uc3m | Universidad Carlos III de Madrid

Grado Universitario en Ingeniería en Tecnologías de Telecomunicación

Curso 2020/2021

Trabajo Fin de Grado

"HERRAMIENTA PARA OBTENCIÓN DE DATOS DEL SISTEMA DE PUBLICIDAD ONLINE DE LINKEDIN"

Ángel Merino Hernández

Tutor

Ángel Cuevas Rumín

Leganés, 24 de junio de 2021

RESUMEN

La publicidad online es un gran negocio sustentado en la compraventa de datos, donde empresas de redes sociales como Facebook o LinkedIn ofrecen anuncios personalizados en función de las características de sus usuarios.

En este trabajo se diseña una herramienta software capaz de obtener información de manera automatizada y masiva de la plataforma de anuncios de LinkedIn que, dado un conjunto de características que definen a un público determinado dentro del conjunto de sus usuarios, informa del tamaño de este, el coste y rendimiento estimados de la campaña publicitaria y una segmentación adicional del público en función de una característica elegida, por ejemplo, los años de experiencia profesional. Toda esta información es relevante obtenida de forma masiva porque permite conocer el valor de cada conjunto de datos, obteniendo el coste de dirigirse a diferentes públicos específicos en tiempo real. El módulo implementado se integrará en una herramienta multiplataforma para cuantificar el valor de mercado de los datos. Además, estos datos permiten estudiar fenómenos sociales, en el caso de LinkedIn relacionados con el mercado laboral; por ello, para completar el trabajo se aborda un caso de uso de la herramienta implementada en el que se analiza la brecha de género por áreas laborales y países en todo el mundo.

Los resultados demuestran la usabilidad de la herramienta para este tipo de estudios y revelan que los países menos desarrollados tienden a mostrar una mayor brecha de género, al menos dentro del conjunto de profesionales que utilizan LinkedIn.

Palabras clave: Publicidad online; Valor económico; Datos masivos; Crawling; LinkedIn.

DEDICATORIA

A todos los que me han acompañado y apoyado durante la elaboración de este trabajo y durante estos últimos cuatro años, tanto en la cercanía como en la distancia, y sobre todo a mis padres. Gracias.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ABREVIATURAS

API	Application Programming Interface
BLS	Bureau of Labor Statistics
CPC	Cost Per Click
CPM	Cost Per Mile
CSRF	Cross-site request forgery
DNS	Domain Name System
DVTMP	Data Valuation Tool from the Market Perspective
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
JSON	JavaScript Object Notation
LCM	LinkedIn Campaign Manager
LDC	LinkedIn Data Crawler
TCP	Transmission Control Protocol
URL	Uniform Resource Locator
URN	Uniform Resource Name

1. INTRODUCCIÓN Y OBJETIVOS

1.1. INTRODUCCIÓN

La publicidad online tiene una gran importancia hoy en día para el mundo empresarial. Al navegar por Internet, es prácticamente imposible no toparse con anuncios. Sin ir más lejos, cada vez que se hace una búsqueda en Google, los primeros resultados, aunque estén relacionados con la búsqueda, son anuncios: los propietarios de la web, producto o servicio que se ha buscado o que está relacionado con la búsqueda han pagado para que el enlace a su página aparezca ahí, como primer resultado, a una persona potencialmente interesada en acceder a ella y que puede acabar comprando el producto.

Esto es solo un ejemplo, pero existen otros muchos, en especial las redes sociales, que se asemejan a una gran plaza llena de pantallas publicitarias, con anuncios camuflados como si fueran una publicación más, en forma de imagen, vídeo, mensaje y muchas otras posibilidades. Esta mimetización de la publicidad o contenido patrocinado con el contenido regular hace que los usuarios pongan mayor atención en lo que se les está intentando vender. Esto, unido a que los anuncios que cada usuario ve están casi siempre personalizados en base a sus datos personales, incluyendo muchas veces los relacionados con su personalidad, gustos e intereses, hace que la publicidad online sea una de las formas más eficientes y eficaces de conectar vendedores de productos o servicios con los clientes más interesados en ellos. Es más eficiente porque es más barato que la publicidad tradicional y además los anuncios están más integrados en el contenido que consume el usuario, siendo más fácil retener su atención en el anuncio; y a la vez es más eficaz porque se dirige de forma específica a los usuarios que decide el anunciante y no a todos.

En efecto, lo que hace la publicidad online tan atractiva es la posibilidad de llegar a un público objetivo muy concreto, que cumple una serie de características elegidas por el anunciante. Las redes sociales que ofrecen este servicio muestran los anuncios a los usuarios seleccionados en base a los datos de su perfil (edad, sexo, aficiones, etc.). Cuando un anunciante desea lanzar una campaña publicitaria, dispone de interfaces en las que selecciona los parámetros del público objetivo, con múltiples opciones de segmentación que dependen de los datos que esa red social tiene de sus usuarios. Por ejemplo, Facebook tiene muchos datos sobre los intereses de sus usuarios a nivel personal, en cambio LinkedIn tiene datos más relacionados con los intereses laborales de cada

usuario, las características de la empresa en la que trabaja y su perfil profesional. La elección de un perfil concreto combinando todas estas facetas resulta en infinitas posibilidades que dependerán de las estrategias de marketing demográficas y psicográficas de cada anunciante. Al elegir este público objetivo, generalmente las plataformas muestran una estimación de rendimiento y coste: número de visualizaciones, de clics, coste por visualización o por clic y coste total de la campaña, entre otros. Esta información es muy útil a la hora de planificar las campañas, diseñar, revisar y corregir continuamente la estrategia, y prever los resultados mejorando así la planificación de producción o la previsión de demanda del servicio.

Sin embargo, este sistema tiene múltiples inconvenientes desde sus inicios que siguen hoy en día sin resolverse de forma definitiva. La falta de transparencia es uno de ellos: los precios de cada perfil varían constantemente, se basan en sistemas de puja, a menudo controlados por sistemas automáticos o semiautomáticos en los que la transparencia es muy limitada de cara a los anunciantes. Además, los usuarios en su gran mayoría no son conscientes del gran beneficio económico que estas empresas tecnológicas generan a través del tratamiento de sus datos personales. En los últimos años este tema ha adquirido mucha importancia a ojos de las entidades reguladoras y gobiernos, y desde el año 2018 existe un Reglamento Europeo de Protección de Datos, de obligado cumplimiento en toda la Unión que persigue aumentar la consciencia sobre el tema, defender a los usuarios, establecer los mecanismos para que estos puedan ejercer fácilmente sus derechos respecto a sus datos personales, y aumentar la responsabilidad y la implicación de estas grandes empresas tecnológicas en las buenas prácticas y la trasparencia.

Aún así, para avanzar en esos objetivos es necesario el desarrollo de herramientas y aplicaciones que permitan arrojar luz sobre estas cuestiones, sobre el valor real de la publicidad online, para aumentar la competencia y reducir la opacidad del sistema. En concreto, el sistema de publicidad de la red social LinkedIn está actualmente menos estudiado que otros, y no existen herramientas que potencien la transparencia de su sistema y sus precios como las desarrolladas, por ejemplo, para Facebook. Además, los datos que maneja esta red social están completamente ligados a los perfiles laborales de los usuarios, y son muy amplios y detallados, convirtiendo a su plataforma de publicidad en una potencial fuente de datos fiables del mercado laboral a nivel mundial, útiles para la realización de estudios sociales que ayuden a revelar dinámicas que los datos tradicionales obtenidos mediante encuestas no pueden revelar.

Si conseguimos automatizar el proceso de búsqueda de poblaciones y ejecutarlo masivamente con distintas combinaciones, tendremos acceso a ingentes cantidades de datos útiles para analizar múltiples fenómenos sociales relacionados con la economía y el empleo. En este sentido, en el presente trabajo desarrollará una herramienta de este tipo dirigida al sistema de campañas de LinkedIn, con los objetivos que se describen en el siguiente apartado.

1.2. OBJETIVOS

El objetivo del presente trabajo es diseñar e implementar un módulo software que sea capaz de obtener de forma automática la información que ofrece la plataforma de publicidad online de LinkedIn, y que puede ser útil en múltiples áreas de investigación, desde el estudio del valor económico de la publicidad online hasta la elaboración de estudios sociológicos enmarcados en el mercado de trabajo.

Lo que diferencia a LinkedIn de otras redes sociales es, precisamente, su orientación al mundo laboral. Por ello, los datos que manejan y que ofrecen como valor añadido en su plataforma de publicidad online están en su mayoría relacionados con esta faceta, y una herramienta capaz de obtener esos datos de manera automática sería una fuente de datos muy novedosa, clave para avanzar en las dos líneas de investigación mencionadas y potenciar su desarrollo, y que tendría además la capacidad de abrir otras líneas de investigación basada en este tipo de datos en función de la evolución de la herramienta y la variedad de datos que pueda ofrecer en el futuro.

En concreto, los tres tipos de datos que se busca obtener a través de el módulo software a diseñar son los siguientes:

- Número total de usuarios que componen una audiencia: el tamaño de la audiencia es el número aproximado de usuarios que cumplen determinadas condiciones o conjunto de características (ubicación o ubicaciones, características profesionales, demográficas o de la empresa en la que trabajan). Se pretende obtener este dato para cualquier audiencia definida por un conjunto de características como el descrito de forma automática a través de el módulo software a diseñar. Con esta información se pueden hacer estudios sobre la masa laboral e investigar fenómenos que se dan en ella, por ejemplo, cómo cambia la cantidad de trabajadores que se dedican a un área laboral en función de su género o su edad.

- Desglose de la audiencia en grupos: información sobre el porcentaje de individuos de la audiencia que pertenece a cada grupo, de entre una lista que incluye opciones de desglose por función laboral, por años de experiencia y otras. Estos datos son muy útiles para estudiar la distribución de una característica concreta en diferentes audiencias, por ejemplo, para estudiar cómo se distribuye el nivel de experiencia en determinados grupos de trabajadores y cómo cambia esa distribución en función del área laboral.

- Parámetros de coste de la campaña: son diversos parámetros que definen el coste de lanzar una campaña publicitaria dirigida a una audiencia. Estos datos, que pueden incluir el número de visualizaciones esperado, el coste diario, semanal, etc., son útiles si se busca determinar el valor de los datos que en este caso LinkedIn tiene disponibles a través de los usuarios de su red social, y por tanto cuantificar el valor económico o precio que los usuarios pagan por utilizar sus servicios en forma de cesión de datos personales. Al existir otras plataformas de publicidad online similares de otras redes sociales, como Facebook, estos datos podían también ser comparados con los obtenidos de las demás plataformas para determinar quién ofrece los mejores precios o las mejores cotas de alcance y visualizaciones para una determinada audiencia.

El desarrollo de este módulo software pretende a su vez dar cumplimiento en primera instancia a dos subobjetivos: Por un lado, integrar el módulo como parte del proyecto europeo PIMCity, del que se dan más detalles a continuación; por otro, efectuar un análisis propio de la brecha de género en el entorno laboral en todo el mundo con datos obtenidos utilizando en módulo.

El módulo software desarrollado se integrará como parte del proyecto europeo PIMCity, del programa H2020 (nº de proyecto 871370), cuyo objetivo es “asegurar que los ciudadanos, las empresas y las organizaciones están informadas y pueden hacer un uso ético y respetuoso de los datos personales” [1]. El proyecto busca la capacidad de seleccionar, clasificar y evaluar información de interés desde el punto de vista de la privacidad y la gestión de los datos personales para ciudadanos y empresas, y en él se diseñan y prueban herramientas y mecanismos que puedan incrementar la consciencia de

los usuarios sobre estos temas, para facilitar esa labor didáctica que logre que la sociedad esté bien informada sobre el uso de datos.

Sus principales objetivos, recogidos también en su página web, son los siguientes:

- Hacer a los usuarios conscientes de la importancia de la privacidad y la gestión de datos personales.
- Ayudar a los ciudadanos que ya son conscientes de lo anterior mediante herramientas útiles para obtener información relevante sobre estos temas.
- Contribuir a la creación de un espacio europeo de datos personales.

En concreto, este trabajo se enmarca en las herramientas de valoración de datos desde la perspectiva del mercado. El módulo del proyecto que se encarga de esto (DVTMP: Data Valuation Tools from the Market Perspective) aprovechará las plataformas de publicidad online más populares para obtener el valor monetario de cientos o miles de audiencias (perfiles de usuarios) diferentes, lo que permitirá a cualquier gestor de información de producto (en este caso el producto son los propios datos sobre las audiencias) tener una estimación suficientemente precisa del valor de las audiencias para poder competir en el mercado de venta de datos basados en audiencias. [2]

Los objetivos de ese módulo del proyecto son los siguientes:

- Obtener el valor de los datos de audiencias de Facebook, Instagram y LinkedIn.
- Procesar, curar y almacenar los datos.
- Proporcionar acceso a los datos a través de una API.

De este modo, el módulo desarrollado en este trabajo se integrará como parte del DVTMP proporcionando toda la implementación correspondiente al acceso a la plataforma de LinkedIn, la obtención de los datos y su presentación en un formato compatible y legible por el resto de las piezas de software que compongan la herramienta final.

El segundo subobjetivo que se persigue con el desarrollo del módulo software está relacionado con la segunda parte de este trabajo, y es demostrar su gran valor social utilizando sus capacidades para obtener datos sobre un fenómeno con gran impacto en la sociedad actualmente: la brecha de género en el mundo laboral. Se estudiará este fenómeno entre países de todo el mundo en función del área laboral de los trabajadores.

Este caso de uso se desarrolla en el capítulo 5 de este trabajo. Para ello, se tendrá que tratar la escalabilidad de la herramienta, es decir, evaluar su capacidad de obtener datos de forma masiva, haciendo las modificaciones pertinentes para adaptarla a esa necesidad como se describe en la sección 4.5.

2. CONTEXTO PREVIO

Este trabajo persigue poder realizar un uso automatizado de la plataforma de publicidad online de LinkedIn, por lo que es fundamental entender su funcionamiento, los servicios que ofrece y la variedad de opciones que incluye.

Para ello, primero se describirá cómo se puede crear una cuenta para empezar a usar la plataforma, después se describirán las diferentes opciones publicitarias disponibles, las opciones de segmentación, la información que se requiere para lanzar la campaña publicitaria, y el significado de cada uno de los términos que forman las previsiones que LinkedIn muestra al definir un grupo objetivo. Esta información es la que se pretende recopilar de manera automática con el sistema que se desarrolla en el trabajo.

2.1. CREANDO UNA CAMPAÑA PUBLICITARIA

La plataforma de anuncios de LinkedIn está accesible a cualquiera que tenga una cuenta en esta red social. En la URL linkedin.com/campaignmanager/login, se inicia sesión y, si no se ha utilizado nunca la plataforma de publicidad con esa cuenta aparece una pantalla como la siguiente:

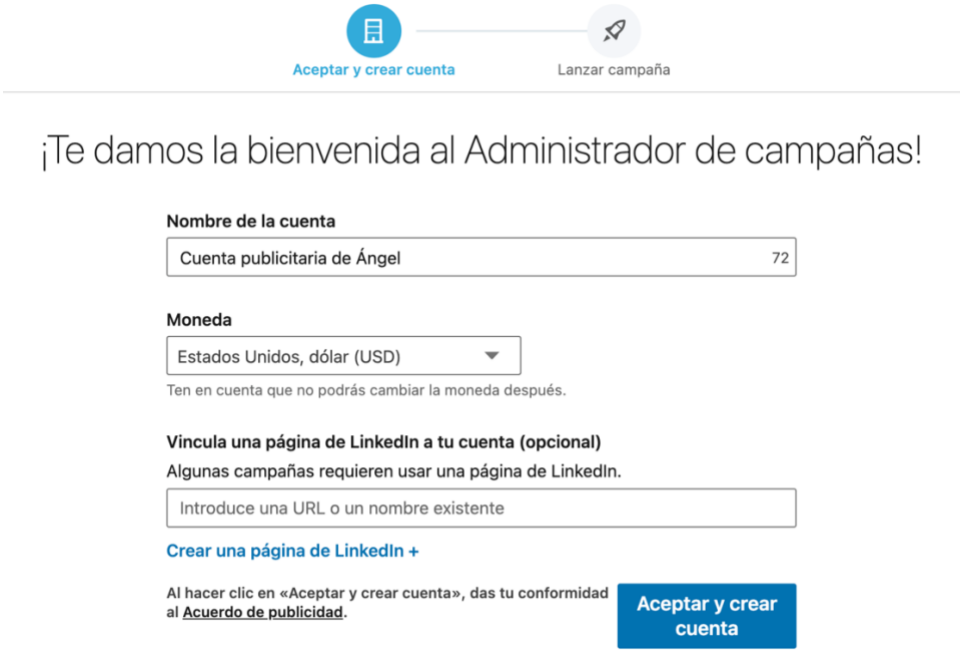

Fig. 2.1. Página de bienvenida al administrador de campañas publicitarias. [3]

Dentro de la cuenta de LinkedIn, hay que crear una nueva cuenta del administrador de campañas publicitarias, seleccionar la divisa que queremos utilizar, y vincular una página de LinkedIn a la cuenta para utilizar en las campañas. Como se indica en la imagen, y como se verá más adelante, para algunos de los diferentes tipos de campañas publicitarias es obligatorio vincular una página de LinkedIn. Si se vincula una página a la cuenta publicitaria, se usará en todas las campañas que se lancen con la cuenta. Una vez rellenos los campos se procede a crear la cuenta, lo que implica dar la conformidad con el Acuerdo de Publicidad de LinkedIn [4].

A continuación, pide elegir el grupo de campañas, si no se tienen grupo creados, se puede crear un nuevo grupo o utilizar el grupo por defecto. Se le puede dar un nombre para diferenciarlo de otros grupos que se puedan crear más adelante.

Los grupos sirven para agrupar las campañas y trabajar con múltiples proyectos publicitarios de forma organizada, gestionando en un mismo grupo las que compartan ubicación, presupuesto, o alguna otra característica. Existe un límite máximo de 1000 campañas por grupo. Para gestionar campañas que tienen distintas audiencias objetivo o distintas estrategias de marketing, se pueden utilizar grupos diferentes. En cada grupo se puede establecer un presupuesto global para todas sus campañas, y también activar o pausar el grupo, activando o pausando así todas sus campañas a la vez. Esto permite tener un mayor control sobre el coste y la programación de las campañas. Es importante saber que, una vez iniciadas las campañas, no se pueden mover de grupo.

En el grupo por defecto se agrupan las campañas existentes antes del lanzamiento de la función de grupos en administrador de campañas. También es el grupo al que por defecto se asignan las campañas si no se selecciona otro. En el grupo por defecto no se pueden cambiar los ajustes de presupuesto y programación, sino que de forma fija se activan las campañas de forma continuada a partir de su creación. [5]

2.2. OPCIONES DE OPTIMIZACIÓN

Tras asignar la campaña a un grupo, aparece la ventana de configuración. En primer lugar, se da a elegir entre diferentes objetivos según las necesidades del anunciante, divididos en tres categorías: conocimiento, percepción y conversiones. En cada categoría se ofrecen distintos objetivos:

- Conocimiento:
    - Conocimiento de la marca: sirve para dar a conocer los servicios, productos y marca del negocio a más gente.
- Percepción:
    - Visitas al sitio web: sirve para generar tráfico a un sitio web o a una página o evento de LinkedIn. La campaña se mostrará a personas que tienen más probabilidades de hacer clic en los anuncios.

    - Interacción: sirve para potenciar la interacción social con el contenido del anunciante en LinkedIn (páginas y eventos), y aumentar el número de seguidores de la empresa. El anuncio se muestra a las personas que tienen más probabilidades de seguir al anunciante e interactuar con su página.

    - Visualizaciones de vídeo: sirve para potenciar las visualizaciones de un anuncio en vídeo. La campaña se mostrará a las personas que tengan mayor probabilidad de ver el vídeo.
- Conversiones:
    - Generación de contactos: sirve para obtener más posibles clientes de calidad. Permite integrar directamente una plataforma de generación de contactos, incluyendo de forma automática la información del perfil de LinkedIn. La campaña se mostrará a las personas que tienen más probabilidad de rellenar un formulario de generación de contactos.

    - Conversiones en el sitio web: sirve para obtener más compras del producto o altas en el servicio que se oferta. La campaña se mostrará a las personas que tienen más probabilidad de realizar las acciones que le interesan al anunciante, que se configuran y monitorizan a través del menú de seguimiento de conversiones.

    - Solicitudes de empleo: sirve para promocionar oportunidades laborales en la empresa anunciante. Permite recibir más solicitudes de trabajo, porque la campaña se muestra a las personas que tienen mayor probabilidad de ver y hacer clic en los anuncios de empleo.

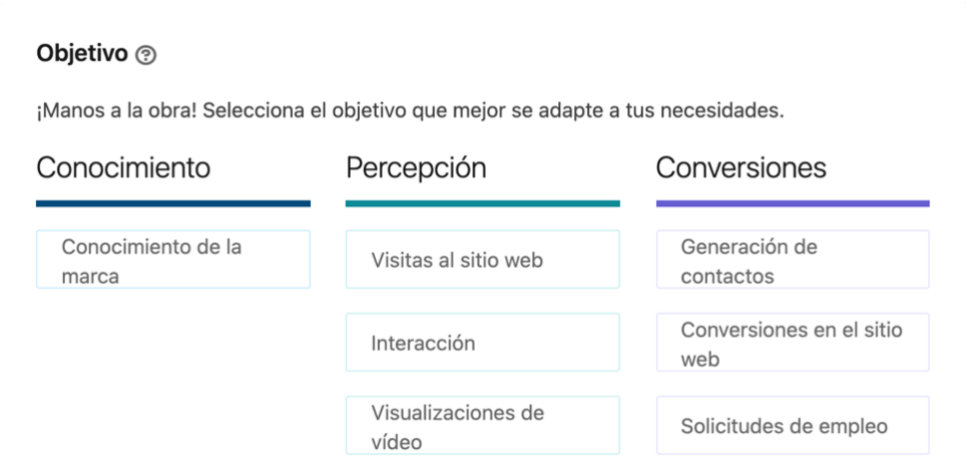

Fig. 2.2. Objetivos de campaña. [3]

2.3. OPCIONES SEGMENTACIÓN

El siguiente paso es establecer un criterio de segmentación para seleccionar el público objetivo de la campaña. El objetivo es aprovechar los datos de los perfiles de LinkedIn para lanzar la campaña solamente hacia personas con un perfil determinado, por ejemplo, público en Madrid que trabaje en el área de ingeniería.

Para seleccionar los criterios de segmentación, el gestor de campañas ofrece una interfaz en la que lo primero que se debe especificar es una ubicación, esto puede ser un país, una región o una ciudad. Es obligatorio introducir al menos una ubicación, aunque se pueden introducir varias, de este modo la campaña se lanzaría al público de todas ellas. También se permite excluir ubicaciones, esto es útil si, por ejemplo, se pretende lanzar la campaña al público de toda una región excepto su ciudad principal. Después, permite seleccionar el idioma en el que el público tiene configurado su perfil, por defecto se escoge inglés para llegar a todos los usuarios.

Para acotar más la audiencia, se ofrece una interfaz de selección de características de perfiles profesionales compuesto por una barra de búsqueda y dos opciones de segmentación:

- Públicos: permite llegar a cuentas y contactos conocidos y se basa en la carga manual de listas que contienen empresas o contactos, en la búsqueda de público similar al que ya tiene la página del anunciante, o en segmentaciones adicionales dentro de ese mismo público.

- Características del público: permite seleccionar directamente características de la audiencia, como su edad, género, industria en la que trabajan, aptitudes, cargos y otras.

Para este trabajo interesa detenerse en la opción características del público porque, como se verá más adelante, es donde se seleccionan los criterios de segmentación que requiere el proyecto en el que se enmarca este trabajo.

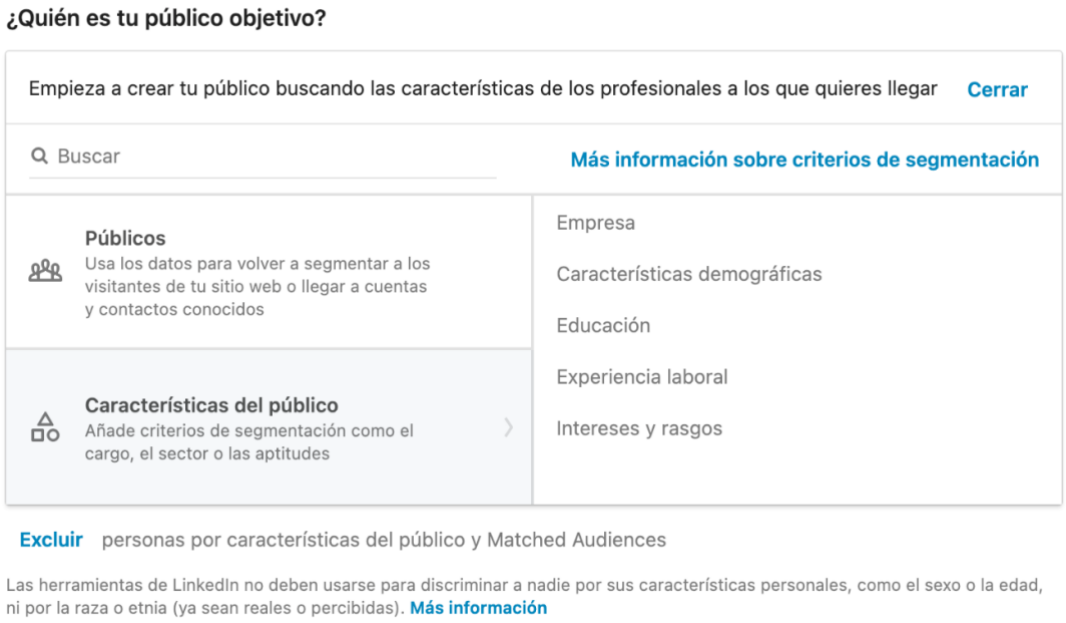

Fig. 2.3. Características del público ofrecidas por LinkedIn Campaign Manager. [3]

Es importante tener en cuenta que algunas de las opciones están limitadas y tienen un conjunto finito de elección en la interfaz, como por ejemplo el género (hombre o mujer), mientras que otras, como las aptitudes, son conjuntos tan grandes que se sirven de una interfaz de búsqueda. Ésta devuelve las aptitudes recogidas en la plataforma que mejor se ajusten al término buscado, por ejemplo, si se busca *programador*, el sistema devuelve:

1. Programador de TI
2. Programador web
3. Programador de sistemas
4. Programador de videojuegos
5. Programador de proyectos
6. …

Existe una barra de búsqueda global en la que, si se busca algún término, cada respuesta incluirá la categoría a la que pertenece (cargos, aptitudes…), y barras de búsqueda dentro de cada una de las categorías que admiten búsqueda en la que solo se devuelven resultados pertenecientes a estas.

A continuación, se exponen las opciones de segmentación más importantes que ofrece el gestor de campañas y la forma en que se clasifican en su interfaz. Se omiten algunas ya que hay listas de opciones seleccionables que son demasiado largas y no son relevantes para este trabajo.

