Pontificia Universidad JAVERIANA Bogotá

Ingeniería de Sistemas

PROPUESTA PARA PROYECTO DE GRADO

TÍTULO

Observatorio de Demanda Laboral en Tecnología en Latinoamérica

ESTUDIANTE(S)

Nicolas Camacho Alarcon

Documento	Celular	Correo Javeriano
cc. 1000942178	3175714599	camachoa.nicolas@javeriana.edu.co

Alejandro Pinzon Fajardo

Documento	Celular	Correo Javeriano
cc. 1052411260	3115369454	alejandro_pinzon@javeriana.edu.co

Daniel Alfredo Vidal De León

Documento	Celular	Correo Javeriano
cc. 1002248377	3002186741	vidal.da@javeriana.edu.co

DIRECTOR

Ing. Luis Gabriel Moreno Sandoval

Documento	Celular	Correo Javeriano	Trabajo
cc.	3016278993	morenoluis@javeriana.edu.co	Pontificia Universidad Javeriana; Profesor Temporal Departamento de Sistemas

1 Visión global

1.1 Antecedentes, problema y solución propuesta

1.1.1 Descripción de la problemática u oportunidad

El sector tecnológico en América Latina ha experimentado un crecimiento significativo en la última década, impulsado por procesos de transformación digital, expansión del comercio electrónico, adopción de inteligencia artificial y migración de servicios a la nube (Echeverría & Rucci, 2022). Este dinamismo ha incrementado la demanda de profesionales especializados en áreas como ciencia de datos, ingeniería de software, ciberseguridad, machine learning e inteligencia artificial (Campos-Vázquez & Martínez Sánchez, 2024; Rubio Arrubla, 2024). Sin embargo, la información sobre estas transformaciones sigue siendo fragmentaria y mayormente retrospectiva, basada en encuestas, censos o reportes institucionales que no capturan la evolución dinámica del mercado laboral.

Estudios recientes han demostrado que el uso de datos masivos extraídos directamente de portales de empleo permite caracterizar con mayor granularidad y actualidad la demanda de habilidades (Lukauskas et al., 2023; Cárdenas Rubio et al., 2015). No obstante, en América Latina estos enfoques están poco implementados, con escasa articulación entre actores públicos, privados y académicos, y con grandes diferencias entre países en cuanto a disponibilidad de datos, estándares de ocupaciones y cobertura digital (Echeverría & Rucci, 2022).

Existe entonces una oportunidad concreta: construir un sistema automatizado, robusto y replicable que permita analizar en tiempo casi real la evolución de las habilidades tecnológicas requeridas en la región, utilizando técnicas modernas de procesamiento de lenguaje natural (NLP), extracción de entidades, embeddings semánticos y algoritmos de agrupamiento.

La selección de Colombia, México y Argentina como casos de estudio responde a tres razones principales. Primero, son países con alto volumen de publicaciones laborales digitales, lo que facilita el scraping y el entrenamiento de modelos robustos (Rubio Arrubla, 2024; Martínez Sánchez, 2024; Aguilera & Méndez, 2018). Segundo, cuentan con estudios previos que explican parcialmente esta problemática, aunque con enfoques no automatizados o sin un pipeline completo de procesamiento (Cárdenas Rubio et al., 2015; Campos-Vázquez & Martínez Sánchez, 2024). Y tercero, representan distintas realidades económicas, territoriales y de madurez digital, lo que permite validar la portabilidad del sistema propuesto a contextos latinoamericanos diversos.

1.1.2 Formulación del problema

Actualmente, no existe una herramienta automatizada en español que permita extraer, estructurar, analizar y visualizar de forma periódica la evolución de habilidades tecnológicas en el mercado laboral de América Latina. Las metodologías existentes en el ámbito académico o institucional suelen depender de enfoques manuales, encuestas o bases de datos cerradas, lo que limita su frecuencia de actualización, granularidad y representatividad (Echeverría & Rucci, 2022; Cárdenas Rubio et al., 2015).

Además, muchas de las soluciones internacionales basadas en NLP fueron desarrolladas para el idioma inglés y no consideran los retos particulares del español latinoamericano, como el uso

frecuente de anglicismos, expresiones mixtas o abreviaturas propias del sector tecnológico (López et al., 2025; Nguyen et al., 2024). Esto dificulta la correcta identificación y agrupación de habilidades, y obstaculiza el análisis semántico con herramientas de clasificación o embeddings multilingües.

1.1.3 Propuesta de solución

Se propone diseñar e implementar un observatorio de demanda laboral tecnológica para América Latina, basado en un pipeline modular que integre las siguientes etapas:

1. Scraping de portales de empleo abiertos como Computrabajo, Bumeran, elempleo.com, entre otros, priorizando los sitios con alto volumen y cobertura nacional (Aguilera & Méndez, 2018; Rubio Arrubla, 2024).

2. Extracción inicial de habilidades mediante técnicas de *Named Entity Recognition* (NER) adaptadas al español y expresiones regulares (regex), aplicadas sobre títulos, descripciones y requisitos de las vacantes (Aito, s.f.; Vásquez-Rodríguez et al., 2024).

3. Enriquecimiento semántico y depuración de habilidades utilizando *Large Language Models* (LLMs) preentrenados, como GPT o T5, para completar y normalizar las habilidades detectadas (Nguyen et al., 2024; Razumovskaia et al., 2024).

4. Vectorización semántica mediante modelos de *embeddings* multilingües como E5, BETO o fastText español, para obtener representaciones densas de cada habilidad o perfil laboral (López et al., 2025; Kavas, Serra-Vidal & Wanner, 2024).

5. Reducción de dimensionalidad mediante técnicas como UMAP, y posterior clustering con algoritmos robustos a ruido como HDBSCAN, para identificar grupos de habilidades o perfiles emergentes (Lukauskas et al., 2023).

6. Visualización macro de resultados a través de gráficos estáticos y reportes interpretables que faciliten la validación cualitativa por expertos, sin necesidad de construir dashboards interactivos ni portales web (Rubio Arrubla, 2024).

Este sistema buscará ser ejecutable localmente, modular, eficiente y ético, con documentación clara y código versionado.

1.1.4 Justificación de la solución

La solución propuesta responde a las debilidades identificadas en los estudios actuales sobre demanda de habilidades en Latinoamérica, aportando una alternativa automatizada, escalable y científicamente sólida. A diferencia de enfoques manuales o herramientas genéricas desarrolladas para otras regiones, este sistema estará adaptado al idioma español, a las expresiones híbridas típicas del sector tech, y a la

estructura irregular de las vacantes latinoamericanas (Echeverría & Rucci, 2022; Aguilera & Méndez, 2018; Martínez Sánchez, 2024).

Además, al integrar técnicas modernas como embeddings multilingües, clustering basado en densidad, y limpieza semántica con LLMs, el observatorio permitirá agrupar habilidades emergentes que no estén explícitamente listadas, facilitando la detección temprana de tendencias y brechas (López et al., 2025; Lukauskas et al., 2023). Esto es especialmente útil para instituciones educativas, gobiernos y profesionales que necesitan tomar decisiones formativas, laborales o políticas basadas en evidencia actualizada.

