<CODE>

Observatorio de demanda laboral en América Latina

Nicolas Francisco Camacho Alarcón
Alejandro Pinzón Fajardo

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
SYSTEMS ENGINEERING PROGRAM
BOGOTÁ, D.C.
2025

<CODE>

Observatorio de demanda laboral en América Latina

Author(s):

Nicolas Francisco Camacho Alarcón
Alejandro Pinzón Fajardo

UNDERGRADUATE FINAL PROJECT REPORT PERFORMED IN ORDER TO ACCOMPLISH ONE OF THE REQUIREMENTS FOR THE SYSTEMS ENGINEERING DEGREE

Director

Ing. Luis Gabriel Moreno Sandoval

Juries of the Undergraduate Final Project

Ing. <Name of the jury >

Ing. < Name of the jury >

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
SYSTEMS ENGINEERING PROGRAM
BOGOTÁ, D.C.
<Month>,<Year>

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
SYSTEMS ENGINEERING PROGRAM

President of the Pontificia Universidad Javeriana

<Name of the President of the University>

Dean of School of Engineering

<Name of the Dean>

Head of the Systems Engineering Program

<Name of the head of the program>

Head of the Systems Engineering Department

<Name of the head of the department>

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

GRATITUDE

Write a message if you feel gratitude for someone who has supported the development of the project. Your family, your partner, your friends, your principal, teachers, etc.

CONTENT

ABSTRACT

El desajuste entre las habilidades demandadas por el mercado y la oferta formativa en Latinoamérica dificultaba decisiones de política, academia y empresa. Este proyecto abordó el problema construyendo un observatorio automatizado que recolectó avisos de empleo multi-portal y multi-país, escalable hacia ~600.000 registros. Se integraron spiders (Scrapy/Selenium con anti-detección), una base PostgreSQL con pgvector y un pipeline de extracción/normalización de habilidades (NER/regex con apoyo LLM) alineadas a ESCO. El sistema generó indicadores, consultas y visualizaciones reproducibles, entregando evidencia comparable por país, sector y tiempo para orientar currículos, formación y estrategias de talento.

I- INTRODUCCIÓN

Los mercados laborales latinoamericanos evolucionan con rapidez y publican sus vacantes en portales heterogéneos, con formatos dispares, vocabularios no estandarizados y alta volatilidad (los avisos desaparecen o cambian con frecuencia). Esta fragmentación dificulta medir, con evidencia objetiva y comparable, qué habilidades técnicas y digitales están siendo demandadas por país, sector y momento del tiempo. El proyecto **Observatorio de Demanda Laboral para América Latina** responde a ese vacío mediante un sistema automatizado que captura, estructura y analiza anuncios de empleo a escala.

La solución integra un **pipeline** modular de ocho etapas: Scraping multifuente y multipaís (Colombia, México, Argentina) con spiders robustos y medidas anti-detección; Normalización y limpieza; Extracción de habilidades combinando NER, patrones regex y apoyo LLM; Alineación a la taxonomía **ESCO**; Generación de **embeddings** multilingües (modelo E5); Reducción dimensional (UMAP); Clustering (HDBSCAN) para descubrir familias de perfiles; y Visualización y reportes. La infraestructura técnica se apoya en **Python/Scrapy**, **PostgreSQL + pgvector** y **Docker**, con registro y monitorización de extremo a extremo.

Este documento guía al lector desde el contexto y la motivación hasta los resultados y conclusiones. Presenta: I) antecedentes y trabajos relacionados; II) arquitectura del sistema y orquestación; III) adquisición y modelado de datos; IV) métodos de extracción y normalización de habilidades; V) componentes de representación y análisis; VI) evaluación y métricas; VII) hallazgos y visualizaciones; VIII) consideraciones éticas y limitaciones; y IX) conclusiones y trabajo futuro.

II- DESCRIPCIÓN GENERAL

1. Oportunidad y problema

1.1. Contexto del problema

El mercado laboral en América Latina se encontró, durante la última década, en una compleja encrucijada definida por la confluencia de dos fuerzas a menudo contrapuestas: una acelerada transformación digital y la persistencia de desafíos estructurales, como una elevada informalidad laboral y brechas de capital humano (Echeverría & Rucci, 2022). La pandemia de COVID-19 actuó como un catalizador sin precedentes, intensificando la adopción de tecnologías y, con ello, la demanda de competencias digitales, al tiempo que exponía la vulnerabilidad de los mercados de trabajo de la región (Azuara et al., 2022). Este dinamismo generó el riesgo de que la automatización y la digitalización, de no ser gestionadas estratégicamente, pudiesen exacerbar las desigualdades existentes, conduciendo a una mayor polarización y segmentación social (Echeverría & Rucci, 2022).