- Empresa:
    - Categoría empresarial:
        - Empresas más innovadoras del mundo según Forbes
        - Fortune 100 (todo el mundo)
        - Fortune 1000 (solo EE.UU.)
        - …
    - Contactos de la empresa (ofrece barra de búsqueda)
    - Nombres de empresas (ofrece barra de búsqueda)
    - …
- Características demográficas:
    - Edad:
        - De 18 a 24
        - De 25 a 34
        - De 35 a 54
        - Más de 55
    - Sexo:
        - Hombre
        - Mujer
- Educación:
    - Disciplinas académicas (ofrece barra de búsqueda)
    - Instituciones educativas (ofrece barra de búsqueda)
    - Titulaciones (ofrece barra de búsqueda)
- Experiencia laboral:
    - Aptitudes (ofrece barra de búsqueda)

    - Años de experiencia (seleccionable un rango desde un año hasta más de doce años)
    - Cargos (ofrece barra de búsqueda)
    - Funciones laborales:
        - Administración
        - Arte y diseño
        - Atención al cliente
        - …
    - Niveles de experiencia
        - No remunerado
        - Formación
        - Principiante
        - …
- Intereses y rasgos:
    - Grupos de LinkedIn (ofrece barra de búsqueda)
    - Intereses de los miembros:
        - Intereses de productos:
            - Sistemas de gestión de aprendizaje (LMS)
            - Software de ciberseguridad
            - Software de colaboración
            - …
        - Intereses generales:
            - Arte y entretenimiento
            - Ciencia y medio ambiente
            - Economía y finanzas
            - …
    - Rasgos del miembro:
        - Ascenso reciente
        - Cambios de empleo reciente
        - Contribuidores recientes
        - …

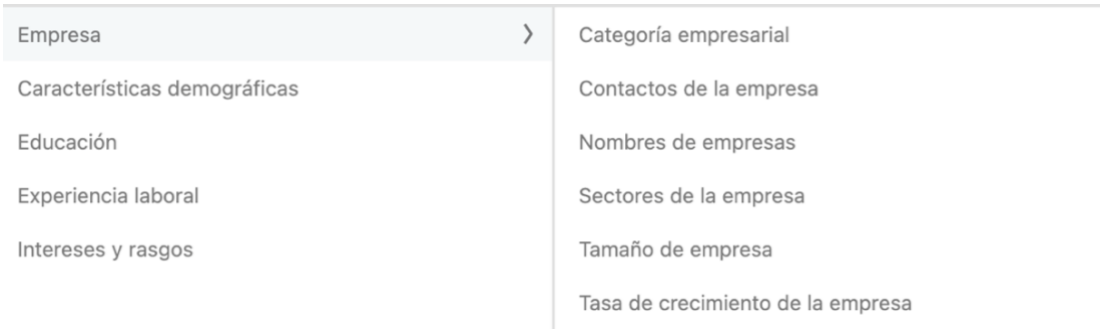

Fig. 2.4. Subcategorías dentro de la categoría de empresa. [3]

Esta interfaz permite hacer una combinación lógica de elementos, equivalente a combinar características mediante sentencias de suma y multiplicación lógicas, o como se llama en la plataforma, acotar más el público (AND) y seleccionar más características (OR).

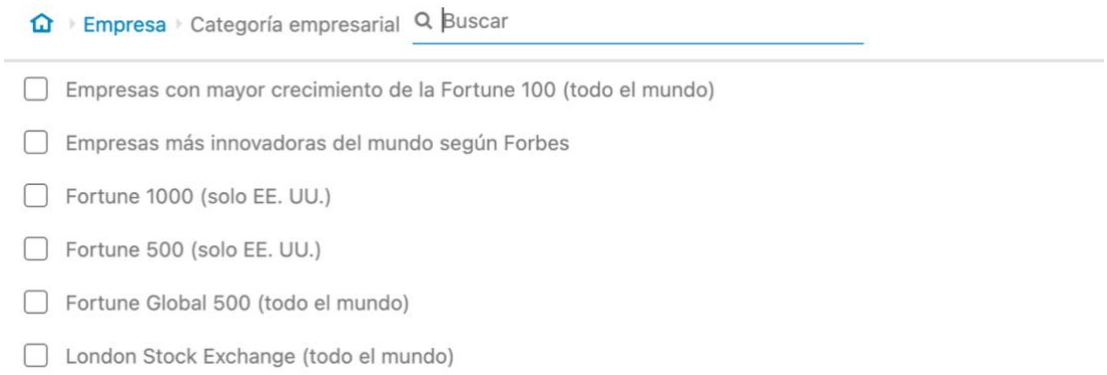

Fig. 2.5. Opciones de categoría empresarial. [3]

Inmediatamente después se puede habilitar o deshabilitar la expansión de público, esto es, si se quiere llegar a personas similares al público objetivo o exclusivamente al público que coincide con las características seleccionadas de forma exacta. En principio, esto aumentaría el tamaño total de la audiencia, sin embargo, en las pruebas que se han realizado no se ha observado nunca una variación en el número de miembros al activar o desactivar esta opción.

A continuación, se explicarán los datos y estimaciones que muestra la plataforma una vez se ha seleccionado una audiencia objetivo.

2.4. ESTIMACIONES DE COSTE Y RENDIMIENTO

Una vez seleccionada la audiencia objetivo, la plataforma muestra una serie de estimaciones y datos que son de utilidad al anunciante para hacerse una idea del alcance y el coste que tendrá su campaña. Esto es lo que la plataforma llama *pronóstico de resultados*.

En primer lugar, se ofrece el tamaño total de la audiencia objetivo (número de usuarios que encajan con el perfil seleccionado mediante las características del público). A continuación, se ofrece un pronóstico de gasto y rendimiento para un día, siete días y treinta días. Normalmente se utiliza como referencia el pronóstico para treinta días, pero los tres están disponibles para su visualización. El pronóstico incluye los siguientes datos:

- Presupuesto económico para el período: muestra una estimación (con cota inferior y superior) del dinero que costaría mantener la campaña activa durante el período temporal correspondiente (un día, siete días o treinta días). Cabe aclarar que en este punto aún no se ha seleccionado un presupuesto, por lo que la estimación está basada en el presupuesto diario por defecto de la plataforma, que son cincuenta dólares al día. El presupuesto de la campaña cambiará posteriormente al establecerse un presupuesto diario personalizado.
- Impresiones en el período: indica, con una cota inferior y una superior, el número de veces que el anuncio se mostrará a la audiencia objetivo durante el período seleccionado.

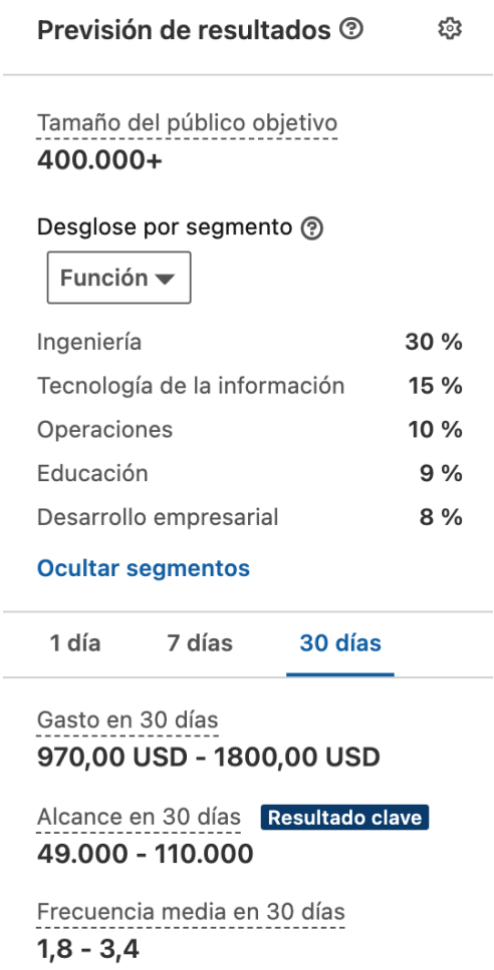

Fig. 2.6. Previsión de resultados. [3]

- CTR: porcentaje de clics respecto a número de impresiones, esto es, el tanto por ciento de impresiones que resultarán cliqueadas por los usuarios. También se expresa mediante cota inferior y superior.

- Número de clics en el período: muestra una cota inferior y superior de los clics esperados en el período de tiempo seleccionado.

Respecto a estos datos, la plataforma informa de que son estimaciones y no garantizan ningún nivel concreto de rendimiento. [6]

Además, se ofrece lo que la plataforma denomina *desglose por segmento*, en el que la plataforma divide la audiencia total según el criterio elegido, a saber: función laboral, nivel de responsabilidad, años de experiencia, tamaño de la empresa, sectores e intereses. Si se escogiese, por ejemplo, un desglose por sectores sobre la audiencia España (ubicación) + programación (aptitudes), el resultado sería el siguiente:

- Servicios y tecnologías de la información: 21%
- Software: 7%
- Enseñanza superior: 5%
- Desarrollo de programación: 4%
- Ingeniería industrial o mecánica: 4%

Se puede ver como los porcentajes no suman el 100%, esto es porque solo se muestran los 5 sectores con mayor porcentaje. También podría darse el caso de que los porcentajes sumasen más del 100%, en ese caso sería porque un mismo grupo de usuarios pertenece, a la vez, a varios sectores de los mostrados. [7]

2.5. FORMATO DEL ANUNCIO

El siguiente paso en la configuración de la campaña es la elección del formato del anuncio. En función del objetivo seleccionado al inicio de la configuración de la campaña (conocimiento de la marca, visitas al sitio web, etc.) se tendrá disponible un conjunto distinto de formatos. La lista completa de formatos se describe a continuación:

- Anuncio con una imagen: anuncios de una sola imagen que aparecerán en la *feed* de noticias de LinkedIn.
- Anuncio en carrusel: anuncios con dos o más imágenes que aparecerán en la *feed* de noticias de LinkedIn.

- Anuncio en vídeo: anuncios con un vídeo que aparecerá en la *feed* de noticias de LinkedIn.
- Anuncio de texto: se muestran en la columna derecha o en la parte superior de la página de LinkedIn.
- Anuncio para destacar: se personaliza con los datos del perfil para promocionar la oferta del anunciante a través de la versión de ordenador.

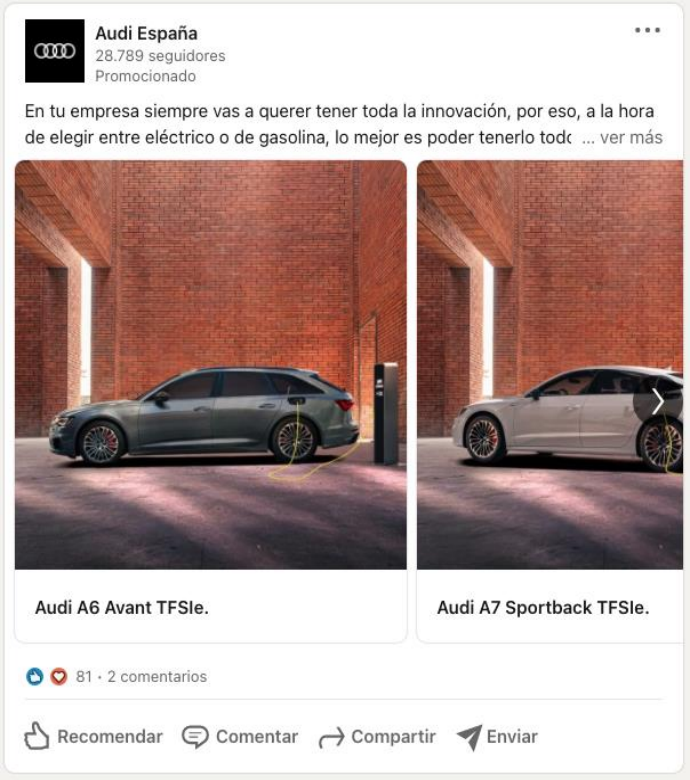

Fig. 2.7. Ejemplo de anuncio en carrusel. [3]

- Anuncio para obtener seguidores: similar al anterior, pero para promocionar la página del anunciante.
- Anuncio por mensaje: aparecen en el buzón de los mensajes de LinkedIn del usuario.
- Anuncio en conversación: similar al anterior, aparece en la conversación.
- Anuncio de evento: anuncia un evento de LinkedIn y aparece en la *feed* de noticias.
- Anuncio con un empleo: anuncio con un solo empleo que se muestra en la *feed* de LinkedIn.
- Anuncio de empleo: anuncio personalizado con los datos del perfil del usuario para promocionar empleos a través de la versión de ordenador.

Como se ha dicho, no todos los formatos están disponibles para todos los objetivos de campaña. A continuación, se describe cuáles están disponibles para cada objetivo:

- Conocimiento de la marca: todos excepto anuncio por mensaje, anuncio con un empleo y anuncio de empleo.
- Visitas al sitio web: todos excepto anuncio para obtener seguidores, anuncio con un empleo y anuncio de empleo.

- Interacción: anuncio con una imagen, anuncio en carrusel, anuncio de vídeo, anuncio para obtener seguidores, anuncio en conversación y anuncio de evento.
- Visualizaciones de vídeo: anuncio de vídeo.
- Generación de contactos: anuncio con una imagen, anuncio en carrusel, anuncio de vídeo, anuncio por mensaje, anuncio en conversación.
- Conversiones en el sitio web: todos excepto anuncio para obtener seguidores y anuncio de evento.
- Solicitudes de empleo: anuncio con un empleo, anuncio de empleo, anuncio con una imagen, anuncio para destacar.

En el siguiente apartado se verán las configuraciones que permiten establecer un presupuesto automatizado, programar el lanzamiento de la campaña y establecer el modelo y parámetros de puja.

2.6. PRESUPUESTO, PROGRAMACIÓN Y MODELO DE PUJA

Una vez elegido el tipo de anuncio, el siguiente paso es configurar el presupuesto y la programación temporal de la campaña. El grupo de campañas puede tener alguna configuración establecida para que sus campañas se pongan en marcha a partir de cierta fecha, en ese caso la plataforma nos informa de ello.

Para configurar el presupuesto existen diferentes opciones: se puede fijar un presupuesto diario, un presupuesto total, o ambos. El presupuesto se fija en dólares. Sin embargo, la plataforma nos advierte de que, en el caso del presupuesto diario, el gasto diario real puede variar, es decir, mediante el presupuesto diario no se fija un límite de gasto diario: cuando los anuncios tengan la posibilidad de obtener más resultados, se podrá superar en hasta un 50% el presupuesto diario, y otros días se gastará menos dinero, pero, según la información que ofrece LinkedIn en sus páginas de ayuda, “durante una semana completa (de lunes a domingo) nunca se superará el presupuesto diario multiplicado por siete. Si tiene un calendario establecido, con una fecha de inicio y finalización, el gasto total de la campaña no superará la cantidad igual al presupuesto diario multiplicado por el número de días de la programación.” [8]

Las opciones de programación temporal de la campaña dependen de la opción elegida para fijar el presupuesto:

- Si se ha escogido fijar el presupuesto diario, se podrá elegir entre activar la campaña de forma continuada (indefinida), o establecer las fechas de inicio y finalización. El gasto máximo total será entonces el presupuesto diario fijado multiplicado por el número de días que durará la campaña.

- Si se ha escogido fijar el presupuesto total, se deben establecer las fechas de inicio y finalización.

- Si se ha elegido fijar tanto el presupuesto diario como el total, se deberá establecer una fecha de inicio de la campaña, y estará activa hasta que se agote el presupuesto total.

El siguiente paso es fijar la estrategia de puja. La publicidad online en general, también en otras plataformas, se rige por sistemas de puja en los que los anunciantes ofrecen un precio (una puja) por unidad publicitaria. Esta unidad publicitaria normalmente es un clic en el anuncio o una impresión. De este modo, existen dos métricas generales para pujar: el coste por clic (CPC) y el coste por mil impresiones (CPM). Los anunciantes elijen su objetivo (clics o impresiones) y ofrecen manualmente una puja en estos términos, y después es la plataforma la que automáticamente adjudica cada espacio publicitario de los perfiles de la audiencia seleccionada al anunciante que más haya pujado por ese objetivo. En algunas plataformas existen también otras métricas como coste por vídeo (CPV). Sin embargo, generalmente las plataformas ofrecen sistemas semiautomáticos o completamente automáticos para calcular las pujas, optimizando el presupuesto fijado para obtener la mayor cantidad de unidades del objetivo seleccionado, de forma que el anunciante no tenga que preocuparse por esto.

La plataforma de LinkedIn ofrece esas herramientas de optimización. En primer lugar, se puede elegir un objetivo de optimización. Además de los básicos (por clics o por impresiones) existen otros, que estarán disponibles en función del objetivo de campaña que se eligió. Los objetivos de optimización disponibles para cada objetivo de campaña son los siguientes:

- Conocimiento de la marca:
    - Alcance (recomendado por la plataforma): Optimiza el número de usuarios únicos a quienes se muestran los anuncios. Es diferente a

impresiones, ya que si, por ejemplo, a un mismo usuario se le muestra el anuncio cinco veces durante la campaña, con el objetivo de impresiones se estaría pagando por las cinco, mientras que con este objetivo se pagaría por haber mostrado el anuncio a ese usuario al menos una vez. La desventaja de este objetivo es que no admite puja manual.

    - Impresiones: muestra tantas impresiones como sea posible.

- Visitas al sitio web:
    - Clics en la página destino (recomendado por la plataforma): muestra los anuncios a las personas que tienen más probabilidades de hacer clic en el enlace a la web del anunciante.
    - Impresiones.

- Interacción:
    - Clics de interacción (recomendado por la plataforma): muestra los anuncios a las personas que tienen más probabilidades no solo de hacer clic, sino de efectuar otras acciones como seguir al anunciante en LinkedIn y otras acciones sociales.
    - Impresiones.

- Visualizaciones de vídeo:
    - Visualizaciones de vídeo (recomendado por la plataforma): muestra los anuncios a las personas que tienen más probabilidades de ver el vídeo durante más de dos segundos seguidos. Se considera visualización dos o más segundos seguidos de reproducción mientras en la pantalla aparece al menos un 50% del vídeo. [9]
    - Impresiones.

- Generación de contactos:
    - Contactos (recomendado por la plataforma): muestra los anuncios a las personas que tienen más probabilidades de enviar un formulario de generación de contactos.
    - Clics.

    - Impresiones.

- Conversiones en el sitio web:
    - Conversiones en el sitio web (recomendado por la plataforma): muestra los anuncios a las personas que tienen más probabilidades de llevar a cabo acciones de valor en la web.
    - Clics en la página destino.
    - Impresiones.
- Solicitudes de empleo:
    - Clics en la página destino (recomendado por la plataforma)
    - Impresiones.

Una vez seleccionado el objetivo, se debe especificar la estrategia de puja. Existen tres modalidades, que pueden estar disponibles o no en función del objetivo de optimización seleccionado:

- Máxima difusión: Es la opción completamente automática, en la que la plataforma nos promete obtener los mejores resultados posibles con el presupuesto total que se haya fijado. Para ello utiliza técnicas de *machine learning*. Está disponible para todos los objetivos de campaña, y aunque la campaña de máxima difusión se pueda optimizar para conseguir clics, impresiones, llegadas a usuarios, etc., la campaña siempre será cobrada por impresiones (CPM), y sin poder fijar ninguna puja u objetivo de coste. [10] [11]

- Coste objetivo: Es una opción semiautomática, en la que se fija un coste objetivo (por ejemplo, coste por cada mil impresiones si se ha elegido impresiones como objetivo de optimización) y la puja se ajusta para obtener los mejores resultados dentro del rango de este coste. Permite estabilizar el coste por resultado a la vez que se optimiza el rendimiento. Las pujas de este tipo pueden variar al principio, y se estabilizan tras los primeros cincuenta resultados clave. El coste medio diario nunca será superior a un 30% más que el coste objetivo. Actualmente no está disponible para la generación de clientes potenciales, las conversiones de sitios web y las solicitudes de empleo. [10] [12]

- Puja manual: Permite especificar el importe a pujar por resultado, y el cobro se hará por clics, impresiones u otros criterios en función del objetivo de optimización seleccionado. Otorga pleno control sobre la oferta del anunciante en la subasta, pero si la puja es demasiado baja se obtendrán menos resultados y por tanto no se llegará al público objetivo en la misma medida. La plataforma muestra el rango en el que se sitúan las pujas de los competidores por un público similar para orientar al anunciante en su estrategia. Las pujas de estos competidores pueden variar durante la campaña, con lo que se debería actualizar periódicamente la puja manual para que siga siendo competitiva. Este tipo de puja está disponible en todos los objetivos de campaña, formatos de anuncio y objetivos de optimización. [10] [13]

Una vez seleccionados estos parámetros, la plataforma actualiza las previsiones de coste y rendimiento que se describieron en el apartado anterior. Los siguientes pasos serían relativos a la configuración de los anuncios, donde se procedería a establecer el aspecto visual del anuncio, las imágenes, textos y/o vídeos que lo componen; y confirmar el lanzamiento de la campaña. No se describen estos pasos en profundidad por no ser de interés para el presente trabajo.

2.7. CONCLUSIONES

Se ha descrito en detalle el funcionamiento y las opciones de configuración del gestor de campañas publicitarias de LinkedIn. Estas nociones de la funcionalidad de la plataforma se utilizan más adelante en este trabajo para desarrollar una herramienta software que puede obtener las previsiones de una campaña: número de individuos que componen la audiencia objetivo, coste de la campaña durante un período de tiempo, rendimiento esperado en términos de cantidad de resultados (clics, impresiones, etc.) en el mismo período, y el desglose por segmento de la audiencia.

3. ESTADO DEL ARTE

El estudio de las plataformas de publicidad online y la obtención de datos de estas tiene ya algo de recorrido, con trabajos recientes centrados sobre todo en la plataforma de publicidad de Facebook. En este apartado se expondrán los principales artículos que se han consultado durante la elaboración de este trabajo, en especial de cara a la obtención y análisis de datos de estas plataformas.

3.1. TRABAJOS PREVIOS EN LA PLATAFORMA DE PUBLICIDAD DE FACEBOOK

La plataforma de publicidad de Facebook se ha utilizado para obtener datos de forma masiva y automática desde hace algún tiempo, y en mayor medida que la de LinkedIn, gracias a que los criterios de segmentación que ofrece son de carácter más general (sobre todo intereses de la audiencia, por ejemplo, interés por ciertas películas, grupos artísticos, partidos políticos, etc.), y también por el hecho de ofrecer una API de programación accesible que facilita la construcción de herramientas de obtención automática de datos.

Los datos de audiencias de Facebook se han utilizado en diversos estudios. Por ejemplo, en [14] se utiliza la API del gestor de anuncios de Facebook para obtener las audiencias que tienen interés por determinados elementos considerados sensibles, relacionados con origen étnico, religión, opiniones políticas o sexualidad entre otros temas, y con los datos obtenidos se analiza la prevalencia de estos intereses sensibles en la asignación de anuncios a los usuarios en todo el mundo. En [15], se utiliza la misma API para construir una extensión de navegador que permite a los usuarios conocer una estimación del valor económico que generan durante una sesión de uso de Facebook, basada en el valor comercial de sus datos personales. Otros trabajos han analizado la discriminación que puede darse en los algoritmos que asignan los anuncios a los usuarios [16], la pobreza en el mundo [17] o incluso los flujos migratorios [18]. Esta cantidad de trabajos demuestra que la API del gestor de anuncios de Facebook ha tenido un mayor recorrido en cuanto a utilización en investigación.

Más relacionado con este trabajo interesa especialmente el estudio realizado por David García *et al.* [19], donde se analiza la brecha de género a través de datos masivos. Utilizando la API de anuncios de Facebook, se obtiene el número de usuarios por edad y

género en cada país, cubriendo 217 países y más de 1.400 millones de usuarios. Específicamente, obtienen mediante la API el número de usuarios y DAUs (usuarios activos diariamente) para cada combinación de género y edad, iterando las edades en intervalos de un año desde los 13 a los 64 años y con un último grupo que incluye los usuarios de 65 años o más. Los datos obtenidos están disponibles en un repositorio de Github [20].

Con esos datos, se calcula una medida de división de género. En concreto, el cálculo se realiza tomando, para cada género y edad, un valor estable del número de usuarios activos diarios (obtenido como la mediana del valor diario durante todo el mes de julio de 2015); y tomando como ratio de actividad para cada grupo el cociente entre este valor y la población total de ese grupo en el país correspondiente (con datos del US Census Bureau). El logaritmo del cociente de esta ratio para hombres y mujeres de todas las edades en un mismo país es el Facebook Gender Divide (FGD). A partir de ese valor se obtiene un mapa de calor en el que se puede visualizar qué países cuentan con una diferencia más acusada entre el número de hombres y mujeres que usan Facebook.

Los resultados revelan que las zonas con mayor predominancia de usuarios hombres son la mayoría de los países de África, oriente medio y el subcontinente indio. Además, se elabora un modelo de regresión lineal del FGD en función de cuatro índices de desigualdad de género del World Economic Forum (oportunidades económicas, educación, salud y participación política), llegando a la conclusión de que el factor con el que más se relaciona el FGD es la igualdad en educación (los países con mayor igualdad de género en educación tienen un menor FGD). En cuanto a parámetros no basados en género, se observa que la penetración de Internet está inversamente relacionada con el FGD, y correlacionado con el producto interior bruto. Por tanto, la metodología de este estudio es útil para el análisis de problemas relacionados con la brecha de género, y sus resultados son sólidos y demuestran la validez de este tipo de datos para el objetivo propuesto.

Otro artículo en el que se estudian las similitudes culturales entre países con datos de Facebook [21] tiene especial relevancia para este trabajo, porque se tomará la metodología utilizada para medir las similitudes entre países en el caso de uso expuesto en el capítulo 5, aplicándola a la medida de similitud de la distribución de brecha de género por áreas laborales entre países. La metodología en cuestión utiliza una implementación de agrupamiento jerárquico disponible en un paquete del lenguaje

Python con la distancia coseno entre vectores y usa como criterio de unión el método de Ward de varianza mínima.

3.2. OTROS TRABAJOS SOBRE DATOS OBTENIDOS DE LINKEDIN

También existen algunos trabajos previos de estudio de datos obtenidos de la plataforma de publicidad de LinkedIn, aunque no se ha encontrado ninguno que detalle si se utiliza para ello alguna herramienta de obtención automática, y por tanto la implementación técnica, el diseño o el enfoque de la herramienta utilizada para automatizar el proceso son desconocidos y no es posible efectuar una comparación de la solución con el enfoque propuesto en este trabajo.

En realidad, estos trabajos se centran en la descripción de los datos que se utilizan y el análisis que se hace sobre ellos, y únicamente se menciona que los datos son obtenidos de LinkedIn.

En el primero de ellos, Haranko et al [22] proponen un análisis de la brecha de género laboral en 20 áreas metropolitanas de Estados Unidos, con datos de estas localizaciones segmentados por edad, género, sector empresarial y aptitudes. En concreto utilizan las 17 categorías empresariales que ofrece la plataforma de LinkedIn, 25 aptitudes (la mayoría relacionadas con el desarrollo de software y algunas con el mundo de la empresa y el marketing) y los 4 grupos de edad que proporciona la plataforma para definir todas las combinaciones posibles junto a las localizaciones y el género. Mediante una medida de la brecha de género definida como el cociente entre el número de hombres y la suma de hombres y mujeres, estiman el nivel de igualdad existente en cada aptitud, industria, grupo de edad y localización; y realizan una comparación y validación de los datos a través de datos del *Bureau of Labor Statistics (BLS)* del año 2014.