El sistema también será replicable en otros países de la región, gracias a su arquitectura modular, dependencias abiertas, y uso de portales laborales públicos, con consideraciones éticas y legales adecuadas (Orozco Puello & Gómez Estrada, 2019).

1.2 Descripción general del proyecto

1.2.1 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.

1.2.2 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.

1.3 Entregables, estándares utilizados y justificación

Entregable	**Estándares asociados**	**Justificación**
Documento de diseño técnico y modelo arquitectónico	UML, IEEE 1074	Producto de la Fase 1, establece la arquitectura modular del sistema. Se apoya en UML para modelar flujos y en IEEE 1074 como referencia de ciclo de vida del software.
Dataset limpio, script funcional, cronograma de ejecución	CRISP-DM, robots.txt	Resultado de la Fase 2. Incluye scraping legalmente respetuoso, aplicado de forma sistemática con base en CRISP-DM.
Diccionario de habilidades, embeddings vectoriales, código del módulo	ISO/IEC 25010, CRISP-DM	Entregable clave de la Fase 3. Evalúa funcionalidad y precisión del procesamiento semántico con NLP y embeddings.
Clusters rotulados, visualización exploratoria, notebook de análisis	ISO/IEC 25010, CRISP-DM	Producto de la Fase 4. Se enfoca en la segmentación de perfiles usando técnicas de clustering, evaluando calidad del agrupamiento.
Informe de pruebas, logs de ejecución, gráficos de resultados	ISO/IEC 29110	Documento de la Fase 5. Refleja la validación funcional del sistema usando pruebas ligeras según ISO/IEC 29110, adaptadas a proyectos pequeños.
Documento final reutilizable, instructivo técnico, repositorio del sistema	IEEE 1074, buenas prácticas de documentación	Consolidación de la Fase 6. Resume el sistema completo y permite su replicabilidad y transferencia de conocimiento.

2 Análisis de impacto

Impacto a corto plazo

- **Disponibilidad de una arquitectura técnica replicable**: El proyecto entregará un pipeline modular, escalable y validado para el análisis de demanda de habilidades en español, accesible a universidades, gobiernos o centros de datos laborales.

- **Innovación metodológica en el uso de LLMs en español**: Se aportarán resultados empíricos sobre el uso de modelos como GPT o T5 en tareas de enriquecimiento de datos, corrección de errores de NER y detección de habilidades implícitas, abriendo camino a nuevas líneas de investigación.

- **Reducción de la dependencia conceptual de reportes estáticos**: Se demostrará que es posible construir análisis dinámicos de demanda con scraping + NLP + clustering, generando una alternativa a métodos tradicionales como encuestas o informes anuales.

Impacto a mediano plazo

- **Toma de decisiones informada por terceros**: Actores del ecosistema educativo y laboral podrán reutilizar el sistema propuesto para ajustar programas académicos, estrategias de formación o diagnósticos institucionales, con base en datos reales del mercado.

- **Adaptación de currículos académicos basada en evidencia**: Instituciones educativas podrán incorporar resultados derivados del pipeline o usar su versión adaptada para identificar lagunas formativas y tendencias emergentes en IA, Ciencia de Datos y tecnología.

- **Mejor conexión entre oferta y demanda laboral**: Al proporcionar una estructura replicable de análisis, se reduce la brecha entre lo que enseñan las instituciones y lo que requiere el mercado, permitiendo acciones más precisas para disminuir el desempleo técnico.

- **Fortalecimiento del talento regional**: Profesionales de Colombia, México y Argentina contarán con un marco claro, si el sistema es implementado por terceros, sobre las habilidades más demandadas, fortaleciendo su empleabilidad y adaptabilidad.

Impacto a largo plazo

- **Transformación estructural en políticas educativas y laborales**: Gobiernos, observatorios nacionales y organismos multilaterales podrían utilizar el sistema como insumo para estrategias públicas de formación, reconversión laboral o fomento a la empleabilidad tecnológica.

- **Impulso a una digitalización regional más armónica**: Al permitir análisis comparables entre países hispanohablantes, el proyecto puede contribuir al diseño de estrategias de transformación digital más sensibles a la evolución real del mercado.

- **Reducción de desigualdades estructurales en el acceso a oportunidades laborales**: El enfoque abierto, modular y multilingüe facilita que actores sin grandes recursos técnicos, como regiones no capitalinas o instituciones pequeñas, puedan beneficiarse de los hallazgos y construir sus propios análisis.

3 Proceso

Este proyecto de grado adopta una estrategia metodológica mixta inspirada en el modelo CRISP-DM y en prácticas ágiles derivadas de Scrum. En lugar de aplicar estos marcos de manera estricta, se toma de CRISP-DM la lógica iterativa por fases encadenadas , desde la comprensión del dominio hasta la validación y documentación, mientras que de Scrum se retoman prácticas de revisión incremental y planificación flexible por semanas de trabajo. El enfoque favorece un desarrollo modular, adaptable y basado en resultados verificables, articulando técnicas de scraping, procesamiento de lenguaje natural, embeddings semánticos, clustering y validación de resultados.

3.1 Fase metodológica 1 - Diseño y arquitectura

3.1.1 Método

Esta fase se basa en el paso de comprensión del negocio de CRISP-DM y en el inicio de un primer sprint de planificación según Scrum. Se enfoca en el análisis extensivo del estado del arte y en la definición conceptual y técnica del sistema a construir. Se realiza una revisión crítica de estudios previos, tecnologías disponibles y taxonomías laborales existentes (como ESCO o CIUO) para seleccionar las herramientas y flujos más adecuados para el contexto latinoamericano (Colombia, México y Argentina). Paralelamente, se inician los primeros módulos de documentación estructurada (guía metodológica, esquema de arquitectura) que serán alimentados en cada fase.

3.1.2 Actividades

- Revisión crítica y sistemática del estado del arte (papers, benchmarks, pipelines existentes).
- Definición de módulos y flujos de datos (scraping, NLP, embeddings, clustering, validación).
- Selección preliminar de tecnologías, modelos pre entrenados y fuentes de datos.
- Diseño del pipeline general y planificación por iteraciones.
- Inicio de la documentación OML y del repositorio metodológico..

3.1.3 Resultados esperados

- Diagrama del sistema propuesto (pipeline modular).
- Lista de portales de empleo seleccionados y características técnicas de acceso.
- Decisión justificada de tecnologías y taxonomías a utilizar.
- Documento metodológico inicial con justificación de diseño y planificación preliminar.

3.2 Fase metodológica 2 - Extracción de ofertas laborales

3.2.1 Método

Inspirado en la fase de recolección de datos de CRISP-DM, esta fase implementa un sistema de scraping automatizado de portales de empleo relevantes en español (como elempleo.com, Bumeran, Computrabajo). Se aprovechan herramientas como Scrapy y Selenium para capturar datos dinámicos de forma robusta. Desde Scrum se aplica una lógica de entrega continua, con iteraciones de prueba para validar la estabilidad de los spiders y ajustar el crawler por portal.

3.2.2 Actividades

- Implementación de spiders con Scrapy + fallback con Selenium para contenido dinámico.
- Extracción de datos clave: título, descripción, ubicación, modalidad, requisitos.
- Normalización básica de campos y almacenamiento en base de datos estructurada.
- Validación de calidad del scraping (frecuencia de actualización, duplicados, errores).
- Registro continuo de avances en la documentación metodológica.

