningún intento de reidentificación de las candidaturas que lo componen.

Las prácticas descritas no constituyen por sí mismas un marco formal de cumplimiento normativo, ya que este aspecto excede el alcance académico del proyecto. Sin embargo, se encuentran alineadas con los principios de minimización de datos y limitación de la finalidad establecidos en la normativa sobre protección de datos personales. Estos principios deberían formalizarse mediante una política de manejo de datos antes de cualquier eventual despliegue operativo del sistema en una PYME real.

CAPÍTULO 3

3. Desarrollo

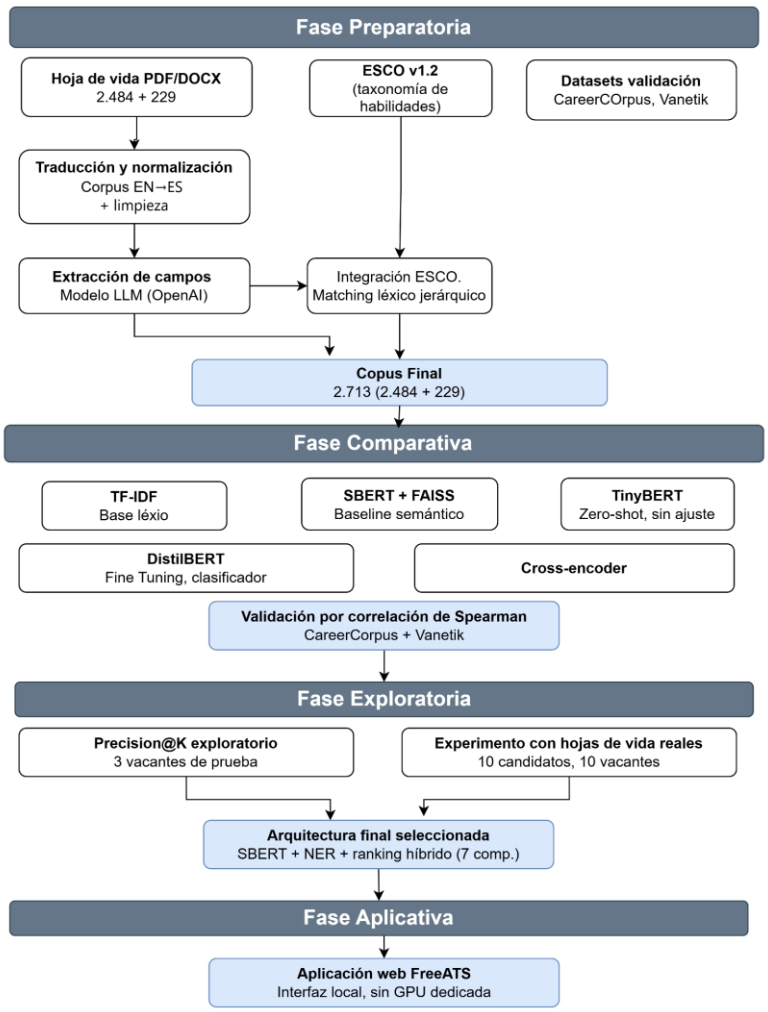
En este capítulo se explica la evolución metodológica y técnica del sistema FreeATS en sus cuatro fases, siguiendo el mismo orden en que se ejecuta el pipeline: preparación de datasets, extracción y normalización del texto, comparación semántica, ranking híbrido, fine-tuning de un clasificador de emparejamiento, validación por correlación de Spearman, extracción de entidades nombradas y entorno de desarrollo. La siguiente figura presenta una vista panorámica de la arquitectura general.

La Figura 3 resume la arquitectura de FreeATS en cuatro fases. En la fase preparatoria se integran 2.484 hojas de vida PDF y 229 DOCX; los documentos se traducen y normalizan, se extraen campos y se alinean las habilidades con ESCO, generando un corpus final de 2.713 registros. En la fase comparativa se contrastan TF-IDF, SBERT con FAISS, TinyBERT y DistilBERT, y se valida el ordenamiento mediante correlación de Spearman con CareerCorpus y el conjunto de Vanetik y Kogan. En la fase exploratoria se evalúa el desempeño mediante Precision@K y diez pares de hojas de vida y vacantes. Esta evidencia conduce a la arquitectura final, compuesta por SBERT, NER y un ranking híbrido de siete

componentes. Finalmente, la fase aplicativa integra estos componentes en FreeATS, una aplicación web local que funciona sin GPU dedicada.

Figura 3

Arquitectura general del sistema FreeATS



3.1.Dataset y roles

Para dar cumplimiento al objetivo del presente proyecto se centró la búsqueda de conjuntos de datos en aquellos que tengan archivos en formatos comunes de hojas de vida como PDF y DOCX. También el trabajo hecho por otros investigadores fueron fuente para

utilizar datos que pueden resultar beneficiosos para la clasificación de las hojas de vida. Con estas pautas en consideración se decidió utilizar los siguientes conjuntos de datos abiertos disponibles para este tipo de proyectos:

3.1.1. Descripción del conjunto de datos: Resume Dataset PDF

Este conjunto de datos abiertos disponible en la plataforma Kaggle reúne un grupo de 2.484 hojas de vida escritas en inglés que incluyen información personal, académica y profesional de cada candidato que fueron publicadas hace 5 años. Las hojas de vida están categorizadas en 24 profesiones (contabilidad, agricultura, arte, diseño, etc.) y cada categoría contiene un grupo de hojas de vida. Aunque todas las categorías cuentan con varias hojas de vida, la distribución entre ellas resulta desbalanceada.

3.1.2. Descripción del conjunto de datos: Resume Dataset DOCX

Este conjunto de datos presenta 229 hojas de vida de diferentes profesionales, aunque la mayoría está relacionada con cargos vinculados al desarrollo de sistemas. A diferencia del primer conjunto de datos, estas hojas de vida se encuentran en formato DOCX. El conjunto de datos fue actualizado hace 6 años; las hojas están en idioma inglés y contienen información personal, académica y profesional. El objetivo principal del proyecto al utilizar este conjunto de datos es aumentar la cantidad de información y contar con un nuevo formato sobre el cual se pueda trabajar, evitando que se convierta en una limitante para el sistema.

3.1.3. Descripción del conjunto de datos: CareerCorpus

A diferencia de los dos conjuntos de datos presentados anteriormente, el CareerCorpus es un conjunto de datos curado que ha pasado por un preprocesamiento para extraer distintos tipos

de información de varios candidatos. Las hojas de vida fueron extraídas y procesadas para obtener información en un formato estructurado, como se muestra en la Tabla 4.

Tabla 4

Campos incluidos en el conjunto de datos CareerCorpus

Campo	Descripción
Domain	Categoría a la cual pertenece la hoja de vida. Incluye seis categorías: Accountant, Apparel, Banking, Finance, Research Assistant y Teacher.
Education	Información sobre la institución educativa y el tipo de estudio realizado.
Skills and Achievements	Habilidades adquiridas y logros relevantes del candidato.
Experience	Información sobre lugares de trabajo, duración de la experiencia y cargos desempeñados.
Job_type	Nivel de antigüedad profesional dividido en tres categorías: Entry-level, Mid-level y Senior-level.
Annotator-1	Calificación entre 0 y 1 otorgada por el primer anotador.
Annotator-2	Calificación entre 0 y 1 otorgada por el segundo anotador.

Para llevar a cabo la evaluación externa exploratoria del ranking, se utilizó CareerCorpus como fuente de puntuaciones humanas preexistentes. El conjunto original contiene 302 perfiles distribuidos en seis categorías ocupacionales: Teacher, Finance, Apparel, Accountant, Banking y Research Assistant. Cada perfil incluye dos puntuaciones independientes de adecuación, denominadas Annotator-1 y Annotator-2. A partir de estos dos valores, se definió como referencia humana el promedio aritmético de las puntuaciones asignadas a cada perfil.

3.1.4. Descripción del conjunto de datos: Data Voto Informado 2026

Con el objetivo de ampliar el conjunto de datos en idioma español, se incorporó a la investigación el dataset alejandrohdo/data-voto-informado-2026. Este recurso, disponible en formato Markdown, reúne 9.149 perfiles de candidatos vinculados a las Elecciones Generales de Perú 2026. La información proviene de registros oficiales del proceso electoral, incluyendo datos sobre candidatos, hojas de vida y organizaciones políticas.

Estos datos fueron recopilados con el propósito de alimentar sistemas de Inteligencia Artificial, en particular aquellos basados en generación aumentada por recuperación. La principal dificultad de este conjunto es que todos los perfiles corresponden a cargos de elección popular, lo que puede generar un desbalance de categorías al combinarlo con otros datasets. Aun así, pese a tratarse de candidatos políticos, resulta valioso para realizar pruebas de extracción y clasificación de habilidades clave en textos redactados en español.

3.1.5. Dataset de emparejamiento vacante–hoja de vida

Conjunto de datos que consta de 66 hojas de vida. Los datos no están categorizados, están anonimizados y se encuentran en formato DOCX. La cantidad de información es pequeña en comparación con los primeros conjuntos de datos pero su valor radica en que los datos fueron organizados en un ranking realizado por humanos, quienes indicaron cuáles hojas de vida son más adecuadas para cada vacante. Este conjunto de datos resulta ser un recurso útil para la investigación en PLN y sistemas de emparejamiento laboral, al combinar hojas de vida reales con evaluaciones humanas de emparejamiento y adecuación frente a vacantes laborales específicas.

3.1.6. Recopilación Propia

Los datos abiertos que se han encontrado hasta el momento pueden funcionar para la investigación, pero también se incorporó una recopilación propia de hojas de vida en español para la fase de validación en condiciones reales. Este proceso se llevó a cabo recolectando hojas de vida de personas conocidas a las cuales se les explicó que el uso de los datos es académico y de investigación. Se enfatizó que la participación es voluntaria y que la solicitud no está vinculada a ninguna oferta laboral ni proceso de selección.

Esta forma de recopilar datos permitió obtener información actualizada, vinculada al entorno local y con mayor representación de las características del mercado objetivo. Además, contribuiría a enriquecer la base de datos, diversificando el tipo de hojas de vida disponibles para su uso, lo cual resulta relevante para el desarrollo y la evaluación de modelos de procesamiento de lenguaje natural aplicados al análisis de hojas de vida.

3.2.Procedimiento de integración y uso de los conjuntos de datos

Para el desarrollo del sistema FreeATS, se definió una estrategia de uso de los datasets organizada por grupo, conformando así tres roles diferenciados: entrenamiento, matching y validación en contexto real. Esta distinción es necesaria dado que los datasets de los que se dispone son heterogéneos respecto a lengua, formato, estructura y nivel de anotación, de modo que asignar a cada uno de los datasets una función concreta dentro del pipeline le permite garantizar la coherencia metodológica, así como evaluar el sistema en condiciones mucho más cercanas al contexto de las PYMEs ecuatorianas en la medida en que sea más factible.

El dataset principal será Resume Dataset PDF, el cual aporta el mayor corpus en el sentido cualitativo y cuantitativo, aportando 2.484 hojas de vida en formato PDF en 24 categorías profesionales; del mismo modo, como Dataset complementario para la etapa de entrenamiento se sumará el Resume Dataset DOCX, que facilita 229 hojas de vida en formato Word (de esta manera se facilita el proceso de aprendizaje aplicado en el pipeline para que los formatos de entrada sean indistintos).

Dado que ambos datasets están en idioma inglés, se aplicará un proceso de traducción automática al español latinoamericano. Para esta etapa se realizará una evaluación entre distintas técnicas y soluciones de traducción basadas tanto en modelos open source como en APIs de modelos de lenguaje avanzados, considerando soluciones como Helsinki-NLP/opus-mt-en-es (enlazable abiertamente con Hugging Face) y servicios de traducción a través de modelos generativos de proveedores como OpenAI, Anthropic u otras empresas similares. La selección final de la estrategia de traducción será determinada mediante la experimentación durante el desarrollo de la investigación, tomando en cuenta métricas de calidad semánticas, revisión de coherencia contextual, uso y preservación de términos técnicos. Con el objetivo de garantizar la calidad de la traducción se realizará una validación manual de los resultados de una muestra aleatoria estratificada para revisar la transferencia de los términos técnicos del ámbito de los recursos humanos. Estos dos datasets traducidos y unificados serán el corpus final o con el que se entrenará el módulo de extracción de habilidades y el clasificador de candidatos.

El dataset de CareerCorpus cumplió una función diferente, ya que se trata de un dataset que cuenta con perfiles estructurados que integran habilidades, experiencia y nivel de antigüedad, y se utilizó como fuente de vacantes y requerimientos de puestos para las pruebas del módulo de emparejamiento semántico. Su valor añadido, a diferencia de los datasets

anteriores, es que dispone de anotaciones humanas de adecuación (Annotator-1 y Annotator-2), de manera que se logró contrastar los scores generados por el sistema con un referente de una evaluación por un experto. Ya que este conjunto de datos cuenta con seis categorías profesionales equilibradas, también se contó con datos para las pruebas de clasificación por perfil.

Para la fase de validación en condiciones reales se emplearán 3 fuentes complementarias. En primer lugar, el dataset de emparejamiento vacante–hoja de vida, que contiene 66 hojas de vida anonimizadas con rankings realizados por evaluadores humanos, será la fuente principal de validación cuantitativa del módulo de ranking al permitir calcular la correlación de Spearman entre los rankings generados por el sistema y los rankings humanos. En segundo lugar, el dataset Voto Informado 2026, compuesto por 9.149 perfiles de candidatos en español en formato Markdown, servirá como corpus de prueba para el pipeline de extracción de texto en español, a partir de su riqueza léxica en español evaluando así la generalización del módulo de extracción de habilidades más allá de su dominio laboral estrictamente profesional; se reconoce que estos perfiles corresponden a candidatos electorales peruanos y no a hojas de vida laborales, por lo que su uso se limita a pruebas de extracción y no al módulo de matching. En tercer lugar, la recopilación propia consistirá en hojas de vida en español recolectadas voluntariamente a través de convocatorias abiertas en LinkedIn, donde se especificará explícitamente el uso con fines académicos, la participación voluntaria y la anonimización de los datos; este subconjunto permitirá validar el sistema en el contexto del mercado laboral ecuatoriano, que es el escenario objetivo del proyecto.

La Tabla 5 resume la función asignada a cada dataset, el idioma en que se encuentra, la cantidad de registros que contiene y la etapa del pipeline a la que contribuye.

Tabla 5

Resumen de función asignada a cada dataset.

Conjunto de datos	Idioma	Volumen	Características principales	Función en el proyecto	Estrategia de procesamiento
Resume Dataset PDF	Inglés (traducido al español)	2.484 hojas de vida	Información personal, académica y profesional distribuida en 24 categorías ocupacionales	Corpus principal para el entrenamiento y procesamiento del sistema	Traducción automática al español
Resume Dataset DOCX	Inglés (traducido al español)	229 hojas de vida	Perfiles profesionales, principalmente relacionados con el desarrollo de sistemas	Ampliación del corpus de entrenamiento y evaluación del procesamiento de documentos en formato DOCX	Traducción automática al español y posterior integración con el Resume Dataset PDF
CareerCorpus (estructurado)	Inglés	302 perfiles	Perfiles estructurados con información sobre educación, habilidades, experiencia, nivel profesional y evaluaciones de dos anotadores	Validación de la extracción y clasificación mediante información estructurada y anotaciones humanas	Utilización directa en inglés y empleo de las puntuaciones de Annotator 1 y Annotator 2 como referencia humana
Dataset de emparejamiento vacante hoja de vida DOCX	Inglés	66 hojas de vida	Documentos anonimizados acompañados de evaluaciones humanas de adecuación frente a vacantes laborales	Validación cuantitativa del ranking generado por el sistema	Uso directo de los documentos y comparación del ranking generado con el ranking humano mediante correlación de Spearman
Data Voto Informado 2026 Markdown	Español	9.149 perfiles	Perfiles de candidatos de las Elecciones Generales de Perú 2026 con información	Pruebas de extracción y evaluación de la generalización del procesamiento de texto en español	Uso directo para evaluar el procesamiento de entidades y extracción de habilidades, sin

Conjunto de datos	Idioma	Volumen	Características principales	Función en el proyecto	Estrategia de procesamiento
			extensa redactada originalmente en español		aplicación del módulo de emparejamiento
Recopilación propia (Archivos varios)	Español	Variable	Hojas de vida recopiladas voluntariamente en el contexto ecuatoriano para fines académicos y de investigación	Validación del sistema en condiciones cercanas al mercado laboral ecuatoriano	Recolección voluntaria, anonimización y procesamiento de documentos reales en español

3.3.Pipeline de extracción y normalización

3.3.1. Corpus principal y traducción

El corpus principal para la fase de entrenamiento lo configuran los datasets Resume Dataset PDF y Resume Dataset DOCX. El primero integra 2.484 hojas de vida para 24 profesiones en formato PDF y el segundo suma 229 más en formato DOCX. La utilización de ambos datasets como corpus principal permite aumentar la totalidad de la información colectada y poder incorporar mayor diversidad estructural en los documentos de entrada, haciendo así menos dependiente a los documentos de un solo formato. Como ambos sets estaban en inglés, se implementa un preprocesamiento de traducción a español latinoamericano; este preprocesado asegura la homogeneidad lingüística del corpus final, facilitando la integración de la solución con recursos léxicos y taxonomías ocupacionales existentes en español.

Este preprocesamiento tiene como riesgo metodológico el “translationese” el cual hace referencia a que el texto traducido puede contener patrones léxicos o sintácticos del texto original, lo que puede introducir sesgos sistemáticos (Vanmassenhove et al., 2021). Sin embargo, para mitigar este riesgo, se comparan cuatro alternativas de traducción, y se

selecciona la alternativa con mejor desempeño según las métricas de calidad lingüística, validada frente a la traducción humana de referencia.

3.3.2. Evaluación de estrategias de traducción

Se realizaron pruebas con tres enfoques distintos de traducción: modelos open source, servicios comerciales y modelos de lenguaje de gran escala (Large Language Models, LLMs), actualmente aplicados en múltiples tareas de PLN (Kumar, 2024). Las alternativas evaluadas fueron Helsinki-NLP/opus-mt-en-es como modelo open source (Helsinki-NLP, 2020), Google Cloud Translation v2 como servicio comercial (Google Cloud, s. f.), y GPT-4o-mini de OpenAI (OpenAI, 2024) junto con Claude Haiku 4.5 de Anthropic (Anthropic, 2025) como representantes de los LLMs.

