instituto
de ingeniería
del conocimiento

20
22

Nº
02

Revista de Casos de Estudio en HR Analytics

Journal of HR Analytics Case Studies

ISSN: 2792-3770 | Instituto de Ingeniería del Conocimiento

06 El uso de IA y PLN para la clasificación de descripciones de puestos de trabajo: Un caso de éxito.

Using AI and NLP in the classification of job descriptions: A successful use case.

ISSN: 2792-3770

La Revista de Casos de Estudio en HR Analytics

Journal of HR Analytics Case Studies

La ***Revista de Casos de Estudio en HR Analytics*** nace con la misión de facilitar el intercambio de conocimiento especializado entre profesionales y académicos en el ámbito de la **analítica de Recursos Humanos**, con el objetivo de mejorar la **efectividad de las organizaciones.** La entidad responsable de esta revista es la **Asociación para el Desarrollo de la Ingeniería del Conocimiento** (ADIC), siendo esta publicación on-line editada por el **Instituto de Ingeniería del Conocimiento** (IIC) con una periodicidad de un número anual.

Objetivo

La revista tiene como **objetivo** principal ser un vehículo para la reflexión y la difusión de las **buenas prácticas, últimos avances ylíneas de investigación** en el ámbito de la analítica aplicada para la toma de decisiones sobre la gestión del capital humano en las organizaciones.

La revista tiene un **carácter científico** y una **vocación divulgativa**, por ello propone artículos fundamentalmente de **carácter aplicado**. Con ellos se pretende que los profesionales de las organizaciones accedan a un conocimiento relevante acerca de cómo otras organizaciones desarrollan HRA. Y, también, acercar a los académicos el conocimiento respecto de cómo se desarrolla HRA en la práctica.

Alcance

El **enfoque de la Revista**, que pretende ser **multidisciplinar**, da cabida (entre otros) a manuscritos que: reflejen **casos prácticos** de aplicación del HRA en las organizaciones; que analicen, comparen y relacionen la utilidad de diferentes **técnicas y/o herramientas** para el abordaje de diferentes objetivos analíticos; que planteen y valoren la efectividad de diferentes **metodologías de trabajo** para el desarrollo de proyectos HRA; que ayuden a entender el **mapa de ruta** por el que transitar desde los niveles básicos del HRA hasta los niveles de excelencia; y que en general ayuden a entender cómo **mejorar la efectividad organizacional** a partir de la analítica de datos referidos a la fuerza de trabajo.

ISSN: 2792-3770

© ADIC - IIC

instituto de ingeniería del conocimiento

Equipo Editorial

La revista está editada por el Instituto de Ingeniería del Conocimiento y tiene los siguientes órganos de gobernanza.

Editor

David Aguado.
Instituto de Ingeniería del Conocimiento.

Editores Asociados

Jesús de Miguel.
Centro de Investigación para la Efectividad Organizacional, Universidad Autónoma de Madrid.

Antonio Delgado.
Universidad Autónoma de Madrid.

María Jesús Belizón.
University College Dublin.

Beatriz Lucía.
Instituto de Ingeniería del Conocimiento.

Delia Majarín.
Telefónica.

Sergio Raja.
Zurich Seguros.

Comité Editorial

Magdalena Nogueira.
Universidad Autónoma de Madrid.

Francisco Abad.
Universidad Autónoma de Madrid.

Carmen García.
Universidad Autónoma de Madrid.

José Manuel de Haro.
Universidad de Alicante.

William Ferrando Durán.
Universidad Javeriana.

Carolina Zúñiga.
Universidad Politécnica Salesiana del Ecuador.

José Carlos Andrés.
Viewnext.

Eduardo Páez.
Cepsa.

Pablo Haya.
Instituto de Ingeniería del Conocimiento.

Álvaro Barbero.
Instituto de Ingeniería del Conocimiento.

Sonia Rodríguez.
Instituto de Ingeniería del Conocimiento.

Celia Martínez.
Instituto de Ingeniería del Conocimiento.

Maite Sáez.
Observatorio de Recursos Humanos y Relaciones Laborales.

Diseño y Maquetación

Nuria Herranz González .
Instituto de Ingeniería del Conocimiento.

Andrés Muñoz Bachiller.
Instituto de Ingeniería del Conocimiento.

ISSN: 2792-3770

Índice de contenidos

01. HRA en la Práctica Profesional: un Campo en Alza. Introducción al Número 2 de la Revista de Casos de Estudio en HR Analytics 03

HRA in Professional Practice: a Rising Field. Introducting the second Issue of Journal of HR Analytics Case Studies.

02. Digitalising strategic workforce planning to enable group wide measurement of workforce capability gaps and risks 13

Digitalización del proceso de la planificación estratégica de recursos humanos para facilitar la medición de capacidades a nivel de empresa

03. Análisis de la Rotación de Personal para Mejorar el Proceso de Toma de Decisiones en una Empresa del Sector TIC: el caso de GFT IT Consulting, S.L.U. 33

Analysing staff turnover to improve the decision-making process in an ICT sector company: the case of GFT IT Consulting, S.L.U

04.How do we continue after the Covid-19 pandemic? Work from home as the new normal and the future of work 70

¿Cómo continuar después de la pandemia? La nueva normalidad del trabajo desde casa y el futuro del trabajo.

05.IImpacto de la Satisfacción con la Comunicación Interna en la Resiliencia de las organizaciones: Análisis de una Empresa de Marketing y Publicidad... 84

Impact of Satisfaction with Internal Communication on Organizational Resilience: Analysis of a Marketing and Advertising Firm.

06.El uso de IA y PLN para la clasificación de descripciones de puestos de trabajo: Un caso de éxito 110

Using AI and NLP in the classification of job descriptions: A successful use case

07. Influencia de la confianza en el rendimiento de los equipos virtuales de trabajo 124

Influence of trust in virtual teams' performance

08. HR Analytics en las pequeñas y medianas empresas españolas 138

Human Resources Analytics in Small and Medium Sized Spanish Businesses

ISSN: 2792-3770

© ADIC - IIC

ISSN: 2792-3770

06

El uso de IA y PLN para la clasificación de descripciones de puestos de trabajo: Un caso de éxito.