Para analizar este fenómeno regional de manera tangible y robusta, este proyecto seleccionó como casos de estudio a tres de las economías más grandes y digitalmente activas de habla hispana: Colombia, México y Argentina. La elección de estos países respondió a tres criterios estratégicos. Primero, su alto volumen de publicaciones de ofertas laborales en portales digitales aseguró la viabilidad de una recolección masiva de datos (web scraping), fundamental para el entrenamiento de modelos de lenguaje robustos (Aguilera & Méndez, 2018; Martínez Sánchez, 2024; Rubio Arrubla, 2024). Segundo, la existencia de estudios previos en cada país, aunque metodológicamente limitados, confirmó la pertinencia del problema y proporcionó una línea de base para la comparación (Cárdenas Rubio et al., 2015; Campos-Vázquez & Martínez Sánchez, 2024). Y tercero, su diversidad en términos de realidades económicas, territoriales y de madurez digital permitió validar que la solución desarrollada fuese portable y adaptable a los distintos contextos que caracterizan a América Latina.

El caso de Colombia sirvió como una ilustración profunda de esta dinámica. El diagnóstico nacional previo al proyecto ya indicaba que el principal cuello de botella para la inclusión digital no era la falta de infraestructura, sino la brecha de capital humano. Específicamente, el "Índice de Brecha Digital" (IBD) del Ministerio de Tecnologías de la Información y las Comunicaciones reveló que la dimensión de "Habilidades Digitales" constituía el mayor componente individual de la brecha en el país. Esta evidencia fue posteriormente corroborada y cuantificada por el análisis empírico de la demanda laboral, el cual demostró que la pandemia generó un cambio estructural y persistente en el mercado. Se encontró que, en los 18 meses poste-

riores al inicio de la crisis sanitaria, las vacantes tecnológicas aumentaron en un 50% en comparación con las no tecnológicas (Rubio Arrubla, 2024). Este cambio no fue solo cuantitativo, sino también cualitativo: se observó una marcada caída en la demanda de herramientas ofimáticas tradicionales como Excel (cuya mención en ofertas cayó del 35.8% en 2018 al 17.4% en 2023) y un surgimiento exponencial de tecnologías especializadas asociadas al desarrollo web y la gestión de datos, como bases de datos NoSQL (12.3%), el framework Django (5.5%) y la librería React (5.3%) para el año 2023 (Rubio Arrubla, 2024).

1.2. Formulación del problema

A pesar de que el contexto del problema —la creciente e insatisfecha demanda de habilidades tecnológicas— estaba claramente identificado, los métodos existentes en la región para analizarlo presentaban limitaciones metodológicas significativas que impedían una comprensión profunda y ágil del fenómeno. Los estudios de referencia en los países seleccionados, si bien valiosos para establecer tendencias macro, se basaron en enfoques de análisis léxico y reglas manuales. En Colombia, el análisis se centró en un sistema de clasificación basado en la Clasificación Internacional Uniforme de Ocupaciones (CIUO), utilizando algoritmos de emparejamiento de texto con tokenización y métricas de similitud basadas en n-gramas (Rubio Arrubla, 2024). De forma análoga, en Argentina, los estudios se concentraron en técnicas de minería de texto con análisis de frecuencias y bigramas para identificar patrones en las ofertas del sector TI (Aguilera & Méndez, 2018). En México, el enfoque combinó datos de encuestas con scraping de portales, apoyándose en el análisis de frecuencia de términos y la creación de tipologías manuales para segmentar las habilidades (Martínez Sánchez, 2024).

La limitación fundamental compartida por estos enfoques es su dependencia de la correspondencia léxica explícita, lo que los hace incapaces de capturar la riqueza semántica del lenguaje. Estos métodos no podían detectar habilidades implícitas (aquellas que se infieren del contexto de un cargo pero no se mencionan directamente), gestionar la ambigüedad del lenguaje informal o el uso de anglicismos técnicos ("Spanglish"), ni identificar clústeres de competencias emergentes que aún no forman parte de taxonomías estandarizadas. La alta variabilidad en la redacción de las ofertas laborales, la falta de estructuras normalizadas y la rápida aparición de nuevas tecnologías hacían que estos sistemas fueran metodológicamente frágiles y requirieran un constante mantenimiento manual (Echeverría & Rucci, 2022; Lukauskas et al., 2023).

En consecuencia, el problema específico que este proyecto abordó fue la ausencia de una herramienta automatizada y de extremo a extremo que, adaptada a las particularidades lingüísticas y estructurales del español latinoamericano, permitiera superar las limitaciones de los análisis léxicos tradicionales. Se identificó la necesidad de un sistema capaz de extraer,

estructurar y analizar la evolución de las habilidades tecnológicas de manera semántica, escalable y con un mayor grado de autonomía, integrando para ello técnicas avanzadas de Procesamiento de Lenguaje Natural (NLP), enriquecimiento contextual con Large Language Models (LLMs) y algoritmos de agrupamiento no supervisado.

1.3. Propuesta de solución

Para dar respuesta al problema formulado, se diseñó e implementó un observatorio de demanda laboral tecnológica basado en un pipeline modular y automatizado, un proyecto enmarcado en las áreas de Ingeniería de Sistemas y Ciencia de Datos. El sistema fue concebido como una solución de extremo a extremo que integró las etapas de recolección, procesamiento, análisis semántico y segmentación de ofertas de empleo publicadas en Colombia, México y Argentina. El objetivo fue crear una arquitectura robusta, replicable y adaptada a las complejidades del contexto latinoamericano, superando las limitaciones de los enfoques puramente léxicos o manuales.

