

**Tipo de documento:** Tesis de maestría

*Escuela de Negocios. Master in Management + Analytics*

# Un enfoque de clasificación no supervisada para un sistema de recomendación de currículums vitae basado en la similitud semántica

Autoría: Alasino, Augusto

Año: 2024

## ¿Cómo citar este trabajo?

Alasino, A. (2024) "Un enfoque de clasificación no supervisada para un sistema de recomendación de currículums vitae basado en la similitud semántica". [*Tesis de maestría. Universidad Torcuato Di Tella*]. Repositorio Digital Universidad Torcuato Di Tella  
<https://repositorio.utdt.edu/handle/20.500.13098/12910>

El presente documento se encuentra alojado en el Repositorio Digital de la Universidad Torcuato Di Tella bajo una licencia Creative Commons Atribución-No Comercial- Sin Derivados 4.0 Argentina ([CC BY-NC-ND 4.0 AR](https://creativecommons.org/licenses/by-nc-nd/4.0/ar/))

Dirección: <https://repositorio.utdt.edu>

UNIVERSIDAD TORCUATO DI TELLA, ESCUELA DE NEGOCIOS  
MASTER IN MANAGEMENT + ANALYTICS

---

**Un enfoque de clasificación no supervisada para un sistema de  
recomendación de currículums vitae basado en la similitud  
semántica**

---

Resumen

En el marco del proceso de transformación digital que atraviesan todas las industrias, esta tesis explora el uso innovador de la inteligencia artificial, con un enfoque en el procesamiento del lenguaje natural (NLP), para mejorar los procesos de reclutamiento dentro de las organizaciones. La investigación se centra en el desarrollo de un sistema avanzado y económicamente eficiente de clasificación y recomendación de candidatos para vacantes laborales, utilizando datos no estructurados extraídos de currículums vitae y descripciones de puestos de trabajo. Los resultados sugieren que es totalmente factible generar una mejora significativa en la eficiencia operativa de los procesos de selección con recursos limitados. El sistema desarrollado permite analizar un enorme volumen de candidatos en cuestión de segundos logrando realizar inferencias con un porcentaje de acierto que promedia el 90%, lo cual permitiría reducir costos y tiempos de contratación, además de minimizar los sesgos humanos en la selección inicial de candidatos. Este trabajo no solo confirma la viabilidad de implementar estas tecnologías en entornos empresariales reales debido a su bajo costo computacional, sino que también sugiere un cambio de paradigma hacia un enfoque más justo y equitativo en el reclutamiento, asegurando que todos los candidatos puedan ser considerados basados en su compatibilidad con el puesto.

Augusto Alasino

Tutor: Luciano Del corro

---

UNIVERSIDAD TORCUATO DI TELLA, ESCUELA DE NEGOCIOS

MASTER IN MANAGEMENT + ANALYTICS

---

**An unsupervised classification approach for a resume  
recommendation system based on semantic similarity**

---

**Abstract**

In the context of the digital transformation sweeping across all industries, this thesis explores the innovative use of artificial intelligence, with a focus on natural language processing (NLP), to enhance recruitment processes within organizations. The research centers on the development of an advanced and cost-effective system for classifying and recommending candidates for job vacancies, using unstructured data extracted from resumes and job descriptions. The findings suggest that it is entirely feasible to generate a significant improvement in the operational efficiency of selection processes with limited resources. The developed system is capable of processing a vast volume of candidates within seconds, achieving inferences with an average accuracy rate of 90%, thus potentially reducing hiring costs and time, while also minimizing human biases in the initial candidate selection. This work not only confirms the feasibility of implementing these technologies in real business environments due to their low computational cost but also suggests a paradigm shift towards a more fair and equitable approach in recruitment, ensuring that all candidates can be considered based on their compatibility with the position.

Augusto Alasino

Thesis advisor: Luciano Del corro

---

# Índice

<b>1. Introducción.....</b>	<b>5</b>
<b>2. Marco teórico y trabajos relacionados.....</b>	<b>6</b>
<b>3. Esquema general.....</b>	<b>8</b>
<b>4. Datos.....</b>	<b>10</b>
4.1. Recolección de datos.....	10
4.1.1. Vacantes laborales.....	10
4.1.2. Currículums vitae.....	12
4.2. Preprocesamiento.....	14
4.3. Etiquetado manual.....	17
4.4. Conjuntos de entrenamiento y prueba.....	19
<b>5. Análisis exploratorio.....</b>	<b>20</b>
5.1. La (no) estructura de los datos.....	20
5.2. Análisis semántico.....	22
<b>6. Metodología.....</b>	<b>29</b>
6.1. Lbl2Vec: Clasificación no supervisada.....	29
6.2. Hiperparámetros.....	33
6.3. Modelos a implementar.....	34
6.4. Método de evaluación.....	35
<b>7. Implementación en clasificación.....</b>	<b>36</b>

7.1. Transformers.....	38
7.2. Lbl2Vec.....	39
7.2.1. Optimización de hiperparámetros en Lbl2Vec.....	40
<b>8. Implementación en recomendación.....</b>	<b>42</b>
<b>9. Aplicación práctica en organizaciones.....</b>	<b>47</b>
9.1. Rango de aplicabilidad.....	47
9.2. Impacto en organizaciones.....	48
<b>10. Conclusiones.....</b>	<b>49</b>
10.1. Puntos destacados.....	49
10.2. Futuras mejoras.....	50
10.3. Conclusión.....	51
<b>11. Apéndice.....</b>	<b>52</b>
11.1. Recomendación utilizando ChatGPT.....	52
<b>12. Bibliografía.....</b>	<b>55</b>

## 1. Introducción

En la era digital actual, el volumen de datos generados en todos los ámbitos de la vida ha experimentado un crecimiento exponencial. Esta realidad no es ajena al sector de recursos humanos, donde la digitalización de procesos se ha convertido en una necesidad imperante. Esta investigación se enfoca en la aplicación de la inteligencia artificial, particularmente en el procesamiento del lenguaje natural (NLP), para innovar en el área de recursos humanos dentro de las organizaciones.

La relevancia de este estudio radica en su propósito de implementar un sistema avanzado de recomendación y clasificación de candidatos para vacantes laborales, aprovechando la riqueza informativa contenida en los currículums vitae y las descripciones de las vacantes. Lo distintivo y desafiante de este proyecto es su enfoque en datos no estructurados, predominantemente en forma de texto sin etiquetas, reflejando así la naturaleza real de la información con la que se encuentran habitualmente los profesionales de recursos humanos. Además, el proyecto asume el reto adicional de trabajar con textos en español, lo que implica enfrentarse a la escasez de conjuntos de datos públicos disponibles en este idioma, en contraste con la abundancia de recursos en idioma inglés.

La motivación detrás de este trabajo surge de la necesidad de agilizar los procesos de reclutamiento en empresas que manejan un alto volumen de candidaturas, así como de optimizar las tareas operativas de los reclutadores. Al automatizar la selección inicial de candidatos, se busca no solo eficiencia operativa para las empresas, sino también garantizar una consideración justa de todos los postulantes, un aspecto crucial en un contexto donde el volumen de aplicaciones puede sobrepasar la capacidad de análisis humano.

En esencia, el sistema propuesto funcionaría como un modelo de clasificación de aprendizaje no supervisado. Se basaría en una colección de currículums vitae no etiquetados y un conjunto de clases definido de antemano según las necesidades del usuario, asignando cada CV a una de las clases predefinidas en función de su similitud con dicha clase. La clave para adaptar esta metodología a tareas de

recomendación consiste en utilizar el puntaje de similitud para generar un ranking dentro de cada clase. Dado que cada clase puede utilizarse para representar una vacante laboral diferente, esto permitiría asignar los CVs a distintas vacantes en función del mencionado puntaje y ordenar los mismos de mayor a menor coincidencia.

El criterio de éxito en esta investigación no solo se establece en la capacidad del sistema desarrollado para generar recomendaciones relevantes, sino que también en su viabilidad en entornos empresariales reales, por lo cual el costo computacional es a su vez un factor muy relevante. Es por esto que el foco está puesto principalmente en evaluar si utilizando técnicas establecidas hace ya varios años, como Doc2Vec, se pueden alcanzar resultados similares en comparación con enfoques más modernos y actuales, pero a su vez más costosos de implementar, tales como Transformers.

Los resultados obtenidos demostraron que la eficacia de un sistema no depende exclusivamente de la adopción de tecnologías de vanguardia. Se observó que herramientas más básicas pueden, en algunos casos, ofrecer mejores resultados que métodos más modernos, y a su vez, esta característica facilita la implementación del sistema en entornos con diferentes niveles de recursos, ampliando su rango de aplicabilidad.

## **2. Marco teórico y trabajos relacionados**

Los conceptos clave para esta investigación se explicarán a medida que se presenten. Sin embargo, para proporcionar un marco teórico adecuado basado en la literatura de referencia, se describirán algunas nociones fundamentales en esta sección.

En primer lugar, es crucial distinguir entre los modelos de lenguaje tradicionales y el impacto de la aparición de los modelos Transformers en el procesamiento del lenguaje natural. Los modelos tradicionales, como word2vec ([Mikolov et al., 2013](#)) y doc2vec ([Quoc V. Le, Mikolov, 2014](#)), están basados en tecnologías de redes

neuronales que funcionan de manera secuencial, procesando una palabra a la vez para predecir el contexto circundante.

Por otro lado, los modelos Transformers, gracias al concepto de "atención" ([Vaswani et al., 2017](#)), permiten al modelo enfocar diferentes partes de una secuencia con distintos grados de importancia simultáneamente. Esto facilita la captura de relaciones complejas en los textos sin depender de la secuencia.

A pesar de las ventajas significativas que ofrecen los modelos Transformers, su alto costo computacional dificulta su aplicación en algunos contextos organizacionales. Las técnicas más tradicionales tienen la ventaja de requerir menos recursos para su implementación, lo cual es crucial para esta investigación, ya que busca resultados aplicables en contextos reales.

Dado que este enfoque se basa en datos no estructurados y no etiquetados, y gran parte de la literatura existente utiliza conjuntos de datos pre armados con información estructurada, nos referimos a trabajos basados en la clasificación no supervisada de textos. Esto es relevante porque las tareas de recomendación pueden abordarse de manera similar, siempre y cuando se obtenga una métrica que permita ordenar las coincidencias de mayor a menor para cada clase.

Existen dos maneras populares de abordar este problema: la clasificación 'Zero-shot' y el enfoque basado en similitud. La clasificación 'Zero-shot' utiliza modelos de lenguaje pre entrenados para clasificar textos en categorías sin necesidad de ejemplos específicos para esas categorías, basándose en conocimiento general y contexto para realizar las predicciones. Aunque este método requiere modelos costosos, es útil en situaciones donde obtener datos etiquetados es impráctico. Por otro lado, los enfoques basados en similitud consisten en calcular una métrica de similitud entre cada texto y las categorías predefinidas, clasificando los textos en función de estas similitudes.