Sus conclusiones son que la brecha de género varía más entre diferentes industrias, algo menos entre aptitudes, y se mantiene bastante estable entre localizaciones, y que tiende con la edad hacia una mayor representación de los hombres. Además, se destaca que el artículo sirve como prueba de concepto del valor que tiene la plataforma de anuncios de LinkedIn para la realización de estudios sobre fenómenos sociales, validando la consistencia de los datos a través de la comparación mencionada con los del BLS, destacando la facilidad de obtención ciertos datos, como los de ciertas aptitudes, que son difíciles de encontrar con otros métodos, y subrayando la mayor granularidad y rapidez

de obtención de los datos en comparación con las fuentes tradicionales (encuestas de empleo).

El segundo artículo consultado sobre el tema es el de Verkroost *et al.* [23], en el que participaron tres de los autores del artículo anterior y donde se analizan también las brechas de género usando datos de LinkedIn; sin embargo, en este caso se centra en el sector de tecnologías de la información y se hace un análisis a nivel de país, comparando algunos de los resultados con los que se dan al analizar los datos obtenidos de la plataforma de Facebook. En este sentido, este trabajo se diferencia del artículo mencionado por dos motivos: en primer lugar, en este estudio no se extrae el valor económico de los datos, que es uno de los objetivos que se persigue en este trabajo dentro de los requisitos de la herramienta que se desarrollará en el proyecto PIMCity, y al que se dará cumplimiento en la solución propuesta en el capítulo 4; por otro lado, en el caso de uso que se presenta en el capítulo 5 se utilizarán tanto las áreas laborales de la plataforma como una clasificación de áreas laborales externa a LinkedIn, que cubre un extenso rango de áreas y que tiene en cuenta múltiples profesiones para cada una, siendo una clasificación exhaustiva, en lugar de centrarse en un ámbito concreto como tecnologías de la información. Además, se realiza un ejercicio de agrupamiento para clasificar distintos países en función de su distribución de la brecha de género en las diferentes áreas.

Por último, se ha valorado la posibilidad de utilizar datos extraídos de LinkedIn disponibles bajo acceso público en el catálogo de datos del Banco Mundial [24]. Este conjunto de datos es parte de una colaboración entre LinkedIn y el Banco Mundial para apoyar a investigadores y gobiernos en la comprensión de un mercado laboral en constante y rápida evolución con datos detallados y dinámicos en el tiempo. Analizan aptitudes, profesiones, industrias y migración de trabajadores, con datos de más de 140 países y muestras anuales entre 2015 y 2019. En concreto, ofrecen sendos archivos con datos útiles sobre las siguientes cuestiones: necesidades de aptitudes en diferentes industrias, penetración de aptitudes, variaciones de nivel de empleo por industrias, y migración de talento. Es un conjunto de datos muy completo, sin embargo, no se puede utilizar para los propósitos de este trabajo por varios motivos. En primer lugar, no incluye datos sobre los valores económicos y previsiones de rendimiento publicitario de las audiencias, que es algo necesario para dar cumplimiento al uno de los objetivos del trabajo en el marco del proyecto PIMCity. En segundo lugar, los datos no se clasifican por

género, de modo que no sirven para realizar el análisis que se propone en el capítulo 5. Por último, los últimos datos disponibles son del año 2019, demasiado desactualizados teniendo en cuenta el dinamismo del mercado laboral y el rápido crecimiento en el número de usuarios de LinkedIn.

Vistos los trabajos previos relacionados, en el siguiente capítulo se expone el proceso de diseño e implementación de la herramienta objetivo de este trabajo y los resultados obtenidos, y posteriormente se efectuará una obtención y análisis de datos en la línea de lo visto en estos artículos a modo de prueba del potencial de la herramienta para la realización de investigaciones de este estilo.

4. LINKEDIN DATA CRAWLER

4.1. REQUISITOS TÉCNICOS Y DISEÑO

El primer paso es abordar el diseño a alto nivel de la herramienta. Para ello hay que hacer una recopilación de funcionalidades y establecer claramente los requisitos de estas.

4.1.1. FORMATO DE LOS DATOS DE ENTRADA

Al no haberse definido aún el formato de entrada que tendrá la herramienta conjunta para diferentes plataformas del proyecto PIMCity, el formato de entrada será un archivo de texto que contenga la combinación lógica de elementos. A los elementos en la misma línea se les aplica la función lógica OR, mientras que a las distintas líneas se les aplica la función lógica AND. Por ejemplo, para la combinación (Leganés OR Colmenarejo OR Getafe) AND (Universidad Carlos III de Madrid OR Redes) se tendría que conformar el siguiente archivo de entrada:

“Leganés” “Colmenarejo” “Getafe”
“Universidad Carlos III de Madrid” “Redes”

Se decide que los elementos en la misma línea estén separados por tabulaciones (archivo en formato TSV), ya que otros separadores como espacio, coma o punto y coma pueden aparecer dentro de los elementos de la combinación.

4.1.2. ARGUMENTOS DE ENTRADA DEL MÓDULO

El módulo debe recibir como argumentos de entrada los siguientes elementos:

- En primer lugar, un fichero de texto con la información de inicio de sesión (correo electrónico en la primera línea y contraseña en la segunda, correspondientes a una cuenta de LinkedIn con una cuenta publicitaria creada).

- En segundo lugar, el fichero de texto que contenga la combinación lógica en el formato descrito en el apartado anterior.

De modo que la forma de ejecutar el módulo será:

```
>node main.js params/login.txt input/input.txt
```

Siendo *main.js* el código principal a ejecutar. En el directorio *params* se almacena tanto el fichero de inicio de sesión como otros ficheros de configuración que se puedan necesitar. En el directorio *input* se almacenan los ficheros de entrada, y en *output* se almacenarán los ficheros de salida resultantes de la ejecución.

4.1.3. TAREAS QUE REALIZAR Y DIVISIÓN EN FUNCIONES

En primer lugar, se quiere que el programa simule el inicio de sesión y navegación de un usuario real, introduciendo los datos de inicio de sesión (usuario y contraseña) en los campos correspondientes, haciendo clic en el botón de iniciar sesión y navegando por la página web hasta obtener, mediante inspección de las cabeceras en las peticiones HTTP que efectúa el navegador durante este proceso, todos los datos necesarios (*cookies*, *tokens* y otras cabeceras) para poder hacer las peticiones específicas requeridas.

Una vez se tienen todas las cabeceras que hacen falta para que el servidor acepte una petición, se necesita una función que formule la petición para obtener la URN de un elemento concreto. Estos elementos pueden ser de diferentes categorías, así que esta función debe permitir elegir de alguna forma el tipo o tipos entre los que se quiere buscar el elemento. Si hay varios tipos posibles para ese elemento deberá poder indicarse una prioridad. Por ejemplo, “Madrid” se podría buscar preferentemente en localizaciones, y en caso de no encontrarse, se podría buscar en Universidades. Normalmente se querrá buscar el elemento en un solo tipo, pero la función debe proporcionar esa flexibilidad de cara a los futuros usos que pueda tener. La lista de tipos se extrae de las diferentes opciones que da la plataforma de publicidad a la hora de definir la audiencia objetivo, seleccionando para la herramienta aquellos que se consideran más interesantes de cara a los objetivos del trabajo, ya que aún no se han especificado las opciones que deben estar disponibles de forma obligatoria en la herramienta global del proyecto PIMCity. Así, las distintas categorías disponibles en la función de búsqueda serán las siguientes:

- Localizaciones
- Aptitudes
- Titulaciones
- Instituciones educativas
- Cargos
- Grupos
- Intereses
- Empresas
- Disciplinas académicas
- Género
- Funciones laborales
- Sectores de la empresa

Alguna de estas categorías no ofrece una interfaz de búsqueda, sino que simplemente muestra diferentes subcategorías predefinidas, que contienen dentro distintas opciones para seleccionar, como por ejemplo *Sexo* dentro de la categoría *Características demográficas*. En estos casos la herramienta no solicita al servidor la URN del elemento, de modo que para simular un comportamiento real se implementará un mecanismo que detecte estas categorías especiales y evite hacer la petición correspondiente, teniendo ya almacenada de antemano la información de estas opciones (género, funciones laborales y sectores de la empresa) para utilizarla en las peticiones en las que se necesite.

Esta función debe además guardar en un archivo un registro de los elementos que se hayan solicitado, que incluya el nombre que LinkedIn da a ese elemento, su URN, su tipo, una lista con los nombres que al buscarse devuelven este elemento (por ejemplo, el elemento Madrid, Comunidad de Madrid, España será devuelto al buscar Madrid, de modo que Madrid debe figurar en esta lista de nombres) y cualquier otra información que sea necesaria para las funciones que tengan que leer este registro. Por ello, se utilizará para implementar el mecanismo mencionado en el párrafo anterior: este registro se inicializará con la información contenida en otro registro fijo, en el que se escribirá la información de las categorías que no efectúan búsqueda de URN, y posteriormente se irá actualizando con la nueva información que se vaya obteniendo durante las distintas ejecuciones de la función.

Tras obtener las URN de los elementos que se quieren buscar, se necesitan funciones que construyan las peticiones de información sobre la combinación lógica de entrada con el formato adecuado. El formato concreto de las peticiones es desconocido a priori, se infiere inspeccionando las peticiones que realiza un navegador cuando se utiliza la plataforma

de anuncios de LinkedIn manualmente para introducir una combinación lógica de elementos.

En este punto se requieren tres tipos de funciones:

- La que nos proporciona la audiencia total. Debe devolver un número entero con la audiencia total.
- La que nos proporciona estimaciones sobre el precio de la campaña, cantidad de impresiones del anuncio, porcentaje de clics respecto a impresiones y clics totales, todo ello en el periodo de 30 días. Se debe poder elegir el tipo de cobro (CPC o CPM).
- La que nos proporciona una segmentación adicional de la audiencia con diferentes grupos dentro de la categoría elegida y el porcentaje de la audiencia que pertenece a cada uno de ellos. Las categorías que deben poder elegirse son funciones laborales, nivel de responsabilidad, años de experiencia, tamaño de la empresa, sectores e intereses.

Todas estas funciones reciben la combinación lógica, construyen la estructura que determina la combinación, y escriben en sendos archivos la respuesta recibida del servidor.

4.2. TECNOLOGÍAS UTILIZADAS

Para lograr construir la herramienta objetivo de este trabajo se podría haber recurrido al uso del API de programación que ofrece Microsoft para la plataforma de anuncios de LinkedIn [25]. Sin embargo, esta API tiene un uso muy limitado y existe mucho control sobre las aplicaciones que tienen acceso a ella. Para poder utilizarla, es obligatorio solicitar el acceso, registrando la aplicación que va a hacer uso de ella en la página para desarrolladores de LinkedIn y dando una descripción detallada de qué información utilizaría y a qué partes del servicio de marketing se necesita tener acceso; para después pasar un proceso de evaluación en el que se comprueba si la aplicación cumple los requisitos para hacer uso de la API. Al ser un proceso tan complejo y obtener una funcionalidad limitada, se opta por buscar alternativas para la arquitectura de la herramienta.

De este modo, se diseña otro modelo en el que se plantean técnicas de *web scrapping* con las que se simulará la navegación por el portal de gestión de campañas y, utilizando información obtenida de esta navegación, se lanzarán peticiones personalizadas al servidor de manera automatizada y se guardará la información de las respuestas.

A continuación, se detallan las tecnologías utilizadas, el propósito de cada una de ellas y por qué se han elegido frente a otras alternativas:

4.2.1. LENGUAJE DE PROGRAMACIÓN JAVASCRIPT

JavaScript es un lenguaje diseñado para su ejecución en el navegador, es decir, para aplicaciones web. Sin embargo, en los últimos años se ha empezado a utilizar en todo tipo de programas que se ejecutan en la máquina local gracias al entorno de ejecución Node.js. Su capacidad de manejar de forma directa archivos de tipo JSON, sin ser necesario programar el código encargado del mapeo del contenido a los atributos de una clase como en otros lenguajes orientados a objetos (por ejemplo, Java), es el motivo principal de utilizar este lenguaje, porque tanto las respuestas del servidor como algunos de los formularios de petición tendrán formato JSON. En JavaScript, el contenido de esos archivos es asignable directamente a una variable del tipo nativo *Object*. Como referencia del lenguaje, además de la documentación oficial [26], se ha utilizado el libro *Up & Going* de Kyle Simpson [27], de la serie *You don't know JS.*

Existían otras alternativas válidas como Python; sin embargo, se ha elegido JavaScript por ser el lenguaje en el que se implementará el resto de los módulos que compongan la herramienta diseñada en el proyecto PIMCity. Además, la ventaja de este entorno es su orientación a eventos asíncronos, que permite crear aplicaciones de red escalables con múltiples conexiones concurrentes de una forma fácil.

4.2.2. PUPPETEER

Para implementar el diseño del módulo se ha utilizado un navegador Chromium integrado en el mismo a través de Puppeteer, que es una librería de Node.js que proporciona una API de alto nivel para controlar el navegador en cuestión, haciendo uso del protocolo DevTools [28]. El navegador se puede usar tanto en modo *headless* (sin interfaz gráfica) como con una ventana de interfaz gráfica en la que se mostrará el contenido de navegación

y todas las opciones de ventana del navegador (como si se tratase de un navegador corriente, pero controlado por el módulo de Node.js que lo lanza).

Este módulo permite controlar los clics, las entradas por teclado y todo tipo de acciones que normalmente se llevan a cabo de forma manual en un navegador, de modo que se pueden construir con él herramientas para pruebas automatizadas de interfaces, obtención automática de información de diversas páginas en internet, obtención de páginas en formato PDF, etc. En este caso, se utilizará para automatizar el inicio de sesión y la navegación mediante clics hasta la interfaz de selección de objetivos del gestor de campañas de LinkedIn. Durante el proceso, haciendo también uso de las herramientas que proporciona Puppeteer, se interceptará el tráfico generado para obtener los parámetros necesarios para construir peticiones idénticas a las generadas por una sesión de navegación manual, estos son el *token* de acceso (en adelante *csrf-token),* el identificador numérico de la cuenta publicitaria (en adelante *account id*) y las cookies de navegación.

4.2.3. CATAPULT WEBPAGE REPLAY

Catapult Webpage Replay [29] es una herramienta escrita en lenguaje Golang que sirve para grabar sesiones de navegación, de modo que se puedan después reproducir y simular que la interacción es con el servidor real cuando en realidad es una grabación previa y no se están lanzando peticiones fuera de la máquina. Se utilizará para grabar sesiones de navegación manual y después reproducirlas cuando se prueban múltiples ejecuciones durante la implementación de la herramienta, para que las respuestas salgan de la grabación y no del servidor real mientras se solucionan errores y se perfecciona la implementación, y así no saturarlo y evitar que se bloquee la cuenta que se utiliza para acceder a la plataforma.

Estas son las principales tecnologías utilizadas para implementar la herramienta. Durante el desarrollo del trabajo se incorporan otras tecnologías, en particular durante la adaptación de la herramienta para que sea escalable a un uso masivo. En los apartados correspondientes se detallan las tecnologías añadidas y el motivo de su incorporación.

4.3. IMPLEMENTACIÓN

En esta sección se explicará la implementación del módulo. Se empezará por mostrar un diagrama del funcionamiento general con los diferentes elementos que componen la solución, y después se explicará detalladamente cada uno de ellos.

El módulo se compone de funciones que implementan cada una de las funcionalidades definidas en la fase de diseño. Estas funciones se llaman desde el código principal y se sigue el flujo de ejecución mostrado en el siguiente diagrama.

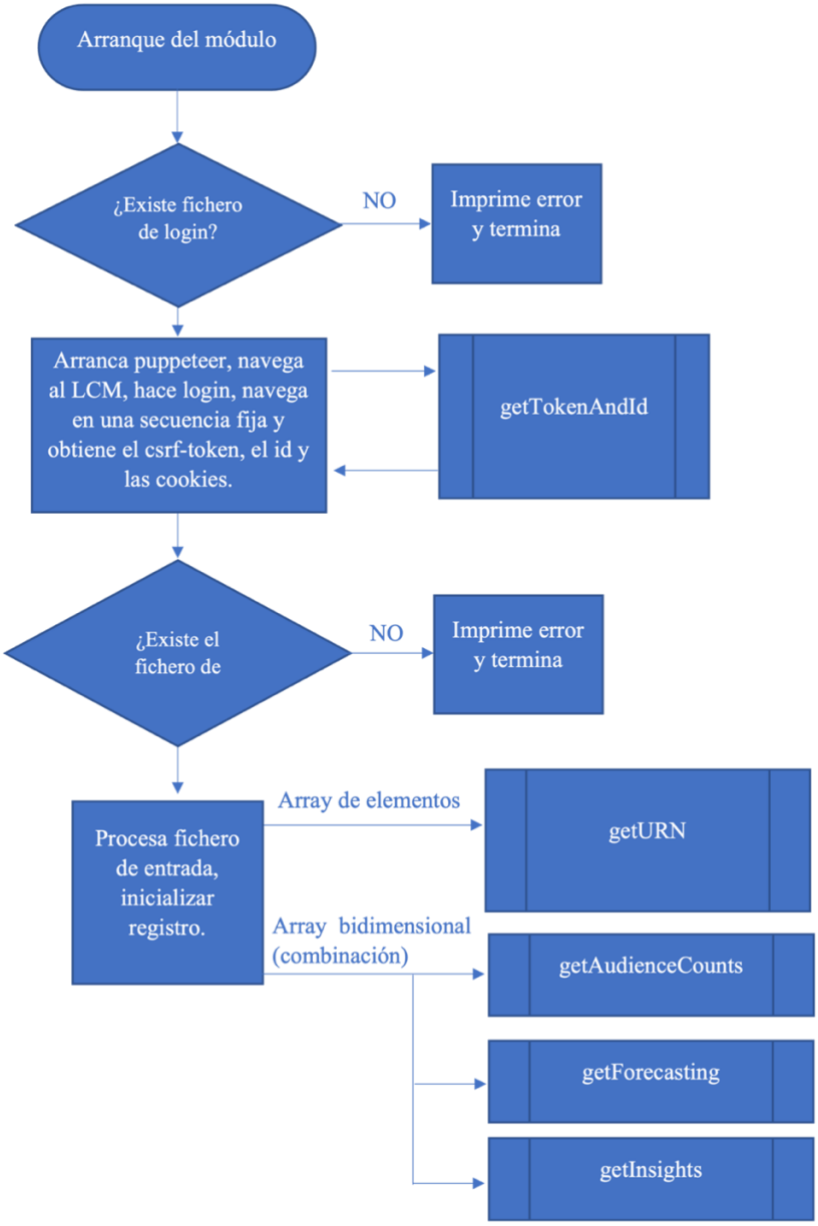

Fig. 4.1. Esquema de las funciones que componen el módulo. Elaboración propia.

4.3.1. OBTENCIÓN DEL CSRF TOKEN Y EL ID DE LA CUENTA

Esta función se encargará de obtener toda la información necesaria para poder conformar peticiones válidas que sean aceptadas y respondidas por el servidor. En particular, se encarga de capturar y almacenar dos elementos imprescindibles en las peticiones: el *id* de la cuenta publicitaria y el *csrf-token*.

CSRF significa *Cross-Site Request Forgery*, por sus siglas en inglés. La inclusión del mecanismo *csrf-token* protege a la aplicación de que se puedan hacer peticiones al servidor desde cualquier sitio de Internet a través del dominio del usuario. Por ejemplo, una página maliciosa podría tener un botón que, al pulsarlo, emita una petición al servicio (en este caso LinkedIn Campaign Manager) que realice cualquier acción sin conocimiento del usuario y haciéndose pasar por él, ya que el origen de la petición es su propia dirección IP.

El *csrf-token* es un número aleatorio generado por el servidor de la aplicación que evita esta práctica, ya que es entregado al cliente tras su autenticación y de forma que solo la aplicación en cuestión lo pueda leer. La aplicación incluye ese número en las peticiones que envía al servidor, y si el número no figura en las cabeceras de la petición o es incorrecto, no se realizará ninguna acción y se devolverá al cliente un error. Por ello, es imprescindible implementar un mecanismo por el que, una vez que en el navegador se ha iniciado sesión, se lean las cabeceras de las peticiones generadas por el cliente al navegar por la interfaz para obtener este número, y también el id de la cuenta publicitaria.

Esta función se implementa interceptando las peticiones que genera el navegador en determinado momento de la navegación por la interfaz del gestor de campañas de LinkedIn. Para ello se utiliza un evento de la página de Puppeteer que se lanza cada vez que la página efectúa una petición, para establecer una interrupción y mirar sus cabeceras.

Al continuar la navegación haciendo clic en diferentes elementos de la interfaz de la aplicación que nos conducen hacia la interfaz del gestor de campañas descrita en el capítulo 2, esta interrupción se lanza para cada petición realizada, lee las cabeceras comprobando la existencia del campo *csrf-token* y, cuando se encuentra, se captura y almacena en una variable global accesible para el resto de las funciones. Por otro lado, se comprueba la URL de cada petición, para detectar la que contenga el id de la cuenta (se busca un patrón preestablecido que se ha definido tras analizar las URLs de forma manual,

viendo la forma en que se incluye este elemento en las direcciones URL a partir de cierto momento de la navegación por la interfaz).

La secuencia de acciones de esta función es la siguiente: al llamarla, se activa la intercepción de peticiones mediante la función de Puppeteer *setRequestInterception.* Después se establece el código que se debe ejecutar cuando se emite el evento *request* (emitido cuando la página hace una petición): se extrae la URL de la petición y las cabeceras HTTP, y se comprueba si contienen el *id* y el *token* respectivamente.

4.3.2. FUNCIÓN PARA LA OBTENCIÓN DE LAS URN

Para implementar esta función se analizan las peticiones que hace el navegador al buscar ítems en las diferentes categorías. Por ejemplo, para la búsqueda *Apple* en la categoría *Empresas*, se genera una petición HTTP de tipo GET a la siguiente URL:

https://www.linkedin.com/campaign-manager-api/campaignManagerAdTargetingEntities?query=Apple&accountId=503797856&facets=List(urn:li:adTargetingFacet:employers)&q=queryAndMultiFacetTypeahead

Se observa que en la URL la categoría se incluye en el campo *facets*, y en este caso es *employers*. Por tanto, se necesita una correspondencia entre las categorías definidas en la fase de diseño y su nombre interno:

- Localizaciones: locations
- Aptitudes: skills
- Titulaciones: degrees
- Instituciones educativas: schools
- Cargos: titles
- Grupos: groups
- Intereses: interests
- Empresas: employers
- Disciplinas Académicas: fieldsOfStudy

De modo que la lista de categorías que debe recibir esta función deberá estar compuesta por uno o más de estos nombres internos.

Además, la lista de nombres debe ser primero filtrada para dejar únicamente los elementos que no hayan sido buscados previamente, esto es, que no estén en el fichero de registro de elementos, del que se hablará más adelante. Para esto se lee el fichero de registro llamado json_urns.json y se recorre la lista de nombres recibida con un bucle *for*, para comprobar si existe en el registro una entrada con el mismo nombre o que contenga ese nombre en su lista de otros nombres. Los elementos que no estén en el registro y por tanto requieran petición al servidor se almacenan en un nuevo array llamado *toSearch*.

Una vez extraídos los elementos a buscar en el servidor se utilizan dos bucles *for* anidados, el primero recorre el array *toSearch* y el segundo la lista de categorías o tipos recibida como argumento de la función. A cada elemento a buscar se le aplica una codificación de tipo *percent encoding* [30] para introducirlo en el campo *query* de la URL de la petición al servidor. También se introduce la categoría en el lugar correspondiente según el ejemplo observado, y se efectúa la petición a través del navegador mediante la función *page.goto(URL)* de Puppeteer, que devuelve la respuesta del servidor.

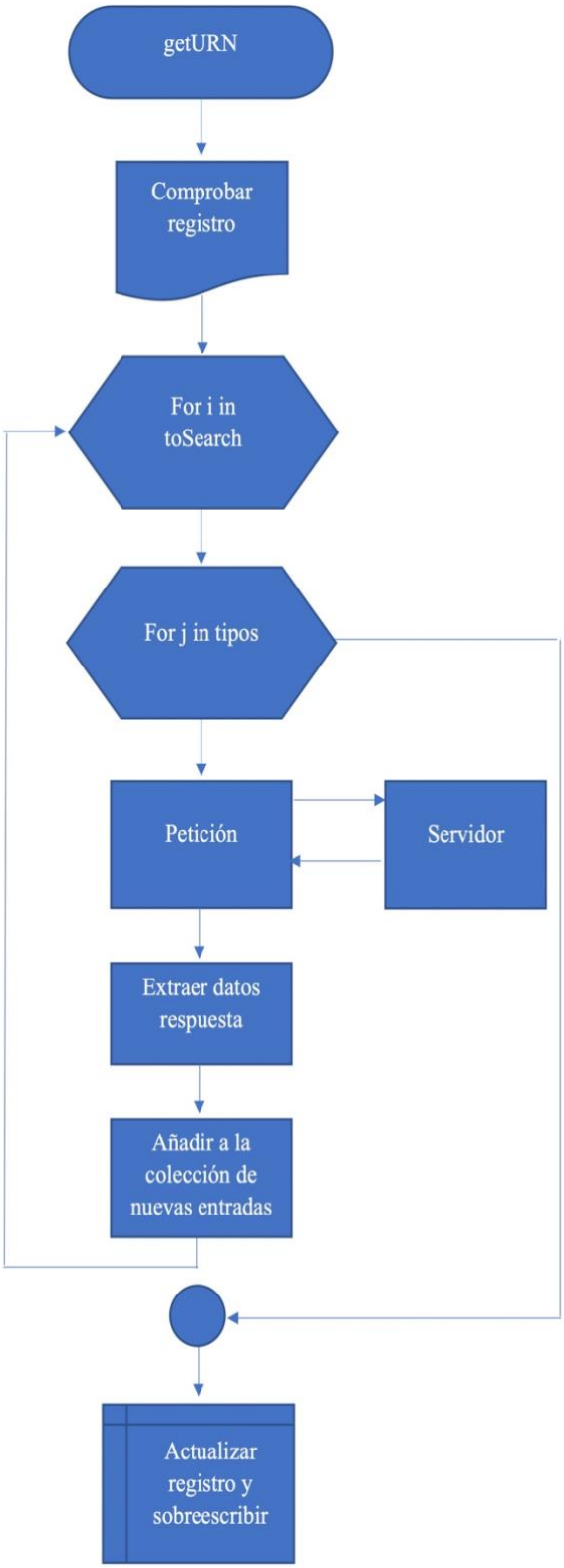

Fig. 4.2. Diagrama de flujo de la función getURN. Elaboración propia.

Con la respuesta en formato JSON, se extraen los datos que deben figurar en el archivo de registro para cada elemento: nombre, URN, tipo y *ancestors*, esto último se explica en el apartado dedicado al archivo de almacenamiento local. Con esta información se añade una entrada a una variable JSON con el mismo formato que las entradas del registro.