3.3.3. Criterios de evaluación de traducción

De acuerdo con Moghe et al. (2025) la evaluación de traducción no se puede reducir a una sola métrica, ya que cada una considera un aspecto distinto de la calidad de la traducción. Por tal motivo, se utilizaron de forma complementaria cinco métricas, como se observa en la Tabla 6.

Tabla 6

Métricas de evaluación de traducción automática

Métrica	Función
BLEU (Bilingual Evaluation Understudy)	Precisión de n-gramas entre la traducción candidata y una de referencia humana (Papineni et al., 2002)
TER (Translation Edit Rate)	Cuenta el mínimo número de ediciones necesarias para pasar de la traducción candidata a la de referencia y que da un valor menor a mejor calidad de la traducción (Snover et al., 2006)

Métrica	Función
chrF	métrica tipo F-score sobre n-gramas de caracteres (Popović, 2015)
chrF++	métrica tipo F-score sobre n-gramas de caracteres y palabras (Popović, 2017)
COMET	Métrica neuronal entrenada para predecir la calidad de una traducción de una forma más similar a la que se realiza desde un juicio humano que basado en las métricas de evaluación de superposición léxica (Rei et al., 2020)

3.3.4. Procedimiento, muestra y configuración de la evaluación de la traducción

Como primer paso para la evaluación, se extrajo una muestra aleatoria de 30 hojas de vida del corpus total (2.484 archivos en PDF y 229 en DOCX), buscando diversidad en torno a formato, extensión y contenido. Para estos 30 casos se realizó una traducción humana de referencia, elaborada de forma independiente y ajena a cualquier proceso automático o uso de herramientas de inteligencia artificial, la cual sirvió como base de comparación para las cinco métricas. El proceso de evaluación fue el mismo para los cuatro métodos: se extrajo el texto original en inglés de cada hoja de vida, se realizó la traducción con cada método, se normaliza las salidas para eliminar inconsistencias de formato y codificación, y por último se calcularon las métricas utilizando el documento completo frente a la traducción de referencia humana (ground truth), en lugar de hacerlo por oraciones, para evitar distorsiones de segmentación.

Con respecto a la configuración establecida para cada método, GPT-4o-mini y Claude Haiku 4.5 se utilizaron mediante sus respectivas APIs, empleando en ambos casos la misma estructura del prompt: traducir al español latinoamericano de forma neutra, conservando la

estructura del documento, los cargos detallados y la terminología técnica, evitando resumir, omitir o sintetizar información. Por otro lado, Google Cloud Translation se utilizó a través de su servicio de traducción convencional, sin configuraciones ni instrucciones adicionales de estilo. En el caso de Helsinki-NLP, al ser un modelo pre-entrenado con una capacidad de procesamiento de texto más limitada, se ejecutó de forma local dividiendo cada hoja de vida en fragmentos más pequeños. Finalmente, la métrica COMET se calculó utilizando en cada caso el texto original en inglés, la traducción obtenida con cada método evaluado y la traducción humana de referencia.

3.3.5. Preprocesamiento y normalización de documentos

Previo al proceso de traducción, los documentos fueron preparados mediante extracción de texto: en el caso de los PDF se emplearon herramientas especializadas (pdfplumber), mientras que los archivos DOCX fueron convertidos a texto plano con python-docx, procurando preservar la originalidad del contenido. Tras esta etapa, se aplicaron técnicas de eliminación de inconsistencias de formato, normalización de caracteres especiales y adaptaciones frente a problemas derivados de la extracción automática.

Una vez obtenido el texto base, se implementó un proceso de segmentación por bloques (chunking), que consiste en dividir cada documento en fragmentos de longitud controlada para facilitar su procesamiento por los modelos. Este procedimiento permite mitigar las limitaciones asociadas a los límites de contexto de ciertos modelos, reducir errores derivados de textos demasiado extensos y adaptar los bloques de entrada a las necesidades de cada sistema. La coherencia lógica de la escritura se asegura mediante mecanismos de reconstrucción final, lo que garantiza que el resultado sea consistente y respete la estructura original de las hojas de vida, al tiempo que se normalizan los espacios en blanco, se eliminan

líneas vacías redundantes, se estandarizan caracteres Unicode y se ajustan saltos de línea o formatos textuales

3.3.6. Estructuración de campos

A partir del texto limpio, se procedió a extraer y estructurar los campos considerados relevantes de cada hoja de vida, entre ellos nombre, ubicación, años de experiencia, empresas, cargos, nivel educativo, certificaciones, idiomas y resumen profesional. Esta extracción se realizó mediante un LLM, específicamente OpenAI GPT-4o-mini, y fue evaluada en términos de cobertura y consistencia. El resultado de esta etapa, junto con la anotación ESCO, se consolidó en un único archivo que sirve como entrada para las etapas posteriores del pipeline.

3.3.7. Integración con taxonomía ESCO

Desde que se obtuvo la traducción y la normalización del corpus, se procedió a la organización de la información extraída de cada hoja de vida. El pipeline incluye las siguientes etapas: texto extraído, texto traducido, texto limpio y texto normalizado, y, por último, formatos estructurados en los que se consolida la información esencial extraída (formación académica, experiencia profesional, lista de habilidades técnicas y una descripción acerca de las competencias declaradas en cada hoja de vida).

Los campos extraídos fueron integrados en la taxonomía ESCO, un marco estandarizado avalado por la Comisión Europea para la descripción de ocupaciones, competencias y cualificaciones profesionales. Este proceso se llevó a cabo en dos fases: por un lado, la construcción de un índice de búsqueda y el marcado de cada hoja de vida

mediante un motor de emparejamiento léxico basado en contenido; por otro lado, la aplicación del mismo motor sobre el corpus estructurado.

En la primera fase, se cargó la base de datos ESCO v1.2, publicada por la Comisión Europea y disponible para descarga en distintos idiomas y formatos (Comisión Europea, 2024). A continuación, se normalizó el texto retirando acentos, signos de puntuación y convirtiendo todo a minúsculas, obteniendo así la normalización del texto. De esta manera se construyó un índice invertido de las habilidades, en tanto que el índice relaciona la etiqueta de habilidad con la preferredLabel, las altLabels y añade un prefijo diferenciador de hasta tres palabras, lo que permite eliminar variaciones terminológicas.

En la segunda fase, el modelo de matching comprobó una a una todas las hojas de vida del corpus procesado y estructurado: todas las competencias detectadas por el pipeline de limpieza y los n-grams de dos y tres palabras extraídos del texto. Los n-grams se filtraron con un vocabulario técnico predefinido (que incluye, por ejemplo, Python, ERP, SQL, logística, etc.) con el fin de reducir al máximo los falsos positivos. Cada término candidato fue evaluado mediante una búsqueda jerárquica de tres niveles: coincidencia exacta (0-100), coincidencia parcial (cuando el término candidato queda englobado en una entrada del índice) y coincidencia inversa (cuando el texto de entrada queda englobado en el término candidato). En los dos últimos niveles, el score obtenido pondera las longitudes relativas de las cadenas y se filtra con un umbral mínimo de similitud de 70 sobre 100. Los resultados se deduplicaron por URI de ESCO y finalmente se ordenaron de mayor a menor según el score.

Por último, para cada una de las habilidades extraídas de las hojas de vida se conservó la información relativa al término original, la etiqueta oficial de ESCO, el URI correspondiente y el mecanismo de emparejamiento empleado para la extracción. El resultado

de esta etapa consolidó el conjunto de habilidades normalizadas bajo el marco de trabajo de ESCO, aplicándose de manera simétrica también a las ofertas de trabajo procesadas, de modo que tanto las hojas de vida como las vacantes quedaron representadas bajo el mismo vocabulario normalizado. En ausencia de esta simetría, el componente de similitud ESCO para el ranking híbrido compararía etiquetas normalizadas del lado del candidato con texto crudo del lado de la vacante, lo que introduciría una asimetría de vocabulario no deseada y reduciría la similitud verdadera.

3.4.Módulo de comparación semántica

Con el corpus final estructurado y anotado, se llevaron a cabo tres métodos de comparación semántica entre las hojas de vida y las vacantes, definidos como los baselines del tercer objetivo específico, así como un cuarto método de clasificación fine-tuning.

3.4.1. Procesamiento de vacantes

Las vacantes del dataset de emparejamiento vacante-hoja de vida fueron traducidas al español latinoamericano y, mediante GPT-4o-mini, se extrajeron los campos estructurados que se muestran en la Tabla 7.

Tabla 7

Campos estructurados extraídos de las vacantes

Campo	Información extraída
Título de cargo	Cargo o posición asociada a la vacante
Sector	Sector profesional de la vacante
Habilidades requeridas	Habilidades consideradas obligatorias
Habilidades deseadas	Habilidades adicionales valoradas
Nivel mínimo de educación	Formación académica mínima requerida

Campo	Información extraída
Certificaciones	Certificaciones solicitadas o valoradas
Idiomas	Idiomas requeridos o deseados

Adicionalmente, cada vacante fue anotada con respecto a la taxonomía ESCO, reutilizando la misma función de anotación aplicada a las hojas de vida, y se exportó en formato JSON y también en .xlsx para su revisión manual.

3.4.2. Generación de embeddings y baseline SBERT

Para cada hoja de vida y cada vacante se generó un texto representativo que concatena el resumen profesional, los cargos, el sector, el nivel educativo, las habilidades ESCO (o las habilidades detectadas) y las certificaciones, el cual fue vectorizado con ayuda del modelo SBERT multilingüe (paraphrase-multilingual-MiniLM-L12-v2) de 384 dimensiones (Reimers & Gurevych, 2019), normalizado L2 para permitir el uso del producto punto como equivalente de la similitud coseno. Con los vectores de las hojas de vida se construyó un índice FAISS (Facebook AI Similarity Search) de búsqueda exacta de producto punto (IndexFlatIP), una biblioteca de búsqueda vectorial de alto rendimiento (Johnson et al., 2019), apropiada para corpus de hasta cientos de miles de vectores.

3.4.3. Baseline TF-IDF

Como baseline léxico se implementó una representación TF-IDF (unigrams y bigrams, máximo 5.000 términos, frecuencia mínima de documento 2) sobre el mismo texto representativo de hojas de vida y vacantes, con similitud coseno. La comparación permite cuantificar el aporte de la comprensión semántica de SBERT frente a la coincidencia léxica,

calculando el sobrelapamiento entre los top-5 candidatos que han recuperado cada método para cada vacante.

3.4.4. Baseline TinyBERT (zero-shot)

Se evaluó un tercer baseline en el que TinyBERT no se encontraba ajustado (zero-shot), aplicando mean pooling sobre las representaciones pre entrenadas para transformar cada documento en un vector, sin incorporar información adicional sobre un dominio específico de reclutamiento. Este baseline evidenció el fenómeno de anisotropía, dado que al no existir un fine-tuning, el score de similitud bruta se infló artificialmente; este efecto se hizo evidente al contrastarlo con la métrica de Precision@K, la cual se mantiene independiente de dicho sesgo de escala.

3.4.5. Síntesis comparativa de los tres baselines

A continuación se resumieron las diferencias conceptuales entre los tres baselines implementados, dado que cada uno de ellos corresponde a un punto distinto dentro del compromiso entre costo computacional y poder de comprensión semántica. El baseline TF-IDF no requiere entrenamiento ni el uso de un modelo preentrenado (ya que se construye sobre el vocabulario del corpus de evaluación), por lo que constituye una alternativa de menor costo computacional y también de menor requerimiento de memoria, a cambio de una comprensión léxica.

El baseline SBERT, en cambio, requiere descargar y cargar en memoria un modelo preentrenado de aproximadamente 470 MB de tamaño, pero no necesita entrenamiento adicional específico del dominio de FreeATS y ofrece una comprensión semántica genuina a costa de un costo computacional moderado. Por último, el baseline TinyBERT, en la

evaluación reportada en este trabajo bajo su configuración zero-shot, comparte el costo de carga de un modelo preentrenado con el baseline SBERT, pero su rendimiento de ranking real fue inferior al de este, pese a presentar un score bruto más elevado, lo que demuestra que la disponibilidad de un modelo preentrenado del tipo transformer por sí sola no garantiza un mejor desempeño en una tarea, debido a las exigencias del fine-tuning asociado.

3.5.Ranking híbrido

El módulo de ranking combinado lee las señales previas y las transforma a score final, que queda mediada por vacante y candidato. La fórmula final suma siete componentes: similitud semántica SBERT (30%), similitud ESCO (25%), coincidencia de cargo (15%), experiencia (15%), educación (10%) e idiomas (5%). Cada componente se normaliza en una escala [0,1] antes de ponderar, como se observa en la Tabla 8.

Tabla 8

Componentes y pesos del score de ranking híbrido

Componente	Peso	Fuente de la señal
Similitud semántica (SBERT)	30%	Producto punto entre embeddings normalizados, recuperación FAISS
Similitud de habilidades ESCO	25%	Coeficiente de Jaccard sobre habilidades normalizadas ESCO
Coincidencia de cargo	15%	Heurística léxica de coincidencia cargo-título
Años de experiencia	15%	Normalización lineal acotada (máx. 12 años)
Nivel educativo	10%	Escala ordinal por nivel académico declarado
Idiomas	5%	Coincidencia binaria de idioma requerido

Durante el desarrollo se detectó un sesgo predecible del componente de similitud semántica (SBERT) y del matching de habilidades ESCO, donde los perfiles de gestión de proyectos eran rankeados en la parte alta en vacantes de desarrollo de software. Para contrarrestar este efecto se incorporaron dos componentes heurísticos: (1) ponderación diferenciada de habilidades en función de la especificidad técnica y (2) un componente de coincidencia de cargo, esto es, un 15% del score final.

3.5.1. Proceso de calibración manual de los heurísticos

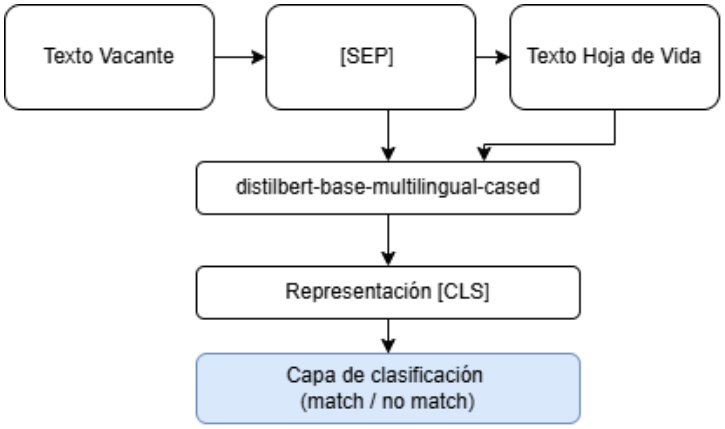
El procedimiento que llevó a determinar los pesos de la tabla 8 fue iterativo y cualitativo. El punto de partida fue una distribución de pesos uniforme entre los siete componentes; a partir de esa variante inicial, el equipo (con el apoyo de una profesional en procesos de selección de personal, cuyo criterio experto sirvió de referencia para juzgar la plausibilidad de los rankings generados) revisó manualmente, sobre el conjunto de vacantes de prueba, aquellos casos en los que el ranking resultante no coincidía con lo que un revisor humano consideraría razonable (el sesgo de sobrevaloración de los perfiles de gestores de proyectos). Ante cada uno de estos casos, se ajustó manualmente el peso relativo de los distintos componentes involucrados y se inspeccionó nuevamente el ranking resultante, repitiendo este ciclo de observación-ajuste.

3.6.Fine-tuning de DistilBERT

Como cuarto enfoque de comparación, como se observa en la Figura 4, se ajustó un modelo distilbert-base-multilingual-cased como clasificador binario de coincidencia entre pares vacante-candidato, dado que se obtiene un par de textos (vacante, hoja de vida) que se concatena con un separador [SEP], el cual es procesado mediante DistilBERT y posteriormente alimentado a una capa de clasificación binaria.

Figura 4

Arquitectura del clasificador de emparejamiento vacante-candidato.



Los pares etiquetados del conjunto de entrenamiento proceden de dos fuentes: pares heurísticos, etiquetados según la coincidencia entre el cargo y la categoría, y pares con ground truth humano (las diferencias y similitudes respecto de los pares heurísticos, etiquetados por coincidencia de cargo-título en el caso de que la vacante y la hoja de vida pertenezcan a la misma categoría). Junto a estos, se incluyeron pares de ground truth humano real, derivados de los juicios de adecuación candidato-vacante contenidos en el dataset de emparejamiento vacante-hoja de vida (Vanetik & Kogan, 2023).

Cabe mencionar que la estructura agregada de Vanetik y Kogan (2023) (por candidatos, no por vacantes) invalidaba la reconstrucción de un ranking total de candidatos para una vacante concreta (por lo que el uso directo de la misma para la validación de Spearman quedaba invalidado). Sin embargo, cada juicio individual del tipo “esta vacante es o no adecuada para este candidato” sigue siendo una etiqueta binaria a nivel de par. Es esa granularidad (la pareja vacante-candidato individual, no la pareja agregada) la que se aprovecha aquí como ground truth humano para el clasificador binario.

El último conjunto de entrenamiento y de evaluación definitiva se compone de 73 pares con etiqueta humana verificada y 1.124 pares con etiqueta heurística. Esta composición responde a la disponibilidad limitada de juicios humanos, pero introduce una asimetría relevante: la mayor parte del aprendizaje y de la evaluación se apoya en etiquetas derivadas de reglas de coincidencia y no del criterio directo de un evaluador. Además, una proporción importante de los negativos heurísticos corresponde a pares no relacionados, cuya separación es menos exigente que la distinción entre candidatos cercanos pero con diferente grado de adecuación. Por ello, el F1 global puede representar una estimación optimista del desempeño frente a escenarios reales; en consecuencia, las métricas se desglosan por fuente de etiquetado y el resultado global no se interpreta como evidencia suficiente de generalización frente al juicio humano.