La solución se materializó a través de un sistema compuesto por módulos secuenciales y cohesivos. El primer módulo consistió en un motor de adquisición de datos que, mediante técnicas de web scraping, extrajo de forma sistemática y ética decenas de miles de ofertas laborales de portales de empleo clave en la región. El núcleo del sistema fue su arquitectura de extracción dual, compuesta por dos pipelines paralelos:

Pipeline A (Tradicional): Implementó un método de extracción basado en Reconocimiento de Entidades Nombradas (NER) utilizando un EntityRuler de spaCy, poblado con la taxonomía completa de ESCO, combinado con expresiones regulares para capturar un baseline de habilidades explícitas de alta precisión.

Pipeline B (Basado en LLMs): Empleó Large Language Models (LLMs) como Llama 3 para realizar una extracción semántica, capaz de identificar no solo habilidades explícitas sino también de inferir competencias implícitas a partir del contexto de la vacante, siguiendo enfoques de vanguardia (Herandi et al., 2024; Nguyen et al., 2024).

Posteriormente, un módulo de mapeo de dos capas normalizó las habilidades extraídas por ambos pipelines contra la taxonomía ESCO. La primera capa realizó una coincidencia léxica (exacta y difusa), mientras que la segunda ejecutó una búsqueda de similitud semántica de alto rendimiento, utilizando embeddings multilingües (E5) y un índice FAISS pre-calculado, inspirado en las arquitecturas de herramientas como ESCOX (Kavargyris et al., 2025). Finalmente, un módulo de análisis no supervisado aplicó una secuencia metodológica de embeddings, reducción de dimensionalidad con UMAP y agrupamiento con HDBSCAN para identificar clústeres de habilidades y perfiles emergentes, un enfoque validado por la literatura para el descubrimiento de estructuras en el mercado laboral (Lukauskas et al., 2023).

1.4. Justificación de la solución

La solución implementada se justificó como una alternativa superior y mejor adaptada para el análisis de la demanda de habilidades en América Latina, ya que abordó directamente las debilidades metodológicas identificadas en los estudios previos. A diferencia de los enfoques basados exclusivamente en reglas léxicas (Aguilera & Méndez, 2018; Rubio Arrubla, 2024) o en el uso aislado de LLMs (Nguyen et al., 2024), la arquitectura de dos pipelines paralelos permitió una validación empírica cruzada: combinó la auditabilidad y alta precisión para habilidades conocidas del Pipeline A con la potencia inferencial y la capacidad de descubrir habilidades implícitas del Pipeline B. Este diseño comparativo proveyó un marco para evaluar objetivamente el rendimiento de los LLMs, en lugar de depender únicamente de su capacidad "black-box".

Técnicamente, el sistema representó un avance significativo en escalabilidad y eficiencia. La implementación de un índice FAISS para la búsqueda semántica de similitud (una mejora sobre la propuesta original de ESCOX) permitió procesar grandes volúmenes de datos a una velocidad órdenes de magnitud superior a las búsquedas en bases de datos vectoriales convencionales, haciendo factible el análisis de todo el corpus recolectado (Kavargyris et al., 2025; Lukauskas et al., 2023). Adicionalmente, el sistema fue diseñado explícitamente para la realidad del español latinoamericano. Este enfoque abordó directamente una limitación crítica de trabajos de vanguardia en LLMs, los cuales se han desarrollado y validado casi exclusivamente sobre datasets en inglés (Herandi et al., 2024), ignorando las particularidades lingüísticas (como el "Spanglish") del dominio tecnológico en la región.

Finalmente, el valor agregado del proyecto residió en su síntesis estratégica de metodologías de vanguardia. El sistema no se limitó a una sola técnica, sino que articuló la cobertura del scraping regional, la potencia de los LLMs ajustados para generar salidas estructuradas (Herandi et al., 2024) , y la capacidad estructuradora del clustering semántico (Lukauskas et al., 2023). Al hacerlo, se desarrolló un observatorio más completo, robusto y metodológicamente transparente que las alternativas existentes, estableciendo una base sólida y replicable para el monitoreo dinámico de la demanda laboral en la región.

2. Descripción del proyecto

El proyecto se concibió como un observatorio automatizado para capturar, normalizar y analizar avisos de empleo en Latinoamérica. Se integraron múltiples portales (CO, MX y AR), se diseñó una base de datos relacional con soporte vectorial, y se implementó un pipeline de extracción de habilidades (NER/regex/LLM) alineadas a ESCO, con generación de indicadores, visualizaciones y reportes. Operativamente, se planificó escalar hasta 600.000 avisos para la defensa, garantizando calidad, trazabilidad y reproducibilidad.

2.1. Objetivo general

Desarrollar un sistema que permita procesar y segmentar la demanda de habilidades tecnológicas en Colombia, México y Argentina, mediante técnicas de procesamiento de lenguaje natural.