Using AI and NLP in the classification of job descriptions: A successful use case.

El presente trabajo ha sido desarrollado con financiación parcial por parte del CDTI.

Nuria Aldama-García
Instituto de Ingeniería del Conocimiento (IIC)
Correo electrónico: nuria.aldama@iic.uam.es
Linkedin/Publons: https://www.linkedin.com/in/nuria-aldama-garc%C3%ADa-6214a9128

Guillem García-Subies
Instituto de Ingeniería del Conocimiento (IIC)
Correo electrónico: guillem.garcia,@iic.uam.es
Linkedin/Publons: https://www.linkedin.com/in/guillemgsubies/

Doaa Samy
Instituto de Ingeniería del Conocimiento (IIC)
Correo electrónico: doaa.samy@iic.uam.es
Linkedin/Publons: https://www.linkedin.com/in/doaa-samy-90354872/?originalSubdomain=es

Raúl Suárez
CEINSA
Correo electrónico: r.suarez@ceinsa.com

Xavier Pérez
CEINSA
Correo electrónico: x.perez@ceinsa.com

Received: 19 abril 2022
Received in revised form: 19 mayo 2022
Accepted: 31 mayo 2022
Available on-line: 28 de octubre 2022

ISSN: 2792-3770

Resumen

El artículo presenta un caso de uso real donde se aplican técnicas de Procesamiento del Lenguaje Natural (PLN) para el desarrollo de un sistema de clasificación de descripciones de puestos de trabajo en el dominio de la gestión de recursos humanos (HR). La metodología de desarrollo se basa en tres fases: 1) Depuración de los datos, 2) Entrenamiento de tres tipos de modelos (modelos clásicos, modelos de lenguaje y modelos jerárquicos) y 3) Evaluación de los diferentes modelos según dos métodos (descripciones individuales y descripciones agrupadas). Se obtienen los mejores resultados con los clasificadores basados en modelos de lenguaje y aplicando el método de agrupación de descripciones.

Palabras clave:
Procesamiento del Lenguaje Natural, PLN, Aprendizaje automático, Modelos de clasificación automática, Inteligencia artificial aplicada al sector HR

Abstract

This article presents a real use case where Natural Language Processing techniques are applied in the development of a classification system for job descriptions in the field of Human Resources Management (HR). The methodology is based on three main stages: 1) Data curation, 2) Training of three different types of classifiers (traditional classifiers, classifiers based on language models and hierarchical classifiers) and 3) Evaluation of the different models according to two approaches (individual job description, grouped job descriptions). The best results are achieved by the classifiers based on language models using the grouped job descriptions.

Keywords:
Natural Language Processing, NLP, Machine Learning, Automatic classification, Artificial Intelligence in HR sector

ISSN: 2792-3770

© ADIC - IIC

1. Introducción

El impacto de las tecnologías avanzadas como la inteligencia Artificial (IA) y la Robótica, es una realidad que está transformando a nivel general el mercado laboral y, también, de manera específica el sector de la gestión de los recursos humanos (HR).

Este impacto se observa tanto en el nivel de las estructuras organizacionales como en nivel de los puestos individuales. Las tecnologías avanzadas están introduciendo nuevos paradigmas que, a su vez, están transformando la manera de realizar el trabajo incluyendo los procesos de la gestión HR como la contratación o la formación, entre otras. Los últimos años han sido un claro testimonio de cómo emergen nuevos puestos de trabajo y desaparecen otros, cómo cambian las estructuras organizacionales y cómo se virtualiza el concepto del entorno físico del trabajo. Varios estudios han analizado este impacto poniendo especial énfasis tanto en las oportunidades como en los desafíos sociales y éticos que implica este impacto (Robert et al., 2020) (Vronis et al., 2021).

Dentro del ámbito de la Inteligencia Artificial (IA), el presente trabajo se centra en el Procesamiento del Lenguaje Natural (PLN) y su papel en transformar algunas tareas del sector HR. El PLN se define como la técnica que simula la capacidad y la inteligencia lingüística humana de comprender (descodificar) y generar (codificar) el lenguaje natural sea en su forma escrita u oral. Para conseguir este objetivo, el PLN se basa en un amplio conjunto de técnicas para el análisis lingüístico de los datos, la extracción de la información, el cálculo de la similitud semántica, la clasificación automática, la agrupación de datos, el análisis de la estructura discursiva, la detección de emociones, el análisis de sentimiento, los sistemas conversacionales, la traducción automática, etc. Optar por una técnica u otra viene determinado por el tipo de problema a resolver y la naturaleza de los datos. Es por esto que el PLN ha ido cobrando mayor importancia en varios sectores por su capacidad y versatilidad en tratar los datos no estructurados y la utilidad de las soluciones prácticas que puede ofrecer a los distintos sectores. Estas soluciones han demostrado su eficiencia asistiendo a los expertos en tareas y sectores que requieren manejar datos no estructurados.

Un claro ejemplo de la aplicación del PLN en el sector de HR, es el uso de agentes conversacionales en la atención al cliente. Sin embargo, existen numerosos ejemplos, como veremos en la sección 1, donde se aplica el PLN en otras tareas de HR como la contratación, la formación o el apoyo a la toma de decisiones.

El presente artículo se centra en la utilización del PLN para apoyar la automatización de uno de los procesos básicos de la gestión HR: el establecimiento de las políticas retributivas. En el establecimiento del salario se utiliza habitualmente la valoración de los puestos de trabajo. Esta valoración se realiza a partir de las tareas y responsabilidades asociadas a los puestos de trabajo y que se encuentran en lo que denominamos la “descripción del puesto de trabajo”. Descripción que, habitualmente, se encuentra realizada en lenguaje natural.

ISSN: 2792-3770

La digitalización de los procesos de gestión de recursos humanos y en especial de los procesos que toman como base la descripción de los puestos de trabajo era una asignatura pendiente debido a la dificultad de homogenizar las descripciones y su posterior aprovechamiento para otros procesos.