Tabla 9

Configuración de hiperparámetros del fine-tuning de DistilBERT

Hiperparámetro	Valor	Justificación
Modelo base	distilbert-base-multilingual-cased	Soporte multilingüe, ~134M parámetros, viable en CPU
Optimizador	AdamW	Estándar en fine-tuning de Transformers, corrige decaimiento de peso de Adam (Loshchilov & Hutter, 2019)
Tasa de aprendizaje	2e-5	Valor bajo típico para fine-tuning, evita olvido catastrófico
Longitud máxima de secuencia	256 tokens	Balance entre cobertura del texto y costo computacional
Warm-up	10% de los pasos totales	Estabiliza el entrenamiento en las primeras iteraciones
Épocas	3 (configurable)	Configuración por defecto
Batch size	8 (configurable)	Ajustado a la memoria disponible en CPU
Criterio de selección de checkpoint	Mejor F1 en desarrollo	Evita sobreajuste al conjunto de entrenamiento

Como se observa en la Tabla 9, el proceso de entrenamiento se lo llevó a cabo con un optimizador AdamW, tasa de aprendizaje de 2e-5, longitud máxima de secuencia de 256 tokens y un warm-up del 10% del total del número de pasos; el modelo que produzca el mejor F1 sobre el conjunto de desarrollo se conservará como el checkpoint final. La evaluación final se hace sobre un conjunto de prueba que se ha reservado, desglosando las métricas respecto a su fuente de etiquetado (heurística vs. ground truth humano), con el objetivo de evidenciar el riesgo de circularidad que tendría reportar un único F1 global dominado por los pares heurísticos, y además se hace una tabla comparativa final con respecto a TF-IDF, SBERT, TinyBERT y el módulo NER.

3.7.Validación mediante correlación de Spearman

La validación cuantitativa del módulo de ranking se lleva a cabo a través de dos fuentes independientes y con una metodología diferente para cada una de ellas.

Con el dataset de Vanetik y Kogan (2023) se contrasta el ranking generado por FreeATS para cada vacante con el ranking original humano. La exploración de su estructura y su implementación durante las pruebas propias de carga evidenció que este dataset se presentaba para ser interpretado como un ranking por candidato (es decir, qué vacantes son las que le convienen a un candidato) y no por vacante como se preveía en el anteproyecto, con lo que su transposición directa a un escenario de ranking por vacante (la que resuelve FreeATS) no es válida sin una reformulación; una transposición permite obtener correlaciones cercanas a cero.

Usando CareerCorpus (302 perfiles en 6 categorías equilibradas), se generan vacantes sintéticas y se calcula la correlación de Spearman entre el score de adecuación que genera FreeATS y la puntuación media por dos anotadores humanos (Annotator-1, Annotator-2). De

esta manera, CareerCorpus se convierte en la fuente principal de validación al ser su estructura de ranking por vacante aplicable a la tarea que resuelve FreeATS.

La creación de vacantes sintéticas a partir de CareerCorpus subsana una limitación de este dataset: dado que se trata de un dataset de hojas de vida individuales anotadas con un score de adecuación general por categoría ocupacional que a la vez no incluye descripciones de vacantes específicas contra las cuales probar un sistema de emparejamiento, la metodología subsana esta limitación generando para cada una de las seis categorías balanceadas del dataset una descripción de vacante representativa a partir de los campos estructurados (Domain, Education, Skills and Achievements) que están presentes en el corpus en el cual el texto se encuentra, de la forma que la vacante sintética que se obtiene es coherente con los términos y las expectativas de la categoría ocupacional a la que pertenece, introduciendo requisitos artificiales ajenos al dominio de los datos originales.

La correlación de Spearman entre el score de FreeATS y la media de los scores de los dos anotadores humanos se lleva a cabo sobre las seis categorías por una parte y de forma agregada sobre el conjunto total a partir de la misma estructura de reporte que la empleada por Chowdhury et al. (2026). Este diseño permite en otro trabajo futuro establecer si el rendimiento del sistema es homogéneo entre categorías ocupacionales o, por el contrario, que haya categorías donde el ranking híbrido se acerca más a la evaluación experta que en otras.

Adicionalmente, la interpretación de la magnitud que puede alcanzar un coeficiente de correlación de Spearman exige de contexto: no hay un umbral que separa una correlación "aceptable" de una "inaceptable", dado que la magnitud esperable depende del ruido que se asuma inherente al fenómeno que se quiere medir. En particular, para la adecuación candidato-vacante, el mismo techo humano reportado por Chowdhury et al. (2026). $\rho = 0.59$

entre dos anotadores expertos evaluando el mismo conjunto de candidatos aparece como la medida más adecuada para contextualizar los resultados que pueda obtenerse por un sistema automatizado sobre ese dataset, porque deja establecer el límite superior de acuerdo que se podría esperar aun sin la existencia de componente algorítmico.

3.8.Módulo de extracción de entidades nombradas (NER)

Al mismo tiempo que se construyó el módulo de comparación semántica, se desarrolló un modelo NER específico de dominio para extraer del texto de la hoja de vida los tipos de habilidades (SKILL), cargo (JOB_TITLE) e instituciones (ORG), como alternativa o complemento a la extracción asistida por modelo de lenguaje.

3.8.1. Preparación de datos de entrenamiento (silver-standard)

De acuerdo con Rebholz-Schuhmann et al. (2011) y Wissler et al. (2014) cuando los recursos para anotación manual son limitados, se implementa el enfoque silver-standard. Debido a la ausencia de datasets con hojas de vida en español que cuenten con anotaciones manuales de acceso público, se generó el conjunto de anotaciones automáticamente.

Por lo tanto, las anotaciones del entrenamiento se generan automáticamente localizando en el texto limpio de cada hoja de vida las posiciones exactas de las habilidades ya identificadas vía ESCO, las habilidades que se detectan en la fase de limpieza, los cargos extraídos y la propia institución educativa indicada. La localización se lleva a cabo por la búsqueda insensible a mayúsculas y acentos, con eliminación de solapamientos priorizando los términos más largos y específicos.

Al heredar los conjuntos de entrenamiento, desarrollo y prueba el mismo procedimiento de etiquetado, las métricas obtenidas reflejan la capacidad del modelo para reproducir dicho proceso de anotación.

3.8.2. Entrenamiento y evaluación

El modelo se entrenó desde el tokenizador en blanco para español (spacy.blank("es")) mediante el comando spacy train, ejecutado en CPU y con un número máximo de pasos configurables. La evaluación final se realizó sobre el conjunto de prueba calculando precisión, exhaustividad (recall) y valores de F1 (globales e individuales por tipo de entidad) mediante el Scorer nativo de spaCy, contrastando los resultados con el umbral objetivo del proyecto ($F1 \geq 0.80$). La distribución del corpus anotado para el entrenamiento del modelo NER se lo observa en la Tabla 10.

Tabla 10

Distribución del corpus anotado para el entrenamiento del modelo NER

Conjunto	Proporción	Semilla	Formato
Entrenamiento	80%	42	spaCy DocBin
Desarrollo	10%	42	spaCy DocBin
Prueba	10%	42	spaCy DocBin

3.9.Herramientas y entorno de desarrollo

El pipeline se desarrolló en Python, sin necesidad de GPU dedicada, con el objetivo de ejecutarlo en equipos de recursos limitados. Se emplearon Transformers (Wolf et al., 2020), Sentence-Transformers (Reimers & Gurevych, 2019), FAISS (Johnson et al., 2019), spaCy (Honnibal et al., 2020) y scikit-learn (Pedregosa et al., 2011) para los módulos principales, junto con la API GPT-4o-mini de OpenAI en la traducción y extracción de

campos. Otras librerías como pandas, NumPy, openpyxl, pdfplumber y python-docx se usaron en tareas auxiliares de procesamiento y manejo de datos. A continuación, en la Tabla 11 se resumen los principales componentes y su rol en el pipeline.

Tabla 11

Componentes principales del entorno de desarrollo

Componente	Herramienta / biblioteca	Rol en el pipeline
Lenguaje	Python 3.x	Implementación de todos los scripts del pipeline
Extracción de texto	pdfplumber, python-docx	Lectura de hojas de vida en PDF y DOCX
Modelos de lenguaje	transformers, torch	DistilBERT, TinyBERT (carga y fine-tuning)
Embeddings de oración	sentence-transformers	Modelo SBERT multilingüe
Búsqueda vectorial	faiss-cpu	Índice IndexFlatIP para recuperación de candidatos
Extracción de entidades	spaCy	Entrenamiento y evaluación del módulo NER
Baseline léxico y métricas	scikit-learn	TF-IDF, similitud coseno, métricas de clasificación
Manejo de datos	pandas, numpy, openpyxl	Procesamiento tabular y exportación a Excel
Traducción y extracción asistida	API de OpenAI (gpt-4o-mini)	Traducción de corpus y extracción de campos estructurados
Requisito de hardware mínimo	8 GB de RAM, CPU estándar	Sin GPU dedicada, ejecutable en equipo de oficina

Previo al procesamiento de los datasets principales, se ejecutó un diagnóstico de recursos del sistema que estima el tamaño de batch y el número de workers recomendado en función de la memoria RAM disponible, con el fin de garantizar que el pipeline completo pueda ejecutarse en un equipo estándar con un mínimo de 8 GB de RAM.

3.10. Reproducibilidad, control de versiones y organización del repositorio

Conforme a los principios de reproducibilidad, el repositorio organiza los scripts del pipeline siguiendo una estructura de carpetas que refleja las etapas del cronograma: data/raw para los datasets originales sin procesar; data/processed para los artefactos intermedios generados en cada etapa (traductores, limpiados, estructurados, con embeddings, anotaciones ESCO); models/ para los checkpoints entrenados de spaCy y DistilBERT; y logs/ para los registros de ejecución y métricas de cada script. Esta organización, definida desde el primer script del pipeline, permite identificar en todo momento la etapa del flujo de trabajo a la que pertenece una ejecución, sin necesidad de revisar el código fuente que lo generó.

Cada script correspondiente al pipeline sigue, por otro lado, un patrón de diseño concreto: verifica al inicio de su ejecución que existen los artefactos de entrada requeridos; procesa los datos de entrada; persiste sus resultados junto a un resumen de ejecución emitido por consola y, en las etapas de evaluación, a un archivo de log estructurado como un JSON. Este patrón evita que el pipeline se ejecute en un orden erróneo o sobre datos inconsistentemente analizados.

3.11. Selección de la arquitectura final

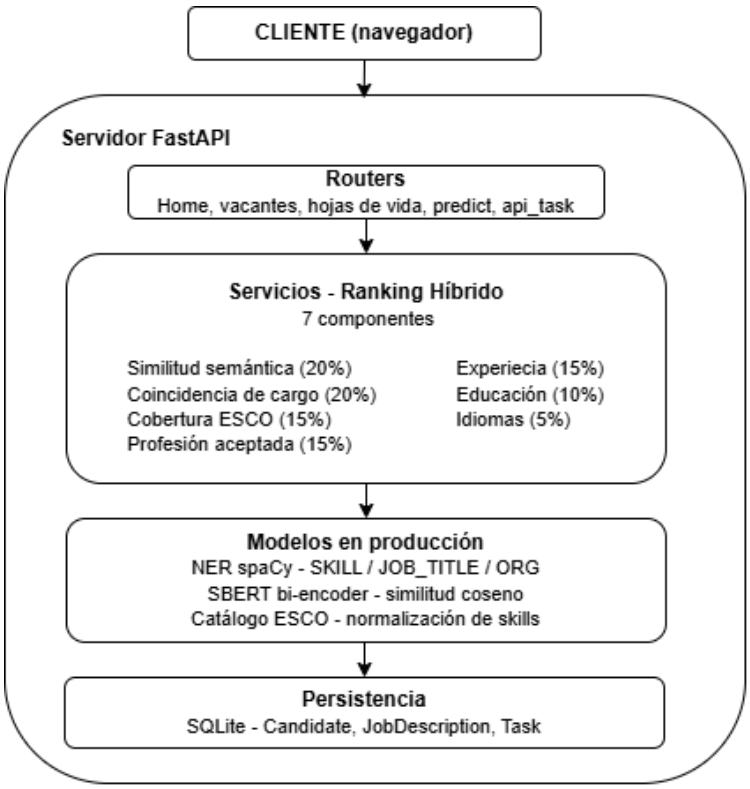
El proceso descrito anteriormente se utilizó para la preparación, extracción, comparación y validación. Por su parte, la sección presente indica cómo estos componentes se integran en la arquitectura final a ser utilizada de forma interactiva (un profesional de recursos humanos registra una vacante, carga hojas de vida y consulta un ranking desde su interfaz web local), aprovechando en su implantación artefactos ya generados del modelo NER, del catálogo ESCO y del bi-encoder SBERT validados, y adaptando su ejecución para procesar cada par candidato-vacante bajo demanda.

3.11.1. Arquitectura del modelo final

Como se observa en la Figura 5, la arquitectura final combina un bi-encoder SBERT con el módulo NER y un ranking híbrido de siete componentes ponderados. Esta configuración se establece a partir de dos alternativas experimentadas: bi-encoder SBERT con similitud coseno y clasificador binario DistilBERT. Además, se evaluó como parte del diseño de la arquitectura a cross-encoder zero-shot operando bajo el esquema retrieve-and-rerank.

Figura 5

Arquitectura final del módulo de matching (bi-encoder SBERT + módulo NER + ranking híbrido ponderado).



A pesar de que el método DistilBERT obtuvo el F1 más alto entre las alternativas evaluadas en pruebas aisladas, no fue la alternativa seleccionada. En las siguientes subsecciones se detallan los criterios metodológicos independientes al valor de las métricas

que sustentan esta omisión; la evidencia cuantitativa se reporta en la sección de análisis de resultados.

Es necesario precisar que la preparación del corpus y el entrenamiento de los modelos constituyen un proceso de investigación que sí depende de conexión a internet y de un servicio comercial en la nube. En específico se refiere a la API de OpenAI (GPT-4o-mini), para la traducción del corpus y para la extracción de campos estructurados. Mientras que la consideración de que FreeATS no depende de OpenAI ni de ningún otro servicio en la nube se refiere, sólo a las operaciones que realiza la PYME usuaria en tiempo de uso o inferencia, a través de la interfaz web local: cargar una hoja de vida, registrar una vacante y consultar un ranking.

3.11.2. Criterios de decisión frente a las alternativas evaluadas

1) Validez de la métrica de origen.

El primer criterio de decisión se enfoca en el resultado F1 de DistilBERT con respecto al resto de candidatos. El conjunto de entrenamiento y de prueba del clasificador está dominado por etiquetas heurísticas (94% de los pares), lo cual afecta las pruebas con datos distintos al conjunto señalado, dado que el modelo fue optimizado con etiquetas fijas y no frente al criterio de un reclutador humano. En la literatura, este problema se conoce como "fuga de etiqueta por proxy": el modelo reporta métricas que reflejan más la fidelidad del sistema en procesos de anotación automática; sin embargo, no demuestra una capacidad real de generalización en la tarea de emparejar candidatos con vacantes (Ratner et al., 2016). Bajo esta premisa, la métrica F1 global reportada no constituye, por sí sola, una evidencia ni un argumento sólido para respaldar su uso.

Por otro lado, el cross encoder sí fue evaluado con métricas independientes de las heurísticas de etiquetado, la correlación de Spearman contra CareerCorpus y el MRR de self retrieval, cuyos resultados fueron inferiores a los obtenidos por el bi encoder. En este contexto, la problemática radica en la transferencia del modelo: al ser zero-shot y no haber sido ajustado para el dominio de emparejamiento vacante-hoja de vida, el método cross-encoder no resulta adecuado para transferirse a este contexto sin un fine-tuning propio.

2) Generalización a hojas de vida reales.

El segundo criterio de decisión es conceptual y anticipa el comportamiento del sistema en pruebas con vacantes y hojas de vida reales. Un modelo ajustado como DistilBERT se entrena para aprender patrones específicos del corpus de entrenamiento, tales como la redacción, la longitud y las estructuras de las hojas de vida del dataset. De este modo, las métricas de desempeño se centran en hojas de vida con estructuras similares; sin embargo, en pruebas con hojas de vida reales, el rendimiento dependerá de cuánto se asemejan a la estructura utilizada durante el entrenamiento.

En un caso real, un usuario de recursos humanos utilizará un conjunto de hojas de vida con formatos y vocabularios potencialmente distintos a los del corpus de entrenamiento, precisamente el escenario en el que el modelo clasificador experimenta un deterioro en sus métricas y un desajuste frente a reglas explícitas que no formaron parte de su entrenamiento.

3.11.3. Diseño de evaluación del ranking con documentos reales y vacantes sintéticas

Para esta investigación se dividió la evaluación en dos casos de exploración con objetivos diferentes. En el primero, mediante la recolección propia, se evaluaron diez hojas de vida reales y vacantes derivadas de esos mismos documentos como una prueba

exploratoria de recuperación controlada. Debido a que cada vacante comparte información con su hoja de vida de origen, se denominó a dicho documento CV fuente y no “candidato ideal”.

En el segundo caso de exploración se evaluó el ranking de FreeATS con CareerCorpus y vacantes construidas independientemente del dataset. Para ello se crearon seis perfiles ocupacionales a partir de ESCO. En cada ejecución, FreeATS ordenó a los candidatos del mismo dominio y se comparó sus resultados con el promedio de dos anotadores mediante Precision@1/@3/@5/@10.

La creación de vacantes sintéticas resaltó una limitación adicional del corpus principal: tanto CareerCorpus como el resto de conjuntos de datos abiertos asocian cada categoría ocupacional a un conjunto acotado y homogéneo de perfiles, lo que dificulta representar situaciones reales de reclutamiento en las que una vacante laboral puede considerar distintos perfiles profesionales (por ejemplo, una vacante de científico de datos puede aceptar tanto ingenieros en sistemas como estadísticos o economistas). Para reflejar este punto, se incorporó al formulario de vacantes y al cálculo del ranking híbrido el campo “profesiones aceptadas”, mediante el cual el usuario puede especificar más de una profesión para la vacante que está buscando.