2.2. Objetivos específicos

- Construir un estado del arte exhaustivo para comparar trabajos existentes en el ámbito de observatorios laborales automatizados y técnicas de procesamiento de lenguaje natural en español.
- Diseñar una arquitectura modular, escalable y reutilizable para el observatorio laboral automatizado, fundamentada en las mejores prácticas identificadas en el estado del arte.
- Implementar e integrar técnicas de inteligencia artificial para la identificación, normalización y agrupación semántica de habilidades tecnológicas en ofertas laborales en español.
- Validar el desempeño y la robustez de la arquitectura y los modelos propuestos mediante métricas cuantitativas y estudios empíricos.

2.3. Entregables, estándares y justificación

Entregable	Estándares asociados	Justificación
Repositorio de código (spiders, orquestador, pipelines)	PEP 8/257/484; Conv. Commits; SemVer	Mantenibilidad, legibilidad y control de versiones.
Esquema BD y migraciones (PostgreSQL + pgvector)	Normalización (3NF); SQL best practices	Integridad, trazabilidad y soporte a consultas vectoriales.
Spiders y configuración de scraping	Polite crawling (delays/retries); manejo anti-bots	Captura estable a escala y resiliencia ante cambios UI.
Orquestador CLI + scheduler	CLI UX (Typer); jobs idempotentes	Operación reproducible, programable y auditable.
Módulo de extracción/normalización de habilidades	ISO/IEC/IEEE 29148 (requisitos); ESCO	Consistencia semántica y comparabilidad entre países.

Embeddings y análisis (E5, UMAP, HDBSCAN)	Procedimientos reproducibles; semillas fijas	Descubrimiento de patrones y replicabilidad experimental.
Datasets consolidados (CSV/JSON) + diccionario de datos	Esquemas declarativos; control de versiones	Consumo externo y verificación de calidad.
Documentación técnica y de proyecto (SRS, SPMP, VFP, manuales)	IEEE 1058 (plan de proyecto); 29148 (requisitos)	Alineación con buenas prácticas y transferencia de conocimiento.
Reportes y visualizaciones (PDF/PNG/CSV)	Principios de visualización; metadatos	Comunicación clara de hallazgos a públicos no técnicos.
Plan de operación y mantenimiento (Docker/monitoring)	Buenas prácticas Docker/Logging	Despliegue consistente y observabilidad del sistema.

III- CONTEXTO DEL PROYECTO

1. Background

Para comprender el diseño y la justificación de la solución desarrollada, fue necesario fundamentar el proyecto en una serie de conceptos clave provenientes de la ingeniería de sistemas, la ciencia de datos y, fundamentalmente, del Procesamiento de Lenguaje Natural (NLP). Estos conceptos no actúan de forma aislada, sino que se articulan en un flujo metodológico que va desde la adquisición de datos brutos hasta la generación de conocimiento estructurado sobre el mercado laboral.

El punto de partida del observatorio fue la recolección de datos a gran escala desde fuentes web públicas. Esta tarea se realizó mediante Web Scraping, una técnica de extracción automatizada de información desde el código HTML de las páginas web (Orozco Puello & Gómez Estrada, 2019). En el contexto del mercado laboral, esta técnica ha demostrado ser fundamental para obtener datos de alta frecuencia y granularidad directamente de los portales de empleo, superando las limitaciones de las encuestas y los reportes institucionales, que suelen ser retrospectivos y de baja periodicidad (Cárdenas Rubio et al., 2015; Rubio Arrubla, 2024).

Una vez extraído el contenido textual de las ofertas laborales, el siguiente paso fue prepararlo para el análisis computacional. Esto implicó una serie de técnicas de preprocesamiento de texto, comenzando con la Tokenización, que consiste en segmentar el texto en unidades mí-

nimas o "tokens" (generalmente palabras o signos de puntuación) (Nguyen et al., 2024). Posteriormente, se aplicó la Lematización, un proceso que reduce las palabras a su forma base o raíz gramatical (lema), permitiendo agrupar variaciones morfológicas de un mismo término (por ejemplo, "programar", "programando" y "programado" se unifican bajo el lema "programar") (Echeverría & Rucci, 2022). Este paso es crucial para estandarizar el vocabulario y reducir la dispersión de los datos antes del análisis.

Con el texto limpio y normalizado, el núcleo del desafío consistió en la extracción de habilidades. Para ello, se empleó un enfoque híbrido. Por un lado, se utilizaron Expresiones Regulares (Regex), un lenguaje de patrones sintácticos que permite identificar y extraer secuencias de texto muy específicas, como nombres de tecnologías o certificaciones con formatos predecibles (Lukauskas et al., 2023). Por otro lado, y como método principal, se aplicó el Reconocimiento de Entidades Nombradas (NER), una técnica de NLP diseñada para identificar y clasificar entidades en un texto, como nombres de personas, lugares o, en este caso, habilidades y competencias (Herandi et al., 2024). El NER permitió pasar de una búsqueda basada en reglas a un sistema capaz de reconocer habilidades en contextos gramaticales complejos.