Existe literatura relativamente reciente que implementa y compara estas alternativas ([Schopf et al., 2022](#)) utilizando conjuntos de datos diversos, y concluye que los enfoques basados en similitud generalmente producen resultados más precisos en este tipo de tareas. Por lo tanto, nos centramos en trabajos relacionados que proponen métodos no supervisados utilizando esa metodología. Una muestra de ello

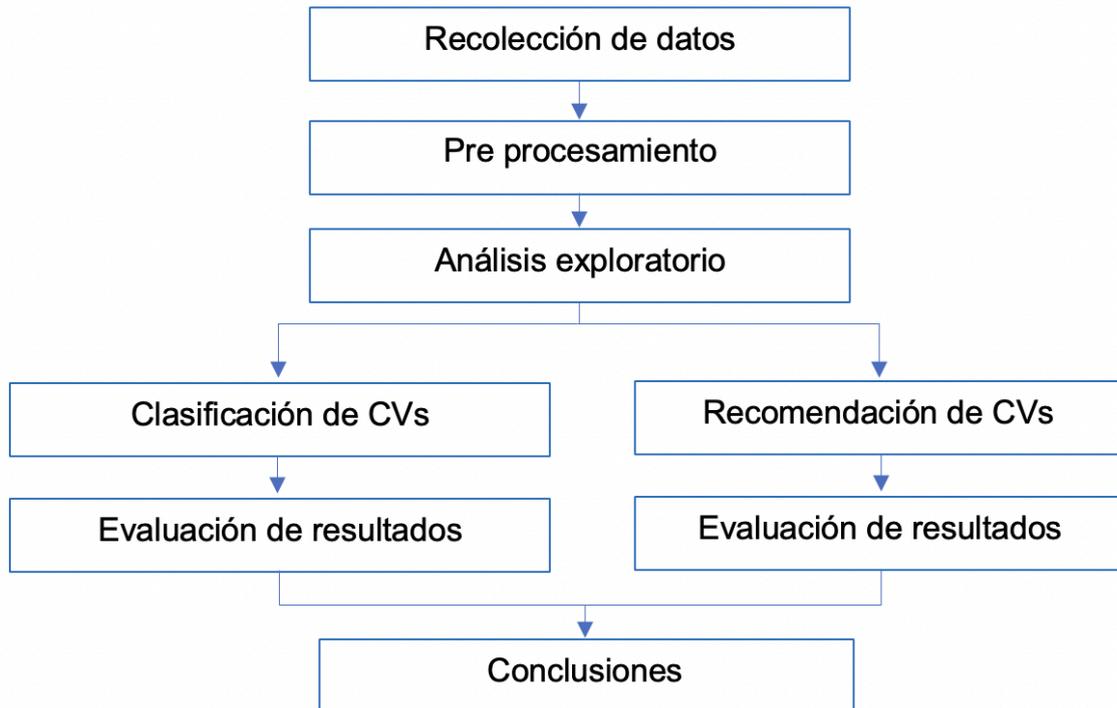
es un modelo de espacio vectorial neuronal ([Gysel et al., 2018](#)) que aprende representaciones de documentos de manera no supervisada y genera rankings basados en la similitud semántica con una consulta iniciada por un usuario. Aunque esto está relacionado con nuestro propósito, difiere en que no se pretende utilizar una consulta, y además, conlleva un gran costo computacional, ya que requiere GPUs (Unidades de Procesamiento Gráfico) para su entrenamiento. Por otra parte, existe una técnica denominada lbi2vec ([Schopf et al., 2021](#)), que permite clasificar documentos en base a su similitud semántica con una serie de palabras clave especificadas de antemano, utilizando técnicas de modelado como doc2vec. En este caso, no se realiza ningún tipo de ranking, pero más adelante veremos cómo se puede adaptar para nuestro propósito.

Dado que la aplicabilidad práctica en diversos contextos organizacionales es fundamental para esta investigación, se buscará trabajar con modelos de bajo costo computacional, como el mencionado doc2vec, y comparar su rendimiento con modelos Transformers pre entrenados, es decir, que ya han sido entrenados con grandes conjuntos de texto para aprender representaciones generales del lenguaje sin realizar ningún tipo de fine-tuning. El objetivo es alcanzar resultados similares a los obtenidos utilizando tecnologías de vanguardia, pero con menores costos.

### **3. Esquema general**

Con el fin de proporcionar una visión estructurada y secuencial de la metodología empleada en este trabajo de investigación, se ilustra cada componente del proceso en el siguiente diagrama (Figura 1).

Figura 1: diagrama de flujo



Este esquema visual presentado en la Figura 1, permite una representación sistemática del progreso y las interdependencias entre las fases del estudio. Comienza con la etapa de recolección de datos y la estrategia empleada para dicha tarea. A continuación, se aborda el preprocesamiento y la depuración de estos datos para su posterior análisis. Esta fase es seguida por un análisis exploratorio que facilita la comprensión de las propiedades y las posibles limitaciones del conjunto de datos. Luego, se procede al entrenamiento e implementación de modelos para tareas de clasificación y recomendación, junto con la correspondiente evaluación de sus resultados, y finalmente, se culmina con la síntesis de conclusiones finales y evaluación de la viabilidad práctica en contextos organizacionales.

## 4. Datos

### 4.1. Recolección de datos

El núcleo de cualquier sistema de procesamiento del lenguaje natural, especialmente aquellos basados en el análisis de similitud semántica entre documentos, reside en la calidad de los datos con los que se alimenta. Por este motivo, es crucial que la estrategia de recopilación de dichos datos sea meticulosa.

La recopilación de datos para este estudio se divide en dos segmentos principales que debieron ser abordados de manera diferente: currículums vitae y descripciones de vacantes laborales.

El propósito final de esta estrategia consistió en formar un corpus lingüístico especializado en el vocabulario y las expresiones propias del ámbito laboral, lo cual se encuentra contenido tanto en el texto proveniente de las descripciones de tareas realizadas por candidatos a lo largo de su experiencia laboral, como en los requerimientos detallados en las publicaciones de vacantes de trabajo. Se pretendía capturar la riqueza y la especificidad del lenguaje utilizado en estos contextos, posibilitando no solo que los modelos aplicados entiendan el significado general, sino que también sean capaces de interpretar las sutilezas y los matices característicos del discurso laboral, donde es común encontrar una amalgama de términos de diferentes lenguas o el uso de nombres de tecnologías y herramientas populares que pueden ser ajenas al idioma español.

#### 4.1.1. Vacantes laborales

En total, se recabaron 3840 descripciones de posiciones vacantes distintas. La Tabla 1, a continuación, muestra la distribución de estos documentos según su fuente de origen.

*Tabla 1: Cantidad de vacantes laborales por cada fuente*

<b>Fuente</b>	<b>Cantidad</b>
ZonaJobs	2203
Bumeran	1637
<b>Total</b>	<b>3840</b>

Las vacantes laborales se obtuvieron mediante un proceso de scraping o extracción de datos en portales de empleo, específicamente en Bumeran y ZonaJobs, dos de los sitios más prominentes en el mercado laboral argentino. Utilizando el lenguaje de programación Python, se desarrollaron programas que permitieron extraer la descripción de una gran cantidad de ofertas de trabajo publicadas en las plataformas, navegando por las mismas de manera automatizada. Esta metodología permitió asegurar una recolección diversa y representativa de vacantes, abarcando una amplia variedad de sectores y posiciones.

En la Tabla 2 se presenta una muestra del conjunto de datos correspondiente a vacantes laborales en su estructura original.

*Tabla 2: Muestra del dataset de vacantes laborales*

<b>Descripción de vacante</b>
Para calificar para este puesto, debe poseer excelentes habilidades en el idioma...
En este momento, nuestro equipo está en busca de un profesional de marketing c..
El Analista de Negocios de Ventas (SBA) brindará soporte al equipo de COMEX p..
El principal objetivo de la posición es a través de habilidades técnicas y analíticas..
Tenemos una oportunidad para vos estamos buscando Business Intelligence Dev..
Nos encontramos en búsqueda de un perfil Senior de Data Science para trabajar...

Estamos en la búsqueda de un Consultor para desarrollo e implementación de m...

Como Data Engineer en nuestro equipo de BI corporativo, desempeñarás un pap...

Buscamos incorporar Delivery Engineer para sumarse a nuestro equipo de Servic..

Estamos en la búsqueda de Analista Semi Sr. de Reporting & Busines Intelligenc...

Como podemos observar, la información se presenta en un formato por definición no estructurado, motivo por el cual requiere una metodología no tradicional para su tratamiento.

#### 4.1.2. Currículums vitae

La metodología empleada para la recolección de los datos en este caso se caracterizó por su enfoque un tanto más manual y meticuloso en comparación con el empleado para la obtención de descripciones de vacantes. Se buscó abarcar un amplio espectro de currículums vitae y garantizar un estándar de calidad en los datos. Uno de los criterios fundamentales para asegurar la calidad de los mismos consistía en que cada CV debía contener texto descriptivo suficiente de su experiencia laboral y tareas realizadas, aspecto que es habitual en la práctica real de búsqueda laboral pero no necesariamente está presente en los perfiles públicos de internet, que constituyen nuestra principal fuente de datos. Este texto descriptivo es crucial, ya que proporciona el contexto necesario para que los modelos de procesamiento del lenguaje natural sean capaces de captar el significado del mismo y poder establecer relaciones semánticas con otros documentos. Además, era imprescindible que todo el texto estuviera en español. Esto garantiza la uniformidad del lenguaje en el corpus, facilitando el entrenamiento de modelos específicos para este idioma, y también refuerza la relevancia de los datos para contextos de habla hispana, asegurando que las inferencias y análisis derivados sean más precisos y aplicables en estas configuraciones lingüísticas.

Al igual que con las vacantes laborales, la Tabla 3 detalla cómo se compone el corpus de CVs según su fuente de origen.

*Tabla 3: Proporción de CVs por cada fuente*

<b>Fuente</b>	<b>Proporción</b>
LinkedIn	82%
Contactos	4%
IA generativa	14%

En total se consiguieron 1120 currículums vitae. Una porción significativa de los mismos, equivalente al 82% del total, se originó de la descarga de perfiles públicos en LinkedIn, seleccionando específicamente aquellos que cumplan los criterios mencionados previamente, garantizando así la calidad del conjunto de datos. Un segmento adicional, correspondiente al 4% de los CVs, fue recabado mediante simples solicitudes directas a contactos y conocidos interesados en colaborar con el proyecto sin importar el perfil o nivel de experiencia de los mismos, proporcionando una capa adicional de autenticidad y diversidad. Finalmente, para complementar la colección y asegurar una representación exhaustiva de competencias y trayectorias profesionales, se generó un 14% de los currículums utilizando tecnologías de inteligencia artificial generativa<sup>1</sup>, los cuales cuentan con datos personales ficticios pero su contenido es indistinguible al de currículums auténticos en términos de calidad y detalle. La razón por la cual se incorpora inteligencia artificial para generar CVs adicionales se debe a la necesidad de incorporar perfiles con experiencia en rubros bien definidos, dado que la trayectoria laboral en muchas personas suele ser variada en cuanto a diferentes áreas, y si bien esto aporta riqueza textual al corpus, puede llegar a dificultar la tarea posterior de evaluar los resultados de los modelos. Incluso para un experto en selección de personal categorizar este tipo de perfiles puede resultar una tarea compleja.

---

<sup>1</sup> En específico, se utilizó ChatGPT para crear CVs ficticios, basados en la estructura de documentos reales empleados a modo de ejemplo.

Este enfoque cuidadoso compuesto por múltiples fuentes, si bien no es el ideal para recabar un volumen masivo de datos dada su característica parcialmente manual, permitió alcanzar una cobertura amplia y representativa, y a su vez adaptar el corpus a las necesidades específicas del caso de estudio en cuestión.

Al igual que en el caso de vacantes laborales, la Tabla 4, permite visualizar una muestra del conjunto de CVs en su estructura original.

*Tabla 4: Muestra del dataset de currículums vitae*

<b>Currículum vitae</b>
Analista de Marketing con más de 5 años de experiencia en estrategias de marke..
Soy desarrolladora BI y analista de datos con una sólida y vasta experiencia en el..
Control de Gestión - Controller - Administración & Finanzas - Sector Seguridad - T..
Ingeniero de Machine Learning con más de 8 años de experiencia en el diseño, d...
Community Manager con experiencia en la gestión de redes sociales y creación d..
Estudiante entusiasta de Recursos Humanos buscando una oportunidad de pasa...
Soy una persona responsable, comprometido con las tareas que se me asignen y..
Analista de Logística con experiencia en la gestión eficiente de la cadena de sumi..
Administrador de Bases de Datos con experiencia en el diseño, implementación y..
Poseo una amplia trayectoria dentro de RR.HH. Me motiva desarrollarme profesio..

## **4.2. Preprocesamiento**

Para abordar la diversidad y riqueza de los CVs, es esencial implementar una serie de procedimientos destinados a limpiar y estructurar los datos brutos, facilitando su posterior análisis y la extracción de información valiosa.