La incorporación de este campo implicó recalibrar los pesos originales del rating híbrido para incluir un séptimo componente, 'Score de profesión'. De tal manera, que los componentes y sus respectivos pesos finalmente quedan como se observa en la Tabla 12.

Tabla 12

Componentes y ponderaciones del ranking híbrido de FreeATS

Componente	Peso
Score SBERT	20%
Score de cargo	20%
Score ESCO	15%
Score de años de experiencia normalizados	15%
Score de profesión	15%
Score de educación	10%
Score de idiomas	5%
Total	**100%**

3.11.4. Consideraciones de eficiencia y restricciones de producto

El alcance de la solución establece que el sistema debe correr localmente, en CPU, sin infraestructura dedicada, lo que condiciona qué arquitecturas de matching resultan viables para el producto final.

- **Latencia:** El bi-encoder calcula por separado los embeddings de las hojas de vida y de la vacante. Los embeddings de las hojas de vida pueden precomputarse y luego compararse de forma vectorizada con el embedding de la vacante. Aunque el costo escale en función del número de candidatos, este proceso sigue siendo mucho más eficiente que ejecutar DistilBERT o un cross-encoder zero-shot para cada par hojas de vida-vacante. Esta diferencia resulta clave puesto que el ranking debe generarse en tiempo real sobre todos los candidatos disponibles.
- **Huella de memoria:** Cargar un modelo adicional de 260 MB (DistilBERT) o 470 MB (cross-encoder) en el módulo de recursos de la aplicación implica un

costo de arranque y de memoria que no se justifica, ya que no aporta mejoras medibles frente al bi-encoder que ya está en uso.

- **Interpretabilidad:** El análisis de brechas mostrado al reclutador en la aplicación web requiere que el score final sea una suma clara y verificable de componentes como similitud semántica, cargo, ESCO, experiencia, educación e idiomas. En cambio, un clasificador de caja negra como DistilBERT devuelve un único score sin desglose, siendo incompatible con el requisito de usabilidad definido en los objetivos específicos.

Estas tres aristas se agregan al análisis macro, en conjunto con las métricas de generalización como pilares de la arquitectura final del proyecto. Teniendo como resultado los piplines enlistados en la Tabla 13.

Tabla 13

Inventario completo de scripts del pipeline FreeATS

Script	Fase	Función
00_setup_project.py	1	Estructura de carpetas y verificación de dependencias
01_download_datasets.py	1	Descarga de los seis datasets desde sus fuentes originales
02_inventory_check.py	1	Inventario y verificación de integridad de datasets
03_diagnostico_pretrad.py	1	Diagnóstico de recursos del sistema (RAM, batch size)
04_translate_openai.py	1	Traducción del corpus con GPT-4o-mini (motor seleccionado)
04_translate_helsinki_alt.py	1	Traducción alternativa local con Helsinki-NLP
05_clean_normalize.py	1	Limpieza y normalización de texto extraído
05b_extract_fields.py	1	Extracción de campos estructurados vía GPT-4o-mini

05b_evaluador_completo.py	1	Evaluación de cobertura de la extracción de campos
06_esco_integration.py	1	Integración y anotación contra la taxonomía ESCO
07_merge_corpus.py	1	Fusión de campos y anotación ESCO en corpus_final.json
08_process_vacancies.py	2	Traducción y extracción de campos de vacantes
09_generate_embeddings.py	2	Generación de embeddings SBERT de hojas de vida y vacantes
10_build_faiss_index.py	2	Construcción del índice vectorial FAISS
11_ranking_hibrido.py	3	Cálculo del score de ranking híbrido ponderado
11a_preparar_corpus_validacion.py	3	Preparación de vacantes sintéticas (CareerCorpus)
11a_b_preparar_career_corpus.py	3	Preparación complementaria de CareerCorpus
11b_validar_spearman.py	3	Validación de Spearman contra el dataset de Vanetik y Kogan (2023)
11c_validar_spearman_careercorpus.py	3	Validación de Spearman contra CareerCorpus
12_prepare_ner_data.py	2	Generación de anotaciones silver-standard para NER
13_train_ner.py	2	Entrenamiento del modelo NER con spaCy
14_evaluate_ner.py	2	Evaluación del modelo NER (precisión, recall, F1)
15_baseline_tfidf.py	2	Baseline de matching léxico TF-IDF
15b_baseline_tinybert.py	2	Baseline de matching TinyBERT zero-shot
16_prepare_finetuning_data.py	2	Preparación de pares etiquetados para fine-tuning
17_finetune_distilbert.py	2	Fine-tuning del clasificador DistilBERT
18_evaluate_distilbert.py	2	Evaluación final y tabla comparativa de métodos

CAPÍTULO 4

4. Análisis de resultados

4.1. Resultados de la evaluación de traducción

Los resultados que arrojan las cuatro alternativas evaluadas: Helsinki-NLP/opus-mt-en-es (open source), Google Cloud Translation v2, GPT-4o-mini de OpenAI y Claude Haiku 4.5 de Anthropic sobre una muestra estratificada del corpus, calculados mediante las cinco métricas se muestran en la Tabla 14.

Tabla 14

Comparativo de métricas de calidad de traducción por método.

Método	BLEU ↑	TER ↓	chrF ↑	chrF++ ↑	COMET ↑
Claude Haiku 4.5	32.6013	344.0752	76.3986	70.1118	0.7932
Google Cloud Translation	28.6607	368.4997	76.5026	69.5915	0.7957
Helsinki-NLP (opus-mt-en-es)	24.4747	369.6319	72.1620	65.3503	0.7185
OpenAI (GPT-4o-mini)	33.2390	349.3292	76.8532	70.5236	0.7890

↑ *valor más alto es mejor;* ↓ *valor más bajo es mejor.*

De acuerdo con los resultados obtenidos ninguno de los métodos analizados domina las cinco métricas simultáneamente, lo que implica que para seleccionar entre uno u otro se pondera el conjunto de evidencia y no un solo indicador. Al respecto, GPT-4o-mini obtiene el mejor resultado en tres de las cinco métricas (BLEU = 33.2390, chrF = 76.8532, chrF++ = 70.5236), situándose en la segunda posición, aunque con diferencias marginales, en las dos restantes: TER (349.3292 frente a 344.0752 de Claude Haiku 4.5 (que es el mejor en esta métrica)) y COMET (0.7890 frente a 0.7957 de Google Cloud Translation, que es el mejor en esta métrica). Helsinki-NLP, la única alternativa local y sin coste por token, obtiene los

valores más bajos en las cinco métricas, siendo evidente la pérdida de rendimiento en la métrica COMET (0.7185), que es la más cercana a la del juicio de calidad humano, lo cual coincide con que es un modelo neuronal de propósito general, sin ninguna optimización en relación con el registro semiformal y el vocabulario técnico propios de una hoja de vida.

Con base en los resultados, se optó por OpenAI, GPT-4o-mini como motor de traducción del corpus. La elección se basó en que este motor obtuvo el mejor desempeño agregado entre las cinco métricas consideradas (incluyendo las métricas de superposición de caracteres / chrF, chrF++, que son relevantes a la hora de evaluar la preservación de la terminología técnica y los nombres propios que tienen protagonismo en las hojas de vida) y que las diferencias obtenidas con respecto a la mejor alternativa en TER y COMET en términos relativos son menores (1.5% y 0.8%, respectivamente), mientras que la ventaja de Helsinki-NLP respecto a sus costos operativos (ya que se puede ejecutar localmente sin hacer uso de una API de pago) no compensa una brecha de calidad que es mucho más importante, sobre todo en COMET. La decisión fue también validada revisando de nuevo una muestra aleatoria estratificada del 5% del corpus traducido, teniendo en cuenta la correcta transferencia de los términos técnicos del ámbito de los recursos humanos. En el repositorio se encuentra también la implementación alternativa sobre Helsinki-NLP como la ejecución local del sistema, es decir, el que no depende de servicios en nube, y se incluye para garantizar la exhaustividad metodológica.

4.2.Resultados de los módulos de PLN

4.2.1. Módulo de extracción de entidades (NER)

El modelo NER entrenado con la metodología definida para el proyecto fue evaluado sobre el conjunto de prueba reservado (10% del corpus anotado, no visto durante el entrenamiento). Los resultados globales y por tipo de entidad se presentan en la Tabla 15.

Tabla 15

Resultados del módulo de extracción (NER) por tipo de entidad

Entidad	Precisión	Recall	F1
SKILL	0.9809	0.9626	0.9717
JOB_TITLE	0.9119	0.7688	0.8343
ORG	0.6935	0.5375	0.6056
Global	0.9551	0.8905	0.9217

Nota. Evaluado contra silver standard .

El F1 global de 0.9217 supera el umbral objetivo declarado en el cronograma del proyecto (F1 ≥ 0.80), lo que es consistente con la hipótesis planteada en el proyecto. Es importante destacar que se evalúa contra un silver-standard.

De acuerdo con los resultados del módulo de extracción (NER) por tipo de entidad, SKILL obtiene el mejor desempeño (F1 = 0.9717), lo cual es coherente con que estas menciones provienen del catálogo normalizado ESCO utilizado tanto para generar las anotaciones de entrenamiento como para evaluarlas. JOB_TITLE obtiene un desempeño intermedio (F1 = 0.8343), con un recall más bajo (0.7688) que su precisión, sugiriendo que el modelo es conservador para esta categoría. Se podría atribuir esto por la mayor variabilidad

léxica de los títulos de puesto frente a las habilidades. ORG presenta el desempeño más bajo (F1 = 0.6056), consistente con que las instituciones educativas y empresas mencionadas en una hoja de vida son nombres propios de alta variabilidad y baja frecuencia individual, lo que dificulta su generalización con un volumen de entrenamiento moderado.

4.2.2. Desempeño y comparación de métodos de matching semántico

En la Tabla 16 se presenta el desempeño individual de los cuatro métodos de matching semántico implementados: los tres baselines planteados inicialmente: TF-IDF, SBERT, TinyBERT zero-shot y el clasificador fine-tuning DistilBERT.

Tabla 16

Desempeño reportado por método de matching semántico

Método	Tipo	Precisión	Recall	F1 / Score
TF-IDF	Baseline léxico	N/A	N/A	0.2735
SBERT	Baseline semántico	N/A	N/A	0.7256
TinyBERT (zero-shot)	Baseline transformer	N/A	N/A	0.9670*
DistilBERT (fine-tuning)	Sistema (OE4)	0.9412	0.9412	0.9412

** Score bruto inflado por anisotropía, no comparable directamente con los demás métodos. DistilBERT evaluado sobre un test set de 121 pares.*

El salto de F1/score entre TF-IDF (0.2735) y SBERT (0.7256) cuantifica de forma directa el aporte de la comprensión semántica frente a la coincidencia léxica, confirmando la hipótesis del proyecto. Sin embargo, el score de 0.9670 obtenido por TinyBERT en zero-shot no debe leerse como evidencia de que este modelo supera a SBERT: por el sesgo de escala que se analiza a continuación, este valor está inflado por anisotropía. Es decir, los

embeddings de un transformer sin fine-tuning ocupan una región angosta del espacio vectorial, elevando artificialmente cualquier similitud coseno.

Para obtener una comparación libre de ese sesgo de escala, se calculó Precision@5 (P@5) promedio para cada método, una métrica de ranking relativo que no depende de la magnitud absoluta del score. En la Tabla 17 se presenta este comparativo, con un ground truth basado en coincidencia de sector_industria.

Tabla 17

Comparación de métodos de matching semántico bajo una métrica común (Precision@5)

Método	P@5 promedio	F1 / Score
B1 - TF-IDF	0.733	0.2735
B2 - SBERT puro	0.800	0.7256
S1 - Híbrido SBERT + ESCO	0.800	N/A
B3 - TinyBERT (zero-shot)	0.333	0.9670*
S2 - DistilBERT (fine-tuning)	0.733	0.9412

** Score bruto inflado por anisotropía que se recomienda comparar por P@5, no por el score bruto.*

Bajo esta métrica de ranking relativo, TinyBERT zero-shot obtiene el P@5 más bajo de los cinco métodos evaluados (0.333), lo cual confirma que su score bruto elevado (0.9670) es engañoso y no se traduce en mejores recuperaciones reales: es el caso que ilustra por qué evaluar un transformer sin fine-tuning por su score de similitud, sin una métrica de ranking complementaria, puede llevar a conclusiones erróneas. SBERT puro (B2) y el híbrido SBERT+ESCO (S1) obtienen el P@5 más alto (0.800) y empatado entre sí, un hallazgo que se analiza con mayor detalle: en esta prueba exploratoria, el componente ESCO no aporta una mejora medible sobre SBERT puro. DistilBERT con fine-tuning (S2) obtiene un P@5 de 0.733, igual al de TF-IDF (B1) bajo esta métrica particular, aun cuando su F1 de clasificación

binaria (0.9412) es mayor. Lo anterior parece una aparente contradicción que se explica porque P@5 mide calidad de ranking (orden relativo) mientras que F1 mide precisión de clasificación binaria sobre pares individuales, dos tareas relacionadas pero no idénticas.

A modo de detalle, se presenta el desglose de Precision@5 y Precision@10 por vacante de prueba para los tres sectores evaluados en esta ronda exploratoria.

Tabla 18

Precision@K por vacante de prueba y método

Vacante	B1 TF-IDF	B2 SBERT	S1 Híbrido	B3 TinyBERT	S2 DistilBERT
V01 Tecnología	P@5=1.000 / P@10=0.700	P@5=0.400 / P@10=0.700	P@5=0.800 / P@10=0.800	P@5=1.000 / P@10=1.000	P@5=0.800 / P@10=0.700
V02 Finanzas	P@5=1.000 / P@10=1.000	P@5=1.000 / P@10=1.000	P@5=1.000 / P@10=0.700	P@5=0.000 / P@10=0.000	P@5=1.000 / P@10=1.000
V03 Salud	P@5=0.200 / P@10=0.300	P@5=1.000 / P@10=1.000	P@5=0.600 / P@10=0.800	P@5=0.000 / P@10=0.000	P@5=0.400 / P@10=0.700

Nota. Ground truth basado en coincidencia de sector_industria.

El detalle por vacante confirma que ningún método domina de forma consistente en todos los sectores: TF-IDF obtiene P@5 perfecto en tecnología y finanzas, pero baja en salud (P@5 = 0.200); TinyBERT zero-shot obtiene P@5 perfecto en tecnología, pero cero en finanzas y salud; SBERT puro es el más estable de los cinco (nunca por debajo de 0.400). Dado el tamaño reducido de esta muestra (tres vacantes).

4.2.3. Ranking híbrido y caso límite documentado

El ranking híbrido combina similitud semántica (30%), similitud ESCO (25%), coincidencia de cargo (15%), experiencia (15%), educación (10%) e idiomas (5%). Durante las pruebas exploratorias se documentó un caso límite ilustrativo del componente de coincidencia de cargo: un perfil identificado como "RaviBurra - Certified PM / DevOps", con

un título profesional que combina gestión de proyectos (PM) y una mención técnica (DevOps), fue rankeado en primer lugar para una vacante de desarrollo backend, desplazando a candidatos con perfiles técnicos más alineados al puesto.

Este caso evidencia el límite estructural de un heurístico basado en reglas léxicas para el componente de coincidencia de cargo: al estar basado en reglas léxicas, no captura casos ambiguos donde el título usa abreviaturas o combina roles técnicos y de gestión (ej. "PM DevOps"). Una extensión robusta requeriría un modelo de clasificación de cargos entrenado sobre un corpus etiquetado, o el uso de la taxonomía de ocupaciones de ESCO.

4.2.4. Validación mediante correlación de Spearman

En la Tabla 19, se observa un resumen de los resultados de los análisis de correlación de Spearman obtenidos a partir de los dos conjuntos de datos de evaluación definidos para el proyecto, junto con las dos configuraciones de pesos del ranking híbrido contrastadas durante el desarrollo.

Tabla 19

Correlación de Spearman por fuente de validación

Fuente	Configuración	Spearman ρ	Estado / interpretación
Vanetik (66 hojas de vida)	6 componentes	0.015	Hallazgo metodológico: la transposición de un ranking por candidato a un ranking por vacante no es válida
CareerCorpus (302 perfiles)	6 componentes	0.232	Configuración preliminar, sin el componente de profesión
CareerCorpus (302 perfiles)	7 componentes	0,2596	Configuración final: cuatro de seis categorías significativas ($p < 0,05$)

El primer resultado obtenido sobre la fuente de Vanetik y Kogan (2023) ($\rho = 0{,}015$) debe interpretarse con cautela, sin asociarlo a una falla del ranking híbrido ni considerarlo evidencia de desempeño: esta fuente está organizada como un ranking por candidato y no por vacante, por lo que no puede trasladarse de manera literal. Por lo tanto, se excluye como evidencia y se lo considera un aporte metodológico, útil para futuros trabajos en los que se emplee esta fuente. En este sentido, la fuente estructuralmente compatible en relación con el ranking de vacantes es CareerCorpus, sobre la cual se concentra el resto del análisis.

Con respecto a CareerCorpus, se evaluaron dos configuraciones sucesivas del ranking. La primera combinación, con seis componentes, obtuvo un resultado de $\rho = 0{,}2325$, mientras que la de siete componentes (que incorpora el componente de profesión descrito en la sección 3.10) alcanzó $\rho = 0{,}2596$, elevando de tres a cuatro las categorías estadísticamente significativas. Dicho valor corresponde a la arquitectura final implementada en el proyecto. Dado que el promedio macro agrega comportamientos heterogéneos, sobre el último valor de ρ obtenido se calcularon el valor p y el desglose por categoría. En la Tabla 20 se observa el resultado de la correlación de Spearman por categoría en la configuración de siete componentes.