Para superar las limitaciones de la extracción de menciones explícitas, el proyecto incorporó el uso de Large Language Models (LLMs). Estos modelos de lenguaje a gran escala, como GPT o Llama 3, entrenados sobre corpus masivos de texto, poseen capacidades de razonamiento contextual que permitieron abordar desafíos más complejos (Herandi et al., 2024). A través del diseño de prompts específicos (Prompt Engineering), fue posible guiar a los LLMs para realizar tareas de enriquecimiento semántico, como la distinción entre habilidades explícitas (mencionadas textualmente) e implícitas (inferidas del contexto del cargo) (Nguyen et al., 2024). Esta capacidad fue fundamental para obtener una visión más completa de los perfiles demandados.

Una vez extraídas y normalizadas, las habilidades debían ser representadas de una forma que permitiera su análisis cuantitativo. Para ello, se utilizaron Embeddings Semánticos, que son representaciones vectoriales (numéricas) de palabras o frases en un espacio de alta dimensionalidad. La propiedad fundamental de estos embeddings es que la distancia entre dos vectores en ese espacio refleja la similitud semántica entre los textos que representan (Kavas et al., 2024). Dado que las ofertas laborales en América Latina a menudo contienen términos técnicos en inglés ("Spanglish"), fue crucial el uso de Embeddings Multilingües, modelos entrenados para que textos con el mismo significado en diferentes idiomas tengan representaciones vectoriales cercanas en el mismo espacio semántico (Echeverría & Rucci, 2022).

Finalmente, para descubrir patrones y estructuras latentes en el conjunto de datos, se aplicó un pipeline de análisis no supervisado. Debido a que los embeddings son vectores de muy alta dimensionalidad (ej. 768 dimensiones), lo que dificulta la efectividad de muchos algoritmos

(la "maldición de la dimensionalidad"), primero se aplicó una técnica de Reducción de Dimensionalidad como UMAP (Uniform Manifold Approximation and Projection). UMAP es un algoritmo no lineal que reduce el número de dimensiones preservando tanto la estructura local como global de los datos, lo que lo hace superior a métodos lineales como PCA para visualizar relaciones semánticas complejas (Lukauskas et al., 2023). Sobre los datos ya en un espacio de baja dimensionalidad, se aplicó un algoritmo de Clustering basado en densidad como HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). A diferencia de métodos como K-Means, HDBSCAN no requiere que se especifique el número de clústeres de antemano y es capaz de identificar grupos de formas arbitrarias y, crucialmente, de separar los puntos que no pertenecen a ningún grupo como "ruido" (Lukauskas et al., 2023). Esta secuencia metodológica, inspirada en la literatura de vanguardia, fue la que permitió la identificación automática de "ecosistemas de habilidades" y perfiles laborales emergentes.

2. Context Analysis

El desafío de extraer, analizar y comprender la demanda de habilidades a partir de fuentes de datos no estructuradas, como las ofertas de empleo en línea, ha sido abordado desde múltiples frentes en la literatura académica y aplicada. Si bien el objetivo es común —traducir texto en conocimiento accionable sobre el mercado laboral—, las aproximaciones metodológicas varían significativamente en su complejidad, escalabilidad y profundidad semántica. Para posicionar adecuadamente la contribución de este proyecto, fue necesario realizar un análisis crítico de las soluciones existentes a nivel global, las cuales se pueden agrupar en tres grandes líneas de trabajo:

- Enfoques Regionales en América Latina: Un conjunto de estudios pioneros en la región que validaron el uso de técnicas de web scraping para la recolección de datos, pero cuyo análisis se fundamentó en métodos de Procesamiento de Lenguaje Natural (NLP) tradicionales, como el análisis léxico y el emparejamiento basado en reglas.

- La Frontera de la Extracción con Large Language Models (LLMs): Investigaciones de vanguardia a nivel internacional que exploraron el uso de modelos de lenguaje a gran escala, tanto en modalidades de prompting (sin re-entrenamiento) como de fine-tuning (con re-entrenamiento), para lograr una extracción de habilidades con mayor capacidad semántica e inferencial.

- Pipelines Semánticos y Descubrimiento No Supervisado: Arquitecturas de análisis completas que, más allá de la extracción, integran embeddings semánticos, técnicas de reducción de dimensionalidad y algoritmos de clustering para descubrir patrones y perfiles laborales emergentes directamente desde los datos.

El siguiente análisis demostrará que, si bien cada una de estas líneas ha aportado herramientas y hallazgos valiosos, ninguna de ellas, de forma aislada, resolvía de manera integral los desafíos metodológicos, geográficos y lingüísticos que presenta el mercado laboral tecnológico en América Latina. Esta fragmentación en el estado del arte fue la que justificó la necesidad de una solución sintética y adaptada, como la que se desarrolló en este proyecto.

2.1. Enfoques Regionales: Caracterización del Mercado con Métodos Léxicos

La primera línea de trabajo relevante para este proyecto comprende un conjunto de estudios pioneros desarrollados en América Latina. Estos trabajos fueron fundamentales porque validaron el uso de portales de empleo en línea como una fuente de datos rica y de alta frecuencia para el análisis del mercado laboral, pero se caracterizaron por emplear metodologías de procesamiento de texto basadas en análisis léxico, frecuencias de términos y reglas manuales.