Una vez tenemos esta variable que contiene todas las entradas nuevas, para cada una de ellas se comprueba que no haya ninguna variable en el registro con la misma URN, si es así se añade la entrada completa al registro, si ya existe un elemento con la misma URN, se añade el nombre por el que se llegó a esa respuesta su lista de otros nombres.

Para acabar se sobrescribe el registro actualizado en el archivo json_urns.json.

4.3.3. ARCHIVO DE ALMACENAMIENTO LOCAL DE LAS URNS

Todos los elementos encontrados con la función del apartado anterior se deben guardar en un registro local, que como se ha dicho anteriormente se llamará json_urns.json. Como indica su extensión, tendrá formato JSON [31], donde el nombre de cada elemento será el nombre que LinkedIn da a este elemento, y su valor será a su vez otro objeto JSON, que incluirá su tipo, su URN, un array con otros nombres para este elemento y un array llamado *ancestorList* en el que se guardarán una serie de URNs que acompañan a algunos de los elementos, por ejemplo, a las localizaciones, en la respuesta a la petición de su URN. Estos elementos serán necesarios al componer la estructura de las siguientes peticiones como se verá en los próximos puntos de este apartado.

La función getURN consulta este archivo para comprobar si existen las entradas antes de realizar peticiones al servidor (bien por coincidencia de la cadena recibida con el nombre del elemento o con alguna de las cadenas en el array *otros nombres*), y en el caso de que no existan, tras hacer las peticiones pertinentes comprueba si los elementos recibidos son iguales a alguno ya existente en el registro, en cuyo caso actualiza la lista *otros nombres* con la cadena que recibió al principio, o bien en el caso de que sean elementos nuevos actualiza el registro para incluirlos.

Las siguientes funciones leerán los datos de este registro para efectuar sus propias peticiones formadas por una estructura que plasma la combinación lógica de elementos a su entrada. Estas peticiones requieren todos estos datos para poder formar sus estructuras de forma que el servidor acepte las peticiones y devuelva respuestas.

A continuación, se expone una muestra del registro en cuestión:

```
{
  "Private Investigator":{
    "otherNames":["Private Detectiv"],
    "urn":6016,
    "type":"title",
    "ancestorList":[]
  }
  "Male":{
    "otherNames":[]
    "urn":"MALE",
    "type":"gender",
    "ancestorList":[]
  }
  "Tahiland":{
    "otherNames":["Tailandia"],
    "urn":"105146118",
    "type":"geo",
    "ancestorList":["102393603"]
  }
}
```

4.3.4. OBTENCIÓN DE LA AUDIENCIA TOTAL DE UNA COMBINACIÓN

Esta función solicita al servidor la audiencia total correspondiente a la combinación lógica que recibe como parámetro de entrada. También recibe la página de Puppeteer y el *id* de la cuenta publicitaria.

Lo primero que se necesita para implementar esta función es analizar una petición de este tipo hecha por el navegador durante una navegación normal en la que se selecciona cierta combinación lógica. Por ejemplo, para la combinación *España AND (Node.js OR JavaScript) AND Programador Web*, en el gestor de campañas de LinkedIn se escoge como localización *España*, después se solicita que la audiencia tenga una de las dos aptitudes (*Node.js* o *JavaScript*), y después mediante la opción *Acotar más el público* se añade la característica *Programador Web* buscándola dentro de la categoría *Cargos*. Al seleccionar el último elemento se produce la petición que se buscaba: es de tipo POST y modifica el fichero *campaignManagerAudienceCounts* en el servidor. A continuación, se analiza el formato del formulario de la petición:

```
q=targetingCriteria&cmTargetingCriteria=(include:(and:List((or:List((facet:(urn:urn%3Ali
%3AadTargetingFacet%3AinterfaceLocales,name:Idiomas%20de%20la%20interfaz),segments:List(
(urn:urn%3Ali%3Alocale%3Aen_US,name:Ingl%C3%A9s,facetUrn:urn%3Ali%3AadTargetingFacet%3Ai
nterfaceLocales)))))),(or:List((facet:(urn:urn%3Ali%3AadTargetingFacet%3Askills,name:Apti
tudes),segments:List((urn:urn%3Ali%3Askill%3A218,name:JavaScript,facetUrn:urn%3Ali%3AadT
argetingFacet%3Askills),(urn:urn%3Ali%3Askill%3A18276,name:Node.js,facetUrn:urn%3Ali%3Aa
dTargetingFacet%3Askills)))))),(or:List((facet:(urn:urn%3Ali%3AadTargetingFacet%3Alocatio
ns,name:Ubicaciones),segments:List((urn:urn%3Ali%3Ageo%3A105646813,name:Espa%C3%B1a,face
tUrn:urn%3Ali%3AadTargetingFacet%3Alocations,ancestorUrns:List(urn%3Ali%3Ageo%3A10050691
4))))))),(or:List((facet:(urn:urn%3Ali%3AadTargetingFacet%3Atitles,name:Cargos),segments:
List((urn:urn%3Ali%3Atitle%3A1359,name:Programador%20web,facetUrn:urn%3Ali%3AadTargeting
Facet%3Atitles)))))),exclude:(or:List()))&withValidation=true
```

Se puede ver que la combinación de elementos se plasma en el campo *cmTargetingCriteria*, con un campo inicial que indica el idioma de la interfaz, y después, los elementos que se combinan con OR entre sí van dentro del mismo campo *or:List*, mientras que una combinación AND se indica con dos campos *or:List* uno a continuación del otro.

Otro punto importante es que la categoría de cada elemento tiene diferentes identificadores dentro de la petición. Por ejemplo, las localizaciones: en una parte del formulario se identifican con la palabra *locations*, después con el nombre en español Ubicaciones, y por último justo antes de indicar la URN, se refiere a esta categoría como *geo*. Lo mismo pasa con el resto de las categorías, en el ejemplo anterior se pueden ver los casos de *Aptitudes* y *Cargos*. Por este motivo se establecen varias listas con cada una de las nomenclaturas, y un diccionario con los nombres en español correspondientes.

La lista *array_in* recoge la nomenclatura que se utiliza primero en la estructura correspondiente a un elemento, *array_out* recoge la nomenclatura final, y el diccionario *dict_in* establece la equivalencia entre cada categoría de *array_in* y su nombre en español para incluirlo en la petición.

Tras declarar estas constantes, se carga la información del registro de URNs (json_urns.json) y se almacena en una variable. Seguidamente se comprueba si la combinación de elementos recibida como entrada contiene al menos una ubicación, esto es una condición necesaria que impone la plataforma de anuncios de LinkedIn para mostrar la información. Para ello, se comprueba para cada uno de los elementos si la entrada del mismo nombre en el registro (o bien la que contenga ese nombre en su lista *otros nombres*) tiene como tipo *geo* (ubicación). Si ninguna entrada es una ubicación, la función muestra por consola un error y finaliza.

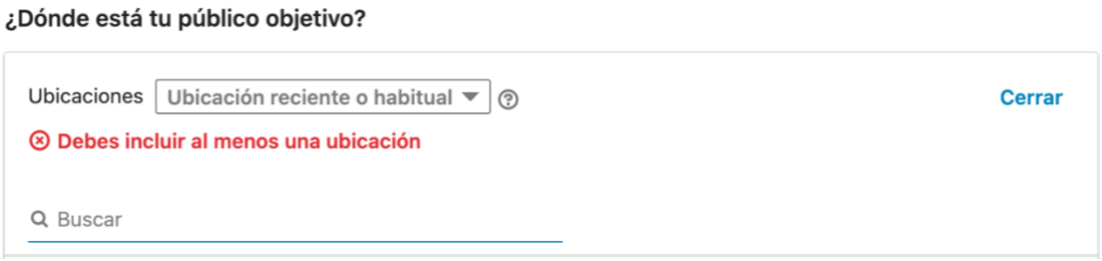

Fig. 4.3. LinkedIn exige al menos una ubicación entre los criterios de búsqueda. [3]

Una vez comprobada la existencia de una ubicación en la combinación, se ejecuta un algoritmo que traduce esa combinación a un formulario como el visto anteriormente que represente esas relaciones lógicas (AND y OR). Este algoritmo se extrae a una función auxiliar llamada getTargetingCriteria porque se utiliza tanto en esta función como en la de obtención del presupuesto. El formulario se almacena en una variable a la que se va añadiendo nuevo texto hasta completarlo. En primer lugar, se añade la primera parte, que es común a cualquier formulario de este tipo, hasta antes del primer *or:List*. Después, se establecen diferentes variables plantilla para diferentes casos:

- Primer elemento dentro de una combinación AND
- Elemento intermedio de una combinación AND
- Elemento intermedio de una combinación AND con un tipo diferente al elemento anterior.
- Último elemento de una combinación AND
- Otras tantas equivalentes a las anteriores para el caso en el que el elemento contenga *ancestors* (su lista de *ancestors* no esté vacía)

En cada iteración del algoritmo a lo largo de la matriz de elementos se introduce en cada plantilla la información correspondiente a cada elemento, y después, en función de si es el primero de la fila actual, el último o un elemento intermedio, y de si tiene *ancestors* o no, se concatena al formulario la plantilla correspondiente. De esta forma se va completando el formulario con el formato adecuado, y al llegar al final de la iteración de la matriz de elementos combinados se concatenan los paréntesis de cierre y el resto del formulario. No se implementa por el momento la posibilidad de excluir determinadas audiencias dentro de la audiencia especificada, por lo que la *or:List* del campo *exclude* siempre se añade vacía.

Una vez se tiene el formulario completo, se declara una variable objeto (JSON) para listar las cabeceras de la petición y se establecen sus valores. Las *cookies* se obtienen con la función getCookies, el *csrf-token* se obtiene de la variable global en la que la función GetTokenAndId lo debe almacenar previamente y el *id* de la cuenta publicitaria se le proporciona a esta función como parámetro tras haberlo obtenido igualmente de la función GetTokenAndId.

Tras completar las cabeceras se efectúa la petición, se lee el JSON recibido como respuesta y se obtiene la audiencia total, contenida en el campo *count* del primer elemento del array *elements*. Al JSON de la respuesta se le añade un elemento llamado *requested* cuyo valor será la combinación lógica sobre la que se solicitó la audiencia total en formato (X AND Y AND …) OR (…) donde X, Y serían los elementos de la combinación. El JSON modificado se transforma a texto y se escribe en un fichero llamado *counts.json*, y para finalizar la función devuelve el valor de la audiencia total.

4.3.5. OBTENCIÓN DEL PRESUPUESTO

La función para la obtención del presupuesto tiene un funcionamiento similar a la anterior. El campo *cmTargetingCriteria* es el mismo, de modo que el proceso de construcción de este elemento se extrae a la función auxiliar getTargetingCriteria que es llamada por ambas funciones.

Además de la combinación lógica, el formulario de esta petición incluye diversas opciones para la confección del presupuesto: tipo de anuncio (una sola imagen, varias imágenes, video, etc.), tipo de resultado por el que se paga o cobra (CPC, CPM, etc.) y otros. En la implementación de la función que reproduce esta petición se van a fijar estos parámetros a unos valores determinados para que los resultados obtenidos para diversas audiencias se puedan comparar. Además, se necesita poder elegir el tipo de cobro entre CPC o CPM. Para ello, se fijan los parámetros de la siguiente forma:

• adFormats: indica el tipo de anuncio que se solicita, pudiendo ser varios en una lista. Se fija la lista con un único elemento, el correspondiente a anuncio de una sola imagen (STANDARD_SPONSORED_CONTENT).

• runSchedule: indica el período para el que se quiere obtener el presupuesto. Solo se especifica la fecha de comienzo, que será la fecha actual (hora del sistema). Esto equivale a seleccionar que la campaña esté activa de forma continuada.

• cmTargetingCriteria: es la combinación lógica en el mismo formato que en la función anterior (getAudienceCounts). Como el formato del texto es común, la generación de este a partir de la combinación lógica expresada en el array bidimensional que reciben las dos funciones se extrae a la función auxiliar mencionada.

• ObjectiveType: indica el objetivo de la campaña seleccionado. Se fija a visitas al sitio web (WEBSITE_VISIT) porque admite tanto cobro por clics como por impresiones, que es lo que se busca.

• costType: indica el tipo de coste utilizado. Tomará el valor que se especifique al llamar a la función, a elegir entre CPC y CPM.

• unitCost: indica una estimación media del coste por unidad. Esta estimación se obtiene con otra petición al servidor que se detalla más adelante. Tomará el valor recomendado que se obtenga.

• dailyBudget: presupuesto diario para la campaña. Por defecto es de 50 dólares. También se puede fijar el presupuesto total añadiendo el campo totalBudget con el mismo formato. En este caso se fija a 50 dólares.

• OptimizationTargetType: Tipo de optimización de la puja. En caso de puja manual se fija a ninguno (NONE), que es lo que se hará en la implementación.

• AudienceExpansionEnabled: indica si se habilita la expansión de audiencia para mostrar los anuncios a público con características no iguales pero similares a las especificadas en la combinación lógica. Se fija a *false* (desactivada).

Existe la posibilidad de incluir en un futuro un fichero de configuración en el que se especifique los valores que deben tomar estas variables para tener mayor flexibilidad, por ejemplo, para poder elegir puja automática en lugar de manual (aunque por lo general solo lo soporta uno de los tipos de cobro). Estos cambios se valorarán en función de la evolución de los requisitos para el proyecto PIMCity.

Para poder efectuar la petición de presupuesto, primero hay que efectuar una petición para obtener una estimación del precio por resultado (tipo de cobro), ya sea CPC o CPM. Se dirige al fichero *campaignManagerLimits* del servidor, y su formato es similar al descrito para la anterior.

Se especifica el tipo de puja, la divisa, y el periodo de la campaña. El tipo de puja será también el especificado al llamar a la función (CPC o CPM). De esta petición se obtiene

una respuesta con los límites máximos y mínimos de la puja, así como una sugerencia de puja basada en estimaciones de LinkedIn sobre los precios que se están pagando en ese momento para un público similar. En la respuesta, el valor del campo *midBid* dentro de *bidSuggestion* es la referencia que se utiliza para calcular el presupuesto por defecto, por tanto, se incluirá dentro de la petición principal en el campo unitCost.

De esta forma, se tiene una función a la que se le debe aportar en su entrada la combinación lógica de elementos (array bidimensional), el *id* de la cuenta publicitaria obtenido con la función getTokenAndId y el tipo de cobro que se desee, soportando CPC y CPM, que no devuelve ningún valor, y que escribe la respuesta (previsiones de rendimiento diario, semanal y mensual) en el fichero *forecasting.json*. En el apartado de resultados se puede ver un ejemplo del fichero resultante.

4.3.6. OBTENCIÓN DEL DESGLOSE POR SEGMENTO

La función para obtener la segmentación del público tiene una implementación diferente a las anteriores. El formato del formulario en este caso es JSON, de forma que su construcción no se efectúa con la función auxiliar común a las dos funciones anteriores.

Siguiendo la nueva estructura, se implementa un algoritmo de construcción similar al explicado para las peticiones anteriores. El código de este algoritmo se escribe dentro de la propia función, ya que ninguna otra función de las implementadas utiliza este formato para expresar la combinación lógica.

El parámetro clave en esta función es el criterio de segmentación, que se incluye a la entrada de la función y puede ser uno de los siguientes: *jobFunctions* (función laboral) *seniorities* (nivel de responsabilidad), *yearsOfExperience* (años de experiencia), *staffCountRange* (tamaño de la empresa), *industries* (sectores de la empresa) e *interests* (intereses).

Por tanto, la función *getInsights* recibe la combinación lógica de elementos (array bidimensional), el *id* de la cuenta publicitaria obtenido con la función getTokenAndId y uno de los grupos de segmentación que se han enumerado, no devuelve ningún valor, y que escribe la respuesta (número de personas y porcentaje respecto al total de la audiencia que corresponde a cada categoría) en el fichero *insights.json.* En el apartado de resultados se puede ver un ejemplo del fichero resultante.

4.4. PRUEBAS

Durante el desarrollo del módulo se hicieron diferentes pruebas para verificar que el funcionamiento de la herramienta era el correcto en todos los casos. El proceso de mejora y corrección de errores se llevó a cabo por un procedimiento de prueba y error: en un primer momento, se implementó el algoritmo de generación del formulario para que funcionara con un primer ejemplo de archivo de entrada. Una vez se consiguió hacer funcionar el módulo en ese caso, se empezaron a probar diferentes casos de archivos de entrada, con más elementos en cada línea, más líneas, combinaciones con diferente número de elementos en cada línea, etc.

En la primera fase, los errores fueron sobre todo relacionados con la construcción de las peticiones y las cabeceras de estas (el servidor devolvía error porque faltaba alguna cabecera o porque no tenían la forma correcta), aunque existieron otros. Por ejemplo, se detectó un problema en la codificación de los nombres de cada elemento en el formulario (que es de tipo *percent encoding* [32]), ya que la función utilizada para tal fin no trataba todos los caracteres especiales que se debían codificar en el caso de este servicio. La solución fue añadir un reemplazo manual como el descrito en [33] para ajustarse a la especificación RFC 3986 [34], la cual establece más caracteres a ser codificados que los que implementa la función encodeURIComponent.

Una vez solucionados estos errores, en la segunda fase se detectaron los relacionados con la construcción del formulario. Para ello, cuando el servidor devolvía un error, se analizaba el formulario enviado y se comparaba con la forma que debería tener para ser aceptado, mirando manualmente el formulario que se envía desde una sesión en el navegador cuando se introduce en la plataforma la misma combinación lógica de elementos. En ese caso, se identificaba la diferencia existente entre los dos formularios, se generalizaba el caso estudiando el motivo por el que el formulario correcto tenía esa forma, y se corregía el algoritmo para crear el formulario acorde al nuevo requisito manteniendo la generalidad, es decir, que en función de los factores que condicionen la forma de dicho formulario se componga de forma correcta para todos los posibles archivos de entrada. Mediante la realización sucesiva de estas correcciones, se consiguió finalmente que la herramienta funcionase correctamente para todas las posibilidades de archivos de entrada.

A continuación, se exponen algunas de las pruebas realizadas, para entender el proceso de prueba/error y algunos de los casos que revelaron la necesidad de modificaciones en el algoritmo:

- (Madrid OR Toledo OR Mojácar OR Vizcaya OR San Sebastián) AND (Médico). Aquí se prueban las combinaciones OR con más de tres elementos y los términos con un solo elemento (Médico), las letras con tilde y los elementos que contienen espacios.

- (Munich OR Lyon OR Brujas) AND (Personal trainer OR Wedding planner). Con esta combinación se prueban palabras en inglés y se verifica que la función de búsqueda las soporta y devuelve el resultado correspondiente en el idioma de la interfaz de la sesión del navegador, en este caso español. También se prueban las combinaciones OR con tres elementos y con dos elementos.

- Los Ángeles AND Apple. Es una combinación básica con una ubicación y un elemento adicional para comprobar que el funcionamiento básico también funciona.

- (Valdastillas OR Denia OR Berna OR Tallín OR Minsk OR London OR Maldivas) AND (Fontanería OR Cirujano OR Director OR Gerente OR Geriatra OR Anestesista). Aquí se comprueba el funcionamiento con muchos elementos en cada función OR.

Con estas pruebas se termina de verificar que todo funciona bien, y se obtiene así una versión definitiva del código interno de la herramienta en cuanto a la composición de las peticiones al servidor, para poder usarlas como *caja negra* de aquí en adelante. En todo momento se sigue comprobando que la herramienta funciona correctamente, tanto en el análisis de escalabilidad que se explica en el siguiente apartado como en el caso de uso descrito en el capítulo 5.

4.5. ESCALABILIDAD

Una vez se ha implementado la herramienta y habiendo comprobado su funcionamiento para una sola combinación en cada ejecución, es el momento de afrontar la posibilidad de que en una misma sesión se efectúe un gran número de peticiones para poder extraer información de muchas combinaciones de características diferentes. Esto implica una serie de cambios en la implementación para soportar esta capacidad de peticiones masivas y la tolerancia a fallos (fundamentalmente por fallos en la conexión con el servidor, como pueden ser bloqueos o reseteo de conexión a nivel TCP).

Para ello, se tendrán que efectuar cambios en la distribución del código, que hasta ahora se englobaba en un mismo archivo con extensión *.js*; y en el formato de entrada, ya que los datos a extraer de forma masiva no estarán en el formato establecido inicialmente, sino en listas de elementos separadas y probablemente en diferentes formatos cada una, y parte del módulo tendrá que iterar estas listas y formar todas las combinaciones posibles. Así mismo, se establecerá un mecanismo por el que se pueda guardar el estado de la ejecución (hasta qué lugar de cada lista de elementos llegó la ejecución antes de pararse o lo que es lo mismo, el valor de los iteradores de los bucles *for* que recorrerán las listas de elementos en el momento en el que por cualquier causa se interrumpa la ejecución).

Además, se utilizará una máquina remota para mantener la ejecución en marcha sin importar el tiempo que pueda tardar en recoger una gran cantidad de datos y se utilizará un sistema que permita mantener la ejecución en marcha incluso después de desconectarse de la máquina remota. Por último, se probará un mecanismo de recuperación automática de la ejecución y se evaluarán sus ventajas e inconvenientes.

4.5.1. ENCAPSULAMIENTO DEL CÓDIGO EN UN MÓDULO DE NODE

La primera decisión que se toma para iniciar esta etapa de escalado del código a un funcionamiento masivo es encapsular el código de las funciones descritas en el capítulo 4 en un módulo de Node.js. Un módulo es un fragmento de código que contiene funciones y objetos de JavaScript que pueden ser utilizados desde otros módulos que se ejecuten con Node.js.

La ventaja de encapsular las funciones implementadas en este trabajo en un módulo es la posibilidad de utilizarlas desde varios programas diferentes (por ejemplo, desde el código inicial que probaba la funcionalidad con una sola combinación y utilizando el formato de

entrada establecido, y desde otros que implementen su utilización masiva a través de bucles que iteran diferentes archivos de entrada para formar combinaciones lógicas con sus contenido) manteniendo una única versión del código de las funciones, de modo que se evitan duplicidades de código y además se pueden implementar cambios, mejoras o eliminar errores en las funciones de forma más fácil, rápida, y haciendo disponible el cambio en todos los módulos a la vez.

Para crear un módulo de Node.js se ha utilizado el tutorial [35], en el que se detallan los pasos a seguir para tal fin. Una de las herramientas que permite la gestión de módulos Node.js es *npm* (Node Package Manager), que viene incluida con el entorno. A través de este gestor de paquetes se pueden descargar módulos de uso común como los descritos en la sección *Tecnologías utilizadas.* En este caso, en lugar de descargar e instalar un módulo, se va a crear un módulo local y se va a instalar para que después pueda ser usado de la misma manera que los demás módulos.

En primer lugar, se mueve todo el código de las funciones implementadas a un archivo separado, que se llamará *index.js*, bajo el directorio *ldc.* El nombre *index.js* es el que se da por defecto al archivo de código principal en un módulo Node.js. En este directorio se inicia *npm*, para que posteriormente el módulo se pueda importar desde otros programas, utilizando el comando *npm init -y*. Con ello se creará el archivo *package.json*, que incluye información sobre el módulo: nombre, versión, descripción, archivo principal, etc. A continuación, se muestra el archivo generado:

```
{
  "name": "ldc",
  "version": "1.0.0",
  "main": "index.js",
  "dependencies": {
    "fs": "^0.0.1-security",
    "request": "^2.88.2"
  },
  "devDependencies": {},
  "scripts": {
    "test": "echo \"Error: no test specified\" && exit 1"
  },
  "keywords": [],
  "author": "",
  "license": "ISC",
  "description": ""
}
```

Se puede ver cómo incluye también los módulos necesarios para que el código del módulo se pueda ejecutar y la versión de estos (el símbolo ^ indica que se requiere la versión especificada o una superior), en este caso son *fs* para la lectura/escritura de archivos y *request* para comunicarse con el servidor.

Para que la funciones se puedan llamar desde otros programas al importar el módulo, se deben incluir unas sentencias especiales al final del archivo *index.js* utilizando la palabra clave *exports*. Solo se harán visibles las funciones que implementan las funcionalidades de obtención del *token* y el *id* (getTokenAndId), de búsqueda (getURN), de obtención de información sobre audiencias (getAudienceCounts, getForecasting y getInsights) y de inicialización del registro con las URN que no se obtienen con la función de búsqueda (initRegister). El resto de las funciones existentes en el archivo son auxiliares y solo se utilizan dentro del código de las que sí se harán visibles. El código añadido es el siguiente:

```
exports.getTokenAndId = getTokenAndId;

exports.getURN = getURN;

exports.getAudienceCounts = getAudienceCounts;

exports.getForecasting = getForecasting;

exports.getInsights = getInsights;

exports.initRegister = initRegister;
```

Una vez hecho esto, hay guardar el módulo local como dependencia y vincularlo. La vinculación es importante porque hará que cualquier cambio en el código del módulo se haga disponible instantáneamente en los otros módulos que lo utilizan. Para ello, en el directorio *ldc* se ejecuta la siguiente secuencia de comandos:

```
>sudo npm link

>cd ../modules

>sudo npm link ldc
```

Con ello queda configurado el módulo para poder llamar a las funciones desde cualquier módulo contenido en el directorio *modules* utilizando la sentencia *const ldc = require('ldc')* al principio del código.

4.5.2. FORMATOS DE ENTRADA

La escalabilidad de la herramienta se probó efectuando peticiones masivas con combinaciones del estilo País AND Empresa AND Aptitud, utilizando para ello una lista de países, una lista de empresas y una lista de aptitudes. Las listas estaban el formato CSV, pero con ligeras diferencias. Por ejemplo, la lista de países contenía la abreviatura de dos letras y el nombre completo del país, cada elemento entre comillas simples y separados por una coma (por ejemplo, 'es','España'). En cambio, la lista de empresas contenía el nombre de la empresa y su URN extraída previamente separadas por punto y coma, al igual que la lista de aptitudes.

Con esta variedad de formatos de entrada, se hace necesario el uso de dos módulos de Node.js para el manejo de archivos de tipo CSV, TSV, etc. llamados csv-parse (para la lectura y mapeo a objeto JavaScript de los ficheros) y csv-stringify (para transformar un objeto JavaScript a formato CSV o TSV y escribirlo en un fichero). Estos módulos proporcionan una API para manejar este tipo de archivos, dando flexibilidad para elegir el delimitador de elementos que tiene cada fichero (en este caso coma o punto y coma, y en caso de ficheros TSV una tabulación), el carácter de escape (comillas dobles o ninguno), y la posibilidad de especificar si la primera línea del fichero contiene los nombres de las columnas (en ese caso, la clave cada elemento del objeto JavaScript creado a partir del fichero será el nombre de la columna, y su valor un *array* que contiene los elementos de esa columna). De esta forma, se tiene flexibilidad para adaptar la herramienta a cualquier fichero de entrada que se tenga que utilizar.

4.5.3. FICHERO DE ALMACENAMIENTO DEL ESTADO DE LA EJECUCIÓN

Dado que para probar la escalabilidad se van a hacer peticiones masivas controladas por bucles for anidados que recorren las listas de elementos, para guardar el estado de la ejecución si esta se interrumpe se guardará el valor que tenían los iteradores de esos bucles, y se proporcionarán los mismos a la siguiente ejecución mediante sus argumentos de entrada, para que al ejecutar de nuevo el programa continúe por donde iba en lugar de volver a empezar desde el principio.