Tabla 20

Correlación de Spearman por categoría en la configuración final de siete componentes

Categoría	n	Spearman ρ	Valor p	Acuerdo entre anotadores (ρ)	% del acuerdo humano	¿Significativa? (α = 0,05)
Teacher	49	0,4253	0,0023	0,5599	76 %	Sí
Finance	50	0,3667	0,0088	0,5947	62 %	Sí
Apparel	50	0,3664	0,0089	0,8879	41 %	Sí
Banking	50	0,2800	0,0489	0,4605	61 %	Sí

Accountant	51	0,0825	0,5651	0,3334	25 %	No
Research Assistant	51	0,0368	0,7975	0,6365	6 %	No
Promedio macro	**301**	**0,2596**	**N/A**	**0,5788**	**45 %**	**4 de 6**

El desglose muestra un comportamiento desigual. En cuatro de las seis categorías la correlación con el juicio experto es positiva y significativa, y alcanza entre el 41 % y el 76 % del acuerdo observado entre los anotadores humanos; en Accountant y Research Assistant no es distinguible de cero. Ese acuerdo también varía por categoría (de $\rho = 0,8879$ en Apparel a $\rho = 0,3334$ en Accountant), de modo que allí donde el criterio de referencia es poco consistente queda limitada la correlación máxima alcanzable por cualquier sistema. A partir de ello, se determinan dos limitaciones del diseño: El componente de profesión se definió mediante palabras clave de cada ocupación, procedimiento más favorable que una verificación real sobre una hoja de vida nueva; y comparación con los modelos originales ($r = 0,83$ a $0,86$) es contextual, pues corresponde a modelos de regresión evaluados con Pearson y bajo otro esquema.

Los resultados permiten afirmar que el ranking híbrido captura parcialmente la señal de adecuación reconocida por evaluadores humanos en la mayoría de las categorías, pero no que reproduzca de forma fiable el criterio experto ni que esté validado para decisiones de selección. Con ello, preliminarmente y de forma técnica se cumple el objetivo 4: mientras el módulo NER y los baselines del objetivo específico 3 cuentan con evaluación completa y trazable sobre el corpus final, la evidencia con respecto al ranking se enfoca en una evaluación exploratoria sobre un conjunto de datos externos. Para una validación más concluyente es necesario más vacantes independientes y evaluación humana externa con criterio de reclutadores, líneas que se retoman en el apartado de trabajo futuro.

4.2.5. Prueba de generalización con Voto Informado 2026

De acuerdo con el alcance definido, el dataset Voto Informado 2026 se utilizó para evaluar la capacidad de generalización del módulo NER sobre texto en español nativo, fuera del dominio de hojas de vida laborales. Sobre una muestra de 300 perfiles, el 62% contuvo al menos una entidad reconocida por el modelo, con un promedio de 1.73 entidades por perfil, distribuidas de forma desigual entre tipos de entidad: SKILL = 0.07, JOB_TITLE = 0.18 y ORG = 1.48 entidades promedio por perfil.

Esta distribución es consistente con la naturaleza del dominio de origen de los perfiles (candidaturas electorales, no hojas de vida laborales): las menciones de organizaciones (partidos políticos, instituciones) son abundantes, mientras que las menciones de habilidades técnicas y cargos laborales en el sentido en que el modelo fue entrenado son escasas. La interpretación de este resultado es de cobertura moderada: el módulo NER generaliza bien al español nativo en cuanto a su capacidad de segmentación y reconocimiento léxico, pero no generaliza al dominio electoral en cuanto al tipo de entidades de interés, dado que el modelo fue entrenado sobre vocabulario ocupacional.

4.2.6. Discusión de los resultados

La arquitectura final (bi-encoder SBERT + NER + Ranking Híbrido Ponderado) es la que tiene mayor evidencia acumulada para el caso de uso analizado. Dadas las dos formas de validación independientes que se utilizan (correlación de Spearman y MRR de self retrieval en la fase comparativa, y el experimento de ranking con hojas de vida reales en la fase exploratoria) esta arquitectura muestra un buen desempeño, presenta una generalización más adecuada al tipo de entrada que se le hará en producción y satisface las restricciones de

eficiencia e interpretabilidad que se han definido para el producto, aun no siendo la que tiene el mejor F1 de manera aislada sobre el test set comparativo.

El hallazgo metodológico más importante del proyecto es el realizado en el contraste entre TinyBERT zero-shot (el modelo que más se alejaba de la arquitectura propuesta) y el resto de los métodos sobre los que se realizó la evaluación. TinyBERT fue el que tuvo el score bruto de similitud más alto de los cuatro baselines (0,9670) y, a la vez, el Precision@5 más bajo (0,333). Evaluar un transformer preentrenado solamente por la magnitud de su score de similitud, sin contrastar con una métrica que sea relativa al ranking, puede llegar a aportar resultados encontrados de lo que puede llegar a ser el desempeño del modelo. Este resultado se alinea con la literatura sobre anisotropía en representaciones contextuales y se transforma en una aportación metodológica aplicada, además de estadística, del experimento, dado que avisa de cómo no se debe evaluar un modelo de similitud semántica más allá de la que se ha realizado en el presente caso, es generalizable a otros sistemas de PLN de poco recurso.

El módulo de extracción de entidades (NER) mostró un buen rendimiento (F1 = 0.9217), destacándose en la categoría SKILL (F1 = 0.9717), que es la más relevante para el propio objetivo del sistema, que era identificar habilidades en hojas de vida. Este resultado ratifica que se puede llegar a un extractor de entidades competitivo entrenando con una anotación silver-standard basada en ESCO y modelos de lenguaje, que es la adecuada en contextos donde no es posible hacer una anotación manual a gran escala (Rebholz-Schuhmann et al., 2011; Wissler et al., 2014).

El módulo de comparación semántica muestra cómo el fine-tuning de DistilBERT alcanzó un F1 = 0.9412 (F1 +0,13 en comparación con baselines zero-shot con TF-IDF, SBERT y TinyBERT), lo que demuestra el valor del fine-tuning específico de tarea sobre el

uso de embeddings generales para el problema de matching candidato-vacante. La elección del motor de traducción GPT-4o-mini se ven corroboradas por la evaluación comparativa: el valor de la métrica COMET con respecto a Helsinki-NLP cobra relevancia puesto que todo el pipeline opera sobre el texto que ha sido ya traducido, es decir, la calidad de la traducción determina la calidad del sistema a través de end-to-end.

La ambigüedad en el caso PM/DevOps es representativa del valor adicional de contar con una taxonomía ocupacional estructurada como ESCO en oposición al uso de heurísticos de coincidencia léxica; mientras estos últimos no permiten distinguir familias ocupacionales relacionadas y diferentes, la jerarquía de ESCO sí lo permite, por lo que cobra valor la inclusión de ella misma como un componente central de la arquitectura y como línea prioritaria de refinamiento.

Se validó la arquitectura del proyecto a partir de dos fuentes de evidencia independientes: la comparación de baselines por sector y la validación con hojas de vida reales dentro de un enfoque exploratorio propio de un primer despliegue con recursos limitados. Extender el número de los escenarios evaluados y consolidar la validez estadística es necesario para una segunda iteración para que las tendencias detectadas entre distintos métodos puedan ser validadas con mayor solidez; de la misma manera, el módulo NER fue evaluado bajo el enfoque silver-standard, incorporar la muestra de validación humana en futuros trabajos permitiría confirmar el rendimiento alcanzado.

4.3.Consideraciones de rendimiento computacional

En la Tabla 21 se observan los resultados sobre el costo computacional relativo de cada etapa del pipeline, como complemento de las métricas de calidad.

Tabla 21

Costo computacional relativo por etapa del pipeline

Etapa	Recurso crítico	Escala de costo relativo
Traducción con GPT-4o-mini	Llamadas a API externa	Medio, condicionado por la latencia de red y el costo por uso
Extracción de campos asistida	Llamadas a API externa	Medio, con un patrón similar al de traducción
Integración ESCO	CPU	Bajo, mediante búsqueda léxica local
Generación de *embeddings* SBERT	CPU	Bajo a medio, con una sola pasada por el corpus
Búsqueda vectorial FAISS	CPU y memoria RAM	Bajo para el tamaño actual del corpus
Fine-tuning de DistilBERT	CPU o GPU	Alto; es la etapa más costosa del pipeline
Entrenamiento NER	CPU	Medio a alto

La etapa de mayor costo computacional es el fine-tuning de DistilBERT, tanto por el tamaño del modelo (aproximadamente 134 millones de parámetros) como por el número de pasos de entrenamiento necesarios sobre CPU en ausencia de GPU dedicada. Esta observación refuerza la pertinencia de mantener, junto al enfoque fine-tuning, los tres baselines de menor costo (TF-IDF, SBERT, TinyBERT zero-shot) evaluados como alternativas viables para escenarios de despliegue donde ni siquiera el tiempo de entrenamiento inicial de DistilBERT resulte asumible por una PYME, aun cuando ese modelo obtenga el mejor F1 de clasificación entre los métodos comparados. En contraste, las etapas de inferencia que incluye el tiempo de uso para la generación de embeddings de una

nueva vacante, búsqueda en el índice FAISS y cálculo del score de ranking híbrido tienen un costo computacional menor y son las que determinan la experiencia de uso cotidiana del sistema una vez entrenado, un aspecto que favorece una experiencia ágil en la interfaz web local.

Esta distinción entre costo de entrenamiento (asumido una única vez, de forma centralizada, antes de la distribución del sistema) y costo de inferencia (asumido repetidamente, por cada PYME usuaria en cada consulta de matching) tiene una implicación práctica directa para el modelo de distribución de FreeATS: los checkpoints ya entrenados (el modelo NER y el clasificador DistilBERT) pueden empaquetarse y distribuirse junto con la interfaz web local, de modo que ninguna PYME usuaria del sistema deba asumir por sí misma el costo computacional de entrenamiento documentado; su interacción con el sistema se limitaría a las etapas de inferencia, de menor costo. Esta separación entre una etapa de entrenamiento centralizada y una etapa de inferencia distribuida es consistente con el objetivo general del proyecto de ofrecer una herramienta operable en las condiciones de infraestructura reales de una PYME, y debe mantenerse como requisito de diseño en las siguientes versiones de la aplicación.

4.4.Validación de la arquitectura de matching con hojas de vida reales

4.4.1. *Evidencia cuantitativa de los modelos para la arquitectura*

Entre los tres métodos comparados en aislamiento, el clasificador DistilBERT ajustado obtuvo una métrica F1 global de 0.9383 sobre los datos de prueba, siendo el valor más alto de todos. No obstante, al analizar este resultado según el origen de la etiqueta, el F1 alcanzó 0.9589 en los datos con etiquetas heurísticas (1.124 de 1.197 pares, es decir, el 94% del total), mientras que alcanzó 0.75 en las etiquetas ground truth (73 pares etiquetados por

humanos). Se observa una brecha que sugiere que el modelo aprendió los patrones de la etiquetación automática, lo que indica que el F1 global podría estar sobreestimando la capacidad real frente a evaluaciones realizadas por personas.

Por otra parte, el cross-encoder zero-shot (mmarco-mMiniLMv2), evaluado bajo la metodología retrieve-and-rerank, fue contrastado con métricas independientes de la heurística de etiquetado y obtuvo resultados inferiores al bi-encoder SBERT: la correlación de Spearman contra CareerCorpus disminuyó de $\rho = 0.2596$ (bi-encoder) a $\rho = 0.2491$ (cross-encoder), y el MRR de self-retrieval decreció de 0.496 (bi-encoder) a 0.277 (cross-encoder). Este deterioro en ambas métricas independientes refuerza la idea de que el cross-encoder, al no haber sido adaptado ni ajustado al dominio hojas de vida-vacante, no logra aplicar su entrenamiento en búsqueda web genérica.

4.4.2. Evaluación externa exploratoria con hojas de vida reales y vacantes sintéticas

Se realizó la evaluación de diez pares de vacantes sintéticas con su candidato fuente, ejecutando en paralelo en el aplicativo web los dos métodos de matching semántico experimentados: el clasificador DistilBERT ajustado y el ranking híbrido (bi-encoder SBERT + NER + siete componentes ponderados). Los resultados del ranking generado por el clasificador DistilBERT se los observa en la Tabla 22.

Tabla 22

Posición del candidato en el ranking generado por el clasificador DistilBERT

Candidato	Nombre de la vacante sintética	Posición	Top 1	Top 3	Top 5	Score
CV_JGD.pdf	Senior Data Analyst – Data Product & Growth Analytics	3	No	Sí	Sí	0.0174
CV_CDB.pdf	Lead Data Scientist – Advanced Analytics & AI	6	No	No	No	0.0078
CV_AM.pdf	Product Designer CX UX UI	1	Sí	Sí	Sí	0.1221
CV_AAMC.pdf	AI Lead – Generative AI, Search & Recommendation Systems	4	No	No	Sí	0.0099
CV_AT.pdf	Científico de Datos – Machine Learning Engineering	2	No	Sí	Sí	0.0646
CV_JT.pdf	Abogado Corporativo Especialista en Derecho Minero	10	No	No	No	0.0059
CV_NM.pdf	Analista de Relaciones Laborales y Talento Humano	5	No	No	Sí	0.0087
CV_MELA.pdf	Abogada Corporativa y Asesora de Emprendimientos	8	No	No	No	0.0069
CV_PB.docx	Especialista de Revenue Growth Management y FP&A	8	No	No	No	0.0066
CV_SGV.pdf	Data Analyst – Fraud Analytics & Data Governance	6	No	No	No	0.008

Nota. Evaluado contra silver standard.

Mientras que los resultados generados por el ranking híbrido se observa en la Tabla 23.

Tabla 23

Posición del candidato en el ranking generado por el ranking híbrido

Candidato	Nombre de la vacante sintética	Posición	Top 1	Top 3	Top 5	Score
CV_JGD.pdf	Senior Data Analyst – Data Product & Growth Analytics	6	No	No	No	0.494
CV_CDB.pdf	Lead Data Scientist – Advanced Analytics & AI	1	Sí	Sí	Sí	0.587
CV_AM.pdf	Product Designer CX / UX / UI	1	Sí	Sí	Sí	0.62
CV_AAMC.pdf	AI Lead – Generative AI, Search & Recommendation Systems	2	No	Sí	Sí	0.492
CV_AT.pdf	Científico de Datos – Machine Learning Engineering	1	Sí	Sí	Sí	0.786
CV_JT.pdf	Abogado Corporativo Especialista en Derecho Minero	1	Sí	Sí	Sí	0.719
CV_NM.pdf	Analista de Relaciones Laborales y Talento Humano	1	Sí	Sí	Sí	0.607
CV_MELA.pdf	Abogada Corporativa y Asesora de Emprendimientos	1	Sí	Sí	Sí	0.698

Candidato	Nombre de la vacante sintética	Posición	Top 1	Top 3	Top 5	Score
CV_PB.docx	Especialista de Revenue Growth Management y FP&A	1	Sí	Sí	Sí	0.804
CV_SGV.pdf	Data Analyst – Fraud Analytics & Data Governance	2	No	Sí	Sí	0.534

Los resultados con DistilBERT confirmaron la tesis previamente mencionada: el candidato ideal ocupó el primer lugar en un solo caso (CV_AM.pdf) y llegó a ubicarse en el último lugar en otro (CV_JT.pdf). Además, los puntajes brutos del clasificador fueron más bajos y menos dispersos que los del ranking híbrido, lo que dificulta su interpretación y reduce el nivel de confianza. En contraste, con el ranking híbrido, el candidato ideal alcanzó la primera posición en ocho de los diez casos evaluados, y en los dos restantes (CV_JMG.pdf y CV_SGV.pdf) quedó fuera del Top 1 pero dentro del Top 3.

Tabla 24

Síntesis comparativa de aciertos en Top 1, Top 3 y Top 5 entre el clasificador DistilBERT y el ranking híbrido, sobre los diez pares vacante-candidato ideal evaluados.

Métrica de acierto	DistilBERT	Ranking híbrido
Candidato ideal en Top 1	1/10	8/10
Candidato ideal en Top 3	3/10	9/10
Candidato ideal en Top 5	5/10	9/10

En la Tabla 24 se observa que la prueba independiente confirma la diferencia entre ambos métodos: el ranking híbrido ubicó al candidato ideal en el Top 1 en 8 de 10 casos, frente a sólo 1 de 10 con DistilBERT. En el Top 3, la ventaja fue de 90% frente a 30%, y en el Top 5, de 90% frente a 50%. Esta brecha respalda dos hipótesis: primero, que el alto F1 de DistilBERT en la fase comparativa de los modelos estuvo inflado por la dependencia del etiquetado heurístico; segundo, que un clasificador entrenado sobre el corpus general no generaliza bien a hojas de vida reales, mientras que el enfoque de similitud semántica combinado con reglas explícitas ofrece mayor consistencia.

4.4.3. Evaluación externa exploratoria del ranking con CareerCorpus

Con respecto al diseño experimental de validación mediante correlación de Spearman con CareerCorpus se analizó la sensibilidad del ranking frente a cambios de redacción y se realizó 18 ejecuciones, correspondientes a tres variantes para cada una de las seis categorías. Estas ejecuciones no representan 18 vacantes independientes, debido a que las tres variantes de cada categoría comparten el mismo perfil ESCO.

Tabla 25

Métricas obtenidas con el ranking de FreeATS sobre el conjunto de datos de CareerCorpus

Categoría	N	P@1	P@3	P@5	P@10
Teacher	50	0,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	0,2000 ± 0,0000
Finance	50	0,0000 ± 0,0000	0,0000 ± 0,0000	0,2000 ± 0,0000	0,2000 ± 0,0000
Apparel	50	0,0000 ± 0,0000	0,0000 ± 0,0000	0,2000 ± 0,0000	0,5000 ± 0,1414
Accountant	51	0,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000

Categoría	N	P@1	P@3	P@5	P@10
Banking	50	0,0000 ± 0,0000	0,0000 ± 0,0000	0,2000 ± 0,0000	0,2000 ± 0,0000
Research Assistant	51	0,0000 ± 0,0000	0,0000 ± 0,0000	0,1333 ± 0,0943	0,3000 ± 0,0000
Promedio macro	30 2	0,0000 ± 0,0000	0,0000 ± 0,0000	0,1222 ± 0,0896	0,2333 ± 0,1491

En la Tabla 25 se observa que el ranking final de FreeATS obtuvo valores de Precision@1, Precision@3, Precision@5 y Precision@10 de 0,0000, 0,0000, 0,1222 y 0,2333, respectivamente. En términos generales, los resultados no muestran una concordancia global consistente entre FreeATS y el orden humano. Por el contrario, evidencian un comportamiento heterogéneo según la categoría y una recuperación limitada de los candidatos mejor puntuados. En consecuencia, estos resultados tampoco permiten demostrar eficacia en contratación, equivalencia con un reclutador ni generalización a PYMEs ecuatorianas.