El estudio más completo y reciente en este ámbito fue el de Rubio Arrubla (2024) para el mercado colombiano. Este trabajo construyó una base de datos masiva mediante web scraping del portal elempleo.com para el periodo 2018-2023. Su principal aporte fue la caracterización cuantitativa del impacto de la pandemia, demostrando un cambio estructural en la demanda de habilidades. Metodológicamente, el estudio implementó una tipología propia de habilidades tecnológicas y clasificó las vacantes utilizando un algoritmo de emparejamiento de texto basado en la descomposición de textos en n-gramas y el cálculo de puntajes de similitud contra la Clasificación Internacional Uniforme de Ocupaciones (CIUO) (Rubio Arrubla, 2024). Si bien esta aproximación permitió extraer tendencias valiosas, su dependencia de la coincidencia léxica representó una limitación fundamental, ya que el método perdía eficiencia a medida que aumentaba el número de palabras en los títulos al no poder capturar el contexto general (Rubio Arrubla, 2024).

De forma análoga, el trabajo de Aguilera y Méndez (2018) para el contexto argentino se centró en el sector de Tecnologías de la Información (TI), extrayendo datos de portales como ZonaJobs y Bumeran. Su análisis se apoyó en técnicas de minería de texto, específicamente en el análisis de frecuencias y el uso de bigramas, para identificar las tecnologías y roles más demandados (Aguilera & Méndez, 2018). Sin embargo, para estandarizar el vocabulario informal de las ofertas, los autores tuvieron que construir una lista de palabras clave de forma semi-manual, lo que limita la escalabilidad del sistema y su capacidad para adaptarse a la aparición de nuevas tecnologías no contempladas inicialmente (Aguilera & Méndez, 2018). Para el caso de México, la investigación de Martínez Sánchez (2024) propuso un enfoque innovador al combinar datos de encuestas oficiales con información obtenida mediante scra-

ping. Su análisis se basó en la frecuencia de términos y en una tipología manual para segmentar las habilidades, arrojando luz sobre el desajuste entre oferta y demanda, pero sin incluir un procesamiento avanzado y automatizado del lenguaje natural (Martínez Sánchez, 2024).

En conjunto, estos estudios regionales fueron cruciales para establecer la viabilidad de la recolección de datos, pero, desde una perspectiva metodológica, expusieron una brecha fundamental compartida: su dependencia de la correspondencia léxica explícita. Al basarse en frecuencias de palabras, n-gramas o listas de términos predefinidos, estos sistemas eran metodológicamente frágiles ante la ambigüedad y la variabilidad del lenguaje natural. Más allá de las limitaciones académicas individuales, esta carencia de infraestructura analítica ha sido reconocida a nivel institucional. El Banco Interamericano de Desarrollo (BID) ha señalado la falta de pipelines de análisis modernos y automatizados en la región, destacando que la mayoría de los sistemas existentes, si bien articulan el scraping, todavía se basan en reglas fijas o mapeos manuales y no han incorporado técnicas de embeddings ni de NLP avanzado (Echeverría & Rucci, 2022). Esta constatación institucional refuerza la conclusión de que existía un vacío sistémico: la ausencia de una solución que superara los enfoques léxicos para proporcionar un análisis semántico, dinámico y escalable de la demanda de habilidades en América Latina.

2.2 La Frontera de la Extracción: El Uso de Large Language Models (LLMs)

Paralelamente a los enfoques regionales, una segunda línea de investigación a nivel internacional ha explorado el uso de Large Language Models (LLMs) para superar las limitaciones de los métodos léxicos. Estos trabajos representan la frontera del estado del arte en extracción semántica, mostrando tanto el potencial transformador de los modelos de lenguaje de gran escala como las complejidades prácticas de su aplicación en dominios especializados como el mercado laboral.

Una de las primeras aproximaciones en este campo fue la de Nguyen et al. (2024), quienes investigaron el uso de LLMs de propósito general, como GPT-3.5 y GPT-4, en una modalidad de prompting sin re-entrenamiento (few-shot learning). Su metodología consistió en proporcionar al modelo una instrucción y unos pocos ejemplos de extracción de habilidades dentro del propio prompt. Experimentaron con dos formatos de salida: uno de extracción directa, donde el modelo devolvía una lista de habilidades ("EXTRACTION-STYLE"), y otro de etiquetado, donde el modelo reescribía la oración original encerrando las habilidades entre etiquetas especiales ("NER-STYLE") (Nguyen et al., 2024). Sus hallazgos fueron reveladores: aunque los LLMs no lograron igualar la precisión (medida con el F1-score) de los modelos supervisados tradicionales, demostraron una capacidad superior para interpretar frases sintácticamente complejas o ambiguas, como aquellas donde múltiples habilidades están conectadas por conjunciones (Nguyen et al., 2024). Sin embargo, el estudio también advirtió sobre las

limitaciones inherentes a este enfoque, principalmente la inconsistencia en los formatos de salida y el riesgo de "alucinaciones", donde el modelo genera entidades que no corresponden a habilidades reales (Nguyen et al., 2024).