El proceso de análisis y descripción de los puestos de trabajo es la célula base que permite definir políticas en organización, compensación y talento. A nivel de organización permite detallar las funciones y responsabilidades, analizando los desequilibrios entre diferentes puestos, también permite definir y ajustar los diferentes modelos organizativos. En compensación es la base para la equidad interna, mediante la valoración de puestos, y, por tanto, la semilla de la definición de una estructura salarial equitativa y competitiva. Respecto al talento, las funciones suelen tener correlación con diferentes competencias, ya sean corporativas o específicas del puesto. Un buen sistema de evaluación por competencias debería tener en cuenta las funciones para que el modelo pueda gestionarse en el día a día.

La automatización de este proceso es clave y permite disponer de descripciones ajustadas y actualizadas. Uno de los grandes problemas de las descripciones de los puestos, es la falta de actualización, lo que repercute en su poca utilización en el día a día de la gestión en una organización. Disponer de una herramienta que permita, de manera fácil, intuitiva y concreta actualizar las descripciones por cualquier miembro de la organización es una herramienta de vital importancia para el diseño e implantación de políticas avanzadas de gestión de las personas.

En este contexto, cuando un cliente ofrece las descripciones de sus puestos de trabajo para comenzar el proceso de diseño de la estructura salarial, los consultores expertos en compensación homogeneizan las descripciones con un lenguaje que refleje claramente la dificultad y responsabilidad del puesto. Posteriormente aplican diferentes criterios para obtener una valoración de cada puesto en puntos. Esos puntos permiten clasificar los puestos y asignarles un rango salarial.

Analizando este proceso desde el punto de vista del experto humano, se puede observar que implica cierta complejidad por lo siguiente:

- Primero, en cuanto a la naturaleza de los datos, las descripciones de puestos de trabajo suelen estar en formato de texto libre. Por tanto, se trata de un conjunto de datos no estructurados cuya extensión pueda variar desde descripciones breves de dos o tres líneas hasta descripciones detalladas de dos o más párrafos.
- Segundo, el proceso en sí implica una serie de pasos que detallamos a continuación:
    - Leer el texto de la descripción en un lenguaje que el experto maneja.
    - Analizar la descripción para comprender su contenido y distinguir entre las tareas específicas que distinguen el puesto de trabajo en cuestión de las tareas genéricas que pueden estar en común con otros puestos de trabajo.
    - Inferir la categoría a la que pueda pertenecer esta descripción basándose en el análisis anterior para asignar una clase de la tipología de puestos aplicada.

Automatizar este proceso es factible y práctico porque se trata de un problema que las técnicas avanzadas de PLN y de clasificación automática, hoy en día, pueden abordarlo de forma eficiente, rentable y en tiempo real. Cabe destacar que no se trata de sustituir al experto, sino de asistirle en realizar las tareas más repetitivas de una forma más eficiente, ahorrando así el tiempo para dedicarlo a otros procesos más complejos. Este planteamiento se alinea con las directrices europeas de responsabilidad ética en el ámbito de la IA. Es una solución que se centra en asistir al experto humano y, por tanto, es “Human-centered AI” y, por otro lado, garantiza la implicación del experto en el proceso cumpliendo con el principio de “human-in-the-loop”.

ISSN: 2792-3770

© ADIC - IIC

Este trabajo presenta un caso de éxito, fruto de una colaboración interdisciplinaria entre el equipo del Instituto de la Ingeniería del Conocimiento (IIC) (www.iic.uam.es) y el equipo de CEINSA (http://ceinsa.com/). Ambos equipos con expertos y especialistas en el sector HR, ciencia de datos, PLN y tecnologías SW reunieron esfuerzos para desarrollar una prueba de concepto para un sistema piloto de clasificación de las descripciones de puestos de trabajo. El sistema fue co-creado y evaluado por los expertos de HR de CEINSA y ha demostrado su eficiencia como un punto de partida para asistir a los expertos en esta tarea. Asimismo, este piloto ha puesto de manifiesto los puntos de mejora para futuras aproximaciones.

El artículo se estructura en cinco secciones. La primera sección ofrece un breve resumen de los casos de uso y los procesos del sector HR en que el PLN desempeña un papel relevante. En la segunda sección, se detallan los objetivos del trabajo. La metodología adoptada con la descripción de la naturaleza de los datos y las fases del procesamiento se explican en la tercera sección. Los resultados obtenidos se comentan en la quinta sección y, finalmente, se resumen las conclusiones derivadas de este trabajo.

2. Estado de la cuestión: PLN como herramienta básica para la digitalización de procesos HR

Las aplicaciones del PLN a la industria son innumerables y crecen exponencialmente a medida que pasa el tiempo y la tecnología avanza. Esta realidad constituye una serie de oportunidades para la innovación en el sector, pero a la vez implica unos retos de carácter ético y social. A continuación, se presentan brevemente tres casos de uso y aplicación de PLN en el ámbito de los RRHH junto a los retos que suponen.

PLN en contratación (*recruiting*)

El PLN se utiliza en el ámbito de la contratación para detectar en los currículos de los candidatos información relevante (habilidades, trayectorias profesionales, experiencias laborales, competencias y áreas de interés) para el puesto vacante. Así, a través de la tecnología es posible filtrar a los candidatos que más se adaptan a un perfil concreto reduciendo tiempos y focalizando esfuerzos.

En contratación es cada vez más común el uso de *chatbots* o asistentes conversacionales en la realización de entrevistas preliminares que tienen como objetivo la recogida de datos personales y el contraste de ciertos aspectos concretos para saber si el candidato, de manera general, reúne los requisitos básicos de elegibilidad para el puesto en cuestión.

Otro aspecto de relevancia en el sector es el control de los sesgos inconscientemente aportados por las personas en procesos de selección de personal o en las evaluaciones internas y externas. Aunque la literatura recoge distintas opiniones al respecto (Robert

ISSN: 2792-3770

et al., 2020), la idea general es que la IA y los sistemas basados en PLN pueden ayudar a aislar dichos sesgos llegando a eliminarlos para que la selección de personal sea lo más imparcial posible.