3.2.3 Resultados esperados

- Base de datos actualizada con vacantes recolectadas de los tres países objetivo.
- Spiders funcionales y adaptables a cambios de formato por portal.
- Informe de scraping con cobertura, precisión y errores detectados.
- Revisión de licitud y ética del scraping conforme a estándares locales.

3.3 Fase metodológica 3 - Procesamiento y análisis semántico

3.3.1 Método

Corresponde a la fase de preparación de los datos y modelado en CRISP-DM. En esta etapa se realizan tareas de extracción de habilidades explícitas e implícitas usando un enfoque híbrido: se combinan técnicas clásicas de NER adaptado al dominio laboral en español con razonamiento semántico vía LLMs en modo few-shot. La fase incorpora un proceso de normalización de términos basado en taxonomías estandarizadas (como ESCO), aplicadas en el idioma original sin traducción previa. A través de actividades similares a los sprints de Scrum, se permite iterar sobre los modelos y prompts hasta alcanzar resultados robustos..

3.3.2 Actividades

- Aplicación de NER entrenado o adaptado al dominio laboral en español.
- Curación y ajuste de regex específicas para secciones clave (requisitos, funciones).
- Implementación de prompts few-shot en LLMs (GPT, LLaMA, Claude) para detección implícita.
- Validación y filtrado de habilidades mediante taxonomías laborales estandarizadas.
- Documentación de los prompts, razonamientos y mejoras obtenidas.

3.3.3 Resultados esperados

- Corpus anotado automáticamente con habilidades por vacante.
- Registro de habilidades explícitas (vía NER) e implícitas (vía LLM).
- Vocabulario laboral multilingüe alineado a taxonomías como ESCO o CIUO.
- Informe de cobertura y rendimiento de los métodos de extracción utilizados.

3.4 Fase metodológica 4 - Segmentación de perfiles laborales

3.4.1 Método

Inspirado en la fase de modelado y análisis de CRISP-DM, esta etapa consiste en representar semánticamente las habilidades detectadas y agruparlas en perfiles o clústeres funcionales. Se utilizan embeddings multilingües (como BETO, LaBSE o E5), reducción de dimensionalidad con UMAP y clustering con HDBSCAN. No se realizan visualizaciones aún; el objetivo es estructurar internamente los grupos y patrones emergentes para su posterior evaluación.

3.4.2 Actividades

- Vectorización de habilidades mediante modelos multilingües preentrenados.
- Reducción de dimensionalidad con UMAP para preservar relaciones semánticas.
- Aplicación de HDBSCAN para identificación de grupos latentes de perfiles laborales.
- Revisión y depuración de clusters generados (eliminación de ruido, interpretación manual inicial).
- Actualización de la documentación con resultados intermedios.

3.4.3 Resultados esperados

- Representación vectorial de habilidades por vacante.
- Clústeres de perfiles laborales con características técnicas y semánticas diferenciadas.
- Informe técnico sobre la estabilidad, coherencia e interpretabilidad de los agrupamientos.

3.5 Fase metodológica 5 - Validación técnica y visualización macro

3.5.1 Método

Equivalente a la fase de evaluación de CRISP-DM, esta etapa se enfoca en validar la calidad de las salidas del sistema: tanto a nivel de precisión del reconocimiento de habilidades como de coherencia en los agrupamientos obtenidos. Se aplican métricas cuantitativas (precisión, recall, silhouette score) y validaciones cualitativas. Adicionalmente, se generan visualizaciones macro que permiten evaluar tendencias emergentes, sin llegar a construir dashboards interactivos.

3.5.2 Actividades

- Evaluación de los resultados de extracción con muestras revisadas manualmente.
- Cálculo de métricas para los clusters (densidad, separación, coherencia).
- Generación de visualizaciones macro estáticas (por frecuencia, temporalidad, distribución regional).
- Revisión crítica con usuarios técnicos (Profesores, revisores) para retroalimentación.
- Registro de hallazgos, mejoras y limitaciones detectadas.

3.5.3 Resultados esperados

- Reporte técnico de validación con métricas y gráficos interpretables.
- Visualizaciones macro de perfiles, habilidades, y patrones relevantes.

- Justificación del valor informativo del sistema construido.

3.6 Fase metodológica 6 - Documentación y guía metodológica

3.6.1 Método

Inspirada en la fase final de despliegue de CRISP-DM, esta etapa compila todo el conocimiento generado en las fases anteriores y estructura una guía metodológica que permita replicar, adaptar o escalar el sistema en otros contextos. Se produce documentación técnica, análisis de replicabilidad, y propuestas de mejora futura.

3.6.2 Actividades

- Consolidación de resultados técnicos por fase.
- Estructuración de la guía metodológica completa del observatorio.
- Elaboración de anexos: scripts, configuraciones, prompts, logs de scraping, visualizaciones.
- Preparación del entregable final para sustentación.
- Revisión general con enfoque en claridad, replicabilidad y escalabilidad.

3.6.3 Resultados esperados

- Documento técnico y guía metodológica del sistema desarrollado.
- Repositorio estructurado con código y recursos.
- Validación interna del sistema como solución replicable en Latinoamérica.

4 Aspectos generales del proyecto

4.1 Compromiso de apoyo de la Institución

El presente proyecto de grado, titulado “Observatorio de Demanda Laboral en Tecnología en Latinoamérica”, se desarrolla en el marco de la línea de investigación de la Facultad de Ingeniería de la Pontificia Universidad Javeriana, sin vinculación con una entidad externa como cliente.

Por tal motivo, no se requiere una carta de compromiso adicional. Sin embargo, se declara que la Facultad brinda su apoyo institucional mediante el acompañamiento académico del director de proyecto, acceso a bases de datos, bibliografía, asesoría técnica y espacios adecuados para el desarrollo de las actividades propuestas.

4.2 Derechos patrimoniales

De acuerdo con las políticas institucionales, los derechos morales del presente proyecto de grado corresponden a los estudiantes autores, quienes conservan el reconocimiento de la autoría intelectual.

Los derechos patrimoniales, por tratarse de un trabajo de grado desarrollado íntegramente dentro de la Universidad y en el marco de sus procesos de investigación, pertenecen a la Pontificia Universidad Javeriana. En consecuencia, cualquier producto generado, incluyendo código fuente, documentos o resultados, podrá ser utilizado, replicado o adaptado por la Universidad, conforme a sus fines académicos e investigativos.

5 Marco teórico

5.1 Fundamentos y conceptos relevantes para el proyecto.

Para comprender adecuadamente la propuesta de este proyecto de grado, es necesario abordar los conceptos fundamentales involucrados en el diseño de un observatorio laboral automatizado orientado al análisis de la demanda de habilidades tecnológicas en América Latina. Esta sección organiza dichos conceptos de acuerdo con las etapas del flujo metodológico del sistema propuesto: desde la recolección inicial de datos hasta su representación vectorial, análisis semántico, segmentación y visualización final.

Recolección y almacenamiento de datos

1. Portales de empleo

Son plataformas web donde empresas publican vacantes laborales y profesionales buscan oportunidades. En este proyecto se consideran fuentes como LinkedIn, Computrabajo, Bumeran, ZonaJobs e Indeed, que constituyen insumos primarios para los procesos de scraping y análisis (Aguilera & Méndez, 2018; Cárdenas Rubio et al., 2015).