En primer lugar, se realiza la extracción del texto desde la totalidad de los documentos almacenados en formato PDF. Esta tarea se ejecuta mediante el uso de bibliotecas especializadas en Python, capaces de leer y almacenar de manera programática el contenido textual de estos archivos.

Una vez extraído el texto, el siguiente paso es la eliminación de ruido y normalización del texto. La eliminación de ruido se enfoca en descartar información irrelevante o superflua, como pueden ser símbolos, signos de acentuación, números, stopwords<sup>2</sup> o caracteres específicos resultantes del proceso de extracción de texto que no aporta valor al análisis posterior. Este proceso de limpieza es esencial para reducir la complejidad y mejorar la calidad de los datos, facilitando así las etapas analíticas siguientes.

Por otro lado, la normalización implica convertir el texto a un formato estándar, lo que incluye la homogeneización de mayúsculas y minúsculas y la estandarización de variantes de palabras para asegurar la consistencia a lo largo del corpus. Para esto último, si bien existen técnicas populares como la lematización<sup>3</sup> o el stemming<sup>4</sup>, que permiten automáticamente unificar dichas variantes de los términos en raíces comunes, se decidió realizar de manera manual y específica únicamente para casos puntuales, con el objetivo de agilizar el procesamiento y conservar el texto en una forma legible que facilite el análisis exploratorio de los datos, ya que la implementación completa de estas técnicas no ofreció mejoras significativas en los resultados que compensaran el aumento en la complejidad y el tiempo de procesamiento. De esta manera, se procedió a normalizar de manera selectiva casos problemáticos puntuales. A modo de ejemplo, términos como 'Power BI' (herramienta popular en análisis de datos), se transformaron en 'powerbi', para facilitar que los modelos entiendan ese término como algo conjunto y no dos palabras separadas.

---

<sup>2</sup> Es el nombre que reciben las palabras sin significado como artículos, pronombres o preposiciones. La biblioteca de código abierto "Natural Language Toolkit", la cual contiene más de 500 stopwords para el lenguaje español, fue utilizada para identificarlas y removerlas del corpus.

<sup>3</sup> Consiste en reducir una palabra a su forma base o de diccionario, conocida como lema. Por ejemplo, transformar las palabras "comiendo" y "comía" a su forma base, "comer".

<sup>4</sup> Proceso lingüístico más simple que la lematización, que consiste en la reducción de palabras a su raíz, independientemente de si la palabra resultante existe o no. Por ejemplo, transformar la palabra "Pensamiento" y "Pensando" a la raíz "Pensa".

Finalmente, el último paso en el preprocesamiento es la tokenización, que consiste en dividir el texto en unidades más pequeñas, denominadas tokens, que en nuestro caso consisten en palabras, pero puede aplicarse también frases u oraciones. Esta segmentación es fundamental para la aplicación de procesamiento del lenguaje natural, ya que permite un análisis granular del contenido textual.

Si bien las tecnologías más recientes y avanzadas ya no requieren de varios de estos procedimientos, como la normalización y la tokenización para el procesamiento de texto, ya que los tienen incorporados automáticamente, los mismos facilitan el análisis exploratorio, permitiendo la identificación de patrones y tendencias en el uso del lenguaje dentro del corpus.

Las siguientes tablas, Tablas 5 y 6, permiten visualizar las palabras más frecuentes en el corpus, antes y después del procedimiento de limpieza de texto, ilustrando el impacto que esto genera en el análisis.

*Tabla 5: Palabras más frecuentes previo al procedimiento de limpieza de texto*

<b>Palabras</b>	<b>Ocurrencias</b>	<b>Proporción</b>
de	125187	2,16%
y	48586	0.84%
en	33730	0.58%
la	22251	0.38%
-	21532	0.37%

*Nota: de las cinco palabras más frecuentes, cuatro son stopwords y la última es un caracter ("-").*

La Tabla 5, permite ilustrar el efecto de las stopwords en el análisis de texto, entorpeciendo el análisis del contenido en el corpus. Se observa incluso como el caracter correspondiente a un guión ("-"), es considerado como una palabra adicional, lo cual carece de sentido a los fines prácticos. A continuación, se observa la misma tabla luego del procedimiento mencionado (Tabla 6).

*Tabla 6: Palabras más frecuentes luego del procedimiento de limpieza de texto*

<b>Palabras</b>	<b>Ocurrencias</b>	<b>Proporción</b>
experiencia	6989	0.21%
trabajo	4665	0.13%
gestion	4161	0.12%
analisis	3402	0.10%
desarrollo	3263	0.09%

Esto permite apreciar la importancia del procedimiento de limpieza. Analizar las palabras más frecuentes cobra mayor sentido y permite un primer acercamiento a la terminología empleada en el corpus.

### **4.3. Etiquetado manual**

Un desafío significativo en tareas de clasificación o recomendación en este contexto, radica en la dificultad para evaluar la precisión de sus resultados debido a la ausencia de datos etiquetados. Sin un conjunto de datos predefinido que indique qué currículums vitae son efectivamente los más adecuados para cada vacante, resulta complejo establecer un estándar riguroso y objetivo para validar la calidad de las clasificaciones o recomendaciones generadas. Esta limitación puede impedir la realización de una evaluación cuantitativa convencional, como puede ser porcentaje de recomendaciones relevantes, dado que no está definido un criterio para definir qué se considera relevante.

Para resolver este problema, se decidió en primer lugar clasificar manualmente los currículum vitae del corpus en función del área laboral predominante para el candidato. Este proceso se torna particularmente intrincado en casos donde el candidato ha pasado por diferentes áreas a lo largo de su trayectoria laboral, o incluso cuando carece de experiencia laboral, lo que dificulta catalogar a la persona como parte de un sector específico. Además, la gran diversidad de perfiles dentro

del corpus conduce a una amplia gama de sectores laborales posibles. Por lo tanto, luego de haber realizado la clasificación, con el fin de simplificar el trabajo manual y la evaluación de los modelos debido a la gran cantidad de clases resultante, se optó por agrupar los currículums en clases más abarcativas dentro del corpus y recurrir a la categoría “otros” para los casos conflictivos mencionados anteriormente. Las clases resultantes pueden observarse a continuación en la Tabla 7, la cual a su vez explica el criterio utilizado para definir dichas clases.

*Tabla 7: Palabras clave definidas para cada clase*

<b>Clase</b>	<b>% del total</b>	<b>Observaciones</b>
Tecnología	32%	Incluye trabajos relacionados al desarrollo y mantenimiento de soluciones tecnológicas, junto con la gestión de proyectos relacionados.
Administración, contabilidad y finanzas	15%	Vinculado a roles de gestión, análisis financiero/contable y tareas administrativas.
Legales	9%	Asesoramiento legal tanto ejerciendo de manera independiente, como dentro de organizaciones.
Marketing y comercial	9%	Vinculado a la ejecución de estrategias para promover productos o servicios, trabajar en la visibilidad de marca y la generación ventas.
Recursos Humanos	6%	Incluye reclutamiento, formación y desarrollo del personal, además de la gestión de políticas internas y tareas de recursos humanos.
Abastecimiento, logística y operaciones	6%	Comprende tareas relacionadas tanto a la planificación y gestión de la cadena de

		suministro, como a la gestión de las operaciones en organizaciones.
Otros	23%	Conformado tanto por los sectores restantes que se repiten con menor frecuencia, como por aquellos documentos que, por su carácter difuso o heterogéneo, no se ajustan fácilmente a categorías específicas.

En nuestro caso en concreto, se pueden emplear estas clases definidas para evaluar la eficacia de diferentes modelos, tanto en tareas de clasificación como de recomendación, analizando por ejemplo, qué porcentaje de los currículums vitae recomendados coinciden con la clase correspondiente a la oferta de empleo. Cabe destacar que los CVs solo pueden pertenecer a una clase con esta metodología, pero esto no impide que en tareas de recomendación los mismos sean recomendados para vacantes de cualquier índole, permitiendo que todos los CVs sean considerados, incluso aquellos con experiencia heterogénea, dado que este enfoque se centra en la similitud semántica entre textos.

#### 4.4. Conjuntos de entrenamiento y prueba

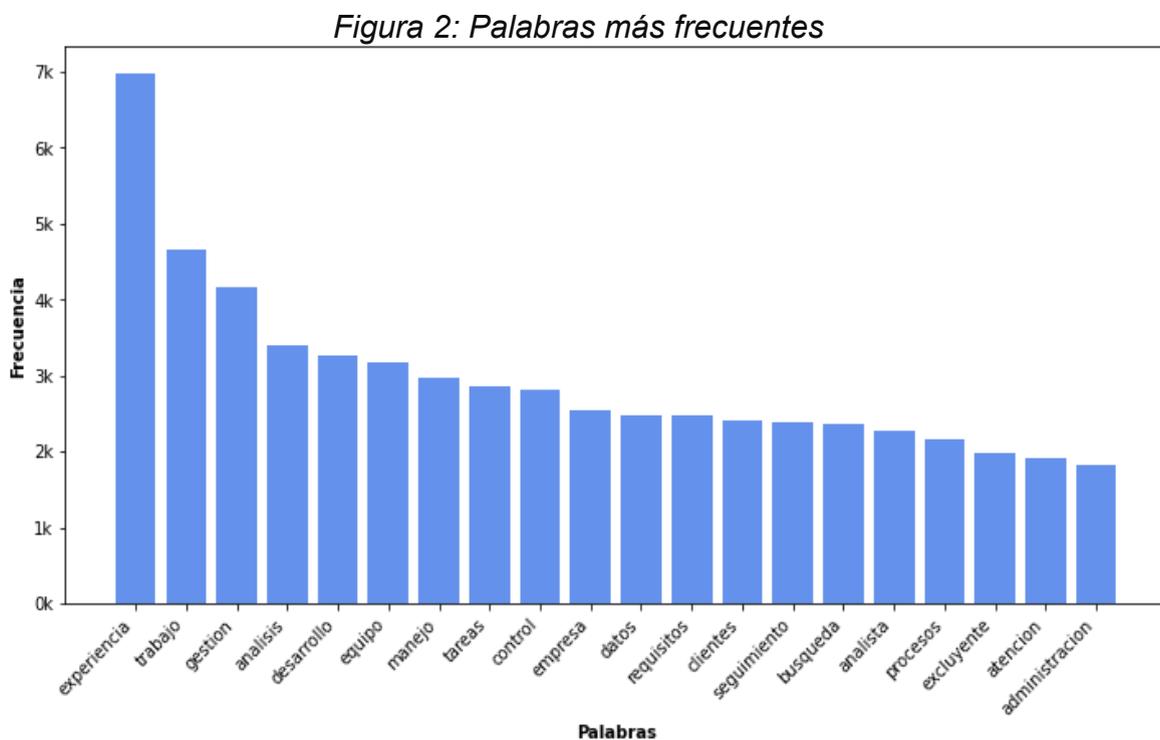
En la fase final de preparación de los datos, se ha establecido que el 30% de los currículums vitae del corpus, seleccionados de forma completamente aleatoria, se destinen al conjunto de prueba. El restante 70% se utilizó para formar el conjunto de entrenamiento. Emplear el conjunto de prueba en el proceso de evaluación, permite medir la capacidad de los modelos de trabajar con CVs a los cuales no había sido expuesto durante el proceso de entrenamiento, reflejando de esta manera situaciones reales de aplicación.

## 5. Análisis exploratorio

### 5.1. La (no) estructura de los datos

El corpus resultante, constituido por 4,960 documentos, los cuales se componen de 1,120 CVs y 3,840 vacantes laborales, acumula un total de 5,784,465 palabras. Esto arroja un promedio de 282 palabras por CV y un promedio de 180 palabras por vacante laboral. Este conjunto diverso es particularmente valioso por su vocabulario especializado en el ámbito laboral, ya que cada uno aporta elementos complementarios para el modelado de texto. Mientras los CVs reflejan las capacidades y experiencias de los potenciales candidatos, las descripciones de las vacantes proporcionan insights sobre las competencias y cualificaciones demandadas en el mercado laboral, así como las responsabilidades asociadas a diferentes roles.