No obstante, existen otras corrientes que resaltan los retos que supone aplicar estas técnicas, sobre todo en cuanto a los aspectos éticos, la trazabilidad y la explicabilidad de la decisión tomada y la legalidad del proceso con la dificultad de asignar la responsabilidad "liability" a lo largo del proceso. En cuanto a los aspectos éticos, estos sistemas tratan datos personales de los candidatos y este tratamiento debería cumplir con el marco legal vigente. Segundo, hasta qué punto un sistema puede decidir sobre un candidato y si los datos con que está entrenado este sistema están libres de sesgos. Por otra parte, la explicabilidad y la trazabilidad de la decisión no siempre es posible, sobre todo cuando se trata de sistemas basados en técnicas de aprendizaje profundo. Por último, la legalidad del proceso se pone en duda al no poder identificar quién es el/la responsable de la toma de decisión.

PLN en formación del personal y carreras profesionales

El PLN sirve como apoyo para poder determinar las áreas formativas de interés para una empresa en las que los empleados han de iniciarse o mejorar, o detectar aquellas áreas de formación que la plantilla tiene cubiertas y en las cuales no hace falta invertir más recursos.

Por otro lado, las técnicas de PLN sirven también como asistencia en el diseño de estrategias de sucesión. Así, las tecnologías aplicadas sobre los perfiles de los empleados de una empresa pueden revelar si dentro de la misma se cuenta con un trabajador que cumple con los requisitos idóneos para gestionar la sucesión de otro trabajador.

PLN en la escucha a los trabajadores

Así mismo, el PLN puede integrarse en las empresas como herramienta de análisis de la información recogida mediante formularios o encuestas a empleados aplicando técnicas de análisis del sentimiento o las emociones, extracción de información, clasificación automática o detección de tendencias.

Escuchando a los trabajadores y analizando sus respuestas se extrae información de gran valor para la empresa que ayuda a focalizar la atención en estrategias de retención de talento, estrategias de fomento del compromiso (*engagement*) del empleado con la empresa o puntos de mejora de satisfacción de la plantilla. Esta escucha activa, además, ayuda a detectar áreas de fricción, tendencias de comportamiento entre los empleados (Vrontis et al., 2021).

PLN en clasificación de puestos

Por último, cabe destacar el problema de clasificación, objeto de este artículo, donde las técnicas de clasificación automática y PLN se emplean para asignar una categoría a una descripción según una tipología concreta. Este tipo de aplicaciones ahorra el tiempo de análisis manual y permite procesar de forma eficiente grandes volúmenes de datos en unos segundos, agilizando así este proceso básico considerado como un punto de partida necesario para otros procesos de planificación y contratación.

ISSN: 2792-3770

© ADIC - IIC

3. Objetivos

El presente trabajo tiene como objetivo desarrollar un sistema de clasificación multi-instancia basado en técnicas avanzadas de PLN y aprendizaje automático con el fin de automatizar el proceso de identificación de la categoría de un puesto de trabajo dada la descripción del mismo. La tipología adoptada es la tipología definida por CEINSA. Se trata de una tipología jerárquica, extensa y no excluyente, es decir, una descripción puede clasificarse bajo más de una categoría. Para alcanzar este objetivo, se plantean unos objetivos específicos que se detallan a continuación:

- Construir un conjunto de datos representativo del problema en cuestión
- Analizar la naturaleza de estos datos
- Entrenar modelos de clasificación automática basados en PLN
- Evaluar y analizar los resultados obtenidos

4. Metodología

El desarrollo del sistema de clasificación se ha llevado a cabo en tres fases. Las dos primeras fases se centran en el análisis y la depuración de los datos para garantizar la calidad. La tercera fase se centra en el entrenamiento del modelo y la selección de los métodos de evaluación más adecuados al problema en cuestión.

Una vez entrenado el modelo, se ha procedido a la evaluación de los resultados de la predicción y, finalmente, la puesta en marcha del servicio de clasificación.

Las tres fases del desarrollo se resumen en lo siguiente

- Fase 1. Descripción y análisis del conjunto de datos. Esta fase se ha llevado a cabo por el equipo técnico en colaboración con los expertos del sector de recursos humanos para entender la naturaleza de los datos.

- Fase 2. Limpieza, depuración y segmentación de los datos. Esta fase se ha llevado a cabo por el equipo técnico compuesto por un científico de datos y lingüistas computacionales especializadas. Las técnicas aplicadas en esta fase son básicamente técnicas de tratamiento automático del texto.

- Fase 3. Aproximaciones y métodos de evaluación de los modelos de clasificación. En esta fase se ha experimentado con diferentes tipos de modelos y se ha optado por diferentes métodos de evaluación. Esta fase se ha llevado a cabo por el equipo técnico compuesto de especialistas en ciencias de datos y PLN.

A continuación, se detalla el trabajo realizado en cada fase.

ISSN: 2792-3770

4.1. Descripción y análisis del conjunto de datos

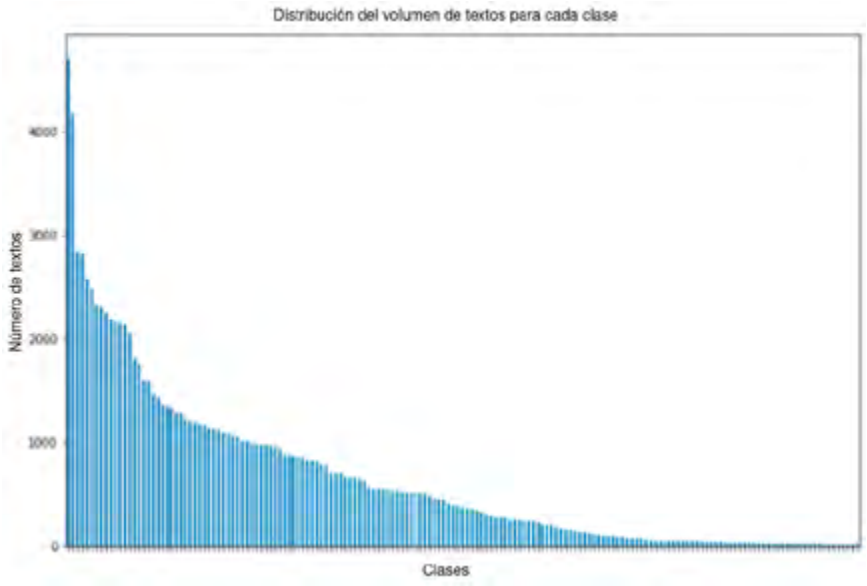
Dado el carácter empírico y técnico de la solución, es imprescindible partir de una muestra de datos reales y representativos. Para ello, los expertos del negocio han proporcionado un conjunto de datos que cuenta con alrededor de cien mil descripciones de puestos de trabajo clasificadas según la tipología de puestos empleada por CEINSA y que comprende más de ciento cincuenta puestos de trabajo. El conjunto de datos consiste en un código de puesto asociado a una descripción. El formato del código de puesto consiste en una letra y un número. La descripción es un texto libre cuya longitud puede variar desde una palabra o una línea hasta numerosos párrafos. A continuación, se pueden visualizar algunos ejemplos genéricos y ficticios del conjunto de datos, ya que por motivos de privacidad de los datos del cliente no se puede incluir datos reales.