4.5.Resultados funcionales de la aplicación web

Para el desarrollo de esta sección se implementó una interfaz web local que sea capaz de funcionar sin infraestructura en la nube, de forma que sea capaz de ejecutarse en condiciones de escasos recursos computacionales y que no requiera del usuario final ningún conocimiento técnico. Las funcionalidades de gestión de las vacantes, seguimiento de las cargas e informes constituyen el soporte operativo necesario para que las tres capacidades requeridas: carga de hojas de vida, consulta de habilidades clave y visualización del ranking; puedan ser usadas en un flujo real de selección de personal.

Se implementó en un ordenador portátil con las siguientes características: 8 GB mínimos de RAM, sin GPU dedicada, y sin conexión alguna a internet en el momento de su

ejecución, confirmando con ello las condiciones de escasos recursos computacionales, así como la independencia de la conectividad, ejecutándose esta última de forma real. A continuación se muestran el flujo completo (carga de hojas de vida, consulta de características claves, ranking de adecuación candidato-perfil) mediante imágenes del sistema en ejecución.

Durante la prueba se trabajó con una vacante de Científico de Datos y se comprobó el flujo completo con archivos PDF y DOCX. La aplicación permitió registrar requisitos, consultar candidatos ya procesados, seguir el estado de nuevas cargas y generar un ranking con doce registros. Cada resultado presentó un puntaje, un detalle de fortalezas y una lista de requisitos que no aparecieron en la hoja de vida.

En la Tabla 26 se observan las funcionalidades y los resultados observados.

Tabla 26

Funcionalidades de la aplicación FreeATS

Funcionalidad	Resultado observado
Gestión de vacantes	Vacante creada y disponible para edición
Carga de hojas de vida	Archivo DOCX aceptado y procesado correctamente
Seguimiento de tareas	Estado actualizado hasta mostrar **Listo**
Listado de candidatos	Búsqueda por nombre y filtro por estado disponibles
Ranking	Doce candidatos ordenados para la vacante seleccionada
Explicación individual	Puntajes parciales y requisitos no encontrados visibles

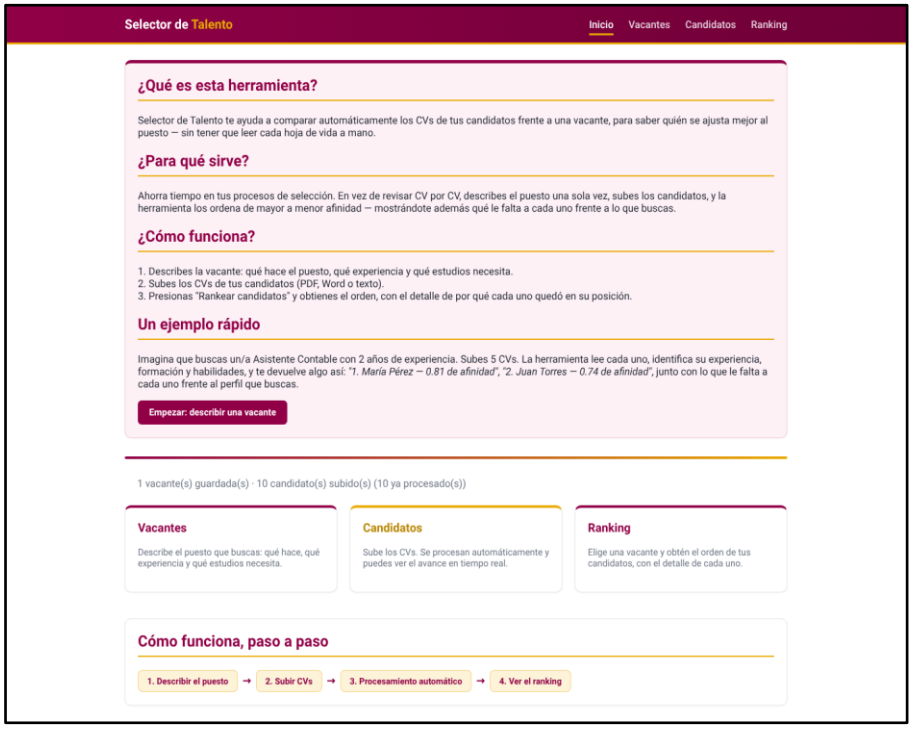
Funcionalidad	Resultado observado
Informe	Opción de descarga disponible desde la pantalla de ranking

4.5.1. Inicio y navegación

Como se observa en la Figura 6, la página de inicio reúne los accesos principales sin sobrecargar al usuario. Desde la barra superior se puede entrar a Inicio, Vacantes, Candidatos y Ranking. También muestra un resumen con el número de vacantes guardadas y candidatos procesados. La sección central presenta el recorrido de uso en cuatro acciones claras: describir el puesto, subir hojas de vida, esperar su procesamiento y consultar el ranking. Esta organización hace que una persona pueda ubicarse rápidamente, incluso si es su primera vez en FreeATS.

Figura 6

Página de 'Inicio' de FreeATS y accesos a vacantes, candidatos y ranking.

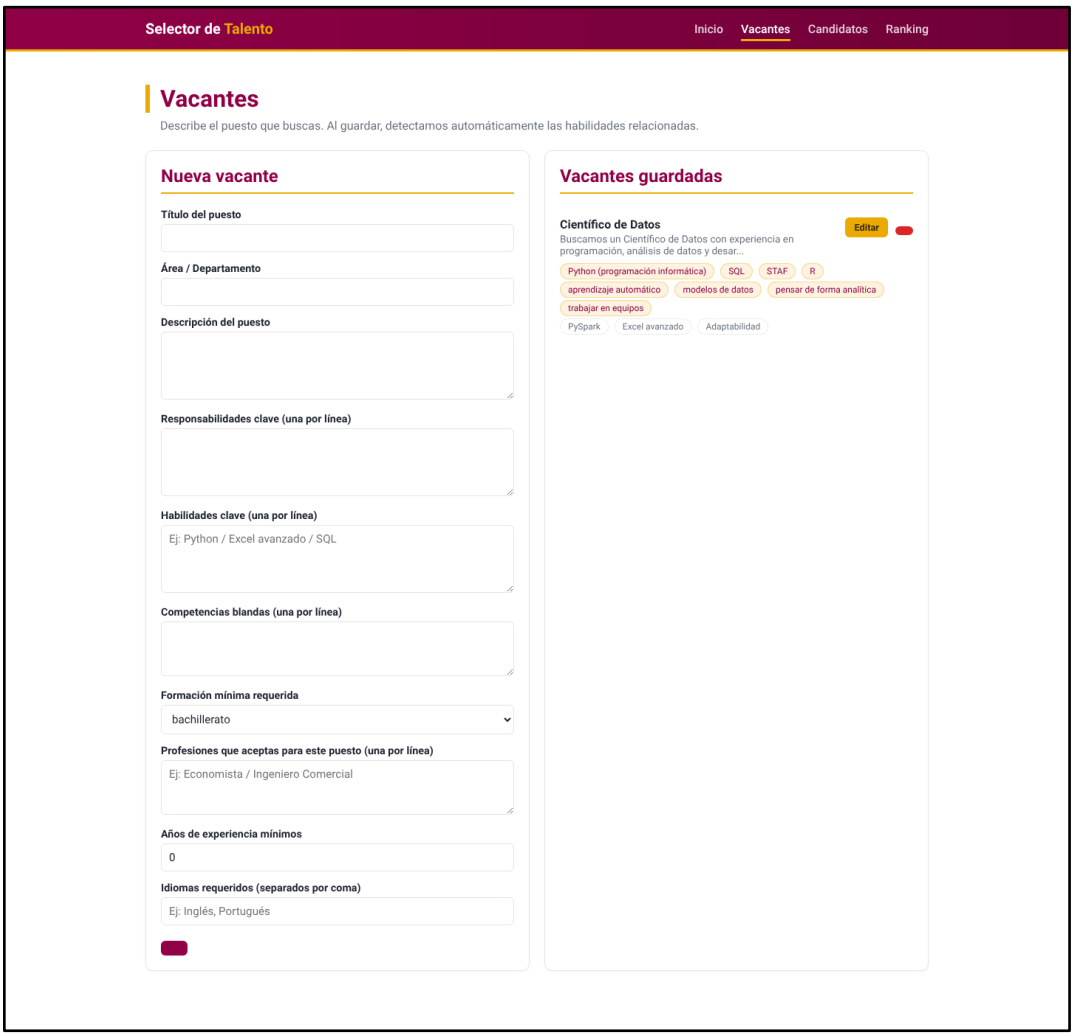


4.5.2. Gestión de vacantes

Como se observa en la Figura 7, la pantalla de vacantes permite registrar la información necesaria para evaluar candidatos: título del puesto, área, descripción, responsabilidades, habilidades, formación, profesiones aceptadas, experiencia e idiomas. En la prueba se utilizó la vacante Científico de Datos. Después de guardarla, el perfil quedó visible en la misma pantalla junto con sus habilidades principales. La vacante también quedó disponible para edición o eliminación, lo que facilita corregir requisitos antes de generar un nuevo ranking.

Figura 7

Formulario de creación y resumen de la vacante Científico de Datos.

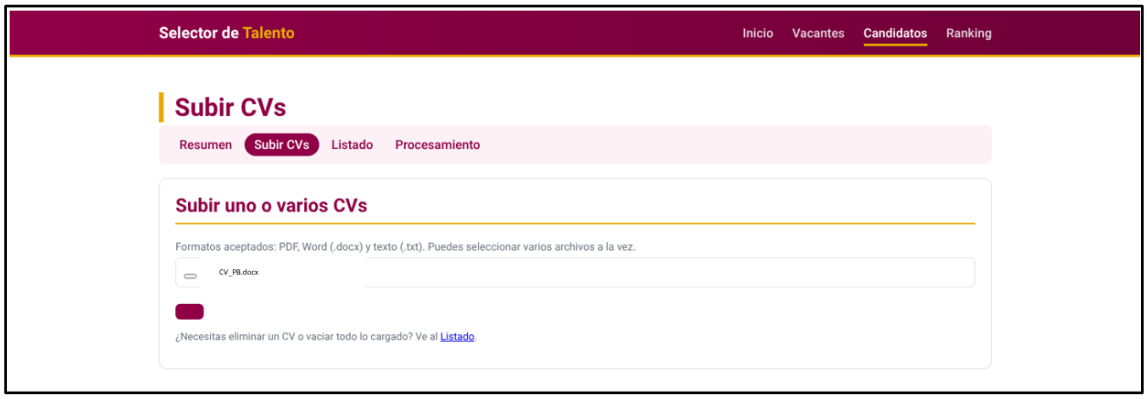


4.5.3. Carga y consulta de candidatos

Como se observa en la Figura 8, FreeATS admite hojas de vida en PDF, DOCX y TXT. Para esta comprobación se utilizó **CV_PB.docx**, evitando el archivo de texto empleado en la prueba anterior. La selección del documento se realizó desde la pantalla **Subir CVs**. El formulario mostró el nombre del archivo antes de enviarlo y mantuvo visibles los formatos permitidos. La carga terminó correctamente y creó un nuevo registro de candidato.

Figura 8

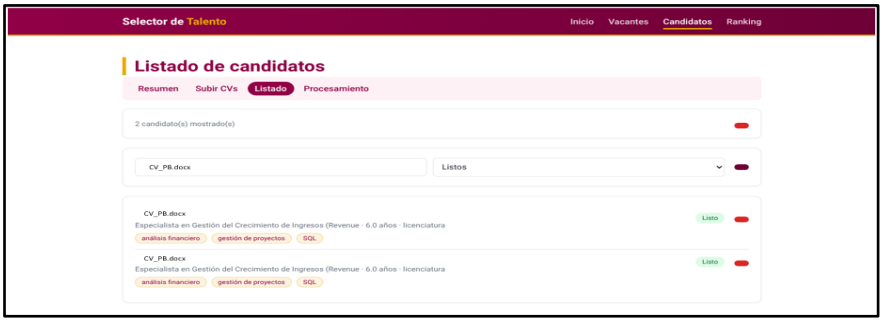
Carga de CV_PB.docx desde la sección de candidatos.



Como se observa en la Figura 9, el listado permitió buscar el archivo por nombre y limitar los resultados a candidatos listos. Esta vista resulta útil cuando existen muchos documentos, porque evita recorrer manualmente toda la lista. También permite eliminar un candidato concreto si se cargó por error o necesita ser reemplazado.

Figura 9

Resultado de la búsqueda de la hoja de vida DOCX después de completar su procesamiento.

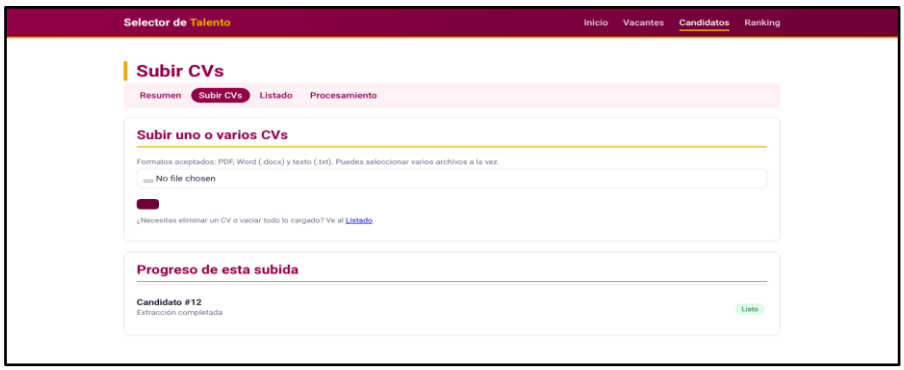


4.5.4. Seguimiento de cargas

Después de subir una hoja de vida, FreeATS muestra el avance de esa carga en la misma pantalla. En la prueba, el registro pasó al estado **Listo** sin necesidad de actualizar manualmente la página, como se observa en la Figura 10.

Figura 10

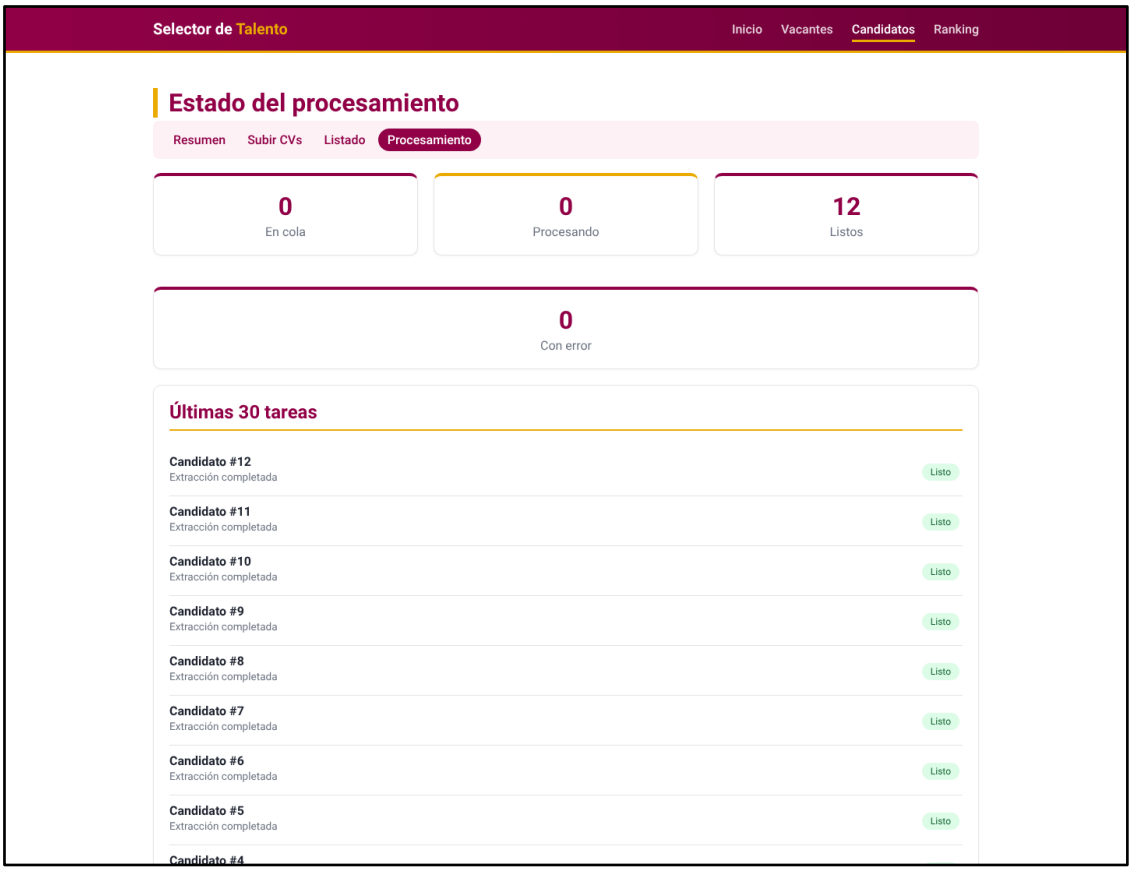
Confirmación de que la hoja de vida DOCX quedó listo para ser utilizado.



La vista general de procesamiento resume cuántos candidatos están en cola, en ejecución, listos o con error, como se observa en la Figura 11. Al finalizar la prueba se observaron doce registros listos, ninguno en cola y ninguno con error. El historial inferior conservó las tareas recientes para facilitar su revisión.

Figura 11

Estado consolidado de las tareas después de completar la prueba.