2. Web Scraping

Técnica de recolección automatizada de datos desde páginas web, utilizando librerías como BeautifulSoup, Selenium o Playwright. Permite extraer de forma estructurada información relevante de las ofertas publicadas (Orozco Puello & Gómez Estrada, 2019).

3. Oferta laboral

Se refiere al anuncio publicado por una organización donde se describe el perfil buscado, incluyendo título del cargo, funciones, requisitos y habilidades deseadas (Rubio Arrubla, 2024).

4. Base de datos relacional (PostgreSQL)

Sistema que organiza los datos recolectados en tablas interconectadas, facilitando su consulta, limpieza y posterior análisis mediante estructuras SQL (Martínez Sánchez, 2024).

5. Normalización de datos

Proceso de limpieza, estandarización y unificación de formatos para reducir ambigüedad, errores y duplicados, y mejorar la coherencia del análisis posterior (Campos-Vázquez & Martínez Sánchez, 2024).

Procesamiento de texto y extracción de habilidades

6. Expresiones regulares (Regex)
Lenguaje sintáctico utilizado para identificar y extraer patrones textuales específicos (como frases que contengan habilidades o requisitos) en grandes volúmenes de texto (Lukauskas et al., 2023).

7. Named Entity Recognition (NER)
Técnica de procesamiento de lenguaje natural (NLP) que identifica y clasifica entidades en un texto, como nombres de habilidades, empresas o tecnologías (Nguyen et al., 2024).

8. Tokenización
Consiste en dividir un texto en unidades mínimas llamadas "tokens" (palabras, signos u oraciones), facilitando el análisis lingüístico automatizado (Nguyen et al., 2024).

9. Lematización
Proceso que transforma las palabras a su forma canónica o raíz gramatical, permitiendo uniformar variaciones morfológicas del lenguaje (Echeverría & Rucci, 2022).

10. Stopwords
Términos frecuentes sin valor informativo (como "de", "por", "la"), comúnmente eliminados en tareas de procesamiento textual (Nguyen et al., 2024).

11. Co-ocurrencia
Medida estadística que indica la frecuencia con que dos o más términos aparecen juntos en un texto, útil para detectar relaciones semánticas (Campos-Vázquez & Martínez Sánchez, 2024).

12. Bigramas y trigramas
Secuencias de dos o tres palabras consecutivas utilizadas para capturar patrones de lenguaje más complejos que las palabras individuales (Aguilera & Méndez, 2018).

Modelado con LLMs y enriquecimiento semántico

13. LLM (Large Language Models)
Modelos de lenguaje de gran escala (como GPT o T5) entrenados sobre corpus masivos, capaces de generar texto, extraer conocimiento implícito y realizar razonamiento contextualizado (Nguyen et al., 2024; Razumovskaia et al., 2024).

14. Prompt Engineering
Diseño estratégico de instrucciones o ejemplos para guiar la salida de un LLM, crucial en tareas de extracción de habilidades o clasificación de ocupaciones (Razumovskaia et al., 2024).

15. Few-shot learning
Habilidad de los LLMs para realizar tareas complejas con pocos ejemplos, lo cual resulta clave cuando se carece de datasets etiquetados masivamente en español (Nguyen et al., 2024).

16. Chain-of-Thought Reasoning (CoT)
Técnica que induce a los modelos a razonar paso a paso, mejorando precisión en tareas como clasificación y desambiguación semántica (Razumovskaia et al., 2024).

17. Infer-Retrieve-Rank (IRR)
Enfoque que primero infiere una entidad, luego recupera candidatos posibles, y finalmente los rankea

con base en relevancia, utilizado para seleccionar habilidades o clasificar ocupaciones (López et al., 2025).

18. Habilidades explícitas vs implícitas

Las primeras están textualmente expresadas ("manejo de Python"), mientras que las segundas deben inferirse por contexto ("implementación de modelos supervisados") (Nguyen et al., 2024).

Representación vectorial y análisis semántico

19. Embeddings semánticos

Representaciones numéricas de textos que capturan similitudes semánticas, permitiendo análisis cuantitativos y clustering. Ejemplos incluyen word2vec, BERT y E5 (Kavas et al., 2025; Vásquez-Rodríguez et al., 2024).

20. Embeddings multilingües

Vectores entrenados para representar texto en múltiples idiomas en un mismo espacio semántico. Son esenciales para manejar contenido mixto español-inglés en ofertas laborales (Echeverría & Rucci, 2022; Razumovskaia et al., 2024).

21. Modelos de lenguaje en español

Incluyen variantes como BETO, MarIA, T5-español, que han sido entrenadas en corpus hispanos y se adaptan mejor a tareas de extracción en este idioma (Nguyen et al., 2024).

22. Espacio vectorial

Marco matemático donde entidades como palabras, frases o documentos son representadas como vectores en un espacio multidimensional (Kavas et al., 2025).

23. Reducción de dimensionalidad (UMAP)

Técnica que transforma espacios de alta dimensionalidad en representaciones más simples, conservando la estructura semántica subyacente para facilitar análisis y visualización (Lukauskas et al., 2023).

Segmentación y visualización

24. Clustering (HDBSCAN)

Algoritmo no supervisado que detecta grupos naturales de observaciones (como habilidades o perfiles laborales) según su similitud semántica, sin requerir número de clusters predefinido (Lukauskas et al., 2023).

25. Evaluación por coherencia semántica

Métrica que mide qué tan bien están agrupadas las instancias similares dentro de un modelo, clave para validar la efectividad del clustering (Vásquez-Rodríguez et al., 2024).

26. Silhouette Score

Indicador que evalúa la calidad de los clusters considerando qué tan cohesionados y separados están entre sí (Lukauskas et al., 2023).

27. Visualización de datos

Proceso de representar información compleja en formatos gráficos o interactivos que permiten interpretar resultados, comunicar hallazgos y apoyar decisiones (Rubio Arrubla, 2024).

28. Python
Lenguaje de programación ampliamente utilizado en ciencia de datos y NLP, por su sintaxis sencilla y librerías especializadas como scikit-learn, spaCy, transformers y pandas (Nguyen et al., 2024).

29. Taxonomía de habilidades (ESCO, CIUO-08, O*NET)
Sistemas jerárquicos y normalizados de clasificación de habilidades y ocupaciones, fundamentales para anclar el análisis a estándares internacionales y mejorar interoperabilidad de los resultados (Cárdenas Rubio et al., 2015; Echeverría & Rucci, 2022).

5.2 Análisis de alternativas de solución

Aunque el presente proyecto busca sentar las bases para el diseño e implementación de un Observatorio Laboral automatizado y adaptable al contexto latinoamericano, con enfoque multilingüe, detección explícita e implícita de habilidades, y visualizaciones analíticas basadas en NLP y clustering, existen líneas de investigación y desarrollos aplicados que abordan parcialmente este problema. A continuación se describen tres alternativas representativas que podrían considerarse soluciones parciales al desafío de extraer y analizar habilidades desde ofertas de empleo en línea.