Tabla 1. Ejemplos genéricos del conjunto de datos

Código_puesto	Descripción
A12	Realizar tareas administrativas. Organizar reuniones
P09	Colaborar activamente con otros departamentos. Monitorizar el rendimiento. Actualizar los recursos.

El análisis del conjunto ha revelado algunos retos significativos para el desarrollo del modelo. Primero, el gran número de puestos que constituye la tipología supone, de por sí, un desafío a la hora de conseguir un buen modelo de clasificación capaz de predecir entre 150 puestos de trabajo. Segundo, el gran desbalanceo de los datos de la muestra, donde algunos puestos tenían menos de una decena de descriptores, mientras que otras tenían miles de descriptores. En el siguiente gráfico de barras se puede observar la extensión de la tipología y el desequilibro en la distribución de las descripciones asociadas a cada puesto.

Figura 1. Distribución de descriptores por puestos de trabajo



ISSN: 2792-3770

© ADIC - IIC

A estos retos, se suma un tercero, dado el hecho de que algunas descripciones o partes de las descripciones se repiten entre los diferentes puestos. Esto se debe a que varias descripciones de puestos de trabajo pueden incluir tareas genéricas y requisitos que están en común con otros puestos. Por ejemplo, la tarea de "Proponer acciones de mejora" aparece en el 25% de las descripciones de puestos incluyendo puestos tan variados como "ingeniero de proyectos", "telefonista", "recepcionista" o "contable". Estas repeticiones dan lugar a ambigüedades que complican el proceso de clasificación a la hora de distinguir entre un puesto y otro

De esta manera, la propia naturaleza de los datos (desbalanceo y descripciones comunes) junto con el gran número de puestos en un problema multiinstancia, han convertido una tarea clásica de clasificación de textos en un proyecto ambicioso, retador y complejo.

4.2. Limpieza, depuración y segmentación de los datos

La revisión, depuración y segmentación de los datos es el punto de partida para el desarrollo del sistema. Cuanto mayor es la calidad de los datos, mejor es el entrenamiento del modelo. En este caso en concreto, a partir de los datos recibidos, se han detectado una serie de fenómenos, algunos tratan de aspectos de formato y otros son aspectos de contenido. De ahí, se han adoptado diferentes estrategias, ya que en algunos casos se ha estimado necesario eliminar los aspectos que pueden considerarse como fuentes de ruido innecesario. Mientras que, en otros casos, se ha optado por mantener estos rasgos porque forman parte del propio contenido.

En primer lugar, se han extraído patrones de aspectos a eliminar como, por ejemplo, duplicados exactos, marcas de formateo automático en los textos o restos de xml o html. A partir de dichos patrones, se ha efectuado una limpieza automática de los datos. Además, se llevó a cabo un proceso de unificación y normalización de los códigos asociados a cada puesto de trabajo.

Por otra parte, se ha decidido mantener algunos aspectos considerados como característicos del dominio, por ejemplo, el uso de términos en inglés. A pesar de que los datos se encontraban en castellano, se ha observado una gran cantidad de anglicismos de uso común en la jerga del dominio (p.ej.: *manager, developer*).

Tras la limpieza y depuración se mantuvieron en torno a 70.000 descripciones del total de 100.000 recibidos. Estas descripciones están asociadas a unos 150 puestos de trabajo.

Una vez depurados los datos, se ha procedido a su segmentación y su adaptación para servir de entrada (*input)* para los modelos de aprendizaje automático. Para el conjunto de entrenamiento se ha contado con un total de 50,000 pares de puesto/descripción y para el conjunto de evaluación, se ha contado con un conjunto de 20,000 pares puesto/descripción.

4.3. Aproximaciones y métodos de evaluación

Aproximaciones.

Con el objetivo de alcanzar los mejores resultados posibles, sin perder de vista la dificultad del caso, se probaron tres aproximaciones para dar solución al problema de clasificación: la aproximación basada en modelos clásicos, la aproximación basada en modelos del lenguaje y la aproximación basada en modelos jerárquicos.

La familia de modelos clásicos está compuesta por vectorizaciones simples del texto (por ejemplo, bolsas de palabras, tf-idf, etc.) seguidos de modelos de clasificación de *machine learning.*