El código del archivo principal se adapta para que los iteradores tomen como primer valor el especificado por los argumentos de entrada de la siguiente forma:

```
for(let i = recover_data_array[0]; i<countries.length; i++){
      var country = countries[i].Country;
      for(let j = 0; j<companies.length; j++){
        if(i == recover_data_array[0] && j == 0) j = recover_data_array[1];
```

Si existen más bucles anidados, como en este caso el que recorre las aptitudes, se procede de la misma manera. Así, cuando se interrumpa la ejecución, se guardará el valor de i, j, y del resto de variables en el fichero; y al ejecutar de nuevo el módulo pasándole como argumento ese fichero, se leerán y se almacenarán en el array *recover_data_array*, de donde se leen los valores en el código mostrado. Si la ejecución había llegado a cierto país y a cierta empresa, la ejecución continuará por el mismo país y empresa, y al pasar al siguiente país será cuando la variable que recorre los países vuelva a empezar desde cero.

De esta manera, en un código que puede tardar bastante tiempo en ejecutarse, se implementa un mecanismo que evita peticiones duplicadas y reduce el tiempo total de obtención de los datos.

4.5.4. MÁQUINA REMOTA, USO DE TMUX Y GESTIÓN DE PETICIONES DNS

Como se ha dicho, una ejecución masiva puede tardar bastante tiempo, del orden de varios días, en obtener todos los datos. Por lo tanto, se hace necesario el uso de una máquina remota en la que se pueda dejar el código ejecutándose, incluso tras cerrar la conexión con esta.

Con este objetivo, se ha creado una máquina virtual sobre la infraestructura de la universidad, a la que se tiene acceso mediante una conexión VPN para acceder a la red interna. Utilizando el comando ssh del terminal, especificando el usuario y la contraseña configurados en la máquina, se puede controlar la misma, aunque sin acceso a interfaz gráfica. La ausencia de interfaz gráfica no es importante, ya que solo se utilizará esta máquina para ejecutar el código cuyo funcionamiento se ha comprobado correcto en la máquina local. Para realizar la transferencia de ficheros desde y hacia la máquina remota se utiliza el comando *scp*.

Sin embargo, al desconectar la sesión ssh, la ejecución se para. Para mantener la ejecución activa se utiliza una herramienta llamada tmux [36], que permite lanzar desde el terminal otros terminales que se pueden desacoplar y quedar en segundo plano, de forma que más adelante se puede acoplar de nuevo y ver la salida generada durante todo el tiempo que haya estado en segundo plano.

Si se inicia una ejecución en uno de estos terminales y se desacopla, se puede desconectar la sesión ssh y el programa seguirá ejecutándose. Al conectarse de nuevo, con el comando *tmux attach* se comprueba el curso actual de la ejecución.

Esto hace que el uso masivo de la herramienta sea mucho más cómodo y factible que si hubiera que tener el programa ejecutándose varios días en una máquina local.

Además, se hace uso de los módulos Node.js *dns* y *dnscache* para guardar en caché la dirección IP del servidor de LinkedIn y no saturar al servidor de DNS con demasiadas peticiones de resolución de direcciones seguidas.

4.5.5. CONTROL DE LA EJECUCIÓN CON CRONTAB

Otra mejora que se explora para automatizar aún más la herramienta en su uso masivo es la posibilidad de que, cuando se para la ejecución, esta se reactive de manera automática. Esto permitiría la automatización absoluta del proceso de recolección de datos masivos.

Para ello, se ha utilizado Crontab, una utilidad de los sistemas Linux que permite configurar la ejecución de tareas escritas en *shell scripts* con cierta frecuencia. En este caso, se quiere comprobar cada minuto si el proceso se sigue ejecutando, y si no es así, ejecutarlo de nuevo, para lo que se elabora el script que se ve en la siguiente imagen.

```
#!/bin/bash
PROCESO=$(ps -e | grep node)
VACIA=“”
echo “$PROCESO”
if [ “$PROCESO” = “$VACIA” ]; then
       node LinkedInDataCrawler/main/main.js …
       echo “Reactivando crawler”
else
       echo “Sigue ejecutando”
fi
```

En el script se inicializa una variable con el resultado de buscar un proceso de Node.js en la lista de procesos, y otra variable con una cadena vacía. Si el proceso no se está ejecutando, la primera variable será una cadena vacía, igual que la segunda. En ese caso

es cuando se ejecuta el comando para lanzar el programa. En el código mostrado se han omitido los argumentos de entrada del módulo.

La ejecución de este script de control se configura en Crontab añadiendo una nueva tarea a la lista con el comando *crontab -e*. Se escribe la siguiente línea para añadir la tarea:

```
* * * * * /home/angelm/LinkedInDataCrawler/main/script.sh
```

Los asteriscos indican que se debe ejecutar cada minuto. Si se quisiera una frecuencia de comprobación menor se podría cambiar, sabiendo que el primer elemento se refiere a los minutos, el segundo a las horas, el tercero a los días del mes, el cuarto al mes y el quinto al día de la semana.

La salida de las ejecuciones de Crontab se recibe en el correo local de la máquina remota, donde se puede consultar el estado del proceso. Además, consultado el archivo de salida donde se escriben los datos que se van recogiendo, se puede comprobar si la ejecución sigue su curso. Con esto, se finaliza el proceso de comprobación de la escalabilidad de la herramienta y adaptación de esta, afirmando que se puede utilizar a nivel masivo.

4.6. RESULTADOS GENERADOS POR EL MÓDULO

A lo largo de este apartado se describirán los resultados referentes a la implementación de la herramienta de obtención de datos de la plataforma de publicidad de LinkedIn, tanto de forma individual como a nivel masivo, y se revisará el cumplimiento de los objetivos descritos al comienzo de este documento.

En primer lugar, la herramienta funciona correctamente cuando se le proporciona una única combinación lógica que define una audiencia. Al ejecutarse las funciones de obtención de información, se obtienen los ficheros correspondientes con el contenido de la respuesta en formato JSON. A continuación, se muestra un ejemplo con cierta combinación: España AND Hombre AND Juez. Notar que el término *hombre* se especifica como *Male* ya que es el nombre que la plataforma establece internamente para esta categoría seleccionable y por tanto es así como figura en el registro de elementos.

```
{
  "elements":
    [
      {"count":800,
        "allowCampaignActivation":true,
        "includesDynamicFacets":false
      }
    ],
  "paging":
      {"count":10,
        "start":0,
        "links":[]
      },
  "requested":"(España) AND (Male) AND (Juez)"
}
```


En el texto superior se muestra la información que obtiene y almacena la función getAudienceCounts. Es la más sencilla de todas, en el campo “count” está el tamaño de la audiencia que se ha solicitado. A continuación, podemos ver la información obtenida con la función getForecasting, que es toda la información que la plataforma mostraría en la previsión de resultados tal y como se describe en el capítulo 2. En la imagen se ha omitido información de las previsiones semanal y diaria, pero también se incluyen en la respuesta recibida. Los parámetros que se obtienen en este caso son el CTR (porcentaje de clics), el gasto previsto y la estimación de resultados (impresiones, clics y *leads).*


```
{
  "elements":
    [
      {"monthly":
        {"ctr":
          {"lowEnd":0.0017,
            "highEnd":0.0033},
        "maxSpend":
          {"highEnd":2300,
            "lowEnd":400},
        "leads":
          {"highEnd":1,
            "lowEnd":0},
        "spend":
          {"highEnd":1800,
            "lowEnd":400},
        "clicks":
          {"highEnd":36,
            "lowEnd":7},
        "impressions":
          {"highEnd":12000,
            "lowEnd":2100}
        },
      "weekly":
        {"ctr":
          {"lowEnd":0.0017,
            "highEnd":0.0033},
        ...

  "paging":
    {"count":10,
      "start":0,
      "links":[]},
  "requested":"(España) AND (Male) AND (Juez)"
}
```

Por último, se puede observar la información que obtiene la función getInsights. En este caso, se ha elegido para el desglose el segmento *funciones laborales*, en el texto del archivo se ve cómo la respuesta incluye, de mayor a menor porcentaje, las funciones laborales en las que se encuadra parte de la audiencia definida en la petición: en este caso, un 99% de la audiencia en cuestión se dedica a temas legales, como es lógico sabiendo que se trata de los usuarios que son jueces, españoles y hombres. El resto de los segmentos identificados siguen la misma estructura, se omiten en el texto para que se pueda ver parte final de la respuesta, que especifica algunos parámetros dirigidos a la forma en que se muestra la información en la plataforma cuando se utiliza desde el navegador, como es el caso del parámetro *viewType*.


```
{
  "value":
    {"audienceInsight":
      {"segmentations":
        [
          {"entityCount":790,
            "entityPercentage":99,
            "segments":
            [
              {"urn":"urn:li:function:14",
                "facetUrn":"urn:li:adTargetingFacet:jobFunctions",
                "name":"Legal",
                "description":"Reach members who work within the Legal profession.",
                "targetable":true}
            ]
          },
          {"entityCount":120,
            "entityPercentage":15,
            "segments":
            [
              {"urn":"urn:li:function:7",
                "facetUrn":"urn:li:adTargetingFacet:jobFunctions",
                "name":"Education",
                "description":"Reach members who work within the Education profession.",
                "targetable":true}
            ]
          },
          ...
      "facets":
        [
          {"urn":"urn:li:adTargetingFacet:jobFunctions",
            "facetUrn":"urn:li:adTargetingFacet:jobFunctions",
            "cantExclude":false,
            "viewType":"SEARCH_AND_LIST",
            "name":"Job Functions",
            "description":"Reach members based on common tasks or activities undertaken within
their job position.",
            "facetsToNotIncludeWithAndWhenExcluded":["urn:li:adTargetingFacet:titles"],
            "facetsToNotIncludeWithAndWhenIncluded":["urn:li:adTargetingFacet:titles"],
            "targetable":false}
        ]
    },
    "totalAudienceCount":800
  },
  "requested":"(España) AND (Male) AND (Juez)"
}
```

Esta información no es en principio imprescindible de cara a los fines de la herramienta, pero como parte de la respuesta, se almacena también, para no perder compatibilidad. La única modificación que se realiza a todas las respuestas es la inclusión de una clave más en el JSON llamada *requested*, que sirve para identificar la combinación de características que identifica a la audiencia que da lugar a estas respuestas.

Esta información se puede leer fácilmente desde cualquier otro módulo, y será utilizada por el resto de los componentes del DVTMP que se está desarrollando en el proyecto PIMCity. Como resultado, se tiene una herramienta plenamente funcional para los propósitos que se presentaron al principio del trabajo: el objetivo de crear un software capaz de recoger datos de manera escalable de la plataforma de anuncios de LinkedIn se ha cumplido, y la herramienta está preparada para ser integrada en el proyecto completo del DVTMP del proyecto PIMCity.

El otro objetivo establecido es recoger información masiva de la plataforma de anuncios para hacer algún análisis estadístico de los mismos, en concreto datos sobre brecha de género en el mundo laboral. En el siguiente capítulo se describe qué datos se han utilizado, cómo se ha realizado su obtención, cómo se han analizado y los resultados y conclusiones a los que se ha llegado tras el análisis.

5. CASO DE USO: ANÁLISIS DE LA BRECHA DE GÉNERO POR PAÍSES

Para probar el verdadero potencial de esta herramienta, se ha ideado un caso de aplicación en el que se pretende recolectar información a escala global, por países, del número de hombres y mujeres que se dedican a ciertas profesiones dentro de determinadas ramas laborales, y a partir de ello calcular la brecha de género existente en cada uno de estos puestos, así como el valor mediano entre todos los que corresponden a una misma rama laboral, como puede ser ingeniería, salud, etc.

5.1. CLASIFICACIONES UTILIZADAS

Para efectuar un análisis de la brecha de género por grupos de profesiones se necesita una taxonomía o clasificación amplia y exhaustiva, pero donde los distintos grupos de profesiones no sean demasiado detallados, de forma que se pueda tener una visión global de unos 20 o 30 sectores de profesiones, como pueden ser salud, cuidados, ingeniería, negocios, computación, gestión, etc.

Inicialmente se pensó en utilizar fuentes de organismos públicos, como la Clasificación Nacional de Ocupaciones (CNO-11) del Instituto Nacional de Estadística [37], de la que a continuación se puede ver una pequeña muestra:

TABLA 5.1. CLASIFICACIÓN NACIONAL DE OCUPACIONES (CNO-11) [37]

1	Directores y gerentes
A	Directores y gerentes
11	Miembros del poder ejecutivo y de los cuerpos legislativos; directivos de la Administración Pública y organizaciones de interés social; directores ejecutivos
111	Miembros del poder ejecutivo y de los cuerpos legislativos; directivos de la Administración Pública y organizaciones de interés social
1111	Miembros del poder ejecutivo (nacional, autonómico y local) y del poder legislativo
1112	Personal directivo de la Administración Pública
1113	Directores de organizaciones de interés social
…	…

Como se puede apreciar, es demasiado detallada, porque profesiones muy similares quedan desagregadas entre sí, de modo que queda descartada.

Por otro lado, una clasificación menos detallada como la Clasificación Internacional Uniforme de Ocupaciones (CIUO-08) de la Organización Internacional del Trabajo [38] resulta poco actual (es del año 2008) y además contiene un número total de ocupaciones más bajo de lo esperado. A continuación, una muestra de esta:

TABLA 5.2. CLASIFICACIÓN INTERNACIONAL UNIFORME DE OCUPACIONES (CIUO-08) [38]

CÓDIGO CIUO 08	TÍTULO SP
1	Directores y gerentes
11	Directores ejecutivos, personal directivo de la administración pública y miembros del poder ejecutivo y de los cuerpos legislativos
111	Miembros del poder ejecutivo y de los cuerpos legislativos.
1111	Miembros del poder legislativo
1112	Personal directivo de la administración pública
1113	Jefes de pequeñas poblaciones
…	…

Finalmente, buscando otras alternativas, se encuentra que la base de datos gratuita O*NET (Red de Información Ocupacional) [39], mencionada en el artículo *Toward understanding the impact of artificial intelligence on labor* [40]. Esta fuente proporciona diferentes clasificaciones de ocupaciones, en concreto las más adecuadas para este propósito son las diferentes disponibles dentro del apartado *find occupations > browse by job family*. Hay 23 categorías diferentes, en el nivel de detalle que se busca, y para cada una se puede descargar un archivo CSV que contiene suficientes ocupaciones distintas como para tener un buen número total de ellas. Al fusionar todos los archivos se obtienen 980 ocupaciones diferentes. Se escoge esta clasificación por ser la alternativa más adecuada según los objetivos fijados.

Una vez se tiene la tabla con la clasificación de profesiones, hay que procesarla para adaptarla a la terminología propia de la plataforma de anuncios de LinkedIn. Muchas de las entradas de la tabla son en realidad profesiones similares, aunque distintas, separadas por comas o por la palabra “and”, de forma que se separan en diferentes entradas para poder capturar posteriormente la URN de cada una de ellas. Esto hace que el número total de entradas en la lista aumente.

También es necesario eliminar algunos términos, por ejemplo, en entradas que incluyen excepciones, las cuales son eliminadas para facilitar la búsqueda de URNs. Por ejemplo, para *Mobile Heavy Equipment Mechanics, Except Engines*, se elimina *Except Engines*. En la siguiente imagen se muestra una parte de la clasificación depurada. La clasificación completa se puede encontrar en el anexo de este trabajo.

TABLA 5.3. CLASIFICACIÓN DE PROFESIONES.
ELABORACIÓN PROPIA CON DATOS DE O*NET [39]

Category	Occupation
architecture_and_engineering	Aerospace Engineers
architecture_and_engineering	Agricultural Engineers
architecture_and_engineering	Architects
architecture_and_engineering	Architectural Drafters
architecture_and_engineering	Civil Drafters
architecture_and_engineering	Automotive Engineering Technicians
architecture_and_engineering	Automotive Engineers
architecture_and_engineering	Bioengineers
architecture_and_engineering	Biomedical Engineers
architecture_and_engineering	Calibration Technicians
architecture_and_engineering	Cartographers
…	…

Una vez curada la lista, se elabora un código principal que lee cada elemento, utiliza la función *get_urn* para buscarla en LinkedIn, y si obtiene un resultado, lo escribe en otro archivo TSV, en el que cada entrada contiene la categoría, el nombre de la ocupación buscada, el nombre de la ocupación devuelta por LinkedIn y su URN. Cabe señalar que antes de buscar cada ocupación se eliminan su dos últimos caracteres, ya que las palabras están en plural mientras que la búsqueda en la plataforma de anuncios de LinkedIn está diseñada para recibir términos en singular, y si se buscan directamente en plural no devuelve ningún resultado.

Tras esto, se tiene una lista de unas 700 entradas, aunque hay que revisarla para eliminar los posibles fallos a la hora de buscar una ocupación, ya que para algunas ocupaciones LinkedIn puede devolver una sugerencia que en realidad no tiene nada que ver.

Tras eliminar estos gazapos y algunos elementos duplicados, el número total de ocupaciones con su URN es de 609. Con esta lista se puede ya solicitar el número de

hombres y mujeres que desempeñan cada profesión en cada país, utilizando una lista de países en formato TSV extraída de un repositorio de GitHub [41].

TABLA 5.4. CLASIFICACIÓN FINAL Y CORRESPONDENCIA CON ELEMENTOS DE LINKEDIN

Category	Occupation	LinkedIn Name	URN
architecture_and engineering	Aerospace Engineers	Aerospace Engineer	3883
architecture_and engineering	Agricultural Engineers	Agricultural Engineer	22400
architecture_and engineering	Architects	Architect	72
architecture_and engineering	Architectural Drafters	Architectural Drafter	13201
architecture_and engineering	Automotive Engineers	Automotive Engineer	11598
architecture_and engineering	Bioengineers	Biological Engineer	25775
architecture_and engineering	Biomedical Engineers	Biomedical Engineer	4279
architecture_and engineering	Calibration Technicians	Calibration Technician	9222
architecture_and engineering	Cartographers	Cartographer	5910
…	…	…	…

5.2. OBTENCIÓN DE LOS DATOS

Para obtener el número de hombres, mujeres, y el total de personas que se dedican a cada profesión de la clasificación diseñada, se elabora un script que recorre línea a línea el archivo CSV que contiene la lista de países, y a su vez, para cada país, recorre el fichero que contiene la clasificación (arbol.tsv). Para cada combinación país-profesión, se hace lo siguiente:

1. Se llama dos veces a la función getURN, primero para obtener la información del país, y después para obtener la información de la profesión.

```
await getURN(page,[country],accountId,['locations']);

await getURN(page,[occupation],accountId,['titles']);
```

2. Se llama a la función *getAudienceCounts* tres veces: una para obtener la audiencia masculina, otra para obtener la audiencia femenina, y otra para obtener la audiencia total:

```
maleCount = await makeCombination(page,
       [[country],[ocupation],['Male']], accountId);

femaleCount = await makeCombination(page,
       [[country],[ocupation]], accountId);

allCount = await makeCombination(page,
       [[country],[ocupation],['Female']], accountId);
```

3. Los resultados se van escribiendo en otro fichero TSV de salida línea a línea, escribiendo el país, la ocupación, el género (hombres, mujeres y todos), y el número correspondiente, tal y como se muestra a continuación.

TABLA 5.5. DATOS OBTENIDOS COMO SALIDA DEL PROGRAMA

Country	Occupation	Gender	Count
United States	Architect	Male	90000
United States	Architect	Female	39000
United States	Architect	ALL	140000
United States	Mechanical Engineer	Male	270000
United States	Mechanical Engineer	Female	48000
United States	Mechanical Engineer	ALL	340000
United States	Electrical Engineer	Male	150000
United States	Electrical Engineer	Female	25000
United States	Electrical Engineer	ALL	190000
…	…	…	…

Los datos se han obtenido por orden según el tamaño de cada país, empezando por los países que tienen una base de usuarios más grande. La base de usuarios de cada país se ha obtenido con un código similar en el que se solicita la audiencia del país sin incluir ningún criterio de segmentación más.

Observando los datos obtenidos, se verifica que, aproximadamente, solo los 20 países que mayor base de audiencia tienen son los que proporcionan información completa (se obtiene un valor para cada posición de la tabla). En el resto de los países se obtienen valores a cero en muchas de las profesiones, lo que significa que las audiencias son demasiado bajas para que LinkedIn la pueda mostrar (inferior a 300).

Por ello, se obtienen también, mediante un código similar, las audiencias masculinas, femeninas y totales de cada una de las funciones laborales que LinkedIn ofrece en la categoría con el mismo nombre. La lista completa es la siguiente:

- Administración
- Arte y diseño
- Atención al cliente
- Bienes raíces
- Compras
- Consultoría
- Contabilidad
- Control de calidad
- Desarrollo empresarial
- Educación
- Finanzas
- Gestión de productos
- Gestión de programas y proyectos
- Ingeniería
- Investigación
- Legal
- Liderazgo
- Marketing
- Medios de comunicación
- Operaciones
- Recursos humanos
- Servicios militares y de protección
- Servicios sanitarios
- Servicios sociales y comunitarios
- Tecnología de la información
- Ventas

Aunque la mayoría de las categorías laborales de la clasificación elaborada y las que ofrece LinkedIn en esta categoría de segmentación no tienen una equivalencia directa, se pueden utilizar ambos conjuntos de datos para realizar análisis paralelos y comparar sus conclusiones.

El primer conjunto de datos se utilizará en un análisis centrado en los países que tienen información suficiente para obtener estadísticos sobre una cantidad de datos más densa, mientras que el segundo se utilizará para obtener una panorámica de la brecha de género laboral a nivel mundial.

5.3. ANÁLISIS DE LOS DATOS

Para analizar los datos se utilizan scripts escritos en el lenguaje de programación Python, utilizando la librería Pandas, que facilita la carga de los datos y la realización de operaciones con ellos. Para elaborar los scripts se utiliza Google Colab, que es una herramienta en la nube en la que se puede ejecutar código Python 3 por fragmentos,

asignando cada parte del código a una celda diferente, de forma que se tiene un mayor control sobre la ejecución del código, se facilita la realización de diferentes pruebas y se favorece en gran medida la corrección de errores.

Para analizar la brecha de género, se calculará una medida representativa de la misma, que está acotada entre 0 y 1, donde un valor cercano a 1 simboliza un sector en el que hay muchos hombres y pocas mujeres, y un valor cercano a 0 significa que el sector está mayoritariamente compuesto por mujeres. El valor intermedio, 0.5, representa la igualdad entre hombres y mujeres en ese sector laboral. Esta medida de la brecha de género se toma del artículo *Professional Gender Gaps Across US Cities* [22].

$$indicador\ de\ brecha\ de\ género\ (\mathrm{GD}) \equiv \frac{hombres}{hombres + mujeres}$$

El proceso de análisis se enfoca de la siguiente forma:

- En primer lugar, para el conjunto de datos que incluye todos los países, se calcula el indicador de brecha de género para cada categoría en cada país, y se calcula la mediana muestral de todas las categorías para cada país. De esta forma se obtiene un valor mediano del indicador de brecha de género para cada país. Con los resultados se elabora un mapa en el que se visualizan fácilmente las diferencias entre países y zonas del mundo.

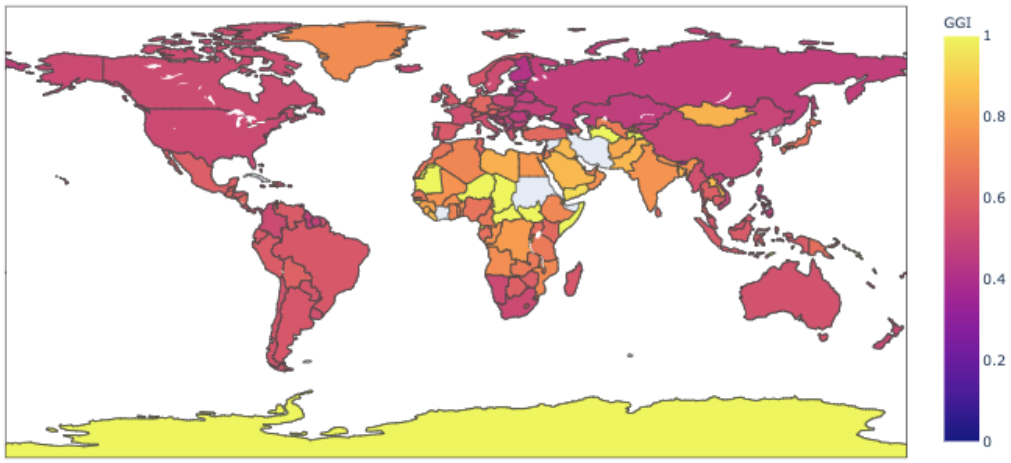

Fig. 5.1. Mapa mundial representando la brecha de género por países. Elaboración propia con datos de LinkedIn Campaign Manager [3].

Se puede ver cómo, por lo general, en los países menos desarrollados es donde existe una inclinación mayor del indicador de brecha de género hacia la sobrerrepresentación de los hombres (tonos naranja y amarillo).

- En segundo lugar, con el conjunto de datos más detallado pero que solo contiene información suficiente para algunos países (se seleccionan los 17 de la lista con mayor población por ser los que tienen un número aceptable de muestras distintas de cero), se calcula para cada país un vector de 23 elementos, uno por categoría, con la mediana del indicador de brecha de género de todas sus ocupaciones. Con estos vectores, se calcula una matriz con la distancia entre cada país, donde cada elemento es la distancia coseno entre los países correspondientes y cada columna y fila corresponde a un país. Los valores de la diagonal principal son, por tanto, 1. A esta matriz se le aplica un algoritmo de agrupamiento (agrupamiento jerárquico implementado en la función *AgglomerativeClustering* del paquete *sklearn.cluster* [42]), para identificar qué países son más similares en su distribución del indicador de brecha de género a lo largo de las diferentes áreas laborales definidas en la clasificación. La clasificación en tres grupos queda de la siguiente forma:

    - Grupo 1: Asia pacífico y Japón, India y Turquía.

    - Grupo 2: Australia, Canadá, Colombia, France, Indonesia, Italia, México, Filipinas, España, Reino Unido, Estados Unidos, Benelux (Bélgica, Holanda y Luxemburgo), y DACH (Alemania, Austria y Suiza).

    - Grupo 3: Brasil.

  En cada grupo, la distribución del indicador por áreas es más parecida entre sí.

- Como último resultado, para el conjunto de datos que incluye todos los países se elabora una tabla en la que se muestra, para cada país, las tres categorías que más se alejan del valor de paridad del indicador (0.5), tanto hacia arriba como hacia abajo. A continuación, se muestra una parte de la tabla. Se puede encontrar la tabla completa en el anexo del trabajo.