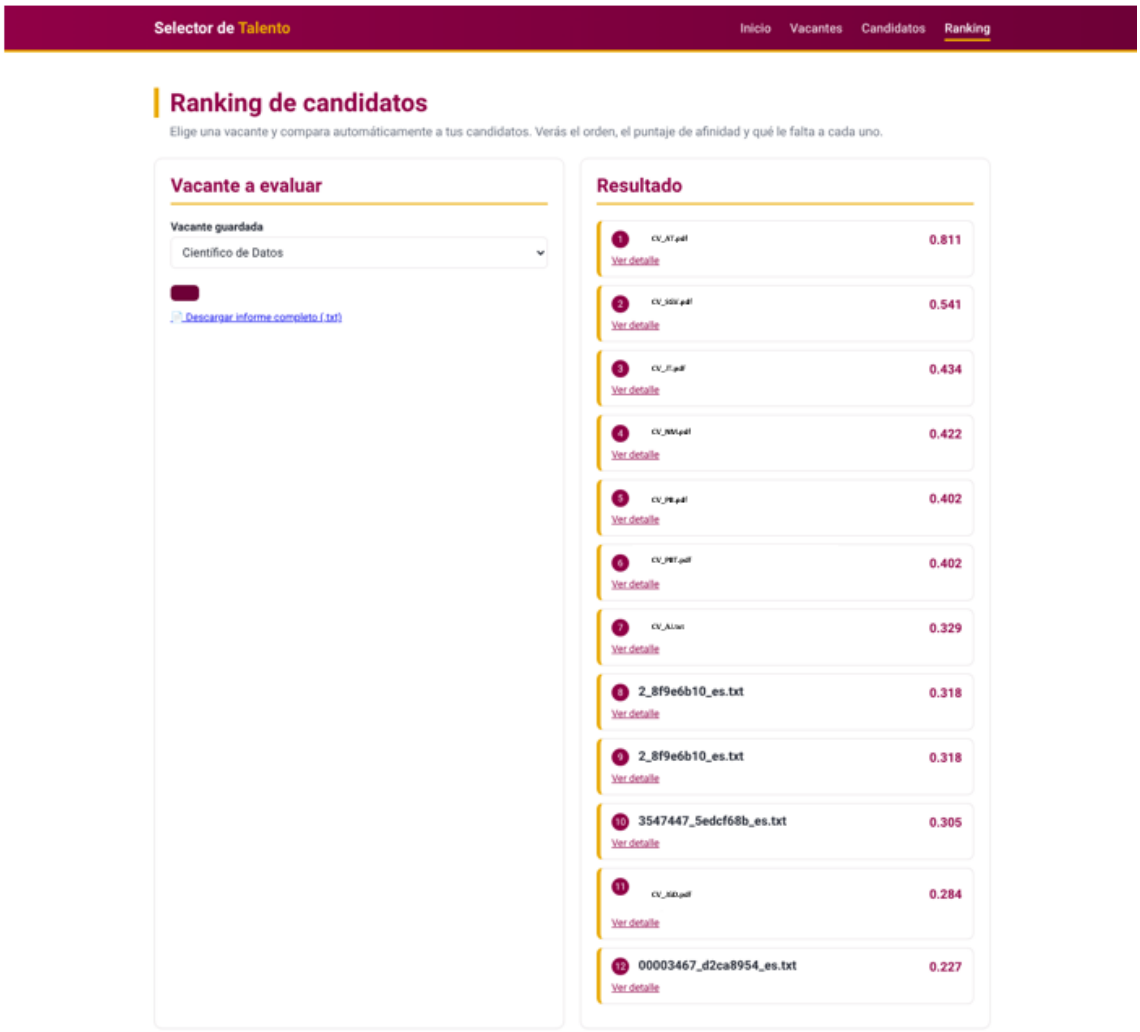


4.5.5. Ranking de candidatos

La pantalla de ranking permite elegir una vacante guardada y comparar todos los candidatos que estén listos. Para **Científico de Datos**, FreeATS ordenó doce registros y asignó un puntaje de afinidad a cada uno. El resultado se presenta como una lista fácil de recorrer. Cada tarjeta incluye posición, nombre del archivo y puntaje. El usuario puede abrir solo los perfiles que le interesan, sin perder la vista general del orden obtenido, como se observa en la Figura 12.

Figura 12

Búsqueda jerárquica de tres niveles contra el índice invertido de habilidades ESCO.



Los tres primeros resultados muestran cómo FreeATS diferencia perfiles con fortalezas distintas. Los puntajes sirven para priorizar la revisión; el detalle ayuda a entender qué conviene comprobar en una entrevista o en una lectura posterior de la hoja de vida. En la Tabla 27 se observa los resultados para el caso de estudio que se está realizando.

Tabla 27

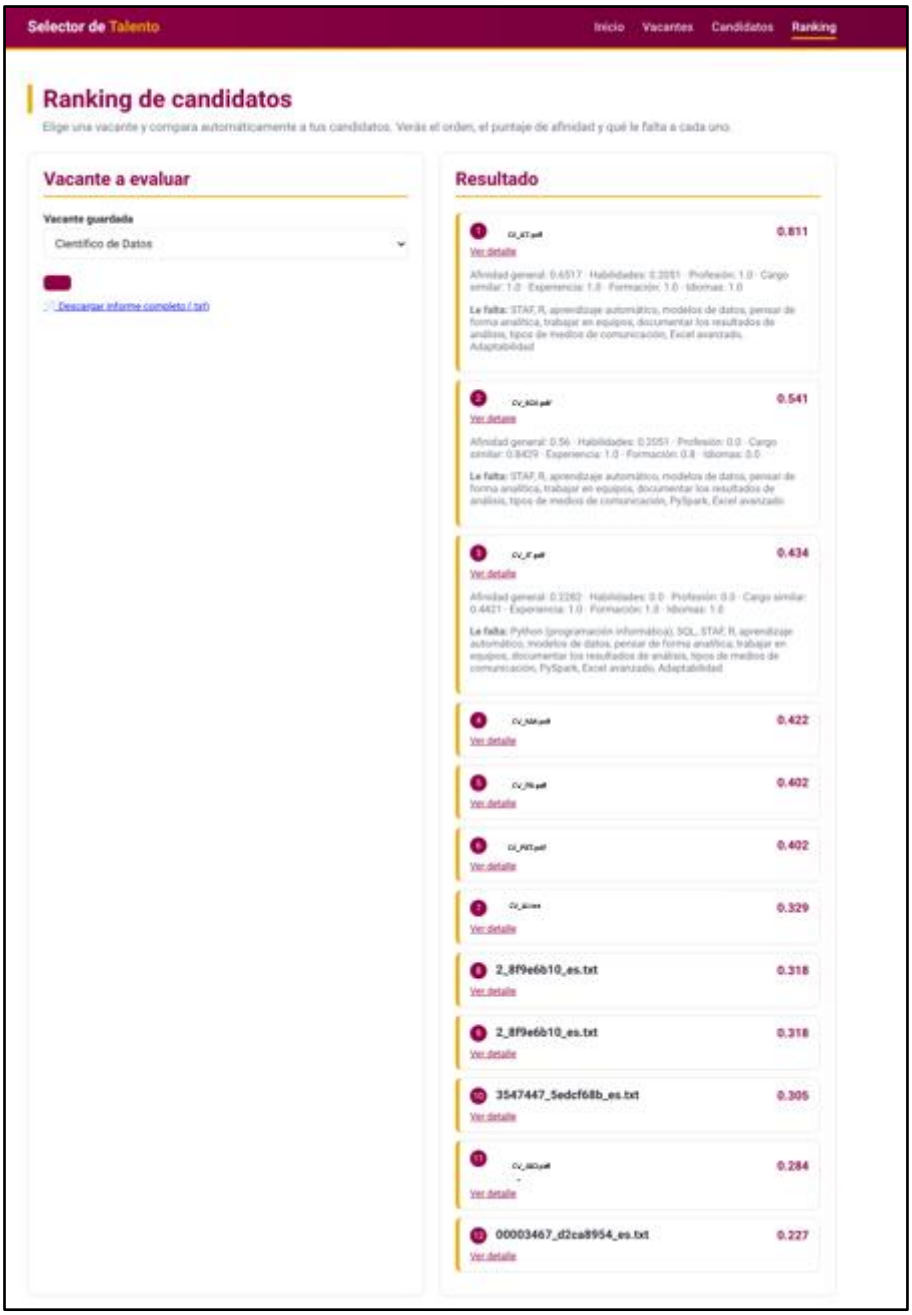
Resultados de mejores candidatos con puntaje e interpretación de resultados

Posición	Candidato	Puntaje	Lectura del resultado
1	CV_AT.pdf	**0.811**	Perfil con mejor resultado general. Destaca en profesión, cargo, experiencia, formación e idiomas. Conviene comprobar habilidades específicas como R, aprendizaje automático, PySpark y Excel avanzado.
2	CV_SGV.pdf	**0.541**	Buen resultado en experiencia y cercanía del cargo. Presenta oportunidades de verificación en formación, idiomas y varias habilidades solicitadas para la vacante.
3	CV_JT .pdf	**0.434**	Cumple bien en experiencia, formación e idiomas, pero tiene menor afinidad con el cargo y no evidencia varias habilidades técnicas requeridas.

En la Figura 13, se muestra el motivo del puntaje, por tal motivo, no debe revisarse de forma aislada. Un candidato puede quedar bien posicionado por su trayectoria general y, al mismo tiempo, requerir validación en herramientas concretas. La tarjeta ampliada convierte el ranking en una guía de revisión y no en una decisión cerrada.

Figura 13

Ejemplos de predicción para los tres candidatos mejor posicionados.



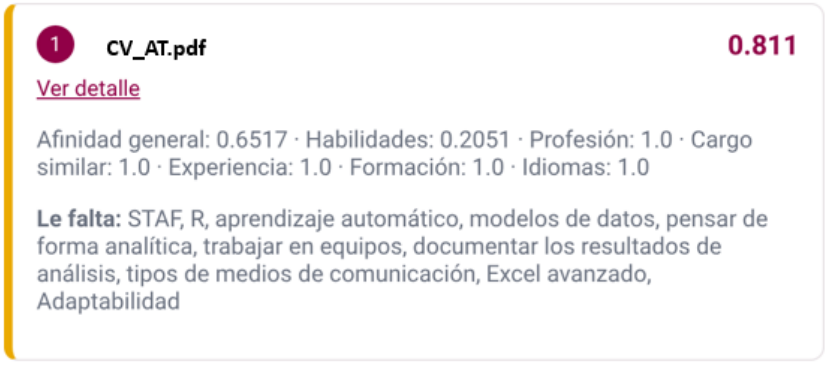
4.5.6. Detalle de resultados e informe

Cada candidato dispone de la opción **Ver detalle**. Como se observa en la Figura 14, al abrirla FreeATS presenta un resumen de los aspectos que aportaron al resultado y una lista de

requisitos que no fueron encontrados en la hoja de vida. Esta información permite preparar preguntas concretas antes de una entrevista.

Figura 14

Tarjeta del candidato mejor posicionado con puntaje, desglose y requisitos no encontrados.



La ausencia de una habilidad en esta lista no confirma que el candidato no la posea. Indica que no quedó reflejada en el documento analizado y que conviene verificarla. Esta distinción es importante para usar el resultado como apoyo durante una subsecuente revisión humana.

La misma pantalla ofrece un enlace para descargar un informe completo en formato de texto. El informe conserva el orden de candidatos y facilita compartir los resultados o incorporarlos como evidencia del análisis.

4.5.7. Resultado de la comprobación funcional

La prueba confirmó que FreeATS cubre el recorrido principal esperado desde la interfaz web:

- La vacante quedó registrada y disponible para reutilización.
- El archivo **CV_PB.docx** fue aceptado y procesado.

- El nuevo candidato apareció en el listado con estado **Listo**.
- El panel de tareas reflejó doce registros completados y ningún error.
- El ranking se generó para todos los candidatos disponibles.
- Los tres primeros resultados mostraron puntajes y explicaciones diferentes.
- La descarga del informe quedó disponible desde la vista de resultados.

En conjunto, la aplicación ofrece una experiencia ordenada y comprensible para trabajar con vacantes y grupos de candidatos. Sus pantallas mantienen visible el estado de cada tarea, permiten localizar documentos específicos y presentan el ranking con suficiente contexto para continuar la revisión de manera informada.

4.6.Comparación ranking humano vs. ranking FreeATS

Como complemento al caso de uso con la vacante **Científico de Datos**, se diseñó una comparación piloto entre el ranking generado por el sistema y el criterio de una profesional de recursos humanos ajena al equipo de desarrollo y distinta a la consultada para establecer los pesos del ranking híbrido. Con el objetivo de tener una primera aproximación empírica a la hipótesis propuesta, referida a la eficiencia frente al filtrado manual se obtuvieron los resultados de la Tabla 28.

Tabla 28

Ranking humano vs. ranking FreeATS para la vacante "Científico de Datos"

Candidato	Ranking humano	Ranking FreeATS
CV_AT.pdf	1	1
CV_SGV.pdf	4	2
CV_JT .pdf	3	3
CV_NM.pdf	5	4
CV_PB.pdf	2	5
CV_PBES.pdf	6	6
CV_AJ.pdf	7	7
98f9e6b10.txt	10	8
98f9e6b10_es.txt	11	9
3547447_5edcf68b.txt	8	10
CV_JGD.pdf	9	11
00003467_d2ca8954_es.txt	12	12

Los cinco candidatos mejor valorados son los mismos en ambas clasificaciones, aunque el orden interno es diferente; de los tres primeros candidatos, también coinciden dos de los tres. La posición exacta coincide en 5 de 12 candidatos. La evaluadora reporta que esta tarea le tomó alrededor de una hora para hacer la revisión manual de las 12 hojas de vida,

mientras que el tiempo de procesamiento del sistema es de aproximadamente un minuto para llevar a cabo la misma tarea.

4.7.Contraste con la literatura revisada

La comparación con la literatura permite ubicar el desempeño de FreeATS frente a enfoques de distinta complejidad. En extracción de información, el F1 de 0.9217 se encuentra en un rango cercano al reportado por Kinger et al. (2024) y Li et al. (2025), aunque las cifras provienen de corpus, idiomas y esquemas de anotación diferentes. Se debe considerar que la comparación debe interpretarse con cautela, dado que el F1 de FreeATS es verificada contra un silver-standard automático y no contra anotación humana independiente. Sin embargo, aun con esa salvedad, el sistema muestra que se puede obtener un desempeño competitivo con una arquitectura liviana (FreeATS alcanza una cifra del mismo orden de magnitud con un modelo NER entrenado desde un tokenizador en blanco, sin arquitecturas de visión por computadora ni modelos generativos de gran escala en la etapa de inferencia).

En matching y ranking, los indicadores operativos reportados por Kiruthiga Devi et al. (2025) no son directamente comparables con Precision@K, F1 o Spearman. Aun así, su trabajo permite contrastar una preocupación común: el equilibrio entre desempeño, dependencia de servicios externos y viabilidad para PYMEs. FreeATS prioriza modelos ejecutables en CPU y datos abiertos, aunque su correlación de ranking todavía se mantiene por debajo del acuerdo humano y del estado del arte reportado por Chowdhury et al. (2026).

En conjunto, FreeATS se acerca a enfoques de mayor escala en la extracción de información, mientras que el emparejamiento y el ranking todavía requieren una validación más amplia y una mejor calibración.

4.8.Síntesis de cifras clave

Como cierre del análisis, en la Tabla 29 se presenta un resumen de las cifras clave presentadas durante la evaluación.

Tabla 29

Síntesis de cifras clave del capítulo 4

Indicador	Valor
F1 global del módulo NER	0.9217
F1 de SKILL / JOB_TITLE / ORG	0.9717 / 0.8343 / 0.6056
Score de TF-IDF	0.2735
Score de SBERT	0.7256
Score bruto de TinyBERT *zero-shot*	0.9670, inflado por anisotropía
F1 de DistilBERT ajustado	0.9412 sobre 121 pares de prueba
Mejor P@5 promedio, SBERT e híbrido	0.800
P@5 de TinyBERT *zero-shot*	0.333
Spearman ρ con Vanetik	0.015
Spearman ρ con CareerCorpus	0.232
Techo humano reportado en CareerCorpus	0.59
Estado del arte reportado en CareerCorpus	0.83–0.86
Cobertura sobre Voto Informado 2026	62 % con al menos una entidad

CAPÍTULO 5

5. Conclusiones y recomendaciones

5.1. Cierre por objetivo específico

Los seis objetivos específicos cuentan con resultados verificables. Los módulos de extracción, normalización y comparación semántica están implementados; el ranking dispone de evidencia exploratoria que todavía debe ampliarse; y la interfaz web ya permite ejecutar el recorrido funcional previsto. En la Tabla 30, se expone el estado de los objetivos específicos planteados.

Tabla 30

Estado de los objetivos específicos

Objetivo	Estado	Evidencia principal
OE1 - Categorías de extracción y ESCO	Implementado	Esquema SKILL/JOB_TITLE/ORG definido y taxonomía ESCO integrada
OE2 - Limpieza y normalización	Implementado	Corpus final de 2.713 hojas de vida procesadas
OE3 - Comparación semántica	Implementado	TF-IDF, SBERT y TinyBERT implementados y comparados
OE4 - Ranking y validación	Implementado técnicamente; evaluación empírica exploratoria	Spearman, Pearson y Precision@1/@3/@5/@10 con CareerCorpus, resultados por categoría y estabilidad ante tres redacciones sintéticas
OE5 - Evaluación de desempeño	Completo para extracción; en consolidación para ranking	F1 de NER = 0.9217 y F1 de DistilBERT = 0.9412
OE6 - Interfaz web local	Completo a nivel funcional	Gestión de vacantes y candidatos, seguimiento de cargas, ranking explicable e informe descargable

5.2.Conclusiones generales

El desarrollo de FreeATS confirma la viabilidad técnica de construir un pipeline de procesamiento de lenguaje natural en español para extraer y comparar información de hojas de vida sin depender de una GPU dedicada. Este resultado evidencia el funcionamiento de la solución prototipo en condiciones limitadas (simulando las condiciones de PYMEs ecuatorianas), mas no todavía en procesos de selección profesionales.

El módulo NER alcanzó un F1 global de 0,9217, superando el umbral objetivo del proyecto establecido en 0.80. La cifra indica que el modelo reproduce con alta fidelidad el esquema de anotación silver-standard generado con ESCO, pero no estima su concordancia con anotadores humanos, por lo que conviene contrastar con una muestra gold-standard elaborada por personas. Aun así, el desempeño fue variado: SKILL alcanzó 0,9717, JOB_TITLE 0,8343 y ORG 0,6056, de modo que la extracción de habilidades es el resultado más sólido y la de organizaciones la que requiere mayor trabajo.