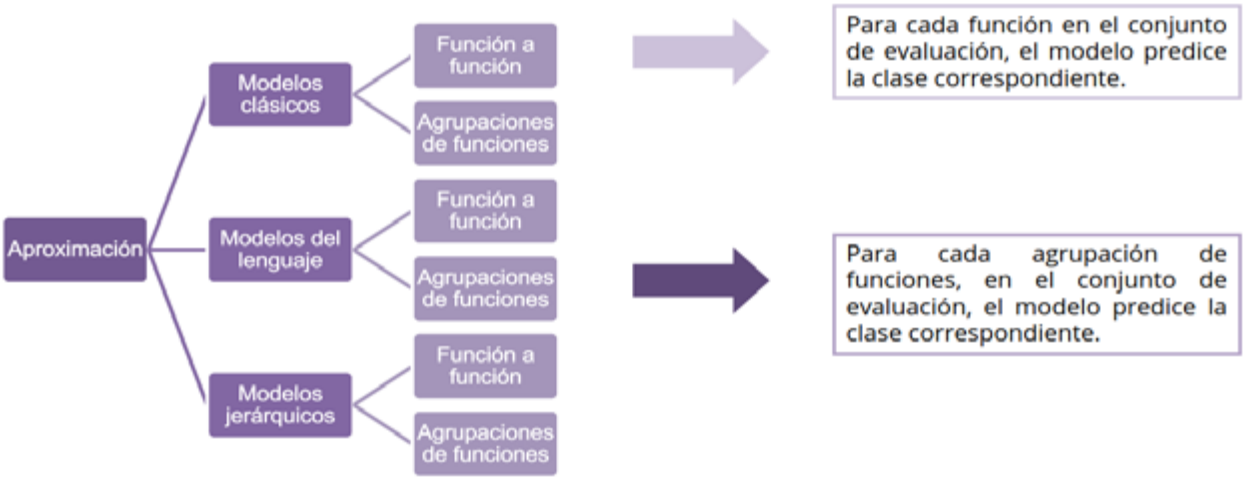
La aproximación de modelos jerárquicos pretende dividir el problema de clasificación entre 150 puestos en problemas de clasificación más reducidos. Esta aproximación funciona de igual manera

ISSN: 2792-3770

que los modelos clásicos, pero tiene en cuenta la estructura de la tipología de puestos, ya que algunos puestos se agrupan bajo puestos más genéricos de clústeres de temática similar. Por ejemplo, existen puestos genéricos como "Administración", "Contabilidad", "Diseño técnico". Cada uno de estos tipos de puestos genéricos, a su vez, tienen subtipos de puestos. De este modo, se ha creado un modelo general para distinguir estos puestos genéricos (con menos número de puestos) y varios sub-modelos que, en función del primer modelo, infieren el subtipo de puesto.

Finalmente, en la familia de modelos del lenguaje, se ha llevado a cabo un proceso de *fine-tuning* o ajuste fino de los parámetros que emplean modelos de lenguaje pre-entrenados, basados en técnicas de *transformers*.

Figura 2. Esquema de familias de modelos y estrategias de evaluación



Métodos de entrenamiento y evaluación.

Teniendo en cuenta las repeticiones frecuentes en las descripciones de los puestos de trabajo, la naturaleza jerárquica de la tipología, así como la utilidad del modelo en escenarios reales, se ha decidido optar por dos métodos de entrenamiento y evaluación de los modelos de clasificación.

En el método de evaluación clásico, cada descripción de un puesto de trabajo se trata de forma independiente e individual. Es decir, los datos de entrenamiento consisten en pares de descriptores individuales junto a su puesto.

En cambio, el segundo método se adapta a la naturaleza jerárquica de los datos y las repeticiones frecuentes entre algunas partes de las descripciones. De ahí, se han agrupado las descripciones de puestos de trabajo que pertenecen a la misma clase jerárquica. De esta manera y en base al conocimiento experto de CEINSA, se han generado grupos de textos que se adecuen a los casos de uso reales. A partir de esos conjuntos de descripciones, el modelo predecía el puesto genérico asociado (modelo multiinstancia).

Se ha entrenado las tres familias de modelos según estos dos métodos y se ha medido el acierto de las tres familias de modelos de aprendizaje según las descripciones individuales y según las agrupaciones de descripciones.

ISSN: 2792-3770

© ADIC - IIC

Métrica de evaluación.

Para la evaluación, se ha utilizado la métrica que se centra en la *top k accuracy score,* es decir, que mide la precisión del modelo sobre las *k* predicciones con mayor probabilidad. La selección de esta métrica viene motivada por el escenario real en el que el sistema se iba a poner en producción y donde los resultados del modelo de clasificación se ofrecen al experto como recomendaciones. Por esta razón, resulta más práctico que el usuario final pueda elegir entre las mejores predicciones del modelo. Según este criterio, se ha decidido realizar la evaluación adoptando una medida de *k=5,* es decir los *top 5 accuracy.*

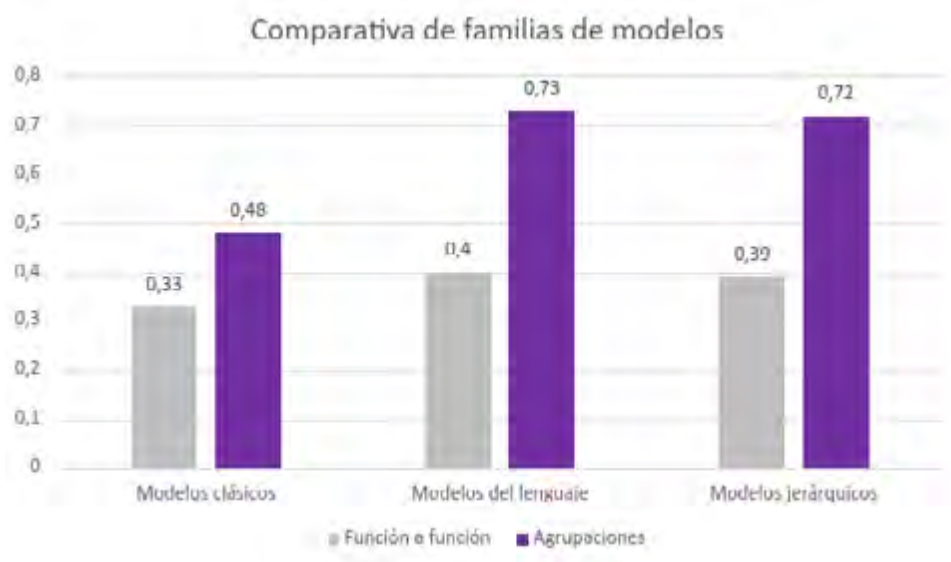
5. Resultados

En la sección anterior, se han descrito las fases del desarrollo y las estrategias seguidas frente a los retos encontrados.