TABLA 5.6. MEDIANA DEL INDICADOR Y CATEGORÍAS CON MAYOR DESIGUALDAD POR PAÍS

Country	GDM	Category 1	GD 1	Category 2	GD 2	Category 3	GD 3
Afghanistan	0,8	Engineering	0.90	Information Technology	0.87	Program and Project Management	0.85
Albania	0,5	Military and Protective Services	0.73	Engineering	0.68	Accounting	0.33
Algeria	0,72	Operations	0.85	Military and Protective Services	0.84	Entrepreneurship	0.84
American Samoa	0,59	Business Development	0.65	Sales	0.60	Education	0.40
Andorra	0,57	Engineering	0.70	Administrative	0.31	Business Development	0.68
…	…	…	…	…	…	…	…

6. MARCO REGULADOR

Este proyecto se encuadra en el ámbito de la publicidad online, y por tanto está relacionado con el almacenamiento y procesado de datos de usuarios de internet, en concreto de la red social LinkedIn. Tanto esta como otras redes sociales que incorporan servicios de publicidad online recogen datos de sus usuarios, tanto de forma explícita si el usuario los introduce durante la creación de su cuenta o posteriormente (sexo, edad, o experiencia laboral en el caso de LinkedIn), como implícita; y los utilizan para ofrecer el servicio de publicidad segmentada. Dicho de otra forma, utilizan el conocimiento de los datos de los usuarios como un producto, y los anunciantes compran ese producto a un precio variable determinado por las pujas.

En esa compraventa entran en juego las regulaciones y aspectos legales de protección de datos personales. En este apartado se revisarán los principales aspectos de la regulación actual en España, la forma en que esta se aplica a las redes sociales que incorporan sistemas de publicidad de este tipo, y cómo afecta a la herramienta desarrollada en este trabajo y a los datos obtenidos con ella.

6.1. REGLAMENTO GENERAL DE PROTECCIÓN DE DATOS

El Reglamento General de Protección de Datos (RGPD o GDPR por sus siglas en inglés) [43] es la normativa que regula los aspectos legales de protección de datos en la Unión Europea. Es de carácter general y de obligado cumplimiento por parte de todos los Estados miembros. Entró en vigor en mayo de 2016, pero no fue de aplicación obligatoria hasta mayo de 2018. Anteriormente, la regulación venía dada por la Directiva 95/46/CE [44], referente a la protección y libre circulación de datos personales, implementada en España a través de la Ley Orgánica de Protección de Datos de Carácter Personal, y derogada por el mencionado Reglamento.

La normativa actual introdujo novedades importantes, sobre todo en lo referente al principio de responsabilidad proactiva: establece que las empresas y organizaciones deben ser conscientes y analizar qué datos tratan, el tipo de tratamiento y su finalidad cumpliendo los preceptos del Reglamento y siendo capaces de demostrar su cumplimiento. Esto afecta de lleno a las empresas de redes sociales como LinkedIn, ya que además el reglamento establece que el mismo se aplica también al tratamiento de

datos de ciudadanos de la Unión por parte de cualquier empresa, incluso si es una empresa situada fuera de la UE.

Además, establece otros principios que deben respetarse a la hora de realizar el tratamiento de datos personales, que aparecen resumidos en la Guía para el Ciudadano de la Agencia Española de Protección de Datos [45], y que se enumeran a continuación:

1. Principio de licitud, lealtad y transparencia: se pretende evitar que los datos sean tratados de forma desleal u ocultando información al interesado sobre los riesgos y las consecuencias del tratamiento obligando a los responsables a ser lo más transparentes posible.

2. Principio de limitación de la finalidad: los datos serán recogidos para unos fines determinados, sin que el tratamiento devenga en incompatibilidad con otros fines. La finalidad del tratamiento debe estar definida de forma clara y permitida por la ley.

3. Principio de minimización de datos: impide que se recojan y traten datos no necesarios para la finalidad solo porque la empresa considere que podrían ser útiles.

4. Principio de exactitud: vela por que los datos personales sean exactos y estén actualizados, y establece que debe haber mecanismos de rectificación y supresión que funcionen con agilidad.

5. Principio de plazo de conservación: los datos solo serán conservados durante el tiempo necesario para llevar a cabo los fines del tratamiento. Tras ello, deberán ser eliminados, o como mínimo suprimirse todos los elementos que permitan identificar al interesado.

6. Principio de integridad y seguridad: se garantizará la seguridad de los datos durante el tratamiento, y se protegerán contra el tratamiento no autorizado y contra pérdida, destrucción o daño accidental aplicando las medidas necesarias para ello.

También establece una serie de derechos explícitos que tienen los sujetos cuyos datos personales sean tratados por cualquier empresa u organización [45]:

1. Derecho de acceso: el responsable del tratamiento está obligado, bajo petición del interesado, a proporcionar una copia de los datos objeto de tratamiento, los fines del tratamiento, la categoría o categorías de los datos, a qué terceras partes se comunicarán, y el plazo previsto de conservación, o en su defecto los criterios que se utilizan para establecer ese plazo. Además, el interesado tiene derecho a limitar u oponerse al tratamiento de sus datos personales, y a conocer el origen de los datos cuando no se han obtenido directamente del interesado y la existencia de decisiones automatizadas. En este último caso el responsable del tratamiento debe informar sobre lógica aplicada y las consecuencias previstas para el interesado.

2. Derecho de rectificación: se podrá obtener a la máxima brevedad posible la rectificación de datos inexactos por parte del responsable del tratamiento y se tiene derecho a que se completen datos incompletos, aportando si es necesario documentación que acredite esta circunstancia.

3. Derecho de oposición: el interesado puede oponerse al tratamiento de sus datos personales cuando el tratamiento se basa en el interés público (por ejemplo, elaboración de perfiles) excepto que se acrediten motivos de fuerza mayor que prevalezcan sobre los derechos del interesado; y cuando la finalidad del tratamiento sea la mercadotecnia directa, por ejemplo, publicidad online basada en perfiles. En este último caso, se comunicará la petición del interesado al anunciante en como máximo diez días hábiles.

4. Derecho de supresión o derecho al olvido: se podrá ejercer cuando los datos ya no son necesarios para el fin establecido, si se retira el consentimiento y no hay otra causa que legitime el tratamiento, si se ha ejercido el derecho de oposición en los términos descritos en él, si los datos se han tratado de manera ilícita, si deben suprimirse por obligación legal establecida por la Unión o los Estados miembros, o si se han obtenido en relación a los servicios mencionados en el artículo 8, apartado 1 del Reglamento (condiciones aplicables al consentimiento del niño en relación con los servicios de la sociedad de la información).

5. Derecho a la limitación del tratamiento: establece, por un lado, las condiciones para solicitar la suspensión del tratamiento (cuando se impugne su exactitud, durante el plazo de verificación, o cuando se haya ejercido el derecho de oposición al tratamiento en base a interés púbico mientras se verifica si los motivos de ese tratamiento prevalecen); y por otro lado las condiciones para solicitar la conservación de los datos (cuando ante tratamiento ilícito el interesado se opone a la supresión y en cambio solicita la limitación de uso o cuando el responsable ya no necesita los datos pero el interesado sí, para reclamaciones).

6. Derecho a la portabilidad de los datos: si el tratamiento se realiza de forma automatizada el interesado recibirá sus datos en un formato estructurado y de uso común para poder transmitirlos a otra entidad.

7. Derecho a no ser objeto de decisiones individualizadas: se garantiza que el responsable y/o terceros no tomen decisiones que afecten significativamente desde un punto de vista jurídico o similar al interesado basándose solamente en el tratamiento de sus datos. Incluye las decisiones individualizadas basadas en perfiles sobre aspectos personales como salud, situación económica, intereses, comportamiento, etc. No es aplicable cuando la decisión se enmarca en la celebración de un contrato entre el responsable y el interesado o se fundamente en un consentimiento previo. En esos supuestos, el responsable debe garantizar al interesado el derecho a la intervención humana, a expresar su punto de vista y a impugnar la decisión.

El ejercicio de los citados derechos debe ser gratuito, responderse en el plazo de un mes (prorrogable a dos meses adicionales o más dependiendo de la complejidad), y realizable directamente o a través de un representante legal o voluntario.

Por último, es importante diferenciar lo que se consideran datos personales de lo que se consideran datos agregados. En el artículo 4 del Reglamento [43], se establece que, a efectos de este, se entenderán como datos personales “toda información sobre una persona física identificada o identificable” y se considera persona física identificable “toda persona cuya identidad pueda determinarse de forma directa o indirecta”, en particular

mediante identificadores como un nombre, un número identificativo, datos de localización "u otros elementos propios de la identidad fisiológica, genética, psíquica, económica, cultural o social de dicha persona". Sin embargo, los datos agregados carecen de cualquier elemento asociado a una persona física identificada o identificable.

El reglamento es de aplicación respecto a datos personales exclusivamente, y hace una única mención a los datos agregados, en su consideración inicial número 162, en el que se establece que el Reglamento debe aplicarse al tratamiento de datos personales con fines estadísticos, y que el fin estadístico implica que el resultado del tratamiento no sean datos personales, sino datos agregados.

A continuación, se estudiará cómo se relacionan estos principios y derechos que establece el Reglamento con la publicidad online y con la herramienta desarrollada en este trabajo.

6.2. CUMPLIMIENTO DEL REGLAMENTO EN LA PUBLICIDAD ONLINE

Una vez se han visto los preceptos principales del reglamento y los objetivos que persigue esta regulación, plasmados en los principios y derechos que se han enumerado, se pretender evaluar el cumplimiento de estos en las herramientas de publicidad online, en particular en el gestor de campañas de LinkedIn.

En primer lugar, LinkedIn recoge datos de sus usuarios en base a las Condiciones de uso [46] de la red social, donde se menciona explícitamente lo siguiente: "acuerdas con LinkedIn que podremos acceder, almacenar, procesar y usar cualquier información y dato personal que facilites de acuerdo con lo establecido en los términos de la Política de privacidad y tus elecciones (incluida la configuración)."

En la Política de privacidad que se menciona [47], se incluye toda la información referente a los servicios a los que se aplica, quiénes son los responsables del tratamiento de los datos, los datos que pueden ser almacenados y tratados, el uso de los datos (finalidad del tratamiento), la política de comunicación de los datos a terceros, y las opciones y obligaciones del usuario respecto a esta obtención, almacenamiento y tratamiento de datos personales, seguridad, transferencia de datos a terceros países, fundamentos legales del tratamiento y otra información, cumpliendo así, al menos sobre el papel, las obligaciones que se derivan de la aplicación del Reglamento General de Protección de datos.

A continuación, se enumera buena parte de los datos que LinkedIn declara que puede almacenar en virtud de su Política de privacidad:

- Datos que proporciona el usuario al crear una cuenta en LinkedIn (nombre, dirección de email, número de móvil, contraseña, y adicionalmente información de pago y facturación; además de datos de perfil como educación, experiencia laboral, aptitudes, fotografía personal, ciudad o ubicación y otros datos aportados al utilizar los Servicios, por ejemplo, al rellenar un formulario, una solicitud de empleo o al responder encuestas).

- Datos personales que pueden ser recopilados sin que el usuario los proporcione de forma directa o datos de terceros (información personal que forme parte de publicaciones, como noticias o logros profesionales; información de contacto; información aportada al utilizar servicios de los clientes y socios de LinkedIn o de sus filiales, incluyendo Microsoft).

- Otra información registrada mediante el uso de los Servicios, mediante cookies, información del dispositivo y la ubicación, mensajes, y al visitar sitios de terceros a través de anuncios en LinkedIn o al iniciar sesión en esos sitios con la cuenta de LinkedIn.

Como se puede ver, algunos de estos datos son los que se utilizan posteriormente en el gestor de campañas de LinkedIn para segmentar las campañas publicitarias, por lo que se utilizan para elaborar perfiles y al mismo tiempo se ofrece esta capacidad de segmentación a los anunciantes como parte del servicio de publicidad. Esto implica que, en cumplimiento del Reglamento, los usuarios tienen derecho a oponerse al tratamiento de los datos mencionados, al menos en la parte del tratamiento que tenga como finalidad la oferta de este servicio de publicidad.

Otro aspecto importante del servicio de publicidad es su posible colisión con la privacidad de los datos personales de cada usuario ya que, tal y como se ha descrito en el capítulo 2 de este trabajo, se puede seleccionar una combinación de características para seleccionar el objetivo de la campaña, de modo que podría ser posible la selección de una combinación que corresponda con un público objetivo de tamaño uno, es decir, una única

persona. Esto supondría un riesgo en materia de seguridad, ya que se podría utilizar la campaña publicitaria como medio para mostrar una información diferente y personalizada a cada persona, lo que se conoce como *nano-targenting*, que puede ser utilizado con propósitos ilícitos como extorsión, estafa, manipulación de la opinión pública, etc.

Sin embargo, LinkedIn ha tomado medidas a fin de intentar evitar esta práctica mediante la inclusión de un límite inferior en el tamaño de la audiencia de una campaña, de forma que si una combinación de características corresponde a un público de tamaño inferior a 300 personas, el servicio no devolverá información sobre el número de personas que componen esa audiencia, ni el pronóstico de rendimiento de la campaña, ni el desglose por segmento de la audiencia, ni por su puesto permitirá el lanzamiento de la campaña.

Esto es un ejemplo del tipo de medidas que estas plataformas pueden tomar para defender la privacidad y seguridad de los datos personales y de los propios usuarios, teniendo en cuenta los posibles riesgos derivados de estas actividades, tal y como les obliga el citado Reglamento.

6.3. CUMPLIMIENTO DEL REGLAMENTO EN ESTE TRABAJO

En cuanto a la herramienta desarrollada en este trabajo, se cumple el reglamento ya que los datos que se recogen no se consideran datos personales, sino datos agregados. Ninguno de los datos recogidos mediante la herramienta incluye información personal, solamente se recogen datos sobre perfiles de varios usuarios, en este caso 300 o más, con características comunes. En ningún momento se dispone de información ligada a una identidad personal.

Los datos concretos que se pueden obtener mediante la herramienta son el número de usuarios que componen un perfil de usuarios determinado, las estimaciones de coste y rendimiento de la campaña publicitaria para ese público y el desglose por segmento. Estos datos van asociados a la definición del perfil, es decir, a la combinación de características que tienen en común los usuarios que lo componen. De este modo, tanto los datos que se pueden obtener con la herramienta como los efectivamente obtenidos y tratados en el caso de uso (capítulo 5 del presente trabajo) son datos agregados, y los resultados del análisis de estos también lo son.

7. ENTORNO SOCIOECONÓMICO

7.1. PLANIFICACIÓN TEMPORAL

A continuación, se detallan las tareas en las que se ha dividido el trabajo, su duración temporal estimada, con las fechas de inicio y finalización, y una línea temporal en la que se visualiza la secuencia de actividades.

Nombre	Fecha de inicio	Fecha de fin
Preparación del entorno de desarrollo	13/7/20	14/7/20
Aprendizaje del grabador de sesiones	14/7/20	15/7/20
Estudio de LinkedIn Campaign Manager (nivel funcional)	16/7/20	22/7/20
Lectura documentación JS y Node.js	23/7/20	5/8/20
Lectura documentación puppeteer	27/7/20	7/8/20
Diseño e implementación del inicio de sesión	28/7/20	10/8/20
Estudio de LinkedIn Campaign Manager (nivel técnico)	11/8/20	17/8/20
Diseño e implementación del módulo	17/8/20	26/11/20
Diseño e implementación de obtención del token	17/8/20	28/8/20
Diseño e implementación de la búsqueda	31/8/20	8/9/20
Diseño del registro e integración en la búsqueda	4/9/20	14/9/20
Función de obtención de la audiencia	15/9/20	28/9/20
Función de obtención del pronóstico	29/9/20	1/10/20
Función de obtención del desglose	2/10/20	15/10/20
Pruebas y solución de errores	16/10/20	26/11/20
Preparación de máquina remota para uso masivo	27/11/20	1/12/20
Adaptación del código para uso masivo	2/12/21	10/12/21
Caso de uso	1/2/21	27/4/21
Obtención de datos	1/2/21	12/3/21
Elaboración de la memoria	1/2/21	12/2/21
Adaptación del código para obtención de datos	15/2/21	26/2/21
Obtención de los datos	1/3/21	12/3/21
Análisis de datos	15/3/21	27/4/21
Consulta documentación Python	15/3/21	19/3/21
Consulta documentación Pandas	22/3/21	26/3/21
Preparación del entorno colab	7/4/21	8/4/21
Curado, procesado y análisis de los datos	9/4/21	22/4/21
Presentación de los resultados	23/4/21	27/4/21
Elaboración de la memoria	28/4/21	22/6/21

Fig. 7.1. Lista de actividades y planificación. Elaboración propia con la herramienta Ganttproject [48].

El trabajo se realiza en jornadas laborales de cuatro horas de lunes a viernes. El tiempo total de trabajo estimado son 776 horas, y entre el inicio y la finalización del proyecto transcurren 339 días naturales.

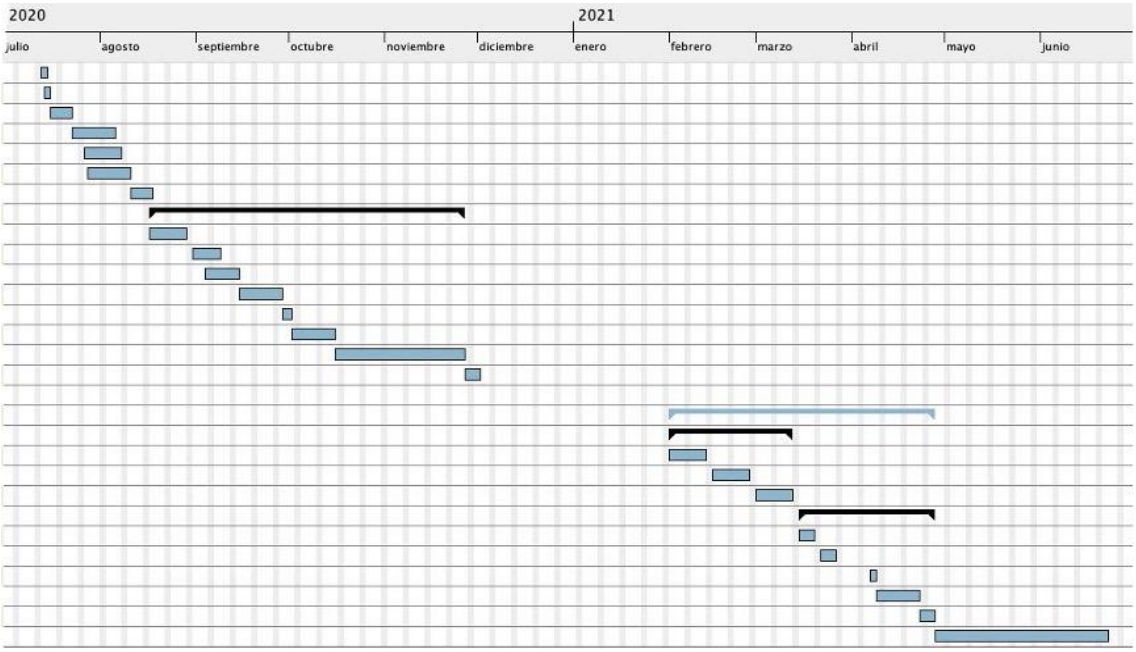

Fig. 7.2. Actividades en línea temporal. Elaboración propia con la herramienta Ganttproject [48].

7.2. PRESUPUESTO DE ELABORACIÓN

7.2.1. PRESUPUESTO DE MATERIALES

En la siguiente tabla se recoge el presupuesto asignado a los recursos materiales necesarios para la realización del proyecto.

TABLA 7.1. PRESUPUESTO DE MATERIALES

CONCEPTO	IMPORTE (€)
Licencia Office 365	126
PC de trabajo	1200
Máquina remota	500
Material de oficina	50
Plataforma Google Colab	0
TOTAL	1876

7.2.2. PRESUPUESTO DE RECURSOS HUMANOS

En la siguiente tabla se muestran los perfiles laborales involucrados en la realización del proyecto, sus horas de trabajo estimadas y el importe total que les corresponde por su participación.

TABLA 7.2. PRESUPUESTO DE RECURSOS HUMANOS

CONCEPTO	HORAS	SUELDO (€/HORA)	IMPORTE (€)
Jefe de proyecto	50	25	1250
Ingeniero	640	12	7680
TOTAL			8930

El coste total del proyecto, sumando los recursos humanos y materiales, asciende a 10836€.

7.3. IMPACTO SOCIOECONÓMICO

El hecho de que se haya desarrollado un Reglamento Europeo que regula el tratamiento de datos personales deja ver la relevancia que está adquiriendo este tema en los últimos años. El módulo software desarrollado en este trabajo y sus futuras aplicaciones tienen una serie de implicaciones económicas y sociales que se analizan en este apartado.

Como se ha dicho, los dos objetivos principales a los que se da cumplimiento en este trabajo a partir del módulo desarrollado son, por un lado, tener finalizado uno de los módulos que van a componer la herramienta final desarrollada en el marco del proyecto PIMCity llamada DVTMP (una herramienta que permite conocer el valor de los datos personales desde el punto de vista del mercado de datos) y, por otro lado, efectuar un análisis a nivel mundial por países de la brecha de género en el ámbito laboral, desarrollado en el capítulo 5 del trabajo.

El conocimiento del valor de los datos de los usuarios en diferentes plataformas y redes sociales es relevante desde el punto de vista económico porque permite comparar las diferentes ofertas o precios de cada una, en un momento dado y para una audiencia en particular, y establecer así un sistema de decisión de la plataforma en la que anunciarse que aumenta la competencia y permite a los anunciantes elegir la oferta con un menor

precio o que mejor se adapta a sus necesidades. Únicamente con el módulo desarrollado en este trabajo se espera tener un impacto importante en el diseño de campañas publicitarias, pudiendo comparar de una forma rápida diferentes opciones de segmentación y públicos objetivos y sus respectivos precios y estimaciones de rendimiento. Al integrar este módulo con el resto de componentes del DVTMP, se podrán hacer estas comparaciones en modo multiplataforma, lo que tendrá un impacto aún mayor en la forma en que se diseñan las campañas publicitarias y en el precio de estas, ya que una mayor competencia puede conducir a una rebaja de los precios.

De hecho, cada vez más servicios se utilizan de forma aparentemente gratuita cuando realmente el precio que se paga son los datos personales de todos y cada uno de los usuarios, de cuyo tratamiento surge una actividad económica, la publicidad online, que tiene un alto beneficio económico para las empresas que lo ofertan. Sin embargo, este sistema carece de transparencia y eso ocasiona la falta de competencia que se pretende subsanar en el marco del proyecto PIMCity.

Mediante herramientas como la desarrollada en este trabajo se abre una oportunidad de potenciar la transparencia económica de este sistema, tanto para las empresas que quieran publicitar sus productos en estas plataformas, como para los usuarios utilizan servicios que implican la cesión de sus datos personales, y que ambos puedan conocer el valor del producto en cuestión (los datos). De esta manera los anunciantes gozarán de mejores ofertas y mejores precios, y a su vez los usuarios podrán ser conscientes del valor económico que generan sus datos. Este valor es en realidad el pago que efectúan a cambio de poder utilizar las redes sociales y otros servicios que ofrecen las mismas empresas que prestan el servicio de publicidad online.

Desde un punto de vista más social, se espera un impacto positivo derivado de los análisis del mercado laboral que se pueden realizar con los datos a los que la herramienta desarrollada proporciona acceso. En este trabajo se ha elaborado un caso de uso centrado en el análisis de la brecha de género a nivel mundial por países, que revela información útil y precisa para abordar este problema social de forma eficaz y realista, y permite una revisión continua de las tendencias y cambios a lo largo del tiempo con tan solo recoger los mismos datos de forma periódica. Esto permite evaluar el resultado de las políticas que se aplican y observar los cambios en la sociedad respecto a este tema.

Sin embargo, existen muchos otros problemas y fenómenos de carácter socioeconómico relacionados con el mercado laboral que se podrán analizar en el futuro utilizando esta herramienta, como los relacionados con la empleabilidad en función de la edad en diferentes sectores, el nivel de adquisición de aptitudes profesionales que tendrán más demanda en el futuro, y además se podrá extender estos análisis a un ámbito más concreto que los países, analizando estos fenómenos por regiones o incluso a nivel local, comparando la situación de diferentes regiones y ciudades con un nivel de granularidad, precisión y cantidad de datos prácticamente imposible de conseguir de otra manera.

Como resultado de estos estudios, se podrían aplicar políticas económicas y sociales destinadas a mejorar el dinamismo laboral y la empleabilidad en las zonas que más lo necesiten, contrarrestando las debilidades de cada zona según la información obtenida, con programas de formación, de inversión y de fomento de diferentes actividades económicas diseñados a partir de información verosímil y específica, que describe la situación concreta de cada zona geográfica, con la posibilidad de realizar una evaluación continuada de los resultados y del retorno de la inversión y llevar a cabo ajustes en las políticas acordes a esta evaluación.

En suma, la herramienta desarrollada tiene un gran potencial para generar valor en la sociedad, y abre una oportunidad para mejorar la transparencia y la competencia en el ámbito de la publicidad online tanto para empresas como para usuarios, e incluso podría utilizarse como ayuda para la aplicación de políticas que fortalezcan la economía y el empleo.

8. CONCLUSIONES Y TRABAJO FUTURO

A lo largo de este trabajo ha quedado claro que los datos personales tienen una gran importancia en el marco de la publicidad online, ya que el mayor atractivo que presenta es la segmentación y elección del público objetivo a través de criterios marcados por esas características de los usuarios.

En este contexto, se ha conseguido desarrollar con éxito una herramienta para la plataforma de publicidad de LinkedIn que permite obtener tanto el tamaño de las audiencias como su valor económico y su desglose por segmentos de forma automatizada, a pesar de no haber contado con una API pública que facilitase la tarea.

Se ha comprobado la escalabilidad de la herramienta y su capacidad para ser usada en obtenciones masivas de datos que permitan realizar estudios de investigación. En particular, se ha estudiado la brecha de género laboral en países de todo el mundo, y se ha observado que la mayoría de ellos tienen una ligera inclinación estadística hacia la predominancia de hombres, si bien esta inclinación es mucho más acusada en países subdesarrollados. Así mismo, se ha observado que existen diferentes grupos de países en función de la distribución del indicador de brecha de género por las distintas áreas profesionales, agrupándose por un lado los países occidentales más desarrollados, y por otros países de menor desarrollo y diferente cultura como son India, China y Turquía. Por último, la obtención de las categorías que presentan una brecha de género más acusada, tanto hacia los hombres como hacia las mujeres, permite analizar el problema desde una perspectiva particular adaptada a la realidad de cada país, y fomentar la igualdad centrándose en las áreas que más lo necesitan.

Este trabajo ha sido desarrollado en el contexto de un contrato como ayudante de investigación en el departamento de Ingeniería Telemática de la Universidad Carlos III de Madrid, donde se continuará con el desarrollo de herramientas avanzadas de medidas en Internet como la presentada en este trabajo para aportar una nueva fuente de datos al campo de las ciencias sociales y para favorecer una integración eficiente entre las áreas de conocimiento de ingeniería y ciencias sociales, avanzando hacia un estilo de investigación multidisciplinar que cada vez es más necesario y más demandado desde el ámbito científico.

La acción más inmediata por realizar es la integración de este módulo con el resto de los módulos software que componen la herramienta multiplataforma del proyecto PIMCity (DVTMP), haciendo las adaptaciones y modificaciones necesarias en función de la evolución de los requisitos del proyecto. La integración con los sistemas de publicidad de otras plataformas hará posible conocer y comparar las políticas de precios para diferentes audiencias, la penetración de distintos grupos en cada red social, y muchos otros datos que permitirán conocer mucho mejor la dinámica de los sistemas de publicidad online.