Luego del procedimiento de depuración de datos, y descartando términos comunes y poco relevantes como nombres de meses, ciudades, entre otros, identificamos las diez palabras más frecuentes en el corpus, ilustradas en la Figura 2:



Se puede observar como la palabra “experiencia”, es por amplia diferencia la más utilizada dentro del corpus, lo cual tiene sentido, dado que es parte fundamental del vocabulario propio del mercado laboral, formando parte de los requerimientos de prácticamente la totalidad de los anuncios de vacantes laborales, y por consiguiente, del vocabulario de los CVs utilizado por los candidatos para describir aptitudes propias. Además, se pueden observar términos que también pueden reflejar tendencias en el mercado laboral, como por ejemplo “análisis”, “equipo” o “datos”, los cuales corresponden a habilidades habitualmente demandadas.

Por otra parte, para proporcionar una visión intuitiva de la distribución de las palabras que conforman los documentos del corpus, presentamos una nube de palabras en la Figura 3, que destaca visualmente los términos más relevantes.

Figura 3: Nube de palabras



En este gráfico, se ilustra de una manera más completa el vocabulario propio del corpus, dado que la frecuencia de aparición de los distintos términos que lo

componen se representa por medio del tamaño de cada palabra. De esta manera podemos visualizar las palabras más frecuentes, tal como se observa en la Figura 2, y a su vez incorporar más términos que resultan relevantes. Esta visualización en particular es sumamente útil, no solo para ilustrar el vocabulario del corpus, sino también para corroborar que la composición del mismo sea útil a los fines del presente proyecto. Se puede apreciar cómo el vocabulario utilizado en la imagen está estrechamente relacionado con el ámbito laboral, lo cual sugiere que la calidad del corpus es adecuada y tiene potencial para alcanzar los objetivos propuestos.

## 5.2. Análisis semántico

El procesamiento del lenguaje natural (NLP) enfrenta el desafío de comprender y analizar el lenguaje humano, lo cual implica abordar tanto su estructura (sintaxis) como su significado (semántica). La distinción entre sintaxis y semántica es crucial para entender cómo las computadoras procesan el lenguaje y extraen información útil de los textos, como los currículums vitae en nuestro estudio.

La sintaxis se refiere a la estructura gramatical de las frases; es el esqueleto del lenguaje que determina cómo se organizan las palabras para formar oraciones coherentes y comprensibles. La semántica, por otro lado, se ocupa del significado de las palabras, frases y textos. En NLP, el análisis semántico busca comprender el significado implícito en el texto, lo que va más allá de la mera estructura de las palabras. Esto implica reconocer la intención del emisor, el contexto en el que se usan las palabras y cómo las combinaciones de palabras generan significados que no son inmediatamente evidentes a partir de los elementos individuales. Dado que este trabajo consiste en desarrollar un sistema capaz de comprender el significado y establecer una similitud entre dos textos, como lo son currículum vitae y vacante laboral, se enfoca principalmente en la dimensión semántica.

Un avance crucial en la representación semántica en NLP ha sido el desarrollo de los *embeddings* de palabras. Los embeddings son representaciones vectoriales de palabras que capturan no sólo la similitud semántica entre términos, sino también las relaciones y matices de significado. Cada palabra se mapea a un punto en un espacio de alta dimensión, donde palabras con significados similares están

ubicadas más cerca entre sí. Este enfoque permite a las máquinas procesar el lenguaje de una manera que refleja más fielmente la comprensión humana, facilitando tareas que implican capturar relaciones semánticas complejas. En el análisis de CVs, los embeddings pueden usarse para evaluar la relevancia de las experiencias y habilidades de los candidatos en relación con las descripciones de los puestos, ofreciendo una comparación semántica profunda que va más allá de la coincidencia superficial de palabras.

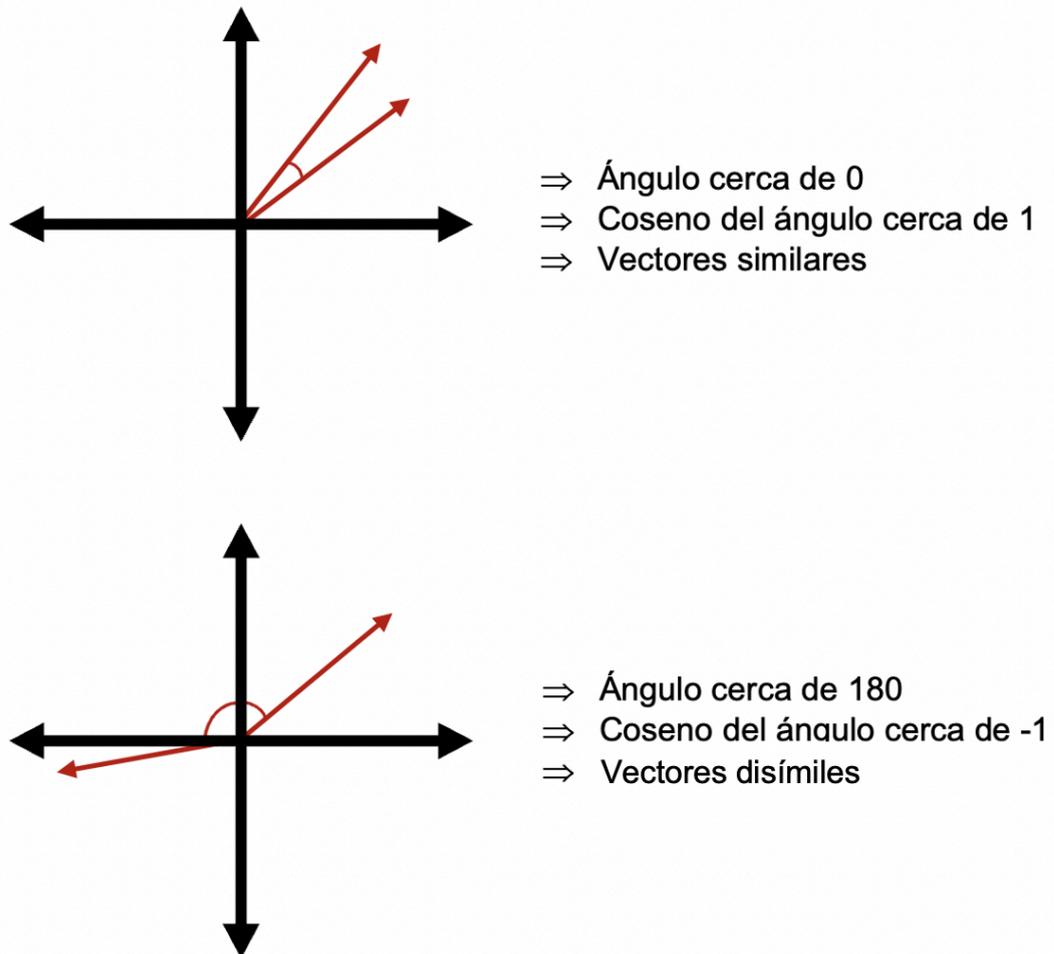
Para explorar las capacidades del análisis semántico en el contexto del procesamiento del lenguaje natural, resulta esencial determinar un método cuantificable para medir la similitud entre documentos, en particular entre currículums vitae y descripciones de puestos. En este trabajo, se adopta la similitud coseno como la métrica principal para este propósito.

La similitud coseno es una medida que consiste en calcular el coseno del ángulo entre dos vectores. En el contexto de NLP, para que pueda ser utilizada, se deben convertir palabras u oraciones en vectores (embeddings). Se calcula a través de la siguiente fórmula:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Esta métrica oscila entre -1 y 1, donde 1 indica una similitud total, 0 ninguna similitud, y -1 una disimilitud total. La Figura 4 a continuación, ilustra visualmente este concepto:

Figura 4: Similitud coseno ilustrada



En la práctica, la similitud coseno se utiliza para evaluar qué tan cercanos están los significados de dos palabras o conjuntos de palabras, basándose en la premisa de que textos semánticamente similares tendrán vectores de palabras que apuntan en direcciones similares en el espacio vectorial.

La elección de la similitud coseno como métrica en este estudio se debe a su efectividad en capturar la proximidad semántica entre conjuntos de palabras, facilitando un enfoque cuantitativo para evaluar el vínculo entre las habilidades y experiencias descritas en los CVs con los requisitos y responsabilidades listadas en las descripciones de las vacantes.

Para ejemplificar cómo los embeddings pueden capturar relaciones semánticas complejas, modelos como Word2Vec<sup>5</sup> se presentan como herramientas valiosas. Al entrenar Word2Vec con nuestro corpus, generamos embeddings que reflejan las interacciones contextuales entre palabras. Esto permite reproducir patrones semántico a través de operaciones aritméticas entre vectores. De esta manera, proponemos resolver la siguiente analogía: ¿que es a ventas, lo que Python (popular lenguaje de programación) es a un programador?

Teóricamente, la relación entre vectores se encuentra a través de la resta de los mismos:

$$f(\text{"python"}) - f(\text{"programador"}) = f(?) - f(\text{"ventas"})$$

Luego, se despeja la incógnita realizando pasaje de términos:

$$f(\text{"python"}) - f(\text{"programador"}) + f(\text{"ventas"}) = f(?)$$

Utilizando el lenguaje Python, implementamos este cálculo para obtener los tres términos con mayor coincidencia para esta analogía, obteniendo los siguientes resultados (Tabla 8):

*Tabla 8: Términos ordenados mayor a menor coincidencia para la analogía*

<b>Palabras</b>	<b>Similitud coseno</b>
relacionamiento	0.48215
métricas	0.44146
negociación	0.44052

Observamos cómo, utilizando el vocabulario provisto en nuestro corpus lingüístico, el modelo es capaz de responder coherentemente a la analogía. Las tres palabras resultantes pueden interpretarse como herramientas propias de la actividad de

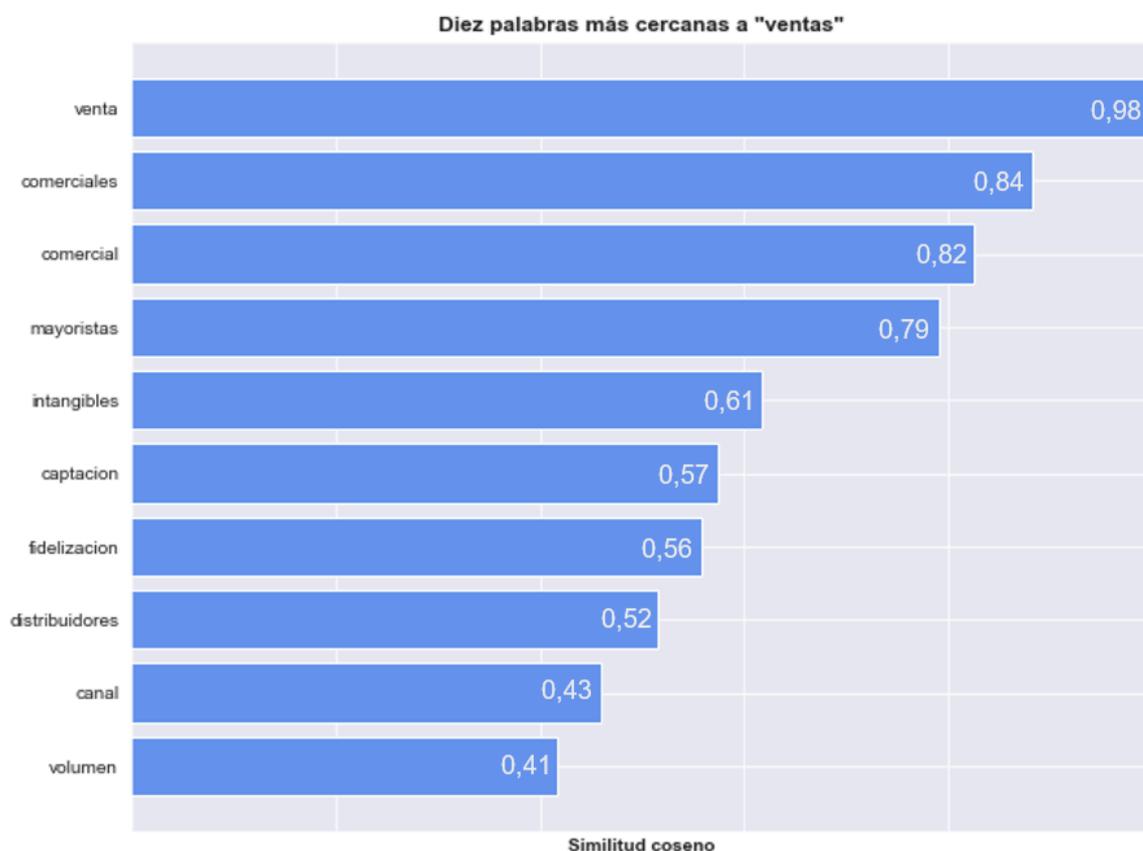
---

<sup>5</sup> Algoritmo desarrollado por Google en el año 2013, que utiliza un modelo de red neuronal que transforma texto en vectores numéricos para aprender asociaciones de palabras relacionadas semánticamente.

ventas, tal cual lo es un lenguaje de programación para un programador. Tomando la palabra con mayor coincidencia, la analogía resultante consistiría en lo siguiente: la capacidad de relacionamiento es una herramienta para los vendedores, equivalente a lo que representa un lenguaje de programación para un vendedor.