Realizado el entrenamiento de las diferentes familias de modelos, se ha llevado a cabo la evaluación de estos modelos de clasificación siguiendo los dos métodos señalados: la evaluación por textos individuales y la evaluación por agrupaciones de textos. La siguiente gráfica muestra las diferencias en los resultados obtenidos entre familias de modelos y los dos métodos de evaluación adoptados. A simple vista se puede apreciar cómo la familia de modelos de aprendizaje basados en modelos del lenguaje obtiene mejores resultados que las familias de modelos clásicos y jerárquicos.

Asimismo, es importante mencionar que la estrategia de agrupación de descripciones de puestos genéricos ha resultado práctica, dado que los modelos de todas las familias experimentaron mejoras sustanciales aplicando la estrategia de entrenamiento por agrupaciones para predecir un único puesto frente a los modelos entrenados con pares descripción/puesto.

Figura 3. Comparativa de resultados en función de la familia de modelos y del método de evaluación



ISSN: 2792-3770

Analizando los resultados de la familia de modelos del lenguaje y con el fin de mostrar la viabilidad de la solución, se han planteado cinco posibles escenarios (A-E) de análisis de los resultados, tal y como se recoge en la tabla 1. Estos escenarios toman en cuenta la frecuencia de los puestos y el número de descripciones por cada puesto. Para cada escenario se han calculado las métricas de acierto del modelo.

El escenario A muestra el acierto del modelo (~0,73) teniendo en cuenta la totalidad de los puestos y descriptores. Los escenarios B y C contemplan unos resultados de entre el 0,75 y ~0,78 considerando el 75% y el 68% de los puestos, respectivamente. En el escenario D, se evalúa el modelo teniendo en cuenta el 25% de los puestos más frecuentes. En este escenario, el modelo predice el puesto correspondiente correctamente con un acierto que asciende a 0,84. Finalmente, en el escenario E, el modelo predice con un acierto de 0,89 el 12% de los puestos (los que cuentan con más de 1000 descripciones asociadas).

ISSN: 2792-3770

Tabla 2. Resultados de familia de modelos de lenguaje

		Resultados de familia de modelos de lenguaje		
Escenario		Métrica	% Puesto/ escenario	% Descriptores/ escenario
A	Datos totales	~0,73	100%	100%
B	75% puestos más frecuentes	~0,75	~75%	~99%
C	Puestos con > 100 descripciones	~0,76	~68%	~98%
D	25% puestos más frecuentes	~0,84	~25%	~66%
E	Puestos con > 1000 descripciones	~0,89	~12%	~43%

© ADIC - IIC

6. Solución final

En base a los resultados obtenidos, se ha desarrollado un servicio API (Application Programming Interface). El servicio integra el modelo de clasificación con los mejores resultados, i.e. modelos del lenguaje con agrupaciones de descriptores.

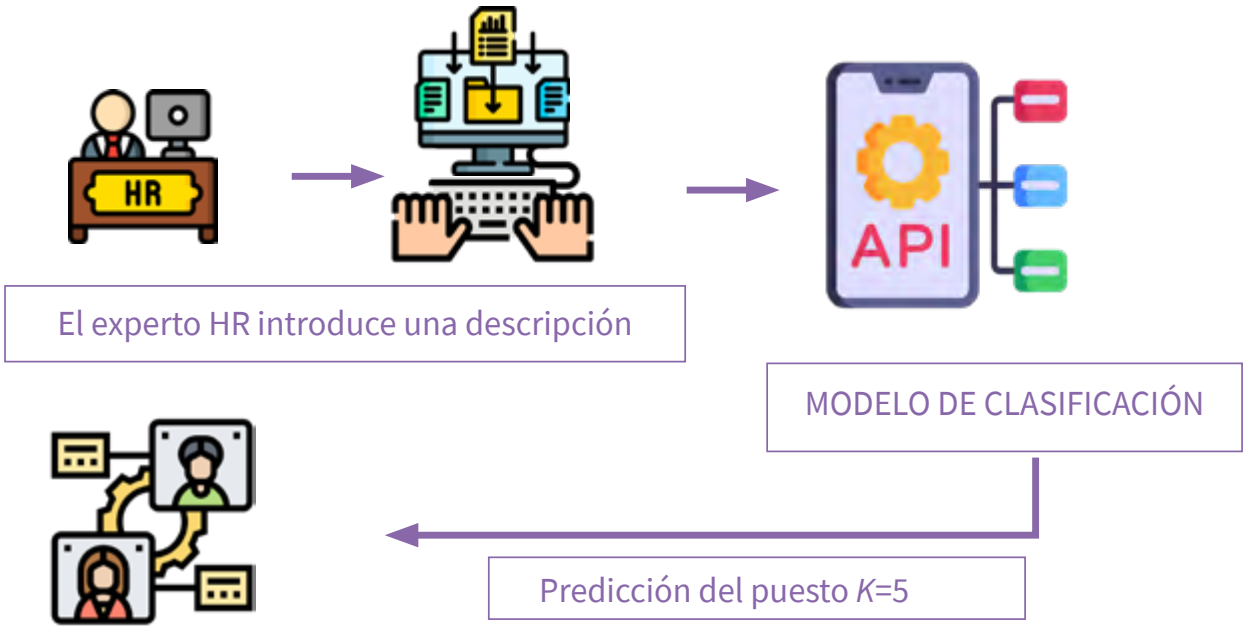
A través de este servicio, cualquier persona con un conocimiento de las responsabilidades de los puestos de una organización, sin ser el experto de HR, puede introducir una descripción y el servicio predice en tiempo real los puestos que se asocian con esta descripción. La predicción de puestos se organiza en orden descendente según la métrica top k accuracy score=5, es decir, que mide la precisión del modelo sobre las 5 predicciones con mayor probabilidad.

Es un servicio online, sencillo e interactivo que se puede consultar en tiempo real ofreciendo de manera práctica y eficiente un apoyo a cualquier empresa y/o entidad, sea cual sea su tamaño y ubicación, en la tarea de clasificación de puestos de trabajo y estimación salarial. El servicio se ha definido para que la consultoría estratégica en compensación sea accesible a empresas de cualquier dimensión (hasta ahora el elevado coste de este servicio, impedía que empresas más pequeñas o startups pudieran utilizar criterios profesionalizados). El servicio permitirá definir los puestos y diseñar el modelo retributivo sin grandes recursos económicos ni dedicación de tiempo ni conocimiento experto.