Tomando estas limitaciones como punto de partida, el trabajo de Herandi et al. (2024) representó la siguiente evolución lógica: el fine-tuning o re-entrenamiento específico de un LLM para la tarea. En su investigación, tomaron el modelo LLaMA 3 8B y lo ajustaron utilizando el dataset de referencia SkillSpan (Herandi et al., 2024). Su principal innovación fue el diseño de un formato de salida estructurado en JSON que no solo extraía la habilidad (skill_span), sino también el contexto textual que la rodeaba. Este enfoque les permitió alcanzar un rendimiento que superó el estado del arte (SOTA), logrando un F1-score total de 64.8%, superior tanto a los modelos supervisados previos como a los LLMs utilizados mediante prompting (Herandi et al., 2024). Más importante aún, su método garantizó la consistencia y la auditabilidad de los resultados, resolviendo uno de los mayores problemas prácticos de los LLMs.

A pesar de su sofisticación técnica, estos estudios de vanguardia comparten una limitación crucial que fue central para la justificación de este proyecto: fueron desarrollados y validados casi exclusivamente en contextos anglosajones y sobre datasets en idioma inglés. El trabajo de Herandi et al. (2024), por ejemplo, se fundamentó íntegramente en el dataset SkillSpan, que, como su nombre indica, contiene únicamente ofertas de empleo en inglés. Esta dependencia del idioma inglés evidenció un claro vacío geográfico y lingüístico en la aplicación de técnicas de NLP avanzadas para el análisis del mercado laboral. En conclusión, si bien los LLMs representan la tecnología de punta para la extracción de habilidades, su aplicación efectiva no es trivial. El prompting simple resulta insuficiente en términos de precisión y consistencia (Nguyen et al., 2024), y las metodologías de fine-tuning de alto rendimiento, aunque superiores, estaban limitadas por la barrera del idioma de los datos de entrenamiento disponibles (Herandi et al., 2024). Esto subrayó la necesidad de un proyecto que no solo aplicara estas técnicas avanzadas, sino que las adaptara y validara para la realidad lingüística y contextual del español en América Latina.

2.3 Pipelines Semánticos y Descubrimiento No Supervisado

La tercera línea de investigación relevante se centra en arquitecturas de análisis completas que van más allá de la simple extracción de entidades para estructurar los datos y descubrir patrones latentes de manera no supervisada. Estos sistemas se enfocan en responder preguntas sobre cómo se agrupan las habilidades y cómo evolucionan los perfiles laborales, en lugar de solo identificar menciones individuales.

El trabajo de Lukauskas et al. (2023) es el pilar fundamental de esta aproximación. Su investigación en el mercado laboral de Lituania propuso y validó empíricamente un pipeline de extremo a extremo que se ha convertido en una referencia metodológica. El flujo comenzaba con la extracción de las secciones de "Requisitos" de las ofertas de empleo mediante expresiones regulares (Regex). A continuación, el texto extraído era vectorizado utilizando un modelo basado en BERT (Sentence Transformers) para generar embeddings semánticos de 384 dimensiones. Conscientes de la "maldición de la dimensionalidad", los autores compararon cinco métodos de reducción de dimensionalidad, concluyendo que UMAP (Uniform Manifold Approximation and Projection) ofrecía los mejores resultados al preservar la estructura local y global de los datos de manera más efectiva que alternativas como PCA o t-SNE. Finalmente, sobre los datos ya reducidos, aplicaron y compararon una batería de algoritmos de clustering, demostrando que HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) fue el más eficaz por su capacidad para identificar clústeres de formas y densidades variables y manejar el ruido de manera robusta. El gran aporte de este estudio fue, por tanto, proporcionar una validación empírica para la secuencia completa Regex → Embeddings BERT → UMAP → HDBSCAN como una metodología de vanguardia para el descubrimiento automático y no supervisado de perfiles laborales coherentes.

En una línea complementaria, enfocada en la estandarización, se encuentra la herramienta open-source ESCOX, presentada por Kavargyris et al. (2025). ESCOX fue diseñada para operacionalizar el mapeo semántico de texto no estructurado contra las taxonomías ESCO e ISCO. Su arquitectura se basa en el uso de un modelo Sentence Transformer pre-entrenado (all-MiniLM-L6-v2) para generar embeddings tanto del texto de entrada como de todas las entidades de ESCO. Posteriormente, calcula la similitud del coseno entre el texto de entrada y cada entidad de la taxonomía, devolviendo aquellas que superan un umbral predefinido. El valor de ESCOX reside en su practicidad, eficiencia y su naturaleza de código abierto, ofreciendo una solución accesible para la estandarización de habilidades. Sin embargo, sus propios autores reconocen la limitación de su enfoque: al ser un método basado en embeddings pre-entrenados sin fine-tuning, su precisión es inherentemente menor que la de modelos más avanzados y especializados, como los basados en arquitecturas Transformer con re-entrenamiento específico para el dominio.