Implementando la técnica de Word2vec nuevamente, podemos explorar cómo los embeddings pueden capturar similitudes semánticas utilizando la métrica de similitud coseno. Tomemos, por ejemplo, el análisis de términos cercanos semánticamente a "ventas" (Figura 5) incluyendo las palabras contenidas dentro de nuestro conjunto de documentos. Las palabras que el modelo Word2Vec destaca por su proximidad semántica ofrecen una oportunidad adicional para evaluar la profundidad textual del corpus. Si Word2Vec no logra identificar términos coherentes, esto podría señalar una insuficiencia en el volumen de texto disponible en nuestro conjunto de datos.

*Figura 5: Diez palabras más cercanas a ventas*



Observamos como los resultados obtenidos, son coherentes en el contexto del ámbito laboral, y guardan relación con la terminología habitual propia del rubro de ventas. Podemos distinguir como los términos más cercanos son similares en cuanto a semántica, y los siguientes son actividades, herramientas o términos relacionados a la actividad. A modo de ejemplo, la aparición del término 'intangibles' se debe a que es un rubro de ventas muy frecuente tanto en las vacantes laborales como en la experiencia de candidatos, y se refiere a la comercialización de productos o servicios que no tienen una forma física concreta.

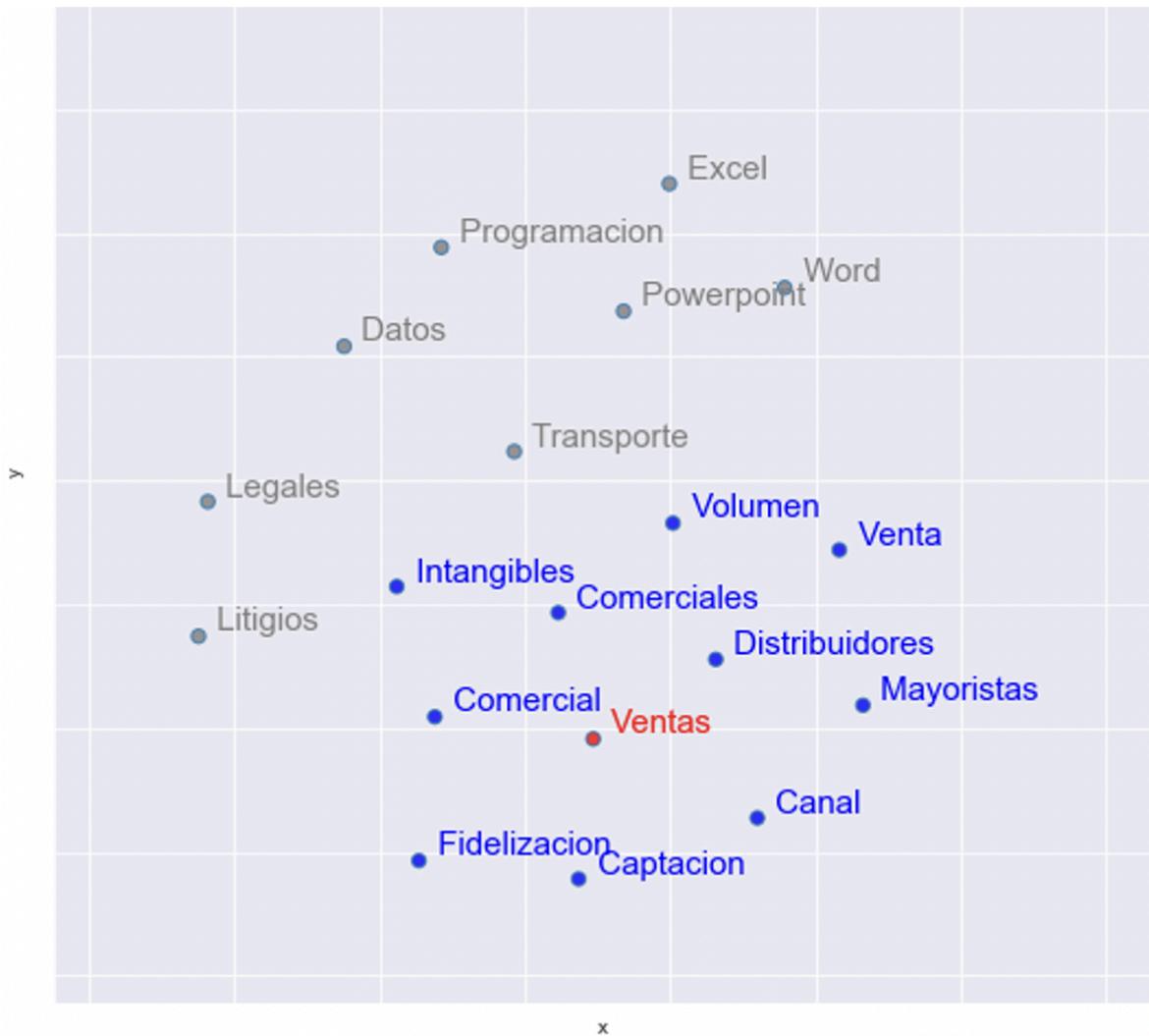
Además de analizar la similitud semántica entre términos, los embeddings generados utilizando Word2Vec nos permiten profundizar en la visualización de estas relaciones a través del algoritmo t-SNE<sup>6</sup>, el cual posibilita reducir su dimensionalidad. El objetivo es transformar los vectores de alta dimensión en representaciones simplificadas bidimensionales, permitiéndonos observar patrones subyacentes entre los vectores en un espacio comprensible.

Para observar como una serie de palabras se agrupan en el espacio vectorial utilizando esta técnica, examinaremos en el siguiente gráfico (Figura 6) las relaciones entre una palabra de consulta, en este caso el término 'ventas' (marcado en rojo), sus palabras más similares en el modelo (en azul), y otras palabras del vocabulario seleccionadas al azar (en gris).

---

<sup>6</sup> t-SNE (T-distributed Stochastic Neighbor Embedding) es un algoritmo diseñado para la visualización de conjuntos de datos de alta dimensionalidad. Los gráficos t-SNE modelan cada objeto de alta dimensionalidad por un punto de 2 o 3 dimensiones, facilitando la interpretación visual.

Figura 6: Visualización t-SNE para el término "ventas"



Esta representación simplificada del espacio vectorial permite visualizar la cercanía semántica entre el mencionado subconjunto de términos dentro del corpus. Las palabras que en el ámbito laboral guardan mayor relación con el término ventas se ubican alrededor del mismo, e incluso términos no directamente relacionados con ventas pero que guardan relación entre sí, como 'Excel', 'Word', 'Powerpoint' y 'Programación', que son herramientas informáticas, o 'Legales' y 'Litigios', se pueden observar cercanos semánticamente.

## 6. Metodología

### 6.1. Lbl2Vec: Clasificación no supervisada

El enfoque metodológico adoptado para el sistema propuesto, consiste en plantear tanto la tarea de clasificación, como la de recomendación, como problemas de clasificación de textos no supervisada basada en similitud semántica. Esta técnica implica evaluar la similitud entre una serie de temáticas definidas previamente, y el conjunto de documentos disponibles en el corpus, asignando a cada documento la categoría con la que muestra mayor similitud.

Para esta tarea se emplea el algoritmo de Lbl2Vec, el cual permite clasificar documentos en base a su similitud semántica con una serie de temas definidos por palabras clave especificadas de antemano. La idea fundamental del algoritmo consiste en que varios términos clave semánticamente similares pueden representar un tema específico. El proceso de entrenamiento consiste en lo siguiente:

- A. Se definen manualmente palabras clave para cada temática de interés. Para esto, se necesita conocimiento del dominio para definir palabras clave que describan los temas y que sean semánticamente similares entre sí en ese contexto.

A continuación se presentan las palabras clave definidas para nuestras clases en la Tabla 9:

*Tabla 9: Palabras clave definidas para cada clase*

Clase	Palabras Clave
Tecnología	software, desarrollo, desarrollador, api, cloud, engineer, scientist, python, java, javascript, programador, script, programación, scrum, sql, informática, it, developer, bi, base, datos, science, funcional, análisis, data, flutter, tecnologías, azure, implementación

Administración, contabilidad y finanzas	administración, presupuestos, finanzas, contabilidad, contable, control, controles, financiero, riesgos, contador, publico, gestion, tesoreria, costos, ingresos, auditoria, financiero, fiscal, impuestos, facturas, administrativo, fiscales, estados, contables, procesos, trámites, auditor
Legales	leyes, jurídico, litigios, pliegos, legal, legales, abogado, normativas, legislación, contratos, penal, demanda, defensa, procurador, paralegal
Marketing y comercial	publicidad, propuestas, competencia, seo, sem, comerciales, marketing, digital, comercial, marca, campanas, redes, sociales, mercado, social, consumidor, ventas, clientes, ejecutivo, cuentas, account, comunicación, negociación, venta, vendedor, activación, cartera, community
Recursos Humanos	reclutamiento, entrevistas, capacitación, beneficios, compensación, people, recursos, humanos, legajos, payroll, sueldos, selección, candidatos, recruiter, personal
Abastecimiento, logística y operaciones	compras, inventario, logística, cadena, planificación, stock, transporte, abastecimiento, operaciones, proveedores, distribución, entrega, materiales, suministros, insumos, supply, chain, entrega, disponibilidad, ruteo, almacenamiento

B. Se crean embeddings de documentos empleando la totalidad del corpus de texto utilizando alguna técnica de modelado, como lo es Word2Vec por ejemplo, pero aplicada a documentos en lugar de palabras.

En la Figura 7 se ilustra gráficamente como los documentos se ubican en el espacio vectorial cerca de otros documentos similares.

Figura 7: Ilustración CVs ubicados en el espacio vectorial



C. Una vez representados los documentos y palabras del corpus en un espacio vectorial, el objetivo del algoritmo es aprender vectores de etiquetas a partir de las palabras clave previamente definidas para cada tema e identificar las similitudes.

En la Figura 8, cada color representa una categoría diferente.

Figura 8: Ilustración CVs ubicados en el espacio vectorial



D. Se detectan e Ignoran los vectores de documentos atípicos para cada tema, considerados *outliers*, en la etapa de entrenamiento.

Los mismos se representan en color rojo en la Figura 9.

Figura 9: Ilustración CVs ubicados en el espacio vectorial



E. Se calcula el centroide de los vectores de documentos como el vector de etiqueta para cada tema.

Los puntos en la Figura 10 representan ese vector de etiqueta.

Figura 10: Ilustración CVs ubicados en el espacio vectorial



F. Se calculan las similitudes, utilizando la métrica de similitud coseno, entre el vector de etiqueta y el vector de documento para cada uno de los mismos en el conjunto de datos. Los documentos se clasifican en el tema con la mayor similitud entre el vector de etiqueta y el vector de documento.