5.2.1 Alternativas de solución e impacto

Alternativa 1: Scraping y análisis léxico descriptivo con reglas manuales (caso Colombia, México y Argentina)

Una línea de trabajo extendida en América Latina consiste en combinar técnicas de recolección de datos mediante web scraping con análisis descriptivo basado en reglas léxicas, tipologías fijas o matching semántico rudimentario. Si bien este enfoque no recurre a modelos de aprendizaje profundo ni representaciones vectoriales, ha sido útil para establecer líneas base de monitoreo laboral.

Uno de los casos más representativos es el trabajo de Rubio Arrubla (2024), quien diseñó un pipeline de extracción sobre el portal elempleo.com en Colombia. Este estudio abarca más de cinco años de ofertas tecnológicas, extraídas mediante técnicas de scraping iterativo. Para la clasificación de habilidades, el autor emplea un esquema basado en la CIUO-08 y una tipología propia, dividiendo las competencias en habilidades generales, especializadas, TIC y de teletrabajo. La identificación de habilidades se realiza mediante tokenización, lematización y matching textual, con una métrica de similitud basada en n-gramas y umbrales de coincidencia. El análisis final presenta tendencias por tipo de habilidad, nivel educativo y sector económico, sin recurrir a embeddings avanzados ni modelos de clasificación automática con LLMs.

De forma similar, Aguilera y Méndez (2018) desarrollaron un sistema de scraping para portales argentinos como ZonaJobs, Bumeran y UniversoBIT. Su estudio se concentra en el sector TI y destaca por utilizar técnicas de minería de texto en R, específicamente bigramas y análisis de frecuencias. La estandarización léxica fue un reto central, dado el uso informal del lenguaje en las ofertas. Los autores construyeron una lista de palabras clave semimanual, segmentando ofertas por tecnología (ej. Java, SQL, Linux) y rol (ej. Analista, Soporte). Un hallazgo clave fue la altísima concentración geográfica de vacantes en Buenos Aires (más del 90%), lo que refuerza la necesidad de visualización territorial en observatorios laborales.

En México, Martínez Sánchez (2024) combinó datos de la ENOE con scraping de portales abiertos, permitiendo comparar la demanda reportada con la demanda publicada en línea. El enfoque se apoya en un análisis de frecuencia de términos y la creación de índices de crecimiento en habilidades digitales, así como en una tipología que segmenta habilidades sociales, cognitivas y técnicas. Aunque no incluye procesamiento avanzado de lenguaje natural, el estudio evidencia el desajuste entre oferta educativa y demanda laboral, justificando la necesidad de observatorios automatizados.

En términos técnicos, estos estudios suelen combinar Scrapy, un framework de scraping asíncrono, con Selenium o scrapy-selenium para acceder a contenido dinámico cargado por JavaScript. Este diseño modular permite definir spiders especializados por portal, región o sector, optimizando el rastreo en paralelo. Los datos recolectados pueden almacenarse en MongoDB, PostgreSQL o archivos planos (CSV/JSON), facilitando su posterior integración con pipelines de análisis o dashboards institucionales.

Estos tres estudios muestran un enfoque efectivo para capturar tendencias macro en demanda laboral, a partir de reglas simples, sin recurrir a modelos complejos. La fiabilidad del análisis depende de la calidad del scraping y de la precisión de los diccionarios empleados.

Impacto técnico: Estos sistemas son fáciles de implementar y mantener. Pueden ejecutarse con lenguajes como Python o R, y son reproducibles por equipos sin formación avanzada en inteligencia artificial. No requieren GPUs ni acceso a APIs costosas, lo que los vuelve apropiados para entornos universitarios con recursos limitados. Sin embargo, no escalan bien ante la ambigüedad o informalidad de las ofertas, y requieren mantenimiento manual constante de listas de términos y reglas.

Impacto social: Permiten una visualización inicial de las tendencias de empleo, siendo útiles para tomadores de decisiones en políticas públicas o para programas de formación técnica. Al ser comprensibles y auditables, estos enfoques generan confianza entre stakeholders no técnicos. Sin embargo, no capturan habilidades emergentes ni estructuras implícitas del lenguaje, lo que puede invisibilizar competencias importantes.

Impacto económico: Reducen costos frente a encuestas laborales tradicionales, pero requieren esfuerzos manuales continuos. No generan automatización de largo plazo ni integración con dashboards interactivos, lo que limita su reutilización institucional o empresarial.

Alternativa 2: Extracción de habilidades mediante LLMs y aprendizaje por instrucciones (casos NLP4HR, CLiC-it, GenAIK)

En años recientes, una nueva línea de investigación ha explorado el uso de modelos de lenguaje de gran escala (LLMs) para la extracción automatizada de habilidades desde ofertas laborales. Estos métodos recurren a técnicas de few-shot learning, prompting estructurado y recuperación aumentada (RAG) para detectar menciones tanto explícitas como implícitas de competencias, con adaptabilidad multilingüe.

El trabajo de Nguyen et al. (2024), presentado en el taller NLP4HR de ACL, es un referente en esta categoría. Los autores replantean la tarea clásica de extracción como una generación de texto supervisada, donde al modelo se le presentan ejemplos dentro del prompt (few-shot in-context learning) y se le solicita extraer las habilidades relevantes. Comparan dos formatos: uno tipo extracción directa (EXTRACTION-STYLE) y otro tipo etiquetado secuencial (NER-STYLE),

utilizando GPT-3.5 turbo. Sus hallazgos muestran que, aunque el rendimiento en F1 es inferior al de modelos entrenados, los LLMs manejan mucho mejor frases sintácticamente complejas o ambiguas. En particular, Nguyen et al. (2024) reportan que el uso de solo cinco ejemplos en el prompt permitió mejorar el F1 en un rango de ~20–28% frente al escenario zero-shot. Además, observaron que los LLMs capturan con mayor precisión listas mixtas de habilidades técnicas y blandas, incluso cuando están embebidas en estructuras narrativas o poco formales.

Complementariamente, Razumovskaia et al. (2024) presentan en CLiC-it un sistema que combina embeddings E5 multilingües con LLMs como Llama-3, aplicando clasificación multilabel de vacantes según la taxonomía ESCO. Las ofertas en español e italiano son vectorizadas con E5, luego las 30 ocupaciones más similares se recuperan vía cosine similarity, y se construye un prompt para el LLM que realiza el etiquetado final. Esta arquitectura híbrida entre recuperación semántica y razonamiento contextual logra una precisión considerable en clasificación de ocupaciones, sin necesidad de entrenamiento supervisado.

Por su parte, López et al. (2025), en el marco del GenAIK workshop, proponen una solución completa basada en knowledge graphs laborales. El sistema integra entity linking con embeddings de tipo fastText y Node2Vec, usando razonamiento Chain-of-Thought (CoT) para validar habilidades inferidas. La extracción se realiza a través de clasificación multietiqueta extrema (XMC) y se ancla a la taxonomía ESCO. Su pipeline multilingüe opera en español, y utiliza prompts en inglés sobre entradas hispanas, estrategia que ha demostrado ser más eficaz que traducir previamente los textos. Este enfoque demuestra ser robusto y preciso, incluso para detectar habilidades implícitas en contextos ambiguos.