En conclusión, el estado del arte al inicio de este proyecto mostraba que ya existían, por separado, pipelines robustos para el análisis no supervisado y el descubrimiento de perfiles (Lukauskas et al., 2023), así como herramientas prácticas para la estandarización semántica (Kavargyris et al., 2025). No obstante, estas capacidades no se habían integrado en una solución única que también incorporara la potencia inferencial de los LLMs de última generación (como los explorados por Herandi et al., 2024) en un flujo coherente. Más importante aún, ninguna de estas arquitecturas avanzadas había sido desarrollada, adaptada o validada para

el contexto específico del mercado laboral en América Latina y las particularidades lingüísticas del español en la región.

2.4 Análisis Comparativo y Valor Agregado de la Solución Propuesta

El análisis del contexto revela un panorama de investigación rico pero fragmentado, donde ninguna solución existente abordaba de manera integral los desafíos del mercado laboral tecnológico en América Latina. Ante esta realidad, el sistema desarrollado en este proyecto no se posicionó como una alternativa incremental, sino como una síntesis estratégica que articuló las fortalezas de las distintas líneas de investigación para crear una solución metodológicamente superior y contextualmente relevante.

El proyecto partió de los aprendizajes de los estudios regionales, adoptando su enfoque en la recolección de datos masivos a través de web scraping como una fuente válida y de alta frecuencia para caracterizar el mercado (Aguilera & Méndez, 2018; Martínez Sánchez, 2024; Rubio Arrubla, 2024). Sin embargo, reemplazó conscientemente su análisis léxico, propenso a errores y de limitada profundidad, con la potencia semántica e inferencial de los Large Language Models (LLMs). Para ello, se inspiró en la investigación internacional de vanguardia, tanto en la exploración del prompting para manejar frases ambiguas (Nguyen et al., 2024), como en la implementación de técnicas de fine-tuning para alcanzar un rendimiento de última generación (Herandi et al., 2024).

Además, la solución no se detuvo en la simple extracción de habilidades, sino que buscó estructurar el conocimiento descubierto. Para lograrlo, integró el robusto pipeline de análisis no supervisado validado empíricamente por Lukauskas et al. (2023), combinando embeddings, UMAP y HDBSCAN para la identificación automática de clústeres de competencias. Crucialmente, todo el sistema fue diseñado desde su concepción para adaptarse a la realidad lingüística y de datos de América Latina, un vacío metodológico dejado por la investigación internacional, que se ha centrado casi exclusivamente en datasets en inglés (Herandi et al., 2024).

Finalmente, el valor agregado más significativo del proyecto residió en su arquitectura comparativa (Pipeline A vs. Pipeline B). Este diseño dual no solo permitió aprovechar lo mejor de los métodos tradicionales y de los LLMs, sino que introdujo un marco de validación empírica que aporta un rigor científico del que carecen muchas aplicaciones prácticas. Al contrastar sistemáticamente un método transparente y auditable (Pipeline A) contra un modelo semántico avanzado (Pipeline B), el sistema no solo generó resultados, sino que también proveyó una medida de la fiabilidad y el valor agregado de cada enfoque, constituyendo una contribución novedosa y completa al campo de los observatorios laborales automatizados.

IV- ANALISIS DEL PROBLEMA

This chapter summarizes the requirements, constraints, functional specification, and any other items described in the requirements specification.

The analysis should include at least the following sections:

1. Requirements

Summarize the main requirements of the problem and detail the most important ones.

2. Constraints

Explain the constraints of the problem, as they were detailed in the requirements specification.

3. Functional specification

Describe the high-level functional specification of the system. The artifacts utilized depend on the specific problem: use cases, BPMN, etc.

V- DISEÑO DE LA SOLUCIÓN

It must present the architecture of the solution and the main artifacts of the detailed design, together with the description of tools and technologies that were selected to solve the problem. It is recommended to justify the selection of tools based on multiple criteria (established during analysis phase) through the use of comparative tables.

For more details, we recommend to reference attached documents that explain every aspect of design (e.g. SDD).

VI- SOLUTION DEVELOPMENT

This chapter must describe the process utilized to create the solution and relate it to the methodology that was specified in the proposal. Additionally, this chapter must also show the final product. For instance, showing screenshots and describing their functions.

VII- RESULTS

Must present the results of the quality control process, according to what was defined in the methodology. For instance, in a software development project, this section should include the results from standard software testing (unit, functional, system, acceptance, etc.). It is important for them to be consistent with the objective of the project and the methodology used for its development.

This chapter must include an analysis of the results obtained, and conclusions from this analysis.

VIII- CONCLUSIONS

1. Impact Analysis of the Project

Explain the impact of the results of this project in the short, medium, and long term. It should explain the impact in all of the relevant stakeholders.

a. *Impact analysis in systems engineering*

b. *Impact analysis in global, economic, environmental, and societal contexts*

2. Conclusions and Future Work

Explain whether the goals were accomplished and why. Future work that should be explained based on the project results.

IX- REFERENCES

This section provides references and literature used for the development of the project. Follow a standard that can be chosen from APA, IEEE, ACM, Michigan or Springer. Once you make the choice, you must use the same standard throughout the document. This section does not count within the 80 pages that are counted for this report.

X- APPENDICES

Place in this section of the document a list of all appendices related to the project. Appendices must be downloadable from the website and should be the same as specified in the proposal.