Esto se representa visualmente en la Figura 11.

Figura 11: Ilustración CVs ubicados en el espacio vectorial



Dado que el algoritmo determina la pertenencia de documentos a temas sin la necesidad de observar etiquetas reales para cada documento indicando la respectiva clase, se trata de un algoritmo de aprendizaje no supervisado.

## 6.2. Hiperparámetros

Previo al entrenamiento del modelo, el algoritmo de Lbl2vec contiene una serie de variables de configuración que deben definirse manualmente para controlar el proceso de aprendizaje. Estas variables, o hiperparámetros, pueden optimizarse de acuerdo a los resultados obtenidos con cada iteración para encontrar una combinación óptima de los mismos.

Dentro del algoritmo de Lbl2vec, los hiperparámetros que fueron seleccionados para ser optimizados, en función de cuán significativa es su influencia en los resultados, son los siguientes:

- “epochs”: Número de iteraciones sobre el corpus.

- “similarity\_threshold”: Solo se utilizan documentos con una similitud mayor a este umbral con respecto a las palabras clave para calcular los embeddings de las etiquetas.
- “min\_num\_docs”: Número mínimo de documentos que se utilizan para calcular los embeddings de etiquetas. Agrega documentos hasta que se cumpla el requisito si el umbral de similitud es demasiado restrictivo.

### 6.3. Modelos a implementar

En el marco del algoritmo Lbl2Vec, se pretende implementar el modelo Doc2Vec y evaluar su eficacia en comparación con modelos pre entrenados Transformers. Doc2Vec es una extensión del modelo Word2Vec mencionado anteriormente, que no solo aprende representaciones vectoriales de palabras, sino también de documentos enteros o párrafos. Esto permite que el modelo genere un vector para cada documento que capta su contexto y contenido semántico global. La elección de Doc2Vec se debe a su simplicidad y menor costo computacional, características que lo hacen atractivo para entornos con recursos limitados. A pesar de su arquitectura más sencilla, Doc2Vec ha demostrado ser efectivo en capturar la esencia semántica de documentos extensos, lo que lo convierte en un candidato viable para tareas relacionadas con similitud de textos.

Los Transformers, por otro lado, representan una arquitectura más avanzada y reciente que Doc2Vec, especialmente eficaz para capturar las sutilezas y contextos complejos del lenguaje, ya que rastrean las relaciones entre los componentes secuenciales del texto, como las palabras de una oración. Utilizando el framework ‘Sentence-Transformers’<sup>7</sup>, se adaptan los modelos Transformers pre entrenados para generar embeddings de oraciones o documentos completos, optimizados para tareas semánticas específicas.

La comparación entre Doc2Vec con modelos Transformers, conocidos por su capacidad más avanzada para comprender matices semánticos complejos, nos

---

<sup>7</sup> Sentence Transformers es una biblioteca basada en los modelos de Transformers que facilita la generación de embeddings de oraciones completas, lo cual es útil para tareas de NLP que requieren comparar semánticamente fragmentos largos de texto.

permitirá determinar hasta qué punto podemos acercarnos a la performance de estas arquitecturas más sofisticadas con menores recursos. Los Transformers, a pesar de su destacada capacidad de análisis semántico, presentan desafíos significativos en términos de requerimientos computacionales y tiempos de entrenamiento, especialmente para conjuntos de datos de gran volumen.

Nuestro objetivo es investigar si utilizando el framework de Lbl2Vec, con su enfoque más directo y eficiente, se pueden ofrecer resultados comparables en la tarea de asociar currículums vitae con descripciones de vacantes laborales. Dado que en esta comparativa no solo se considera la precisión y relevancia de las recomendaciones generadas por ambos modelos, sino que también la viabilidad operativa de su implementación en términos de recursos computacionales y escalabilidad.

#### 6.4. Método de evaluación

Para evaluar la eficacia de nuestros modelos, utilizaremos el F1 score como métrica principal, la cual consiste en la media armónica entre *precision*<sup>8</sup> y *recall*<sup>9</sup>, permitiendo una representación equilibrada del rendimiento de los modelos.

La fórmula del F1 score expresada en términos de *precision* y *recall*, es la siguiente:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

La misma fórmula expresada en términos de Verdaderos Positivos (TP), Falsos Negativos (FN) y Falsos Positivos (FP), consiste en lo siguiente:

$$F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

---

<sup>8</sup> Popular métrica en tareas de clasificación que se refiere a la proporción de predicciones correctas (positivas) entre todas las predicciones clasificadas como positivas. En otras palabras, mide la exactitud de las predicciones positivas del modelo

<sup>9</sup> Métrica que mide la proporción de verdaderos positivos identificados correctamente por el modelo en relación con el total de casos que son verdaderamente positivos en la realidad. Es decir, indica cuántas de las instancias reales positivas fueron captadas por el modelo.

Además, se pondrá un foco especial en discernir entre los F1 score de promedio 'micro' y promedio 'macro' para una comprensión más detallada de los resultados. El promedio 'micro' calcula un F1 score global a partir de la suma total de Verdaderos Positivos (TP), Falsos Negativos (FN) y Falsos Positivos (FP) de todas las clases, insertando estos valores en la ecuación del F1 score para obtener el resultado final.

Por otro lado, el promedio 'macro' calcula el F1 score de forma independiente para cada clase y luego toma el promedio de estos, sin tener en cuenta el balance entre clases. Esto significa que cada clase contribuye por igual al F1 score final, sin importar su tamaño o desbalance.

En resumen, el promedio macro le da la misma importancia a cada categoría, mientras que el promedio micro, le da igual importancia a cada muestra, por lo que este último en clasificación multi clase es equivalente a la métrica de 'accuracy'<sup>10</sup>.

## 7. Implementación en clasificación

Tanto Transformers, como Lbl2Vec se emplearán para agrupar los currículums vitae del corpus, basándose en las palabras clave específicas previamente asignadas a cada grupo definido manualmente. La clasificación se basa en la similitud entre los embeddings, asignando cada CV a la clase cuyas palabras clave descriptivas obtengan la mayor similitud coseno. Como se mencionó anteriormente, dado que en este proceso los modelos no tienen acceso a las etiquetas reales de cada CV, estamos ante una estrategia de clasificación no supervisada. Posteriormente, utilizaremos las etiquetas manualmente definidas de antemano para determinar la precisión de esta clasificación no supervisada, evaluando su efectividad.

En la Tabla 10, se ilustra el procedimiento descrito de clasificación.

---

<sup>10</sup> Métrica que mide la proporción de predicciones correctas (tanto positivas como negativas) realizadas por el modelo sobre el total de casos examinados. Es decir, evalúa el desempeño general.

*Tabla 10: Ejemplo de cómo funciona la clasificación basada en similitud coseno*

CV	Categoría más similar	Mayor puntaje de similitud	Tecnología	RRHH	Legales
A	Tecnología	0.92	0.92	0.56	0.65
B	Legales	0.84	0.56	0.43	0.84
C	RRHH	0.79	0.54	0.79	0.34

*Nota: la tabla real resultante de este proceso contiene seis columnas en total, una para cada una de las clases en cuestión (excluyendo la clase “otros”). En este caso se visualizan sólo tres a modo ilustrativo.*

Se observa cómo se computa la similitud coseno para cada CV frente a todas las clases definidas de antemano, seleccionando aquella que obtiene mayor puntaje de similitud.

Cabe destacar que todo el corpus de entrenamiento se utiliza para entrenar los embeddings, pero luego solo la clasificación se realiza sobre los currículums manualmente categorizados, tanto para el conjunto de entrenamiento como para el de testeo, de manera tal que los resultados puedan ser evaluados en base a la precisión de las clasificaciones.

Para simplificar la evaluación de los modelos en clasificación, CVs pertenecientes a la clase denominada ‘otros’, si bien forman parte del entrenamiento y generación de embeddings, serán descartados en la etapa de clasificación, dado que es una categoría muy abarcativa y ambigua, que no puede definirse con solo una serie de palabras clave.

## 7.1. Transformers

En primer lugar, a través de la plataforma HuggingFace<sup>11</sup>, se seleccionaron algunos de los modelos pre entrenados más populares, priorizando aquellos adaptados para trabajar con textos en español y capaces de llevar a cabo la tarea de detección de similitud semántica entre textos.

El proceso inicia con la creación de embeddings para las distintas clases a partir de las palabras clave seleccionadas, seguido por la generación de embeddings para cada uno de los currículums vitae en el corpus, trabajando por separado los conjuntos de entrenamiento y prueba. Tal como se mencionó previamente, para cada CV se calcula la similitud coseno frente a cada clase, para asignar cada CV a la clase más adecuada.

La tabla de resultados (Tabla 11) presentada a continuación detalla el desempeño de cada uno de los modelos Transformers pre entrenados implementados, tanto en entrenamiento como en testeo.

*Tabla 11: Tabla de resultados aplicando transformers para los conjuntos de entrenamiento y testeo.*

Modelo	Train		Test	
	F1 score Micro	F1 score Macro	F1 score Micro	F1 score Macro
sentence-transformers/LaBSE	0.79	0.69	0.77	0.70
intfloat/multilingual-e5-small	0.71	0.63	0.71	0.64
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	0.70	0.59	0.67	0.58
sentence-transformers/distiluse-base-multilingual-cased-v2	0.56	0.48	0.51	0.50

<sup>11</sup> Conocida plataforma que ofrece una amplia biblioteca de modelos pre entrenados de NLP.

hackathon-pln-es/paraphrase-s panish-distilroberta	0.64	0.47	0.66	0.55
sentence-transformers/distiluse -base-multilingual-cased-v1	0.51	0.47	0.48	0.46

Se observa que aun en el modelo con mejor performance, existe un desbalance entre el F1 score micro y macro, lo cual significa que no logra replicar el mismo nivel de eficacia de manera uniforme en todas las categorías, funcionando de mejor manera para algunas en específico. Cabe destacar que se están usando distintos modelos que han sido pre entrenados con grandes conjuntos de datos de diversa índole, y estos modelos no han sido especialmente ajustados a nuestro conjunto de datos en particular.

## 7.2. Lbl2Vec

Siguiendo el mismo procedimiento, pero esta vez aplicando el algoritmo Lbl2Vec, empleamos el modelo Doc2Vec para realizar las clasificaciones. A diferencia de los modelos Transformers pre entrenados, en este caso entrenamos el modelo Doc2Vec desde cero, utilizando nuestro propio corpus de datos de entrenamiento. Esto es posible debido al reducido costo computacional que esto implica, en comparación con el caso de los transformers.

A continuación, se detallan los resultados obtenidos en la Tabla 12:

*Tabla 12: Tabla de resultados aplicando Lbl2Vec para los conjuntos de entrenamiento y testeo.*

Modelo	Train		Test	
	F1 score Micro	F1 score Macro	F1 score Micro	F1 score Macro
Doc2vec	0.866	0.865	0.844	0.841

Esta metodología no solo es más eficiente computacionalmente, completando el entrenamiento y la inferencia en solo unos segundos, sino que además se puede observar que presenta un rendimiento superior en tareas de clasificación para este caso específico. Además, observamos una mayor proximidad entre los F1 scores micro y macro. Esto indica que el modelo es efectivo tanto en la clasificación precisa de casos positivos a través de todo el conjunto de datos como en el manejo equitativo de las distintas clases. Esta característica es crucial en aplicaciones prácticas donde es fundamental que el modelo no solo exhiba un buen desempeño general, sino que también opere de manera justa y equitativa, especialmente en entornos con una significativa variabilidad en el tamaño de las clases, como ocurre en el contexto de recursos humanos donde algunos perfiles laborales son más predominantes que otros.

### **7.2.1. Optimización de hiperparámetros en Lbl2Vec**

A través de iteraciones con diversas combinaciones de hiperparámetros se puede alcanzar una combinación óptima que maximice la métrica de resultados. Para este proceso, establecimos rangos de valores para cada hiperparámetro, y se seleccionaron de manera aleatoria y automática, una combinación de valores distintos para cada hiperparámetro dentro de los rangos definidos.