Impacto técnico: Estos métodos representan el estado del arte en extracción semántica. Permiten abordar textos en español informal o con lenguaje mixto (spanglish), y extraer competencias no mencionadas literalmente. En contextos universitarios o de formación técnica, pueden utilizarse como prototipos rápidos para cursos avanzados de NLP o como herramientas exploratorias en observatorios laborales sin requerir re-etiquetado masivo. Sin embargo, requieren acceso a LLMs potentes (como GPT-4 o LLaMA-3) y GPUs para ejecución eficiente. Son difíciles de auditar y pueden generar resultados inconsistentes sin mecanismos de validación.

Impacto social: Al captar habilidades implícitas, visibilizan competencias valiosas que no se nombran explícitamente en las ofertas. Esto puede beneficiar a sectores poblacionales con trayectorias laborales no convencionales. También permiten mayor actualización y adaptabilidad, ajustándose a los cambios en el lenguaje laboral.

Impacto económico: Aunque su implementación inicial es costosa, reducen drásticamente la necesidad de anotación manual y escalabilidad limitada. Pueden implementarse en observatorios o empresas para análisis automatizado de miles de vacantes. Sin embargo, la dependencia de proveedores externos (OpenAI, Anthropic, etc.) conlleva riesgos de sostenibilidad y costos recurrentes.

Alternativa 3: Pipelines de análisis semántico y agrupamiento (casos Lituania, BID, OIT)

Una tercera alternativa metodológica combina técnicas tradicionales de extracción con representaciones vectoriales semánticas y algoritmos de agrupamiento (clustering) no supervisado. Esta línea permite no solo identificar habilidades, sino agruparlas en perfiles o clústeres dinámicos, facilitando visualizaciones de tendencias y análisis estructurales del mercado laboral.

El estudio de Lukauskas et al. (2023), publicado en Applied Sciences, es pionero en esta aproximación. El equipo lituano procesó más de 500.000 vacantes mediante un pipeline que comienza con extracción de requerimientos por expresiones regulares (regex) y segmentación por secciones (ej. "Requirements" → "Company Offers"). Luego, los fragmentos relevantes son vectorizados con Sentence-BERT, y se aplica reducción de dimensionalidad con UMAP. Finalmente, se emplea HDBSCAN para identificar clusters de habilidades emergentes. El estudio concluye que esta metodología es capaz de descubrir perfiles laborales y generar descripciones automáticas mediante GPT-4 para cada cluster. Aunque el corpus está en lituano, las herramientas y técnicas utilizadas son multilingües, y la metodología es replicable con modelos entrenados en español.

Desde una perspectiva institucional, el Banco Interamericano de Desarrollo (Echeverría & Rucci, 2022) ha identificado pipelines similares en América Latina, articulando scraping, clasificación por taxonomías (ESCO, CIUO), y visualización dinámica. En países como Uruguay, Paraguay y República Dominicana se han implementado tableros interactivos que monitorean tendencias de habilidades en tiempo real. Sin embargo, la mayoría de estos sistemas se basan en reglas fijas o matching manual, y aún no incorporan técnicas de embeddings ni NLP avanzado.

Un enfoque intermedio es el de la OIT/CINTERFOR (2015), que propone un pipeline desde el scraping hasta el análisis ocupacional, apoyado en bigramas y emparejamiento manual contra clasificaciones internacionales. Aunque más cercano a la práctica regional, este enfoque carece de automatización robusta y no aborda habilidades implícitas ni clustering semántico.

Impacto técnico: Estos pipelines ofrecen robustez y escalabilidad. El uso de embeddings (ej. BETO, LaBSE), UMAP y HDBSCAN permite construir representaciones dinámicas del mercado laboral, detectar nuevas combinaciones de competencias y monitorear perfiles emergentes. Sin embargo, requieren infraestructura de análisis y conocimiento especializado para su correcta configuración y evaluación.

Impacto social: Facilitan la comprensión de cambios estructurales en la demanda de habilidades, beneficiando a instituciones educativas, empleadores y trabajadores. Los dashboards derivados pueden convertirse en herramientas de planificación estratégica, adaptadas a contextos locales. Además, el uso combinado de expresiones regulares y modelos como BERT permite mantener cierto grado de explicabilidad y transparencia, lo que favorece la confianza entre usuarios institucionales o no expertos, en comparación con enfoques completamente basados en LLMs.

Impacto económico: Aunque más costosos en implementación que enfoques léxicos simples, estos pipelines permiten análisis replicables y actualizables, reduciendo costos de investigación a largo plazo. Además, su potencial para caracterizar clústeres emergentes puede ayudar a diseñar programas de formación más alineados con el mercado.

Las tres alternativas exploradas ofrecen enfoques complementarios para abordar el desafío de extraer, representar y analizar habilidades desde ofertas de empleo publicadas en línea. Sin embargo, ninguna de ellas (y ninguno de los estudios asociados) resuelve de forma integral los requerimientos de automatización, adaptabilidad lingüística y escalabilidad técnica que exige un observatorio laboral moderno, especialmente en contextos latinoamericanos multilingües y altamente heterogéneos. Si bien algunos alcanzan a aproximarse a los requerimientos generales de los componentes propuestos, ninguno de esa complejidad publica resultados en el contexto latinoamericano, especialmente en Colombia, México, y Argentina, ni estrictamente enfocados al idioma español.

La primera alternativa, basada en scraping y análisis léxico con reglas manuales, ha demostrado ser efectiva para caracterizar tendencias generales en demanda laboral, particularmente en países como Colombia, Argentina y México. Su principal ventaja radica en su bajo costo de implementación y su claridad metodológica, lo que facilita su adopción por universidades, gobiernos y actores del tercer sector. No obstante, esta línea metodológica presenta limitaciones importantes: su dependencia de listas estáticas de palabras clave impide detectar habilidades implícitas, emergentes o mal redactadas, y su falta de representaciones semánticas profundas restringe la capacidad de generalización del sistema ante nuevas formas lingüísticas o sectores no previamente tipificados. Además, el mantenimiento de los diccionarios y expresiones requiere esfuerzo manual constante y conocimiento experto del dominio.

La segunda alternativa, centrada en el uso de LLMs y aprendizaje por instrucciones, representa el estado del arte en procesamiento de lenguaje natural aplicado al dominio laboral. Estos modelos permiten extraer habilidades complejas con prompts bien diseñados, manejar lenguaje mixto (español-inglés), y operar sin necesidad de entrenamiento supervisado. Su flexibilidad los hace especialmente útiles para detectar competencias blandas, inferencias implícitas y relaciones contextuales. Sin embargo, esta potencia técnica viene acompañada de desafíos importantes: el acceso a modelos avanzados suele requerir infraestructura de alto costo o dependencias de proveedores externos, su comportamiento puede ser opaco y difícil de auditar, y existe un riesgo potencial de sesgos si no se aplican filtros culturales y lingüísticos adecuados. En el contexto latinoamericano, su adopción requiere una cuidadosa adaptación semántica y una evaluación ética del impacto social de los modelos generativos.

La tercera alternativa, basada en pipelines de análisis semántico y agrupamiento no supervisado, ofrece un compromiso interesante entre complejidad técnica, interpretabilidad y escalabilidad. Al integrar técnicas como embeddings multilingües, reducción de dimensionalidad (UMAP) y clustering (HDBSCAN), estos enfoques permiten construir visualizaciones dinámicas de perfiles laborales, detectar clústeres emergentes de habilidades y ofrecer una segmentación más robusta del mercado de trabajo. En comparación con los LLMs, estos métodos son más eficientes computacionalmente, más explicables, y pueden implementarse con modelos preentrenados accesibles en español (como BETO o LaBSE). Sin embargo, requieren una etapa previa de extracción bien resuelta (idealmente por regex o NLP tradicional) y una calibración cuidadosa de parámetros de agrupamiento para evitar la generación de clusters artificiales o incoherentes.