El servicio será evolutivo e irá incorporando nuevas funcionalidades a medida que las organizaciones lo vayan utilizando. Introducimos algunas posibles mejoras en la siguiente sección.

ISSN: 2792-3770

Figura 4. Solución final



7. Discusión

Basándose en los resultados obtenidos, se puede concluir que la aplicación del PLN a tareas del dominio HR resulta viable y práctica. Estas técnicas pueden proporcionar al experto soluciones que les asistan a llevar a cabo su trabajo de forma ágil y eficiente a través de la automatización de algunas tareas. El caso de la clasificación de descripciones de puestos de trabajo, objeto de este artículo, es un ejemplo de estas tareas en el que el PLN proporciona una base sobre la que el experto pueda trabajar reduciendo el tiempo y el esfuerzo invertido.

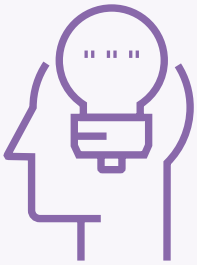
Por una parte, técnicamente se ha comprobado la importancia tanto de la calidad como la cantidad de los datos. Es imprescindible contar con un número mínimo de descripciones para que el modelo de clasificación pueda obtener resultados relevantes. En este caso concreto ese número se sitúa en torno a las 300 descripciones aunque, como se ha demostrado, cuantas más descripciones representativas e informativas del puesto al que pertenecen, mejores son los resultados.

Por otra parte, se ha comprobado cómo los nuevos modelos del lenguaje basados en redes neuronales son mucho más efectivos que las técnicas de PLN más clásicas que se venían usando hasta estos últimos años.

Por último y como pasos futuros para la mejora del modelo, se propone utilizar técnicas concretas de *multi-instance learning* para solventar este problema. Dichas técnicas tratan de mejorar la comprensión que tiene el modelo de una muestra que está compuesta por varios textos distintos, como es el caso de uso que ha sido presentado en este artículo.

Otra posible mejora podría ser el uso de modelos compatibles con el *active learning* para que, al poner en producción el modelo, el usuario pueda darle *feedback* y mejorar tanto el *dataset* como las predicciones a medida que es explotado el modelo.

ISSN: 2792-3770



8. Referencias bibliográficas

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. Human–Computer Interaction, 35(5-6), 545-575.

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. The International Journal of Human Resource Management, 33(6), 1237-1266.

© ADIC - IIC

instituto de ingeniería del conocimiento

INSTITUTO DE INGENIERÍA DEL CONOCIMIENTO (IIC)

PIONEROS EN INTELIGENCIA ARTIFICIAL DESDE 1989

Somos un centro de I+D+i experto en **Big Data e Inteligencia Artificial**. El núcleo, experiencia y trayectoria del IIC gira en torno al análisis de datos.

Nuestra apuesta de valor se basa en el desarrollo de algoritmos y técnicas de análisis a medida, de modo que conformen soluciones tecnológicas altamente adaptadas a las necesidades de cada cliente.

Únete a un equipo joven y dinámico, formado por más de 150 profesionales especializados en tecnologías de vanguardia.
Estamos ubicados en la Universidad Autónoma de Madrid (UAM). Nos nutrimos del mejor talento universitario y somos nexo entre la universidad y la empresa.

Nuestros productos tienen **presencia internacional**: Alemania, Argentina, Australia, Brasil, Colombia, EE. UU., España, Italia, México, Panamá, Paraguay, Perú, Portugal, Reino Unido, Rumanía,Venezuela.

Puedes desarrollar tu carrera profesional como analista, desarrollador o científico de datos en todos los sectores, siendo especialistas en:

ENTORNO BANCARIO | ENTORNO RR.HH. | ENTORNO SEGUROS | ENTORNO DIGITAL | ENTORNO ENERGÍA | ENTORNO SALUD | ENTORNO INTELIGENCIA DE CLIENTE

NUESTROS ASOCIADOS:

Santander | MAPFRE | IBM | UAM Universidad Autónoma de Madrid

TECNOLOGÍA Y ANÁLISIS DE DATOS AL SERVICIO DE RR. HH.

HR ANALYTICS

El análisis de los datos de RR. HH. permite obtener información de valor para una mejor gestión del talento. El IIC aplica **analítica descriptiva y predictiva** para optimizar procesos de selección, predecir el absentismo o la rotación e identificar a los profesionales con más potencial, entre otros proyectos.

EVALUACIÓN DE COMPETENCIAS

Dentro de la **plataforma online eValue,** desarrollamos pruebas objetivas y fiables para evaluar las competencias transversales, el nivel de inglés o las motivaciones de candidatos y empleados. Además de tomar mejores decisiones, se obtienen **datos de calidad** para analizar, por ejemplo, sus necesidades de formación.

ANÁLISIS DE REDES ORGANIZACIONALES

Los proyectos AROS permiten analizar las relaciones de trabajo y las interacciones entre los profesionales. Representadas visualmente en un grafo, se pueden identificar **redes informales, referentes ocultos o cuellos de botella** en la organización, para emprender acciones de mejora.

www.linkedin.com/company/instituto-de-ingenier-a-del-conocimiento---iic

www.twitter.com/IIConocimiento

www.youtube.com/IIConocimiento

INNOVADATA

Instituto de Ingeniería del Conocimiento

C/ Francisco Tomás y Valiente, 11 EPS, Edificio B, 5ª planta UAM Cantoblanco. 28049 Madrid

http://www.iic.uam.es/empleo-iic/

rrhh@iic.uam.es

(+34) 91 497 2323

20
22

instituto
de ingeniería
del conocimiento

www.iic.uam.es

Instituto de Ingeniería del Conocimiento

C/ Francisco Tomás y Valiente, nº 11
Escuela Politécnica Superior (EPS),

Edificio B, 5ª planta
Universidad Autónoma de Madrid (UAM).

28049 Cantoblanco, Madrid

T. (+34) 91 497 23 23