Además, dado el potencial de esta herramienta para la realización de estudios académicos relacionados con el valor de los datos y la dinámica del mercado laboral, se pretende seguir realizando estudios como el presentado en el trabajo, no solo analizando la brecha de género, sino otros fenómenos como la empleabilidad por edades; estudiar los costes de mostrar anuncios a diferentes profesionales en función de su área profesional, sus años de experiencia, el tamaño de las empresas donde trabajan y otras variables. Estos son solo algunos ejemplos de los problemas que pueden ser analizados con la información disponible a través de la herramienta desarrollada.

Los estudios realizados se extenderán poniendo el foco en que la resolución de los datos sea regional en lugar de nacional, para tener una imagen mucho más precisa y diferenciada de los fenómenos estudiados y poder así determinar mejor sus causas, llegar a conclusiones consistentes y alcanzar un nivel de análisis que permita aplicar soluciones específicas a los problemas encontrados. Así mismo, la aplicación de técnicas de agrupamiento jerárquico permitirá conocer qué países o áreas profesionales responden a comportamientos más parecidos, ya sea en cuanto a su brecha de género, en cuanto al coste de mostrar los anuncios a esos grupos, o en cuanto a otras características analizadas.

Otra forma de extender el análisis será el estudio de la evolución de los datos en el tiempo, comparando los mismos datos con periodicidad anual para observar los cambios, las tendencias y la evolución global de los fenómenos que se analizan, obteniendo así un método para evaluar los resultados de las posibles medidas correctivas en el caso de que estas se apliquen.

En resumen, la proyección en el futuro de esta herramienta y otras derivadas de ella es muy prometedora y se pretende aprovechar la experiencia adquirida durante la realización del trabajo para expandir las posibilidades de estudio del mercado laboral, integrando así los conocimientos del mundo de la ingeniería en el campo de las ciencias sociales.

ANEXO A: EXTENDED ABSTRACT IN ENGLISH

I. INTRODUCTION AND OBJECTIVES

Online advertising systems are very popular nowadays in the business and sales world. While browsing the Internet, is virtually impossible not to come across advertisements. For instance, every time you make a search at Google, the first results are related or even are just what you were looking for, but they are ads. This means that the owners of the product or service have paid Google to show you these ads, as you are a person who may be interested in buying the product.

The platforms that often offer this kind of advertising service are social networks. Facebook, LinkedIn, Twitter, or Instagram are plenty of ads pretending to be regular publications, as images, videos, messages, or other forms. This makes users more interested in the advertised product than in the case of regular advertising. If we combine this with the fact that all these ads are personalized based on their personal data, makes this system one of the most efficient ways to reach potential clients.

In fact, the platforms that offer online advertising service generally have interfaces that allow the advertiser to select the kind of users they want to see their ads, which may include age, gender, interests, and others, depending on the kind of social network. For example, the LinkedIn Campaign Manager, which is the platform to manage ads campaigns at LinkedIn, offers a lot of segmentation options related to the job of the user, his professional skills, the characteristics of the company where the user works, and many other. When the advertiser selects a combination of characteristics that define his targeted audience, the platform gives the size of the audience (how much users match the criteria), an estimation of the expected cost of the campaign and performance in terms of visualizations and clicks, as well as a breakdown of the different profiles that comprise the audience regarding the selected criteria, for example, years of experience.

The main goal of this work is to automate the process of retrieving this information for a given audience (defined through a logic combination of characteristics) implementing a software module capable of repeating this process in a massive way. This would allow two different sub-objectives to be met. In the first place, the module would implement the functionality of the LinkedIn platform in a multiplatform tool designed to retrieve this kind of data from various add platforms, in the context of the European H2020 project

PIMCity, which aims to ensure that citizens, companies and organizations are informed and can make an ethical and respectful use of personal data with the building of the Data Valuation Tool from the Market Perspective (DVTMP), a tool with which data market players can know the updated value of data through the price of advertising based on this data. The development of this tool seeks to promote the transparency of online advertising systems. Secondly, we will design and conduct a case of use of the developed module to analyse a current phenomenon in the context of the labour market, the gender gap in different professions and areas of work, worldwide by country. This analysis will be performed using data collected with the developed software, specifically, the number of men and women LinkedIn users that work on each profession in each country.

II. REQUERIMENTS AND IMPLEMENTATION

To understand the functionality of the developed tool, first we describe the options that the LinkedIn Campaign Manager offers when the advertiser designs a campaign. There are different campaign goals to be selected, such as brand awareness, website visits, engagement, video views, lead generation, website conversions and job applicants. Also, there are different kinds of ads: single image, carousel, video, text, conversation, and others. But what is more important, it offers different options to define the target audience: location, gender, age; the industry, name, size, or growth rate of the company where the users work; their education or institutions where they have studied, their job experience (seniorities, titles, skills…) and their interests. These characteristics can be combined with logic operators, for example to reach users living in Spain and that know JavaScript or C++ programming languages, or that know both JavaScript and C++. This kind of logical combination is the definition of the audience that the module will receive as input, and the output will be the size of the public, the estimations of cost and performance expressed in the selected type of pricing (cost per click or cost per thousand impressions) and the breakdown of the audience in groups depending on the selected criteria (job functions, seniorities, years of experience, staff count, industries, or interests).

To develop the software module, we choose the JavaScript programming language and the Node execution engine for its high performance and simplicity in building network-intensive applications and establishing concurrent behaviours. LinkedIn has an API to interact with his Campaign Manager, but the access is restricted (before using it, your

application and his purposes must be approved by LinkedIn), and the functionality is limited. Because of this, the approach to develop the tool is based on an automated browser that is controlled with code. In this browser, the code runs a navigations sequence to log-in and navigate through the interface and get all the information needed to make custom requests to the server identical to those sent by the browser. To achieve this, we implement a function that obtains the account id and the csrf-token, which are the elements that the server needs to be in the requests to reply. The csrf-token is a random number sent by the server that prevents an attacker to impersonate the user to the server. We are not attackers, but we want to perform custom and automated requests, so this function logs-in and navigates the interface of the Campaign Manager with an automated sequence of clicks, investigates the requests sent, among which some will eventually contain the token that was sent by the server and that the client sends in the headers to prove its identity, and the account id.

This function allows us to build the next functionality, that is another function that, given an array of elements (locations, interests, skills, etc.), ask the server the URN (Uniform Resource Name) of each one and saves the answer in a register, along with the full name of the item that may be different to the name searched, and other information that will be used by other functions of the module. This function performs a search, so it returns several elements for each item searched, but we keep the information of the first element returned. To implement this function, we analyse the format of the requests sent when we use the platform manually (the URL and the headers, since is a GET request), and replicate it replacing the part corresponding to the name of the item searched by the item or items received. The register is a JSON file that contains one key for each item and the value is another object that contains the URN, the full name, and other data.

Once we have the items of the logical combination in the register, we build three more functions that implement the following functionalities: the obtention of the audience's size, the obtention of the forecasting (estimation of performance and cost), and the obtention of the audience breakdown into different groups.

The first function implements the obtention of the audience's size. To do that, as we did to implement the function that obtains the URNs, we analyse the corresponding request sent when using the Campaign Manager manually. In this case, the most important element is the form, as is a POST request and the logical combination of elements is expressed inside by a structure of AND and OR operators. Through this we infer the rules

of the structure to express any logical combination, and we implement an algorithm that builds the correct structure depending on the combination given, and for each element introduces the name, URN, and other needed data in the correct place. Once this is done, the function sends the request and, when it receives the reply, returns the audience count and saves the whole JSON received in a file (counts.json).

The function that obtains the forecasting works in a similar way. The structure of the logical combination is the same, but it has other parameters in the form that depend on the optimization goal and the type of ad, but we fix all the values except the one that selects the cost unit (CPC or CPM), that is given to the function as a parameter. Again, this function saves the response in a file, in this case in forecasting.json.

The function that retrieves the audience's breakdown is slightly different, as the form is not an encoded text, but a JSON object. Because of this, instead of reusing the algorithm to build the form, we have adapted it to the new format, but the logic behind it is very similar. In this case the response is written in a file called insights.json

The implementation is not so easy, we carried out a process to test the tool and fix bugs and errors, especially in the algorithms that build the logical combination in the functions, because there are many different cases, for example, in the number of elements combined in the same OR operator.

Once the tool was tested and its correct operation was checked, we started to test the scalability of the tool and to adapt it to a massive usage. To do that, we modified the code to save the state of the execution in a file in case of any nontreated error. This state is, basically, the iterators that traverse the input lists of items to be combined. Also, the code is modified to read this list (in CSV or TSV format) and combine the elements instead of reading a single file where the elements are already combined. This allows us to, for example, given a list of countries, a list of companies and a list of skills, retrieve the information related to all the possible combinations of one country, one company and one skill, that represent the audience living in that country, working in that company, and having that skill at once. This are a lot of possible combinations, and the time that it takes to retrieve all the information can be even several days, so we keep the module running in a remote machine. This remote machine is suited for the purpose with a tool called *tmux*, that allows the module to keep running even when the remote connection is closed.

With this, we run the module with a high number of combinations and verify that is possible to retrieve the information for all of them.

III. CASE OF USE AND LEGAL IMPLICATIONS

At this point, we started to work in the usage case to show the potential of the developed tool: obtain the number of men and women users of LinkedIn that work in certain job of a detailed, classified by area list of jobs. The list is made with data form O*NET [39], and it contains 609 occupations classified in 23 categories. We retrieve the audiences of men and woman that have each occupation of the list in each country (from a list of more than 200), starting with the countries that have more LinkedIn users, but we find that, after the 20 first countries, in many cases the audiences are below 300 users, and in these cases, LinkedIn does not specify the audience count, and this leads to a value of zero in the output table. Because of this, except for the 20 first countries, the resulting table is very sparse, then is not valid for statistical analysis.

To solve this problem, we retrieve the information of the job functions defined by LinkedIn, that include engineering, sales, legal, and others. Although this is not a hierarchical classification, it will be more general and lead to consistent data for more countries. With this list we get a dataset with data for almost all countries in the world, which allows us to compute the gender divide indicator, which is the quotient between the number of men and the sum of men and women, for each job function in each country, and represent the median of this indicator over all job functions for each country in a choropleth map, showed in section 5 of this document.

With the first dataset, we run a clustering algorithm for the first 17 countries, taking one vector per country that contains 23 indicators, one per category, being the median of all job's indicators of that category. The result is that developed countries fall in the same group, while others (India, China, and Turkey) fall in other group, meaning that their gender divide distribution across professional areas is more similar. This may be due to cultural and developmental reasons.

Finally, with the dataset that includes all countries, we provide a table that contains, for each country, its median of the gender divide indicator and the three categories that are further away, from the gender equality (0.5), both over and under it, along with the indicator value for each category. This table can be found can be found in this annex.

With this analysis, we have proved the potential and utility of the developed module and showed some conclusions on gender gap in employment worldwide.

Then we analyse the legal implications of the processing of personal data in Europe, showing the main points of the recently adopted regulation governing this matter. The conclusion is that the tool developed in this work complies with the requirements of the regulation, as the data obtained is not considered personal data, but aggregated data, which is not subject to such limitations.

IV. CONCLUSIONS AND FURTHER WORK

In this work we have successfully developed a tool for LinkedIn's advertising platform that allows us to obtain both the size of audiences and their economic value and their breakdown by segments in an automated way, despite not having a public API to facilitate the task. The scalability of the tool and its capacity to be used in massive data collection to carry out research studies have been tested. In particular, the labour gender gap has been studied in countries around the world, and it has been observed that most of them have a slight statistical bias towards male predominance, although this bias is much more pronounced in underdeveloped countries. It has also been observed that there are different groups of countries according to the distribution of the gender gap indicator by the different professional areas, with the most developed countries on the one hand, and other less developed countries with a different culture, such as India, China, and Turkey, on the other. Finally, we obtain the categories with the widest gender gap, which can help to analyse the problem from a particular perspective adapted to the reality of each country.

This work has been developed in the context of a contract as a research assistant in the Department of Telematics Engineering at the University Carlos III of Madrid, where the development of advanced Internet measurement tools such as the one presented in this work will continue to provide a new source of data to the field of social sciences.

The most important action to be carried out in the future is the integration of this module with the rest of the software modules that make up the multiplatform tool of the PIMCity project (DVTMP), making the necessary adaptations and modifications according to the evolution of the project requirements. The integration with the advertising systems of other platforms will allow to know and compare the pricing policies for different

audiences, the penetration of different groups in each social network, and many other data that will allow a much better understanding of the dynamics of online advertising systems.

Furthermore, given the potential of this tool for academic studies related to the value of data and the dynamics of the labour market, we want to continue carrying out studies such as the one presented in the paper, not only analysing the gender gap, but also other phenomena such as employability by age, focusing on regional rather than national data resolution, in order to have a much more accurate and differentiated picture of the phenomena studied and thus be able to better determine their causes, reach consistent conclusions and reach a level of analysis that allows specific solutions to the problems encountered to be applied.

These studies will be extended by changing the resolution from country level to regional or local level, which will give a more accurate view of the problems and allows to better determine the situation of each zone and to design the appropriate corrective actions.

Another way of extending the analysis will be to study the evolution of the data over time, comparing the same data on an annual basis to observe changes, trends and the overall evolution of the phenomena being analysed, thus obtaining a method for evaluating the results of possible corrective measures if these are applied.

In summary, the future projection of this tool and others derived from it is very promising and it is intended to take advantage of the experience acquired during the work to expand the possibilities of studying the labour market, thus integrating knowledge from the world of engineering into the field of social sciences.

ANEXO B: CLASIFICACIÓN DE PROFESIONES COMPLETA

Category	Occupation
architecture_and_engineering	Architects
architecture_and_engineering	Mechanical Engineers
architecture_and_engineering	Electrical Engineers
architecture_and_engineering	Manufacturing Engineers
architecture_and_engineering	Civil Engineers
architecture_and_engineering	Industrial Engineers
architecture_and_engineering	Engineering Technicians
architecture_and_engineering	Environmental Engineers
architecture_and_engineering	Drafters
architecture_and_engineering	Landscape Architects
architecture_and_engineering	Validation Engineers
architecture_and_engineering	Chemical Engineers
architecture_and_engineering	Materials Engineers
architecture_and_engineering	Aerospace Engineers
architecture_and_engineering	Transportation Engineers
architecture_and_engineering	Marine Engineers
architecture_and_engineering	Biomedical Engineers
architecture_and_engineering	Petroleum Engineers
architecture_and_engineering	Naval Architects
architecture_and_engineering	Nuclear Engineers
architecture_and_engineering	Energy Engineers
architecture_and_engineering	Cartographers
architecture_and_engineering	Mining Engineers
architecture_and_engineering	Electronics Engineers
architecture_and_engineering	Human Factors Engineers
architecture_and_engineering	Surveying Technicians
architecture_and_engineering	Water Engineers
architecture_and_engineering	Engineering Technologists
architecture_and_engineering	Ergonomists
architecture_and_engineering	Calibration Technicians
architecture_and_engineering	Automotive Engineers
architecture_and_engineering	Fire Protection Engineers
architecture_and_engineering	Mechanical Drafters
architecture_and_engineering	Architectural Drafters
architecture_and_engineering	Robotics Engineers
architecture_and_engineering	Civil Engineering Technicians
architecture_and_engineering	Electronic Engineering Technicians
architecture_and_engineering	Electrical Engineering Technicians
architecture_and_engineering	Electro-Mechanical Technicians
architecture_and_engineering	Computer Hardware Engineers
architecture_and_engineering	Electrical Drafters
architecture_and_engineering	Geological Engineers
architecture_and_engineering	Agricultural Engineers

architecture_and_engineering	Mechanical Engineering Technicians
architecture_and_engineering	Mechatronics Engineers
architecture_and_engineering	Bioengineers
architecture_and_engineering	Civil Engineering Technologists
architecture_and_engineering	Photogrammetrists
architecture_and_engineering	Non-Destructive Testing Specialists
architecture_and_engineering	Robotics Technicians
architecture_and_engineering	Solar Engineers
arts_design_entertainment_sports_and_media	Designers
arts_design_entertainment_sports_and_media	Editors
arts_design_entertainment_sports_and_media	Graphic Designers
arts_design_entertainment_sports_and_media	Art Directors
arts_design_entertainment_sports_and_media	Writers
arts_design_entertainment_sports_and_media	Photographers
arts_design_entertainment_sports_and_media	Artists
arts_design_entertainment_sports_and_media	Technical Writers
arts_design_entertainment_sports_and_media	Translators
arts_design_entertainment_sports_and_media	Public Relations Specialists
arts_design_entertainment_sports_and_media	Interior Designers
arts_design_entertainment_sports_and_media	Actors
arts_design_entertainment_sports_and_media	Animators
arts_design_entertainment_sports_and_media	Interpreters
arts_design_entertainment_sports_and_media	Musicians
arts_design_entertainment_sports_and_media	Media Planners
arts_design_entertainment_sports_and_media	Media Directors
arts_design_entertainment_sports_and_media	Singers
arts_design_entertainment_sports_and_media	Music Directors
arts_design_entertainment_sports_and_media	Video Editors
arts_design_entertainment_sports_and_media	Camera Operators
arts_design_entertainment_sports_and_media	News Reporters
arts_design_entertainment_sports_and_media	Fashion Designers
arts_design_entertainment_sports_and_media	Entertainers
arts_design_entertainment_sports_and_media	Choreographers
arts_design_entertainment_sports_and_media	Dancers
arts_design_entertainment_sports_and_media	Court Reporters
arts_design_entertainment_sports_and_media	Commercial Designers
arts_design_entertainment_sports_and_media	Film Editors
arts_design_entertainment_sports_and_media	Referees
arts_design_entertainment_sports_and_media	Athletes
arts_design_entertainment_sports_and_media	Creative Writers
arts_design_entertainment_sports_and_media	Talent Directors
arts_design_entertainment_sports_and_media	Audio Technicians
arts_design_entertainment_sports_and_media	Poets
arts_design_entertainment_sports_and_media	Set Designers
arts_design_entertainment_sports_and_media	Sound Technicians
arts_design_entertainment_sports_and_media	Fine Artists
arts_design_entertainment_sports_and_media	Floral Designers
arts_design_entertainment_sports_and_media	Lighting Technicians

arts_design_entertainment_sports_and_media	Video Technicians
arts_design_entertainment_sports_and_media	Music Composers
arts_design_entertainment_sports_and_media	Broadcast Technicians
arts_design_entertainment_sports_and_media	Umpires
arts_design_entertainment_sports_and_media	Radio Disc Jockeys
building_and_grounds_cleaning_and_maintenance	Janitors
building_and_grounds_cleaning_and_maintenance	Cleaners
building_and_grounds_cleaning_and_maintenance	Housekeepers
building_and_grounds_cleaning_and_maintenance	Exterminators
building_and_grounds_cleaning_and_maintenance	Tree Trimmers
business_and_financial_operations	Accountants
business_and_financial_operations	Business Manager
business_and_financial_operations	Human Resources Specialists
business_and_financial_operations	Financial Analysts
business_and_financial_operations	Buyers
business_and_financial_operations	Auditors
business_and_financial_operations	Loan Officers
business_and_financial_operations	Credit Analysts
business_and_financial_operations	Project Management Specialists
business_and_financial_operations	Compliance Officers
business_and_financial_operations	Benefits Specialists
business_and_financial_operations	Management Analysts
business_and_financial_operations	Investment Analysts
business_and_financial_operations	Fundraisers
business_and_financial_operations	Development Specialists
business_and_financial_operations	Event Planners
business_and_financial_operations	Appraisers
business_and_financial_operations	Compensation Specialists
business_and_financial_operations	Budget Analysts
business_and_financial_operations	Regulatory Affairs Specialists
business_and_financial_operations	Logistics Analysts
business_and_financial_operations	Claims Adjusters
business_and_financial_operations	Business Operations Specialists
business_and_financial_operations	Tax Preparers
business_and_financial_operations	Claims Examiners
business_and_financial_operations	Meeting Planners
business_and_financial_operations	Logistics Engineers
business_and_financial_operations	Labor Relations Specialists
business_and_financial_operations	Real Estate Appraisers
business_and_financial_operations	Fraud Analysts
business_and_financial_operations	Fraud Investigators
business_and_financial_operations	Logisticians
business_and_financial_operations	Cost Estimators
business_and_financial_operations	Financial Specialists
business_and_financial_operations	Purchasing Agents
business_and_financial_operations	Insurance Underwriters
business_and_financial_operations	Revenue Agents
business_and_financial_operations	Retail Buyers

business_and_financial_operations	Tax Examiners
business_and_financial_operations	Credit Counselors
business_and_financial_operations	Personal Financial Advisors
business_and_financial_operations	Financial Examiners
business_and_financial_operations	Tax Collectors
business_and_financial_operations	Business Continuity Planners
business_and_financial_operations	Sustainability Specialists
community_and_social_service	Counselors
community_and_social_service	Social Workers
community_and_social_service	Chaplains
community_and_social_service	Career Counselors
community_and_social_service	Guidance Counselors
community_and_social_service	Social Service Specialists
community_and_social_service	Probation Officers
community_and_social_service	Marriage Therapists
community_and_social_service	Mental Health Counselors
community_and_social_service	Educational Advisors
community_and_social_service	Family Therapists
community_and_social_service	Substance Abuse Counselors
community_and_social_service	Mental Health Workers
community_and_social_service	Rehabilitation Counselors
community_and_social_service	Health Education Specialists
computer_and_mathematical	Software Developers
computer_and_mathematical	Computer Systems Analysts
computer_and_mathematical	Web Developers
computer_and_mathematical	Database Administrators
computer_and_mathematical	Network Administrators
computer_and_mathematical	Web Designers
computer_and_mathematical	Information Technology Project Managers
computer_and_mathematical	Telecommunications Engineers
computer_and_mathematical	Computer Specialists
computer_and_mathematical	Statisticians
computer_and_mathematical	Computer Network Support Specialists
computer_and_mathematical	Interface Designers
computer_and_mathematical	Web Administrators
computer_and_mathematical	Business Intelligence Analysts
computer_and_mathematical	Actuaries
computer_and_mathematical	Database Architects
computer_and_mathematical	Information Security Analysts
computer_and_mathematical	Geographic Information Systems Technicians
computer_and_mathematical	Document Management Specialists
computer_and_mathematical	Software Quality Assurance Analysts
computer_and_mathematical	Clinical Data Managers
computer_and_mathematical	Computer Programmers
computer_and_mathematical	Operations Research Analysts
computer_and_mathematical	Information Security Engineers
computer_and_mathematical	Mathematicians
computer_and_mathematical	Computer System Administrators

computer_and_mathematical	Software Quality Assurance Testers
computer_and_mathematical	Biostatisticians
computer_and_mathematical	Data Scientists
computer_and_mathematical	Data Warehousing Specialists
computer_and_mathematical	Blockchain Engineers
computer_and_mathematical	Penetration Testers
construction_and_extraction	Electricians
construction_and_extraction	Carpenters
construction_and_extraction	Construction Laborers
construction_and_extraction	Plumbers
construction_and_extraction	Construction Inspectors
construction_and_extraction	Building Inspectors
construction_and_extraction	Brickmasons
construction_and_extraction	Pipefitters
construction_and_extraction	Tapers
construction_and_extraction	Roofers
construction_and_extraction	Miners
construction_and_extraction	Energy Auditors
construction_and_extraction	Sheet Metal Workers
construction_and_extraction	Quarry Manager
construction_and_extraction	House Painters
construction_and_extraction	Elevator Mechanics
construction_and_extraction	Concrete Finishers
construction_and_extraction	Drywall Finishers
construction_and_extraction	Pipelayers
construction_and_extraction	Derrickhands
construction_and_extraction	Roustabouts
construction_and_extraction	Boilermakers
construction_and_extraction	Iron Workers
construction_and_extraction	Plasterers
construction_and_extraction	Floor Layers
construction_and_extraction	Glaziers
construction_and_extraction	Paving Operators
educational_instruction_and_library	Teaching Assistants
educational_instruction_and_library	Librarians
educational_instruction_and_library	Tutors
educational_instruction_and_library	Substitute Teachers
educational_instruction_and_library	Education Teachers
educational_instruction_and_library	Elementary School Teachers
educational_instruction_and_library	Art Teachers
educational_instruction_and_library	Curators
educational_instruction_and_library	Music Teachers
educational_instruction_and_library	Archivists
educational_instruction_and_library	English Language
educational_instruction_and_library	Preschool Teachers
educational_instruction_and_library	Kindergarten Teachers
educational_instruction_and_library	English as a Second Language Instructors
educational_instruction_and_library	Social Sciences Teachers

educational_instruction_and_library	History Teachers
educational_instruction_and_library	Business Teachers
educational_instruction_and_library	Middle School Teachers
educational_instruction_and_library	Library Technicians
educational_instruction_and_library	Chemistry Teachers
educational_instruction_and_library	Physics Teachers
educational_instruction_and_library	Drama Teachers
educational_instruction_and_library	Computer Science Teachers
educational_instruction_and_library	Physical Education Specialists
educational_instruction_and_library	Communications Teachers
educational_instruction_and_library	Economics Teachers
educational_instruction_and_library	Secondary School Teachers
educational_instruction_and_library	Foreign Language Teachers
educational_instruction_and_library	Geography Teachers
educational_instruction_and_library	Instructional Coordinators
educational_instruction_and_library	Psychology Teachers
educational_instruction_and_library	Educational Instruction
educational_instruction_and_library	Museum Technicians
farming_fishing_and_forestry	Hunters
farming_fishing_and_forestry	Graders
farming_fishing_and_forestry	Sorters
farming_fishing_and_forestry	Animal Breeders
farming_fishing_and_forestry	Fishers
food_preparation_and_serving_related	Cooks
food_preparation_and_serving_related	Chefs
food_preparation_and_serving_related	Bartenders
food_preparation_and_serving_related	Hosts
food_preparation_and_serving_related	Baristas
food_preparation_and_serving_related	Food Servers
food_preparation_and_serving_related	Dishwashers
food_preparation_and_serving_related	Waiters
food_preparation_and_serving_related	Head Cooks
food_preparation_and_serving_related	Counter Workers
food_preparation_and_serving_related	Dining Room Attendants
food_preparation_and_serving_related	Food Preparation Workers
Healthcare_Practitioners_and_Technical	Registered Nurses
Healthcare_Practitioners_and_Technical	Pharmacists
Healthcare_Practitioners_and_Technical	Therapists
Healthcare_Practitioners_and_Technical	Physical Therapists
Healthcare_Practitioners_and_Technical	Dentists
Healthcare_Practitioners_and_Technical	Occupational Therapists
Healthcare_Practitioners_and_Technical	Pathologists
Healthcare_Practitioners_and_Technical	Pharmacy Technicians
Healthcare_Practitioners_and_Technical	Paramedics
Healthcare_Practitioners_and_Technical	Emergency Medical Technicians
Healthcare_Practitioners_and_Technical	Nurse Practitioners
Healthcare_Practitioners_and_Technical	Licensed Practical Nurses
Healthcare_Practitioners_and_Technical	Physician Assistants