Los rangos de valores a su vez están definidos en función de iteraciones previas y ajustes sucesivos que permitieron acotarlos lo máximo posible. Para los hiperparámetros en cuestión, los rangos son los siguientes (Tabla 13):

*Tabla 13: Tabla de rangos de valores posibles definidos para cada hiperparámetro.*

Hiperparámetro	valor mínimo	valor máximo
similarity_threshold	0.6	0.8
min_num_docs	40	60
epochs	30	40

Para cada combinación de hiperparámetros se entrena el modelo nuevamente y se computa la métrica de resultados, en este caso el F1-score, usando el conjunto de datos de prueba.

En este caso se realizaron 30 iteraciones. A continuación se visualizan las diez con mejores resultados (Tabla 14):

*Tabla 14: Top 10 iteraciones con mejores resultados ordenados de mayor a menor F1-Score*

#	similarity_threshold	min_num_docs	epochs	F1-Score Macro
1	0.649829956	56	37	0.873418622
2	0.616926324	57	39	0.872422584
3	0.601171517	60	38	0.86985091
4	0.622687125	53	36	0.864486772

5	0.693731458	52	40	0.86351283
6	0.660569525	55	37	0.860171657
7	0.682127278	57	39	0.858973868
8	0.601856717	60	38	0.858136411
9	0.645502404	58	37	0.85514067
10	0.755848245	46	31	0.854435514

Se puede observar cómo el proceso de optimización de hiperparámetros permite alcanzar aún mejores resultados, dado que la métrica de evaluación resultante en la combinación óptima, alcanza el valor de 0.87 en F1 Score macro, superando al valor de 0.84 obtenido previamente, en 3 puntos porcentuales.

## 8. Implementación en recomendación

Siguiendo el enfoque establecido en la clasificación, aplicaremos tanto Transformers como Lbl2Vec en el proceso de generar recomendaciones para vacantes laborales específicas, simulando un escenario real de selección de personal. Seleccionaremos seis vacantes representativas, cada una correspondiente a un rubro diferente presente en nuestro corpus de CVs (excluyendo el rubro 'otros'), para ilustrar la aplicabilidad práctica de nuestros modelos en procesos de recomendación.

Para cada una de estas vacantes, tal como se realizó en clasificación, definiremos manualmente un conjunto de palabras clave que capturen la esencia y los requisitos del puesto. Posteriormente, procederemos a entrenar los modelos con el corpus de entrenamiento completo, excluyendo estas seis vacantes en cuestión, y centrando la recomendación exclusivamente en las mismas.

El objetivo es generar cinco recomendaciones por vacante utilizando únicamente el conjunto de datos de prueba, es decir candidatos a los que el modelo no ha tenido acceso durante el entrenamiento. Estas recomendaciones estarán basadas en la proximidad semántica entre los embeddings de los currículums vitae y las palabras clave descriptivas de las vacantes. La evaluación de la relevancia de estas recomendaciones se realizará considerando la congruencia entre el sector de la vacante y el de los currículums vitae recomendados. Por ejemplo, si una vacante pertenece al sector tecnológico y se recomienda un CV también perteneciente a la clase de tecnología, esta recomendación se considerará relevante. Esta medida de relevancia nos proporcionará una métrica clara para poder valorar la efectividad de los modelos en el contexto de recomendaciones de empleo.

Las vacantes seleccionadas para realizar la prueba serán las siguientes (Tabla 15):

*Tabla 15: Vacantes seleccionadas para simular un escenario real de recomendación*

<b>Puesto</b>	<b>Clase</b>	<b>Contenido</b>
Data Scientist Ssr.	Tecnología	¿Te interesa una posición com..
IT recruiter	Recursos Humanos	Estamos en la búsqueda de u...
Abogado Jr.	Legales	Importante empresa de Retai....
Profesional en tesorería y finanzas	Administración, contabilidad y finanzas	Nuestra distinguida empresa, líder en iluminación, se encue...
Analista de compras y logística	Abastecimiento, logística y operaciones	Estamos buscando un Analista de Compras y Logística para u..
Analista de marketing	Marketing y comercial	Estamos buscando Marketing...

En este escenario, hemos elegido el modelo de Transformers que mostró el mejor desempeño en la tarea de clasificación para emplearlo en el proceso de

recomendación y llevar a cabo la comparativa frente a la técnica de Lbl2vec. Posteriormente, el procedimiento es idéntico al utilizado en la clasificación. Se computan los embeddings tanto para las vacantes, por medio de las palabras clave definidas, como para cada uno de los CVs disponibles y luego genera un ranking de candidatos para cada vacante empleando la métrica de similitud coseno. Finalmente, se seleccionan los cinco primeros de cada ranking, obteniendo los candidatos recomendados ordenados de mayor a menor similitud.

Evaluando la relevancia para la totalidad de las recomendaciones en función de la congruencia entre la clase de la vacante y la clase de los currículums vitae recomendados, los resultados obtenidos son los siguientes:

*Tabla 16: Resultados en recomendación*

<b>Modelo</b>	<b>F1 score Macro</b>
Doc2Vec (Lbl2Vec)	0.93
sentence-transformers/LaBSE	0.90

*Nota: utilizamos F1 score macro únicamente, ya que al tratarse de una menor cantidad de documentos no hay diferencia significativa entre micro y macro.*

En primera instancia, a través de las métricas resultantes, se observan en general mejores resultados en el contexto de recomendación que los obtenidos en clasificación y el rendimiento de la técnica Lbl2vec supera nuevamente al obtenido utilizando Transformers. No obstante, debe tenerse en cuenta que en clasificación se ocupaba la totalidad de los CVs disponibles para clasificar, mientras que en este caso solo se están incluyendo los 30 CVs recomendados. Igualmente, esta métrica es útil con fines comparativos, y sugiere además que la gran mayoría de las recomendaciones son en principio relevantes, teniendo en cuenta el criterio explicado anteriormente, es decir congruencia entre la clase de CV y vacante.

Para realizar un análisis más exhaustivo sobre las recomendaciones realizadas, que vaya más allá del criterio de relevancia básico basado en la congruencia entre clases de documentos, se puede examinar la tabla resultante del proceso de

recomendación para una vacante específica a modo de ejemplo, para cada modelo. Esto puede permitir analizar visualmente en detenimiento los perfiles específicos recomendados por cada modelo para identificar diferencias.

En la Tabla 17, a continuación, se pueden observar las recomendaciones correspondientes a una vacante utilizando Transformers.

*Tabla 17: Muestra de la tabla resultante en el proceso de recomendación para una vacante a modo de ejemplo usando transformers*

#	Vacante	Clase - Vacante	Perfil - CV recomendado	Clase - CV	Similitud coseno
1	Data Scientist	Tecnología	Analista de datos	Tecnología	0.51
2	Data Scientist	Tecnología	Analytics consultant	Tecnología	0.48
3	Data Scientist	Tecnología	Data scientist	Tecnología	0.47
4	Data Scientist	Tecnología	Analista de datos	Tecnología	0.46
5	Data Scientist	Tecnología	Analista de datos	Tecnología	0.45

Podemos observar como las recomendaciones son relevantes en términos de congruencia con la clase, pero el perfil de los CVs recomendados no siempre coincide con lo que la vacante busca. Por ejemplo, llama la atención que un perfil 'Data Scientist' sea recomendado en tercer lugar, siendo superado por otros que en principio, si bien están relacionados, tienen menor similitud. Esto no necesariamente es un error, y requiere de un análisis en detalle del contenido de cada CV, ya que las especificaciones de la vacante pueden coincidir con el contenido del CV recomendado, independientemente del rótulo que el mismo tenga. Además, tal como sucede en escenarios reales, los CVs a disposición pueden no coincidir directamente con el perfil buscado, y en esas situaciones se requiere buscar lo más similar posible al escenario ideal.

En la Tabla 18 se puede observar nuevamente la tabla de recomendación resultante para la misma vacante de ejemplo, esta vez utilizando Lbl2vec.

*Tabla 18: Muestra de la tabla resultante en el proceso de recomendación para una vacante a modo de ejemplo usando Lbl2Vec*

#	Vacante	Clase - Vacante	Perfil - CV recomendado	Clase - CV	Similitud coseno
1	Data Scientist	Tecnología	Data Scientist	Tecnología	0.91
2	Data Scientist	Tecnología	Data Scientist	Tecnología	0.87
3	Data Scientist	Tecnología	Data scientist	Tecnología	0.86
4	Data Scientist	Tecnología	Analista de datos	Tecnología	0.85
5	Data Scientist	Tecnología	Data developer	Tecnología	0.85

Se puede observar como el modelo Lbl2vec alcanza mejores resultados también para tareas de recomendación. Comparando visualmente las tablas resultantes para la vacante utilizada como ejemplo, las tres primeras recomendaciones corresponden a perfiles 'Data Scientist', lo cual coincide con el título de la vacante. Tal como se explicó anteriormente, esto no necesariamente implica mejores resultados, ya que para ello hay que analizar el contenido de los CVs de manera detallada, pero si se trata de un indicio que puede sugerir que se trata de recomendaciones más relevantes. Además, al analizar la métrica de evaluación para la totalidad de las recomendaciones, efectivamente los resultados evidencian que la performance es matemáticamente superior en términos de relevancia.

## 9. Aplicación práctica en organizaciones

### 9.1. Rango de aplicabilidad

Los modelos desarrollados en este proyecto, tienen un amplio espectro de utilidad y pueden ser valiosos para distintas clases de entidades. A continuación se detallan los distintos tipos de organizaciones a los cuales este proyecto apunta:

- **Sistemas de Seguimiento de Candidatos (ATS):** Los ATS son plataformas ampliamente utilizadas por grandes empresas para gestionar grandes volúmenes de aplicaciones. Estos sistemas pueden integrar modelos de NLP para mejorar la precisión en la filtración, clasificación y recomendación de candidatos, facilitando la identificación rápida de los perfiles más prometedores y mejorando la eficiencia del proceso de selección.
- **Consultoras de Recursos Humanos y reclutadores independientes:** Dado que son profesionales en parte dedicados a entrevistar y analizar un gran volumen de candidatos para distintas organizaciones, los mismos pueden beneficiarse enormemente de la implementación de este tipo de tecnologías. Pueden utilizar estos modelos tanto como para realizar preselecciones de candidatos, como para realizar análisis profundos sobre el mercado laboral.
- **Empresas con sistema ATS propio:** Aquellas empresas que han desarrollado o adaptado sus propios sistemas para gestionar los procesos de reclutamiento pueden integrar estos modelos para mejorar su funcionamiento y optimizar procesos.
- **Pequeñas y medianas empresas (PYMEs):** Aunque las PYMEs pueden carecer de los recursos de las grandes organizaciones, la implementación de manera local de estos modelos puede generar soluciones rentables para optimizar sus procesos de reclutamiento, reduciendo la carga laboral operativa de los analistas de recursos humanos, y permitiendo tomar decisiones más informadas.

## 9.2. Impacto en organizaciones

La adopción de estos modelos avanzados ofrece un considerable potencial para transformar positivamente las organizaciones, particularmente en procesos de reclutamiento. Su aplicación práctica puede significativamente mejorar y eficientizar una variedad de procesos organizacionales, desde la selección y contratación de personal hasta la gestión y desarrollo del talento existente.