En conjunto, estas alternativas muestran un panorama metodológico rico pero fragmentado. Los enfoques simples basados en scraping y reglas ofrecen confiabilidad y bajo costo, pero escasa adaptabilidad semántica. Los LLMs ofrecen potencia inferencial y aprendizaje contextual, pero imponen barreras de costo, auditoría y replicabilidad. Los pipelines semántico-clusterizados facilitan la estructuración macro de perfiles laborales, pero requieren componentes previos bien diseñados y cierto conocimiento técnico para su ajuste.

Para el contexto de América Latina, marcado por la alta informalidad laboral, la ambigüedad textual de las ofertas, y la coexistencia de múltiples registros lingüísticos (incluyendo spanglish y terminología local), ninguna de estas alternativas resulta suficiente por sí sola. La solución óptima debe articular las fortalezas de cada enfoque: combinar scraping automatizado y modular (Alternativa 1), extracción híbrida e inferencial con LLMs adaptados (Alternativa 2), y representación semántica con agrupamiento dinámico (Alternativa 3). Esta articulación permitiría construir un observatorio

laboral flexible, multilingüe, transparente y escalable, capaz de detectar tanto las macro-tendencias como los matices implícitos de un mercado laboral digital en constante transformación.

5.2.2 Comparación de alternativas

Estudio	Aportes principales	Limitaciones técnicas y conceptuales	Qué podemos aprovechar
Rubio Arrubla (2024)	Scraping masivo en Colombia. Tipología de habilidades (TIC, teletrabajo, especializadas). Matching léxico y CIUO.	No usa embeddings avanzados, LLMs, ni clustering. Extracción explícita dependiente de n-gramas.	Tipología de habilidades adaptada al contexto colombiano. Corpus base útil para comparación temporal.
Aguilera y Méndez (2018)	Scraping en Argentina. Análisis léxico con bigramas. Enfoque geográfico.	Análisis muy dependiente de reglas. Sin vectorización ni detección implícita.	Lecciones sobre desigualdad regional y vocabulario técnico informal en español.
Martínez Sánchez (2024)	Cruce entre encuestas y scraping. Segmentación de habilidades blandas/cognitivas.	Procesamiento manual de términos. Falta de automatización.	Modelo de integración institucional (INEGI + scraping). Fundamento para dashboards nacionales.
Nguyen et al. (2024) – ACL NLP4HR	LLMs en few-shot. Manejo superior de menciones complejas.	No incluye scraping. Requiere LLM potente. Limitado por sesgos.	Estructura de prompting y validación útil para detección implícita.

Razumovskaia et al. (2024) – CLiC-it	Embeddings E5 + LLM para clasificación multilabel. Uso de ESCO.	Asume embeddings preexistentes y datasets estructurados.	Arquitectura híbrida (vectorización + razonamiento) para normalización de ocupaciones.
López et al. (2025) – GenAIK	Entity linking + CoT reasoning + XMC. Extracción multilingüe.	Alta complejidad técnica. Uso de vocabularios no localizados.	Pipeline de última generación replicable con glosarios regionales.
Lukauskas et al. (2023) – Applied Sciences	Regex + BERT + UMAP + HDBSCAN. Segmentación por clústeres. GPT para síntesis de perfiles.	No incluye scraping. Formato lituano. Fuerte dependencia de patrones iniciales.	Pipeline vectorial + clustering útil para construir dashboards e identificar perfiles emergentes.
BID (2022)	Panorama regional. Casos en Uruguay, Paraguay, Colombia, y más.	No menciona técnicas modernas (regex, NLP clásico). No multilingüe.	Justificación institucional del uso de scraping y falta de pipeline moderno.
OIT/CINTERFOR (2015)	Pipeline básico desde scraping hasta análisis ocupacional.	Sin embeddings ni clustering. Manual.	Inspiración estructural para etapas de matching y validación ocupacional.
Puello & Gomez (2019)	Scraping robusto con Scrapy + Selenium. Arquitectura modular.	Sin análisis de lenguaje. Extrae texto sin semántica.	Infraestructura de scraping adaptable a portales latinoamericanos dinámicos.

A partir del análisis detallado de los estudios y soluciones existentes, es evidente que ninguna alternativa actual resuelve de manera integral los desafíos que enfrenta la caracterización de la demanda de habilidades tecnológicas en América Latina. Si bien cada línea de trabajo aporta elementos valiosos, todas presentan limitaciones críticas en términos de cobertura, adaptabilidad lingüística, automatización, o profundidad analítica. Por esta razón, el enfoque propuesto en este proyecto no se posiciona como una cuarta alternativa más, sino como una síntesis estratégica de lo mejor de cada una, articulada específicamente para el contexto de Colombia, México y Argentina, y diseñada para operar en entornos hispanohablantes con alta variabilidad textual y carencia de recursos anotados.

Los estudios latinoamericanos (Rubio Arrubla, Aguilera y Méndez, Martínez Sánchez) ofrecen un punto de partida sólido en términos de scraping regional, segmentación inicial por tipologías y análisis

descriptivo por frecuencias o n-gramas. Su valor reside en su contextualización local y en su aplicabilidad práctica en entornos con infraestructura limitada. Sin embargo, estos trabajos no emplean representaciones vectoriales modernas, no detectan habilidades implícitas, y carecen de mecanismos para identificar combinaciones emergentes de competencias. Tampoco han sido diseñados como pipelines reutilizables ni como sistemas escalables.

Por otro lado, las investigaciones avanzadas centradas en LLMs (Nguyen et al., Razumovskaia et al., López et al.) han demostrado la capacidad de los modelos generativos para identificar habilidades complejas y no explícitas, así como para realizar tareas de clasificación multilabel en múltiples idiomas. No obstante, estos enfoques fueron desarrollados en contextos europeos, con corpus preprocesados, sin scraping propio ni consideración explícita del español latinoamericano. Tampoco abordan directamente los desafíos culturales, léxicos y sintácticos propios de las ofertas en nuestra región, caracterizadas por lenguaje informal, mezclas con inglés técnico y estructuras poco estandarizadas. De hecho, ninguno de estos estudios ha sido validado o adaptado al ecosistema de portales laborales latinoamericanos, lo que limita su aplicabilidad directa.

Finalmente, los pipelines basados en embeddings y clustering (como Lukauskas et al., BID, OIT/CINTERFOR) permiten construir mapas semánticos útiles para análisis macro y segmentación de perfiles laborales. Sin embargo, estos enfoques no incluyen extracción explícita de habilidades ni validación contextual con LLMs modernos, y han sido aplicados principalmente en idiomas distintos al español o mediante clasificaciones manuales. Además, si bien su arquitectura es prometedora, ninguno ha sido implementado de forma completa y automatizada en entornos hispanohablantes, ni en combinación con herramientas de scraping dinámico.