En la comparación semántica, SBERT superó con claridad a TF-IDF, con un score de 0,7256 frente a 0,2735. El comportamiento de TinyBERT sin fine-tuning dejó un aprendizaje importante: su score bruto de 0,9670 parecía superior, pero su Precision@5 fue de 0,333. Estos resultados indican que solo con la magnitud no es suficiente para evaluar un ranking y que debe acompañarse de métricas que midan el orden real de recuperación. Cabe precisar que SBERT y el ranking híbrido empataron en Precision@5 promedio (0,800), por lo que los datos disponibles no son concluyentes para afirmar que los componentes adicionales del híbrido mejoren de forma general al baseline semántico.

En el caso de DistilBERT ajustado, obtuvo una métrica F1 de 0,9412 en clasificación binaria. Aun así, el 94 % de los pares de entrenamiento provino de etiquetas heurísticas y el

F1 descendió a 0,75 en el subconjunto validado por personas. Por tanto, si bien el valor resultante es prometedor, no debe presentarse como una métrica de desempeño en condiciones reales.

El ranking híbrido mostró una correlación de $\rho = 0{,}2596$ con CareerCorpus, positiva y estadísticamente significativa en cuatro de las seis categorías evaluadas, aunque todavía distante del acuerdo entre anotadores humanos y de los resultados reportados por modelos de mayor escala. Este resultado confirma que el enfoque utilizado logra reconocer parte de la señal de adecuación; además, logra señalar los puntos a mejorar: ampliar las vacantes evaluadas, fortalecer el ground truth y optimizar los pesos del ranking.

La interfaz web completa el paso desde un pipeline experimental hacia una herramienta utilizable. Las pruebas realizadas confirmaron la creación de vacantes, la carga y consulta de candidatos, el seguimiento de procesamiento, el ranking con desglose de puntajes y la descarga de informes. Cabe resaltar que el aplicativo no es un sustitutivo del criterio personal de recursos humanos; organiza la información, agiliza la revisión de las hojas de vida y hace visibles los elementos que conviene revisar antes de tomar una decisión.

Con respecto a la eficiencia frente al filtrado manual planteada en la hipótesis, el piloto de la sección 4.7 registró una diferencia clara: aproximadamente una hora de revisión manual frente a un minuto de procesamiento automático, junto con una coincidencia completa en los cinco primeros candidatos. Al tratarse de una sola vacante y una sola evaluadora, el resultado constituye un indicio de ahorro potencial, pero no una estimación general de eficiencia. En conjunto, la hipótesis del proyecto se considera apoyada de forma parcial: la evidencia respalda la viabilidad técnica del prototipo y muestra señales

exploratorias de utilidad, mientras que su eficacia en procesos reales de selección aún está por demostrarse.

5.3.Contribuciones del trabajo

5.3.1. Contribuciones prácticas

La principal contribución práctica es un pipeline reproducible que traduce, limpia, estructura y normaliza hojas de vida en español sobre un corpus consolidado de 2.713 documentos originalmente en idioma inglés. A ello se suma un modelo NER específico del dominio, entrenado sin una campaña previa de anotación manual, y una comparación trazable entre TF-IDF, SBERT, TinyBERT y DistilBERT que permite valorar de manera concreta el equilibrio entre calidad, costo computacional y facilidad de despliegue.

La aplicación web aporta una capa operativa sobre estos componentes. Su interfaz permite que una persona gestione vacantes y candidatos sin interactuar con componentes técnicos o archivos intermedios, y presenta el ranking junto con explicaciones parciales y requisitos no detectados. Esta combinación facilita que el resultado se utilice como apoyo a la revisión, no como una decisión automática cerrada.

5.3.2. Contribuciones metodológicas

El trabajo documenta la anisotropía observada en TinyBERT zero-shot y muestra por qué un score alto puede coexistir con un ranking pobre. También distingue entre un ranking organizado por candidato y otro organizado por vacante, una diferencia que explica el resultado casi nulo obtenido con Vanetik y evita reutilizar ese dataset de forma incorrecta. Finalmente, el desglose de métricas según la procedencia de las etiquetas hace visible el

riesgo de circularidad cuando predominan ejemplos heurísticos sobre ejemplos revisados por personas.

5.4.Recomendaciones de trabajo futuro

5.4.1. Mejorar la clasificación de cargos con ESCO

La taxonomía de ocupaciones de ESCO puede sustituir el heurístico léxico utilizado para los cargos. Una clasificación jerárquica ayudaría a diferenciar un puesto técnico de otro de gestión que solo menciona herramientas técnicas, y reduciría casos ambiguos como el perfil de PM/DevOps observado durante las pruebas.

5.4.2. Incorporar más validación humana

Se recomienda aumentar el número de evaluación humana del módulo NER mediante la selección aleatoria de una muestra de 50 hojas de vida del corpus silver-standard, realizar las anotaciones manuales y calcular el F1 y el coeficiente kappa de Cohen, el cual captura el nivel de acuerdo entre ambas fuentes de anotación (Rebholz-Schuhmann et al., 2011).

5.4.3. Consolidar estadísticamente el ranking

La validación debe repetirse con un conjunto más amplio y diverso de vacantes, acompañado de pruebas de significancia para las diferencias de Precision@K. Los pesos del ranking híbrido también deberían ajustarse sobre datos de validación, en lugar de depender de una calibración manual. Este paso permitiría determinar si ESCO y los demás componentes añaden una mejora consistente sobre SBERT puro.

5.4.4. Evaluar la aplicación con más usuarios

Aunque el recorrido funcional está implementado, queda pendiente una evaluación de usabilidad con personal de recursos humanos de PYMEs ecuatorianas. Estas pruebas deberían observar tiempos de tarea, comprensión de las explicaciones, errores de uso y confianza en el ranking. También sería útil revisar si el informe descargable contiene la información necesaria para documentar una preselección sin fomentar decisiones automáticas.

5.4.5. Ampliar el corpus ecuatoriano

La recopilación de hojas de vida redactadas originalmente en español y procedentes del mercado laboral ecuatoriano permitiría comprobar si los hallazgos se mantienen fuera de un corpus traducido desde el inglés. Además de mejorar la validez externa, este material ayudaría a identificar variaciones locales de cargos, formación y vocabulario profesional que no siempre aparecen en los datasets internacionales.

5.5.Reflexión sobre el enfoque de bajo recurso

Los resultados ofrecen una lectura favorable, aunque matizada, del enfoque de bajo recurso. La extracción de entidades y SBERT muestran que es posible obtener resultados útiles en CPU y con datos abiertos. El ranking, en cambio, recuerda que reducir el costo computacional no elimina la necesidad de buenos datos de validación, etiquetas humanas y una calibración cuidadosa. La brecha frente a modelos de mayor escala no debe ocultarse, pero tampoco invalida la decisión de priorizar una solución que pueda instalarse y utilizarse en condiciones reales de una PYME.

La arquitectura separa el entrenamiento, que concentra el mayor costo, de la inferencia cotidiana. Los modelos ya entrenados pueden distribuirse junto con la aplicación, de modo que el usuario final sólo asume el costo menor de procesar una vacante, consultar el

índice y calcular el ranking. Esta separación es clave para conservar la viabilidad operativa sin renunciar a futuras mejoras del modelo.

Bibliografía

Abhishek, K. L., Niranjanamurthy, M., Aric, S., Ansarullah, S. I., Sinha, A., Tejani, G., & Shah, M. A. (2025). Developing an intelligent resume screening tool with AI-driven analysis and recommendation features. *Applied AI Letters*, *6*(2), e116. https://doi.org/10.1002/ail2.116

Anthropic. (2025). Claude Haiku 4.5 [Documentación técnica]. Recuperado el 10 de julio de 2026, https://www.anthropic.com/claude/haiku

Biea, E. A., Dinu, E., Bunica, A., & Jerdea, L. (2024). Recruitment in SMEs: The role of managerial practices, technology and innovation. *European Business Review*, 36(3), 361–391. https://doi.org/10.1108/EBR-05-2023-0162

Butt, S., Ceballos, H. G., & Madera-Espíndola, D. P. (2026). Tec-Habilidad: Skill classification for bridging education and employment. En L. Martínez-Villaseñor, R. A. Vázquez, & G. Ochoa-Ruiz (Eds.), Advances in Soft Computing. MICAI 2025 (*Lecture Notes in Computer Science*, vol. 16221). Springer. https://doi.org/10.1007/978-3-032-09037-9_25

Chowdhury, M. S., Chowdhury, A. F., Banu, A., & Hossain, R. (2026). CareerCorpus: A comprehensive dataset of annotated resumes. *Data in Brief, 65*, 112567. https://doi.org/10.1016/j.dib.2026.112567

Comisión Europea. (2024). *ESCO dataset – v1.2.0* [Conjunto de datos]. Dirección General de Empleo, Asuntos Sociales e Inclusión. https://esco.ec.europa.eu/en/use-esco/download

Deshmukh, A., & Raut, A. (2024). Applying BERT-based NLP for automated resume screening and candidate ranking. *Annals of Data Science*. https://doi.org/10.1007/s40745-024-00524-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En Proceedings of NAACL-HLT 2019 (pp. 4171–4186). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/N19-1423

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. En Proceedings of EMNLP-IJCNLP 2019 (pp. 55–65). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/D19-1006

Google Cloud. (s. f.). Cloud Translation API. Google. Recuperado el 10 de julio de 2026 de https://cloud.google.com/translate

Helsinki-NLP. (2020). opus-mt-en-es [Modelo de traducción automática]. *Hugging Face*. Recuperado el 10 de Julio de https://huggingface.co/Helsinki-NLP/opus-mt-en-es

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy*: Industrial-strength Natural Language Processing in Python [Software]. *Explosion*. https://doi.org/10.5281/zenodo.1212303

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446. https://doi.org/10.1145/582415.582418

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. *En Findings of EMNLP 2020* (pp. 4163–4174). https://doi.org/10.18653/v1/2020.findings-emnlp.372

Jiechieu, K. F. F., & Tsopze, N. (2021). Skills prediction based on multi-label resume classification using CNN with model predictions explanation. *Neural*

Computing and Applications, 33, 5069–5087. https://doi.org/10.1007/s00521-020-05302-x

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Kinger, S., Kinger, D., Thakkar, S., & Bhake, D. (2024). Towards smarter hiring: Resume parsing and ranking with YOLOv5 and DistilBERT. *Multimedia Tools and Applications, 83*(35), 82069–82087. https://doi.org/10.1007/s11042-024-18778-9

Kiruthiga Devi, M., Moraes, H. A., Arkeshwar, S., & Nikhil Kishore, M. (2025). System (ATS) for revolutionizing recruitment: AI-powered application tracking system for small enterprises. En Proceedings of the International Conference on Intelligent Systems and Digital Transformation (ICISD 2025) (pp. 489–502). Atlantis Press.

Koman, G., Boršoš, P., & Kubina, M. (2024). The possibilities of using artificial intelligence as a key technology in the current employee recruitment process. *Administrative Sciences, 14*(7), 157. https://doi.org/10.3390/admsci14070157

Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artificial Intelligence Review, 57*(10), 260. https://doi.org/10.1007/s10462-024-10888-y

Lan, F. (2022). Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method. *Advances in Multimedia, 2022*, 7923262. https://doi.org/10.1155/2022/7923262

Li, H., Tang, X., Liu, Q., Liu, B., & Zhao, S. (2025). Enhancing intelligent recruitment with generative pretrained transformer and hierarchical graph

neural networks. *Journal of Organizational and End User Computing, 37*(1), 1–24.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. En Proceedings of the 7th International Conference on Learning Representations (ICLR 2019). https://openreview.net/forum?id=Bkg6RiCqY7

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Marliyas, A., Ummah, M. A. C. S., & Gunapalan, S. (2025). The transformation of talent acquisition through artificial intelligence in the context of Industry 4.0: A systematic literature review. *Journal of Business Research and Insights, 11*(2), 1–20.

Moghe, N., Fazla, A., Amrhein, C., Kocmi, T., Steedman, M., Birch, A., Sennrich, R., & Guillou, L. (2025). Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics, 51*(1), 73–137. https://doi.org/10.1162/coli_a_00537

Nikolaou, I. (2021). What is the role of technology in recruitment and selection? *The Spanish Journal of Psychology, 24*, e6. https://doi.org/10.1017/SJP.2021.6

OpenAI. (2024, 18 de julio). GPT-4o mini: Advancing cost-efficient intelligence. OpenAI – GPT-4o mini

Otani, N., Bhutani, N., & Hruschka, E. (2025, April). Natural language processing for human resources: A survey. En Proceedings of NAACL-HLT 2025 (Industry Track) (pp. 583–597). https://arxiv.org/abs/2410.16498

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. En Proceedings of the 40th Annual Meeting of the ACL (pp. 311–318). *Association for Computational Linguistics.* https://doi.org/10.3115/1073083.1073135

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. En *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)* (pp. 1310–1318). PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., & Batet, M. (2022). The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics, 48*(4), 1053–1101. https://doi.org/10.1162/coli_a_00458

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. En Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 392–395). *Association for Computational Linguistics.* https://doi.org/10.18653/v1/W15-3049

Popović, M. (2017). chrF++: Words helping character n-grams. En Proceedings of the Second Conference on Machine Translation (pp. 612–618). *Association for Computational Linguistics.* https://doi.org/10.18653/v1/W17-4770

Ratner, A., De Sa, C., Wu, S., Selsam, D., & Ré, C. (2016). Data programming: Creating large training sets, quickly. A*dvances in Neural Information Processing Systems (NeurIPS), 29*.

Rebholz-Schuhmann, D., Jimeno Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J. B., Baker, C. J. O., Kuo, C.-J., Clematide, S., Rinaldi, F., Farkas, R., Móra, G., Hara, K., Furlong, L. I., Rautschka, M., Neves, M. L., Pascual-Montano, A., Wei, Q., Collier, N., Chowdhury, M. F. M., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J. L., van Mulligen, E., Kors, J., & Hahn, U. (2011). *Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. Journal of Biomedical Semantics, 2*(Suppl 5), S11. https://doi.org/10.1186/2041-1480-2-S5-S11

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. En Proceedings of EMNLP 2020 (pp. 2685–2702). *Association for Computational Linguistics.* https://doi.org/10.18653/v1/2020.emnlp-main.213

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. En Proceedings of EMNLP-IJCNLP 2019 (pp. 3982–3992). *Association for Computational Linguistics.* https://doi.org/10.18653/v1/D19-1410

Saatçı, M., Kaya, R., & Ünlü, R. (2024). Resume screening with natural language processing (NLP). *Alphanumeric Journal, 12*(2), 121–140. https://doi.org/10.17093/alphanumeric.1536577

Sáenz de Viteri, C., García, A., O'Rorke, J., Burbano, J. A., & López, R. (2026). Navigating the labyrinth: Entrepreneurial barriers in Ecuador and Latin America in the digital age. *Economía y Negocios, 17*(1), 75–87. https://www.redalyc.org/journal/6955/695583208004/html/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. En NeurIPS 2019 Workshop on Energy Efficient Machine Learning. https://arxiv.org/abs/1910.01108

Senger, E., Zhang, M., van der Goot, R., & Plank, B. (2024). Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. En Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024) (pp. 1–15). Association for Computational Linguistics.

Seow, W. L., Chaturvedi, I., Hogarth, A., Mao, R., & Cambria, E. (2025). A review of named entity recognition: From learning methods to modelling paradigms and tasks. *Artificial Intelligence Review, 58*, 315. https://doi.org/10.1007/s10462-025-11321-8

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *En Proceedings of AMTA 2006* (pp. 223–231). Association for Machine Translation in the Americas.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.

Sridevi, G. M., & Suganthi, S. K. (2022). AI-based suitability measurement and prediction between job description and job seeker profiles. *International Journal of Information Management Data Insights, 2*(1), 100109. https://doi.org/10.1016/j.jjimei.2022.100109

Sundberg, L., & Holmström, J. (2023). Democratizing artificial intelligence: How no-code AI can leverage machine learning operations. *Business Horizons, 66*(6), 777–788. https://doi.org/10.1016/j.bushor.2023.04.003

Trovão, H., Mamede, H. S., Trigo, P., & Santos, V. (2025). Artificial intelligence in recruitment: A multivocal review of benefits, challenges, and strategies. *Emerging Science Journal, 9*(6), 3458–3485. https://doi.org/10.28991/ESJ-2025-09-06-030

Vanetik, N., & Kogan, G. (2023). Job vacancy ranking with sentence embeddings, keywords, and named entities. *Information, 14*(8), 468. https://doi.org/10.3390/info14080468

Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021, April). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume (pp. 2203-2213).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems (Vol. 30)*. Curran Associates, Inc. https://arxiv.org/abs/1706.03762

Voorhees, E. M. (1999). The TREC-8 question answering track report. En Proceedings of the 8th Text Retrieval Conference (TREC-8) (pp. 77–82). National Institute of Standards and Technology.

Wang, X., Sanders, H. M., Liu, Y., Seang, K., Tran, B. X., Atanasov, A. G., Qiu, Y., Tang, S., Car, J., Wang, Y. X., et al. (2023). ChatGPT: Promise and challenges for deployment in low- and middle-income countries. *The Lancet Regional Health – Western Pacific, 41*, 100905. https://doi.org/10.1016/j.lanwpc.2023.100905

Wissler, L., Almashraee, M., Monett, D., & Paschke, A. (2014, junio). The gold standard in corpus annotation [Ponencia]. 5th IEEE Germany Student Conference, Passau, Germany. https://doi.org/10.13140/2.1.4316.3523

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. En Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2024). GLiNER: Generalist model for named entity recognition using bidirectional transformer. En Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 5364–5376). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.300

Zhang, C., & Wang, H. (2018). ResumeVis: A visual analytics system to discover semantic information in semi-structured resume data. ACM Transactions on Intelligent Systems and Technology, 10(1), Article 8. https://doi.org/10.1145/3230707

Zhang, M., van der Goot, R., & Plank, B. (2023, July). ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. En Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 11871–11890). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.662

Apéndice

Repositorio:

https://drive.google.com/drive/folders/1KtQYXTbESaUtKrKud_z1Bwv3JwAb7Ore