Algunos de los beneficios proporcionados por la implementación de dichas tecnologías son los siguientes:

- **Mayor eficiencia:** La automatización del proceso de revisión de currículums vitae, clasificación y recomendación de candidatos permite a los equipos de recursos humanos procesar un mayor volumen de aplicaciones más rápidamente. Esto reduce significativamente el tiempo necesario para identificar candidatos potencialmente adecuados, acelerando todo el ciclo de contratación y permitiendo liberar recursos para que puedan ser redirigidos hacia iniciativas más estratégicas.
- **Mayor precisión en la selección:** Los modelos pueden realizar una preselección de candidatos, basado en reglas definidas de negocio, para permitir a los analistas de recursos humanos concentrar sus esfuerzos en analizar con mayor detalle los candidatos preseleccionados.
- **Reducción de sesgos:** Los modelos tienen el potencial de mitigar el impacto de los sesgos humanos inconscientes en el proceso de selección, especialmente en las etapas iniciales de revisión. Al utilizar datos y modelos analíticos para tomar decisiones, se puede favorecer una evaluación más objetiva y equitativa. Sin embargo, es importante reconocer que los algoritmos también pueden reflejar sesgos presentes en los datos de entrenamiento, por lo que es crucial implementar y supervisar adecuadamente estas tecnologías.
- **Decisiones basadas en datos:** Con la ayuda de estos modelos, las decisiones de contratación se vuelven más objetivas y basadas en datos. Esto permite a los analistas de recursos humanos tomar decisiones informadas y

respaldadas por análisis concretos, asegurando que las decisiones estén alineadas con las estrategias definidas y objetivos organizacionales.

- Reducción de costos: La optimización del proceso de contratación a través de la automatización y una selección más precisa puede conducir a una reducción significativa en los costos asociados al reclutamiento. Menos tiempo dedicado por parte de los reclutadores y menor número de contrataciones no exitosas, son ejemplos claros de cómo se pueden lograr ahorros significativos en el proceso de selección.
- Democratización de las oportunidades: La capacidad de las máquinas para analizar todas las postulaciones recibidas elimina las limitaciones físicas y temporales a las que un humano estaría sujeto. Esto asegura que cada candidato tenga una oportunidad justa de ser considerado, basado en sus méritos y compatibilidad con el puesto, sin restricciones de capacidad de revisión. Este enfoque equitativo y exhaustivo contribuye a un proceso de selección más justo y abierto.
- Mejora de la Experiencia del Candidato: Procesos de selección ágiles y exhaustivos, permiten ofrecer retroalimentación oportuna a los candidatos en cada etapa del proceso, mejorando significativamente su experiencia general. Una experiencia positiva del candidato es esencial para la reputación de la empresa y puede impactar favorablemente tanto en la cantidad de postulaciones recibidas como en la disposición del candidato a aceptar una oferta de empleo.

## 10. Conclusiones

### 10.1. Puntos destacados

Los resultados más relevantes obtenidos a lo largo de este trabajo se pueden resumir en los siguientes puntos:

- Doc2Vec, utilizando la técnica de Lbl2Vec, supera el rendimiento de modelos pre entrenados Transformers, tanto en clasificación como en recomendación.

- Esta técnica implica un costo computacional considerablemente bajo, ya que solo requiere unos segundos tanto para el entrenamiento como para la inferencia, lo que la hace aplicable en una variedad de contextos.

## 10.2. Futuras mejoras

Para seguir perfeccionando la eficacia y la precisión de los modelos utilizados en este proyecto, se han identificado varias áreas de mejora que podrían ser implementadas en el futuro. Estas mejoras permitirán ampliar las capacidades del sistema y a su vez enriquecer la calidad de las recomendaciones y clasificaciones proporcionadas.

- Obtención de más datos: Uno de los aspectos cruciales para la mejora del modelo es la expansión del conjunto de datos actual. Adquirir más currículums vitae enriquecerá el corpus y permitirá al modelo aprender de un espectro más amplio de ejemplos, lo cual es fundamental para mejorar la generalización y precisión de las predicciones. Esto puede ser resuelto a través del desarrollo de un sistema de scraping, que permita descargar perfiles de LinkedIn y a su vez etiquetarlos de alguna manera para evitar esa parte manual del proceso.
- Uso de análisis sintáctico para extracción de información: Implementar técnicas de procesamiento del lenguaje que se enfocan en la estructura sintáctica de los textos podría permitir detectar patrones dentro de los documentos que permitan la extracción de información específica de los CVs, como la cantidad de años de experiencia, habilidades específicas y datos de contacto de los candidatos.
- Predicción del seniority del candidato: Desarrollar capacidades para estimar el nivel de seniority de los candidatos basándose en su experiencia y habilidades detalladas en los CVs podría significativamente afinar la precisión de las recomendaciones. Esto ayudaría a alinear mejor las vacantes con los perfiles de los candidatos, asegurando que las oportunidades se ofrezcan a aquellos en los cuales además de coincidencia semántica, exista coincidencia entre la cantidad de experiencia específica requerida.

- Fine-tuning de los modelos Transformers: Ajustar los modelos Transformers específicamente para nuestro conjunto de datos es otra mejora potencial. A través del fine-tuning, los modelos pre entrenados pueden ser optimizados para entender mejor las peculiaridades y características únicas de los datos relacionados con el empleo, lo cual podría resultar en un aumento notable en la eficacia del modelo.

### 10.3. Conclusión

Los modelos desarrollados en este estudio no solo han demostrado ser efectivos para las tareas asignadas, sino también eficientes en términos de recursos computacionales. Utilizando tecnologías como Doc2Vec, con varios años desde su descubrimiento, hemos logrado superar en rendimiento a tecnologías de vanguardia como los Transformers, aplicando la técnica de Lbl2Vec. Esto supone un enorme potencial para transformar organizaciones de cualquier tamaño, ya que permite implementar estas técnicas independientemente de los recursos computacionales disponibles, generando rápidamente un impacto positivo.

Además, es importante destacar nuevamente que los beneficios no se limitan solo a las organizaciones. La aplicación de estos modelos también democratiza el proceso de selección laboral, asegurando que todas las postulaciones sean consideradas equitativamente, reduciendo la influencia del azar y la suerte en la consecución de empleo. Asimismo, facilita que las empresas proporcionen retroalimentación de manera ágil a los candidatos, economizando tiempo valioso para los aspirantes cuyas postulaciones no son consideradas para avanzar. Esto subraya cómo la adopción de estas tecnologías en el ámbito de la búsqueda de empleo puede contribuir a un mercado laboral más eficiente y equitativo.

## 11. Apéndice

### 11.1. Recomendación utilizando ChatGPT.

Para profundizar en la comparación del desempeño del modelo con tecnologías avanzadas y de actual relevancia, se procedió a ejecutar la misma tarea de recomendación utilizando ChatGPT-4.

Esto supone ventajas y desventajas. Para una limitada cantidad de candidatos esta herramienta permite además de realizar recomendaciones, explicar el razonamiento<sup>12</sup> detrás de las mismas, lo cual permite que la metodología empleada por del modelo sea más transparente para los usuarios, y a su vez se llegue a conclusiones más precisas.

Por otro lado, debido a las limitaciones en el procesamiento de grandes cantidades de texto que esta herramienta actualmente posee, esta tarea no puede realizarse a gran escala, por lo que únicamente se aplicó para una vacante de las seis utilizadas previamente (para la posición de Data Scientist), y un subconjunto de 100 CVs seleccionados aleatoriamente. Esto demuestra que para casos puntuales puede ser una herramienta muy útil para analistas de recursos humanos en sus tareas cotidianas, pero todavía no permite generar soluciones prácticas a gran escala que puedan ser implementadas en organizaciones.

A continuación se expone el prompt<sup>13</sup> utilizado y los resultados del mismo.

- Prompt:

*Sos un reclutador de talentos y trabajas para una agencia de recursos humanos que se especializa en conseguir candidatos ideales para vacantes laborales de sus clientes.*

*Es tu labor revisar los perfiles de los potenciales candidatos y seleccionar a los más aptos para el siguiente puesto.*

---

<sup>12</sup> En el contexto de inteligencia artificial, esto se denomina "chain of thought" (cadena de pensamiento).

<sup>13</sup> Un "prompt", en el contexto de la inteligencia artificial, es una instrucción o entrada de texto que se proporciona al modelo para que este genere una respuesta o continúe con el texto.

*En los archivos adjuntos verás los potenciales candidatos para el puesto y la vacante a cubrir. Tener en cuenta que antes de cada currículum se indica un número, que permite identificarlos fácilmente.*

*Selecciona los cinco mejores candidatos, ordenados de mayor a menor relevancia para el puesto, y explicame el razonamiento de tus decisiones. Por favor al realizar la selección, indicar el número del currículum seleccionado para poder identificarlo.*

- Respuesta de ChatGPT-4:

*Basándome en la descripción de la vacante laboral y en los perfiles revisados, he seleccionado los cinco mejores candidatos para la posición de Data Scientist Semi Sr., considerando la experiencia relevante, habilidades técnicas, educación, y competencias que mejor se alinean con los requerimientos del puesto. Los candidatos son seleccionados y ordenados de mayor a menor relevancia según su ajuste con la vacante:*

- *Candidato número 71:*
  - *Razonamiento: Este candidato tiene un perfil robusto como Senior Data Scientist con experiencia en análisis exploratorio y desarrollo de modelos predictivos, particularmente en el área de riesgo crediticio, lo que demuestra una sólida comprensión de las estadísticas y la aplicación de la ciencia de datos. Además, tiene experiencia en el desarrollo de modelos de datos, construcción de dashboards y la aplicación de modelos de ML, lo cual es altamente relevante para la posición.*
- *Candidato número 95:*
  - *Razonamiento: Posee experiencia como Data Analyst y Analytics Consultant con habilidades en R, SQL, Big Query, y herramientas de visualización como PowerBI, lo que se alinea con las necesidades de la posición. La experiencia en consultoría y estrategia de medición y análisis de datos añade un valor significativo al perfil.*
- *Candidato número 70:*
  - *Razonamiento: Este candidato muestra una trayectoria sólida en Business Intelligence y análisis de datos, con experiencia en la*

*generación de KPIs y análisis estadístico para la toma de decisiones. La habilidad para crear sistemas de inteligencia de negocio y desarrollar dashboards interactivos en PowerBI, junto con la experiencia en el análisis de datos para optimizar recursos, son competencias valiosas para el rol.*

- *Candidato número 6:*
  - *Razonamiento: Aunque su experiencia se centra más en Business Intelligence, la habilidad demostrada en la migración de reportes de Hyperion a PowerBI y el desarrollo de dashboards son relevantes para la vacante. Además, su experiencia como líder técnico y en la gestión de proyectos de BI sugiere una sólida base en análisis de datos y manejo de herramientas pertinentes.*
  
- *Candidato número 129:*
  - *Razonamiento: Aunque su experiencia previa es más general en el análisis de costos y gestión de proyectos, su educación en Management Analytics y su enfoque en data analytics y visualización a través de Udacity lo convierten en un candidato prometedor que está en el proceso de especializarse en ciencia de datos, lo que podría aportar una perspectiva fresca y enfoque analítico al equipo.*

*Estos candidatos fueron elegidos porque sus habilidades, experiencia y educación se alinean estrechamente con los requerimientos del puesto, incluyendo la extracción y procesamiento de datos, análisis estadístico, uso de lenguajes de programación como Python y R, y experiencia con herramientas de BI como PowerBI.*

## 12. Bibliografía

- Tim Schopf, Daniel Braun & Florian Matthes (2021). Lbl2Vec: An Embedding-based Approach for Unsupervised Document Retrieval on Predefined Topics. Proceedings of the 17th International Conference on Web Information Systems and Technologies WEBIST - Volume 1, 124-132.
- Tim Schopf, Daniel Braun & Florian Matthes (2021). Semantic Label Representations with Lbl2Vec: A Similarity-Based Approach for Unsupervised Text Classification. Lecture Notes in Business Information Processing, 59-73.
- Christophe Van Gysel, Maarten de Rijke, Evangelos Kanoulas (2018). Neural Vector Spaces for Unsupervised Information Retrieval. ACM Trans. Inf. Syst., 36(4).
- Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing (2nd ed.). Prentice Hall.
- Jurafsky, D., & Martin, J. H. (2024). Speech and Language Processing (3rd ed. draft)
- Tim Schopf, Daniel Braun & Florian Matthes (2022). Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches. NLPiR '22: Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval, 6-15.
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, 3111-3119.
- Quoc V. Le, Tomas Mikolov (2014). Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning, 1188-1196.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017). Attention is All you Need. Advances in Neural Information Processing Systems 30.