Frente a este panorama, la propuesta que se presenta en este proyecto se justifica como una solución superior y mejor adaptada a las necesidades reales del análisis de habilidades en América Latina, por las siguientes razones:

1. Integración de extremo a extremo: A diferencia de todas las alternativas revisadas, nuestra propuesta integra en un mismo pipeline automatizado las fases de scraping, extracción, enriquecimiento semántico, clustering y visualización macro. Esto asegura coherencia entre etapas, evita pérdidas de información, y permite ejecutar el proceso de forma periódica y sostenible.

2. Cobertura multirregional y multilingüe: El sistema está diseñado para operar en Colombia, México y Argentina, con spiders configurables para múltiples portales (Computrabajo, Bumeran, elempleo, etc.), y soporte real para textos en español con léxico mixto. Se emplean embeddings multilingües (E5, BETO, fastText) y prompts adaptados al contexto hispanoamericano, lo que supera las limitaciones de todos los estudios centrados exclusivamente en inglés o italiano.

3. Extracción híbrida y adaptable: El sistema combina reglas sintácticas (regex) con NER entrenado en español y validación mediante LLMs en tareas específicas de enriquecimiento (como la desambiguación o la detección de habilidades implícitas). Esto permite mantener eficiencia computacional y al mismo tiempo lograr mayor precisión semántica, sin depender exclusivamente de modelos generativos costosos o difíciles de auditar.

4. Enfoque escalable, modular y replicable: El sistema propuesto puede ejecutarse localmente o en la nube, sus módulos pueden activarse o sustituirse según el caso de uso, y su arquitectura abierta permite replicarse en otras regiones o ampliarse a otros dominios más allá del sector tecnológico.

5. Orientación a la realidad del lenguaje laboral latinoamericano: La propuesta parte del reconocimiento explícito de que en nuestra región el lenguaje de las ofertas es informal, ruidoso, ambiguo y muchas veces híbrido. A diferencia de los estudios europeos o asiáticos, este proyecto no asume estructuras limpias ni corpora idealizados, sino que trabaja directamente con datos reales, y adapta sus componentes para esa complejidad.

En resumen, la propuesta no solo toma lo mejor de las alternativas analizadas, sino que las articula estratégicamente en un flujo coherente, localizado y accionable. Integra la cobertura del scraping regional (Alternativa 1), la potencia inferencial de los LLMs (Alternativa 2), y la capacidad estructuradora de los embeddings y clustering (Alternativa 3). Al hacerlo, llena un vacío metodológico y operativo en el campo de los observatorios laborales automatizados en español, con potencial de escalar, replicarse y generar impacto social, académico y económico en el corto y mediano plazo.

6 Bibliografía

1. Orozco Puello & Gómez Estrada, L. F. (2019). *Proyecto de Grado II – Web Scraping* [Trabajo de grado, Universidad del Sinú Elías Bechara Zainúm, Seccional Cartagena, Facultad de Ciencias Exactas e Ingenierías, Escuela de Ingeniería de Sistemas]. Recuperado de http://repositorio.unisinucartagena.edu.co:8080/jspui/bitstream/123456789/94/1/1.%20Proyecto%20de%20Grado%20II%20-%20WEB%20SCRAPING_FINAL.pdf
2. Rubio Arrubla, J. A. (2024). *Demanda de habilidades tecnológicas: evidencia desde el mercado laboral colombiano* [Tesis de maestría, Universidad de los Andes]. Recuperado de https://repositorio.uniandes.edu.co/server/api/core/bitstreams/ea4ea129-d35e-498c-aa46-028e3d8ffb5e/content
3. Lukauskas, M., Šarkauskaitė, V., Pilinkienė, V., & Stundziene, A. (2023). *Enhancing Skills Demand Understanding through Job Ad Segmentation Using NLP and Clustering Techniques*. ResearchGate. Recuperado de https://www.researchgate.net/publication/370816962
4. Martínez Sánchez, J. C. (2024). *Desajuste en el mercado laboral: análisis de los perfiles de candidatos y las ofertas de trabajo publicadas en internet*. *Revista del INEGI*, (44). Recuperado de https://rde.inegi.org.mx/wp-content/uploads/2024/pdf/RDE44/RDE44_art01.pdf
5. Cárdenas Rubio, J. A., Guataquí Roa, J. C., & Montaña Doncel, J. M. (2015). Metodología para el análisis de demanda laboral mediante datos de internet: caso Colombia [Informe técnico]. Organización Internacional del Trabajo / CINTERFOR. Recuperado de https://www.oitcinterfor.org/sites/default/files/file_publicacion/Metodolog%C3%ADa%20An%C3%A1lisis%20Demanda%20Laboral%20Mediante%20Datos%20Internet%20caso%20Colombia.pdf
6. Campos-Vázquez , R. M., & Martínez Sánchez, J. C. (2024). Skills sought by companies in the Mexican labor market: An analysis of online job vacancies. Estudios Económicos De El Colegio De México, 39(2), 243–278. Recuperado de https://estudioseconomicos.colmex.mx/index.php/economicos/article/view/452/619
7. Aguilera, S. O., & Méndez, R. E. (2018). *¿Qué buscan los que buscan?Análisis De mercado laboral IT en Argentina*. *Revista Perspectivas*, *1*(1), 15–30. Recuperado de https://revistas.ub.edu.ar/index.php/Perspectivas/article/view/39
8. Nguyen K., Zhang M., Montariol S., & Bosselut A. (2024). *Rethinking Skill Extraction in the Job Market Domain using Large Language Models*. En *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)* (pp. 27–42). St. Julian's, Malta: Association for Computational Linguistics. Recuperado de https://aclanthology.org/2024.nlp4hr-1.3/
9. Echeverría L., & Rucci G. (2022). *¿Qué suma la ciencia de datos a la identificación y anticipación de la demanda de habilidades?* Banco Interamericano de Desarrollo. Recuperado de https://publications.iadb.org/es/que-suma-la-ciencia-de-datos-la-identificacion-y-anticipacion-de-la-demanda-de-habilidades
10. Kavas H., Serra-Vidal M., & Wanner L. (2025). *Multilingual Skill Extraction for Job Vacancy–Job Seeker Matching in Knowledge Graphs*. En *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*. Abu Dhabi, Emiratos Árabes Unidos:

International Committee on Computational Linguistics. Recuperado de https://aclanthology.org/2025.genaik-1.15.pdf

11. Vásquez-Rodríguez L., Audrin B., Michel S., Galli S., Rogenhofer J., Negro Cusa J., & van der Plas L. (2024). *Hardware-effective Approaches for Skill Extraction in Job Offers and Resumes*. En *RecSys in HR Workshop 2024*. Bari, Italia. Recuperado de https://publications.idiap.ch/attachments/papers/2024/Vasquez-Rodriguez_RECSYSINHR24_2024.pdf
12. Aito A. (s.f.). *SkillNER: Skill Named Entity Recognition* [Repositorio de código]. GitHub. Recuperado de https://github.com/AnasAito/SkillNER
13. Kavas H., Serra-Vidal M., & Wanner L. (2024). *Enhancing Job Posting Classification with Multilingual Embeddings and Large Language Models*. En *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 440–450). Pisa, Italia: CEUR Workshop Proceedings. Recuperado de https://aclanthology.org/2024.clicit-1.53.pdf