Healthcare_Practitioners_and_Technical	Surgeons
Healthcare_Practitioners_and_Technical	Hospitalists
Healthcare_Practitioners_and_Technical	Chiropractors
Healthcare_Practitioners_and_Technical	Nutritionists
Healthcare_Practitioners_and_Technical	Optometrists
Healthcare_Practitioners_and_Technical	Psychiatrists
Healthcare_Practitioners_and_Technical	Dental Hygienists
Healthcare_Practitioners_and_Technical	Dietitians
Healthcare_Practitioners_and_Technical	Respiratory Therapists
Healthcare_Practitioners_and_Technical	Opticians
Healthcare_Practitioners_and_Technical	Radiologists
Healthcare_Practitioners_and_Technical	Nurse Anesthetists
Healthcare_Practitioners_and_Technical	Pediatricians
Healthcare_Practitioners_and_Technical	Audiologists
Healthcare_Practitioners_and_Technical	Veterinary Technicians
Healthcare_Practitioners_and_Technical	Athletic Trainers
Healthcare_Practitioners_and_Technical	Anesthesiologists
Healthcare_Practitioners_and_Technical	Medical Records Specialists
Healthcare_Practitioners_and_Technical	Art Therapists
Healthcare_Practitioners_and_Technical	Diagnostic Medical Sonographers
Healthcare_Practitioners_and_Technical	Cardiologists
Healthcare_Practitioners_and_Technical	Exercise Physiologists
Healthcare_Practitioners_and_Technical	Nuclear Medicine Technologists
Healthcare_Practitioners_and_Technical	Ophthalmologists
Healthcare_Practitioners_and_Technical	Family Medicine Physicians
Healthcare_Practitioners_and_Technical	Medical Registrars
Healthcare_Practitioners_and_Technical	Patient Representatives
Healthcare_Practitioners_and_Technical	Orthodontists
Healthcare_Practitioners_and_Technical	Neurologists
Healthcare_Practitioners_and_Technical	Radiation Therapists
Healthcare_Practitioners_and_Technical	Veterinarians
Healthcare_Practitioners_and_Technical	Clinical Laboratory Technicians
Healthcare_Practitioners_and_Technical	Music Therapists
Healthcare_Practitioners_and_Technical	Surgical Assistants
Healthcare_Practitioners_and_Technical	Dermatologists
Healthcare_Practitioners_and_Technical	Podiatrists
Healthcare_Practitioners_and_Technical	Surgical Technologists
Healthcare_Practitioners_and_Technical	Acupuncturists
Healthcare_Practitioners_and_Technical	Genetic Counselors
Healthcare_Practitioners_and_Technical	Recreational Therapists
Healthcare_Practitioners_and_Technical	Ophthalmic Medical Technicians
Healthcare_Practitioners_and_Technical	Urologists
Healthcare_Practitioners_and_Technical	Critical Care Nurses
Healthcare_Practitioners_and_Technical	Cytotechnologists
Healthcare_Practitioners_and_Technical	Emergency Medicine Physicians
Healthcare_Practitioners_and_Technical	Psychiatric Technicians
Healthcare_Practitioners_and_Technical	Oral Surgeons
Healthcare_Practitioners_and_Technical	Orthotists

Healthcare_Practitioners_and_Technical	Gynecologists
Healthcare_Practitioners_and_Technical	Medical Dosimetrists
Healthcare_Practitioners_and_Technical	General Internal Medicine Physicians
Healthcare_Practitioners_and_Technical	Cardiovascular Technicians
Healthcare_Practitioners_and_Technical	Health Technicians
Healthcare_Practitioners_and_Technical	Obstetricians
Healthcare_Practitioners_and_Technical	Clinical Nurse Specialists
Healthcare_Practitioners_and_Technical	Acute Care Nurses
Healthcare_Practitioners_and_Technical	Healthcare Practitioners
Healthcare_Practitioners_and_Technical	Hearing Aid Specialists
Healthcare_Practitioners_and_Technical	Orthoptists
Healthcare_Practitioners_and_Technical	Histology Technicians
Healthcare_Practitioners_and_Technical	Naturopathic Physicians
Healthcare_Practitioners_and_Technical	Allergists
Healthcare_Practitioners_and_Technical	Neurodiagnostic Technologists
Healthcare_Practitioners_and_Technical	Histotechnologists
Healthcare_Practitioners_and_Technical	Magnetic Resonance Imaging Technologists
Healthcare_Practitioners_and_Technical	Prosthodontists
Healthcare_Practitioners_and_Technical	Advanced Practice Psychiatric Nurses
Healthcare_Practitioners_and_Technical	Anesthesiologist Assistants
Healthcare_Practitioners_and_Technical	Cytogenetic Technologists
Healthcare_Practitioners_and_Technical	Physical Medicine Physicians
Healthcare_Practitioners_and_Technical	Pediatric Surgeons
Healthcare_Practitioners_and_Technical	Prosthetists
Healthcare_Practitioners_and_Technical	Sports Medicine Physicians
Healthcare_Practitioners_and_Technical	Immunologists
Healthcare_Practitioners_and_Technical	Low Vision Therapists
Healthcare_Practitioners_and_Technical	Dietetic Technicians
healthcare_support	Massage Therapists
healthcare_support	Medical Assistants
healthcare_support	Nursing Assistants
healthcare_support	Medical Transcriptionists
healthcare_support	Dental Assistants
healthcare_support	Phlebotomists
healthcare_support	Physical Therapist Assistants
healthcare_support	Home Health Aides
healthcare_support	Veterinary Assistants
healthcare_support	Pharmacy Aides
healthcare_support	Physical Therapist Aides
healthcare_support	Healthcare Support Workers
healthcare_support	Personal Care Aides
healthcare_support	Occupational Therapy Aides
healthcare_support	Occupational Therapy Assistants
healthcare_support	Speech-Language Pathology Assistants
healthcare_support	Endoscopy Technicians
installation_maintenance_and_repair	Avionics Technicians
installation_maintenance_and_repair	Automotive Body
installation_maintenance_and_repair	Riggers

installation_maintenance_and_repair	Aircraft Technicians
installation_maintenance_and_repair	Maintenance Workers
installation_maintenance_and_repair	Vending Machine Servicers
installation_maintenance_and_repair	Heating Mechanics
installation_maintenance_and_repair	Commercial Divers
installation_maintenance_and_repair	Motorcycle Mechanics
installation_maintenance_and_repair	Wind Turbine Service Technicians
installation_maintenance_and_repair	Millwrights
installation_maintenance_and_repair	Mobile Heavy Equipment Mechanics
legal	Lawyers
legal	Legal Assistants
legal	Judges
legal	Mediators
legal	Judicial Law Clerks
legal	Arbitrators
legal	Title Examiners
legal	Administrative Law Judges
legal	Hearing Officers
life_physical_and_social_science	Psychologists
life_physical_and_social_science	Economists
life_physical_and_social_science	Chemists
life_physical_and_social_science	Environmental Scientists
life_physical_and_social_science	Biological Scientists
life_physical_and_social_science	Clinical Psychologists
life_physical_and_social_science	Microbiologists
life_physical_and_social_science	School Psychologists
life_physical_and_social_science	Quality Control Analysts
life_physical_and_social_science	Occupational Health and Safety Specialists
life_physical_and_social_science	Epidemiologists
life_physical_and_social_science	Urban Planners
life_physical_and_social_science	Archeologists
life_physical_and_social_science	Historians
life_physical_and_social_science	Transportation Planners
life_physical_and_social_science	Foresters
life_physical_and_social_science	Biochemists
life_physical_and_social_science	Environmental Technicians
life_physical_and_social_science	Hydrologists
life_physical_and_social_science	Regional Planners
life_physical_and_social_science	Chemical Technicians
life_physical_and_social_science	Wildlife Biologists
life_physical_and_social_science	Neuropsychologists
life_physical_and_social_science	Food Scientists
life_physical_and_social_science	Molecular Biologists
life_physical_and_social_science	Materials Scientists
life_physical_and_social_science	Physical Scientists
life_physical_and_social_science	Anthropologists
life_physical_and_social_science	Geographers
life_physical_and_social_science	Sociologists

life_physical_and_social_science	Geneticists
life_physical_and_social_science	Geoscientists
life_physical_and_social_science	Bioinformatics Scientists
life_physical_and_social_science	Biological Technicians
life_physical_and_social_science	Social Scientists
life_physical_and_social_science	Plant Scientists
life_physical_and_social_science	Clinical Neuropsychologists
life_physical_and_social_science	Geological Technicians
life_physical_and_social_science	Medical Scientists
life_physical_and_social_science	Hydrologic Technicians
life_physical_and_social_science	Political Scientists
life_physical_and_social_science	Occupational Health and Safety Technicians
life_physical_and_social_science	Astronomers
management	Managers
management	Chief Executives
management	Sales Managers
management	Marketing Managers
management	Operations Managers
management	Human Resources Managers
management	Controllers
management	Engineering Managers
management	Purchasing Managers
management	Training Managers
management	Treasurers
management	Facilities Managers
management	Public Relations Managers
management	Supply Chain Managers
management	Construction Managers
management	Security Managers
management	Compliance Managers
management	Distribution Managers
management	Advertising Managers
management	Financial Managers
management	Promotions Managers
management	Information Systems Managers
management	Benefits Managers
management	Real Estate Managers
management	Regulatory Affairs Managers
management	Transportation Managers
management	Compensation Managers
management	Lodging Managers
management	Clinical Research Coordinators
management	Storage Managers
management	Loss Prevention Managers
management	Entertainment Managers
management	Administrative Services Managers
management	Food Service Managers
management	Fundraising Managers

management	Spa Managers
management	Social Managers
management	Farm Managers
management	Recreation Managers
management	Health Services Managers
management	Fitness Coordinators
management	Education Administrators
management	Community Association Managers
management	Emergency Management Directors
management	Architectural Managers
office_and_administrative_support	Account
office_and_administrative_support	Customer Service Representatives
office_and_administrative_support	Checkers
office_and_administrative_support	Production Clerks
office_and_administrative_support	Tellers
office_and_administrative_support	Shipping
office_and_administrative_support	Executive Secretaries
office_and_administrative_support	Dispatchers
office_and_administrative_support	Executive Administrative Assistants
office_and_administrative_support	Posting Clerks
office_and_administrative_support	Proofreaders
office_and_administrative_support	Couriers
office_and_administrative_support	File Clerks
office_and_administrative_support	Data Entry Keyers
office_and_administrative_support	Measurers
office_and_administrative_support	Typists
office_and_administrative_support	Desktop Publishers
office_and_administrative_support	Telephone Operators
office_and_administrative_support	Payroll Clerks
office_and_administrative_support	Billing Clerks
office_and_administrative_support	Court Clerks
office_and_administrative_support	Word Processors
office_and_administrative_support	Credit Clerks
office_and_administrative_support	Procurement Clerks
office_and_administrative_support	Messengers
office_and_administrative_support	Switchboard Operators
office_and_administrative_support	Inventory Clerks
office_and_administrative_support	Receiving Clerks
office_and_administrative_support	Auditing Clerks
office_and_administrative_support	Freight Forwarders
office_and_administrative_support	Financial Clerks
office_and_administrative_support	Samplers
office_and_administrative_support	Cargo Agents
office_and_administrative_support	Information Clerks
office_and_administrative_support	Public Safety Telecommunicators
office_and_administrative_support	Statistical Assistants
office_and_administrative_support	Order Clerks
personal_care_and_service	Concierges

personal_care_and_service	Hairstylists
personal_care_and_service	Makeup Artists
personal_care_and_service	Tour Guides
personal_care_and_service	Hairdressers
personal_care_and_service	Cosmetologists
personal_care_and_service	Morticians
personal_care_and_service	Group Fitness Instructors
personal_care_and_service	Ushers
personal_care_and_service	Childcare Workers
personal_care_and_service	Barbers
personal_care_and_service	Manicurists
personal_care_and_service	Ticket Takers
personal_care_and_service	Animal Trainers
personal_care_and_service	Residential Advisors
personal_care_and_service	Escorts
personal_care_and_service	Embalmers
production	Apparel Workers
production	Inspectors
production	Testers
production	Machinists
production	Welders
production	Assemblers
production	Bakers
production	Production Workers
production	Fabricators
production	Plant Operators
production	Model Makers
production	Jewelers
production	Patternmakers
production	Cabinetmakers
production	Casters
production	Meat Cutters
production	Coating Workers
production	Power Plant Operators
production	Nuclear Power Reactor Operators
production	Dental Laboratory Technicians
production	Potters
production	Butchers
production	Engravers
production	Refinery Operators
production	Stationary Engineers
production	Molders
production	Upholsterers
production	Water Treatment Plant Operators
production	Solderers
production	Woodworkers
production	Sewers
production	Tool Sharpeners

protective_service	Firefighters
protective_service	Lifeguards
protective_service	Correctional Officers
protective_service	Security Guards
protective_service	Private Detectives
protective_service	Criminal Investigators
protective_service	Records Officers
protective_service	Fire Inspectors
protective_service	Bailiffs
protective_service	Game Wardens
protective_service	Detectives
protective_service	Ski Patrol
protective_service	Flaggers
protective_service	Crossing Guards
transportation_and_material_moving	Packagers
transportation_and_material_moving	Flight Attendants
transportation_and_material_moving	Motor Vehicle Operators
transportation_and_material_moving	Air Traffic Controllers
transportation_and_material_moving	Chauffeurs
transportation_and_material_moving	Stockers
transportation_and_material_moving	Packers
transportation_and_material_moving	School Bus Drivers
transportation_and_material_moving	Airline Pilots
transportation_and_material_moving	Sailors
transportation_and_material_moving	Commercial Pilots
transportation_and_material_moving	Recycling Coordinators
transportation_and_material_moving	Aviation Inspectors
transportation_and_material_moving	Parking Attendants
transportation_and_material_moving	Railroad Operators
transportation_and_material_moving	Locomotive Engineers
transportation_and_material_moving	Order Fillers
transportation_and_material_moving	Shuttle Drivers
transportation_and_material_moving	Dredge Operators
transportation_and_material_moving	Pump Operators
transportation_and_material_moving	Crane Operators
transportation_and_material_moving	Taxi Drivers
transportation_and_material_moving	Ambulance Drivers

ANEXO C: CATEGORÍAS CON MAYOR BRECHA DE GÉNERO EN CADA PAÍS

Country	GDM	Cat. 1	Val. 1	Cat. 2	Val. 2	Cat. 3	Val. 3
Afghanistan	0,8	Engineering	0,9	Information Technology	0,88	Program and Project Management	0,85
Albania	0,5	Military and Protective Services	0,73	Engineering	0,68	Accounting	0,33
Algeria	0,72	Operations	0,86	Military and Protective Services	0,85	Entrepreneurship	0,84
American Samoa	0,59	Business Development	0,66	Sales	0,6	Education	0,4
Andorra	0,57	Engineering	0,71	Administrative	0,31	Business Development	0,69
Angola	0,74	Information Technology	0,88	Operations	0,87	Arts and Design	0,83
Anguilla	0,62	Operations	0,62	Administrative		Arts and Design	
Antarctica		Arts and Design		Business Development		Education	
Antigua and Barbuda	0,51	Education	0,32	Business Development	0,63	Operations	0,58
Argentina	0,56	Engineering	0,79	Information Technology	0,72	Military and Protective Services	0,69
Armenia	0,42	Human Resources	0,19	Administrative	0,27	Engineering	0,71
Aruba	0,51	Business Development	0,67	Operations	0,66	Healthcare Services	0,37
Australia	0,54	Engineering	0,83	Administrative	0,22	Healthcare Services	0,3
Austria	0,6	Engineering	0,87	Information Technology	0,79	Business Development	0,77
Azerbaijan	0,62	Engineering	0,82	Operations	0,77	Information Technology	0,75
Bahrain	0,71	Engineering	0,87	Military and Protective Services	0,84	Operations	0,81
Bangladesh	0,84	Quality Assurance	0,9	Engineering	0,89	Real Estate	0,89
Barbados	0,46	Administrative	0,17	Healthcare Services	0,27	Support	0,27
Belarus	0,46	Accounting	0,08	Human Resources	0,18	Engineering	0,73
Belgium	0,59	Engineering	0,85	Information Technology	0,76	Business Development	0,73
Belize	0,47	Administrative	0,27	Operations	0,68	Business Development	0,66
Benin	0,72	Information Technology	0,83	Engineering	0,83	Legal	0,82
Bermuda	0,46	Healthcare Services	0,25	Education	0,3	Business Development	0,64
Bhutan	0,45	Administrative	0,34	Education	0,35	Healthcare Services	0,39
Bolivia	0,6	Engineering	0,83	Operations	0,75	Information Technology	0,74
Bosnia and Herzegovina	0,51	Military and Protective Services	0,74	Engineering	0,74	Operations	0,69

Botswana	0,59	Engineering	0,81	Business Development	0,71	Arts and Design	0,69
Brazil	0,56	Engineering	0,79	Information Technology	0,77	Military and Protective Services	0,76
British Virgin Islands	0,63	Business Development	0,63	Operations	0,63	Accounting	
Brunei	0,53	Engineering	0,79	Operations	0,7	Information Technology	0,68
Bulgaria	0,49	Engineering	0,77	Human Resources	0,25	Military and Protective Services	0,74
Burkina Faso	0,72	Engineering	0,84	Information Technology	0,83	Operations	0,82
Burundi	0,7	Operations	0,78	Education	0,76	Business Development	0,76
Cambodia	0,58	Business Development	0,66	Engineering	0,65	Consulting	0,63
Cameroon	0,67	Engineering	0,83	Information Technology	0,81	Operations	0,77
Canada	0,52	Engineering	0,8	Administrative	0,22	Healthcare Services	0,27
Cape Verde	0,55	Information Technology	0,75	Operations	0,74	Business Development	0,73
Cayman Islands	0,46	Support	0,32	Healthcare Services	0,32	Education	0,35
Central African Republic	0,58	Operations	0,8	Administrative	0,58	Education	0,57
Chad	0,77	Operations	0,87	Administrative	0,67	Accounting	
Chile	0,55	Engineering	0,78	Military and Protective Services	0,76	Information Technology	0,75
China	0,49	Human Resources	0,29	Administrative	0,32	Engineering	0,64
Colombia	0,52	Engineering	0,79	Military and Protective Services	0,74	Operations	0,71
Comoros		Accounting		Administrative		Business Development	
Congo (DRC)	0,77	Engineering	0,87	Information Technology	0,86	Operations	0,84
Costa Rica	0,57	Engineering	0,81	Military and Protective Services	0,74	Operations	0,7
Croatia	0,48	Engineering	0,79	Military and Protective Services	0,76	Accounting	0,25
Curacao	0,46	Administrative	0,21	Information Technology	0,67	Business Development	0,66
Cyprus	0,51	Engineering	0,81	Administrative	0,25	Military and Protective Services	0,7
Czechia	0,59	Engineering	0,87	Human Resources	0,22	Information Technology	0,76
Denmark	0,53	Engineering	0,84	Healthcare Services	0,23	Military and Protective Services	0,76
Djibouti	0,74	Operations	0,84	Education	0,74	Administrative	0,55
Dominica	0,59	Education	0,35	Business Development	0,61	Operations	0,59
Dominican Republic	0,53	Engineering	0,8	Information Technology	0,73	Administrative	0,29

Ecuador	0,57	Engineering	0,81	Military and Protective Services	0,79	Operations	0,77
Egypt	0,76	Military and Protective Services	0,87	Engineering	0,87	Operations	0,84
El Salvador	0,57	Engineering	0,79	Information Technology	0,75	Operations	0,72
Equatorial Guinea	0,65	Operations	0,86	Administrative	0,44	Accounting	
Estonia	0,45	Human Resources	0,17	Accounting	0,19	Engineering	0,78
Eswatini	0,54	Administrative	0,27	Operations	0,72	Information Technology	0,67
Ethiopia	0,69	Engineering	0,78	Operations	0,77	Military and Protective Services	0,75
Faroe Islands		Arts and Design		Business Development		Education	
Fiji	0,48	Engineering	0,8	Administrative	0,28	Business Development	0,68
Finland	0,42	Human Resources	0,18	Healthcare Services	0,2	Administrative	0,21
France	0,53	Engineering	0,76	Administrative	0,25	Information Technology	0,75
French Guiana	0,5	Administrative	0,26	Operations	0,69	Business Development	0,67
French Polynesia	0,52	Administrative	0,3	Business Development	0,67	Operations	0,61
French-Guadeloupe	0,49	Administrative	0,21	Engineering	0,74	Information Technology	0,74
French-Martinique	0,44	Administrative	0,21	Information Technology	0,73	Engineering	0,71
Gabon	0,64	Engineering	0,82	Information Technology	0,81	Operations	0,78
Georgia	0,41	Administrative	0,27	Accounting	0,3	Support	0,3
Germany	0,62	Engineering	0,85	Military and Protective Services	0,8	Information Technology	0,79
Ghana	0,7	Engineering	0,89	Military and Protective Services	0,82	Information Technology	0,8
Gibraltar	0,59	Business Development	0,72	Operations	0,68	Finance	0,59
Greece	0,55	Engineering	0,82	Military and Protective Services	0,78	Information Technology	0,74
Greenland	0,46	Education	0,46	Administrative		Arts and Design	
Grenada	0,54	Education	0,36	Business Development	0,62	Operations	0,58
Guam	0,48	Military and Protective Services	0,8	Administrative	0,24	Information Technology	0,73
Guatemala	0,61	Engineering	0,85	Operations	0,76	Information Technology	0,76
Guernsey	0,56	Business Development	0,69	Operations	0,67	Information Technology	0,59
Guinea	0,74	Engineering	0,83	Information Technology	0,83	Operations	0,81
Guinea-Bissau		Administrative		Business Development		Community and Social Services	

Guyana	0,49	Administrative	0,22	Military and Protective Services	0,71	Support	0,3
Haiti	0,66	Engineering	0,79	Military and Protective Services	0,76	Information Technology	0,75
Honduras	0,56	Engineering	0,8	Operations	0,72	Information Technology	0,69
Hong Kong SAR	0,53	Engineering	0,75	Information Technology	0,71	Administrative	0,31
Hungary	0,49	Engineering	0,81	Human Resources	0,22	Administrative	0,25
Iceland	0,5	Engineering	0,76	Information Technology	0,67	Healthcare Services	0,33
India	0,75	Military and Protective Services	0,85	Real Estate	0,85	Operations	0,82
Indonesia	0,57	Engineering	0,78	Military and Protective Services	0,73	Operations	0,7
Iraq	0,82	Operations	0,9	Military and Protective Services	0,88	Engineering	0,88
Ireland	0,54	Engineering	0,84	Administrative	0,24	Entrepreneurship	0,71
Isle of Man	0,52	Business Development	0,69	Administrative	0,31	Healthcare Services	0,33
Israel	0,63	Engineering	0,79	Military and Protective Services	0,78	Information Technology	0,76
Italy	0,53	Engineering	0,75	Real Estate	0,74	Military and Protective Services	0,73
Jamaica	0,45	Engineering	0,84	Administrative	0,18	Human Resources	0,23
Japan	0,65	Engineering	0,8	Military and Protective Services	0,8	Business Development	0,75
Jersey	0,47	Administrative	0,24	Operations	0,68	Healthcare Services	0,32
Jordan	0,69	Operations	0,83	Engineering	0,82	Military and Protective Services	0,79
Kazakhstan	0,46	Accounting	0,11	Human Resources	0,18	Engineering	0,77
Kenya	0,63	Engineering	0,87	Military and Protective Services	0,8	Operations	0,73
Kuwait	0,77	Engineering	0,87	Military and Protective Services	0,86	Quality Assurance	0,85
Kyrgyzstan	0,49	Engineering	0,77	Information Technology	0,68	Administrative	0,34
Laos	0,68	Operations	0,73	Business Development	0,68	Education	0,56
Latvia	0,4	Accounting	0,16	Human Resources	0,22	Healthcare Services	0,25
Lebanon	0,55	Engineering	0,8	Information Technology	0,75	Operations	0,71
Lesotho	0,53	Business Development	0,68	Operations	0,67	Administrative	0,45
Liberia	0,8	Information Technology	0,87	Operations	0,87	Community and Social Services	0,81
Libya	0,78	Engineering	0,9	Operations	0,89	Finance	0,87
Liechtenstein	0,78	Business Development	0,81	Operations	0,74	Accounting	
Lithuania	0,43	Engineering	0,85	Human Resources	0,16	Accounting	0,17

Luxembourg	0,61	Engineering	0,82	Information Technology	0,75	Entrepreneurship	0,72
Macao SAR	0,5	Engineering	0,77	Military and Protective Services	0,76	Information Technology	0,7
Madagascar	0,56	Engineering	0,73	Information Technology	0,68	Operations	0,64
Malawi	0,7	Engineering	0,86	Military and Protective Services	0,8	Operations	0,78
Malaysia	0,56	Military and Protective Services	0,77	Engineering	0,76	Operations	0,69
Maldives	0,74	Business Development	0,86	Operations	0,84	Accounting	0,76
Mali	0,75	Engineering	0,82	Military and Protective Services	0,82	Operations	0,81
Malta	0,56	Engineering	0,86	Information Technology	0,74	Business Development	0,72
Mauritania	0,82	Operations	0,87	Business Development	0,84	Education	0,8
Mauritius	0,54	Engineering	0,78	Administrative	0,28	Military and Protective Services	0,71
Mayotte	0,58	Education	0,58	Administrative		Business Development	
Mexico	0,58	Engineering	0,83	Operations	0,75	Information Technology	0,74
Moldova	0,43	Accounting	0,15	Engineering	0,76	Education	0,27
Monaco	0,5	Administrative	0,25	Business Development	0,7	Operations	0,66
Mongolia	0,64	Operations	0,67	Business Development	0,64	Education	0,46
Montenegro	0,43	Engineering	0,74	Administrative	0,31	Healthcare Services	0,33
Morocco	0,66	Military and Protective Services	0,82	Engineering	0,81	Operations	0,8
Mozambique	0,72	Information Technology	0,86	Operations	0,85	Engineering	0,83
Myanmar	0,53	Accounting	0,37	Military and Protective Services	0,63	Information Technology	0,58
Namibia	0,51	Engineering	0,81	Administrative	0,23	Operations	0,7
Nepal	0,75	Engineering	0,85	Information Technology	0,82	Sales	0,82
Netherlands	0,62	Engineering	0,86	Information Technology	0,75	Administrative	0,26
New Caledonia	0,51	Engineering	0,77	Administrative	0,25	Business Development	0,7
New Zealand	0,51	Engineering	0,82	Administrative	0,23	Healthcare Services	0,29
Nicaragua	0,61	Engineering	0,79	Operations	0,77	Information Technology	0,7
Niger	0,79	Operations	0,86	Business Development	0,81	Education	0,81
Nigeria	0,66	Engineering	0,83	Military and Protective Services	0,79	Information Technology	0,74
North Macedonia	0,43	Accounting	0,26	Human Resources	0,27	Military and Protective Services	0,69

Northern Mariana Islands	0,65	Operations	0,65	Accounting		Administrative	
Norway	0,55	Engineering	0,8	Information Technology	0,74	Entrepreneurship	0,71
Oman	0,74	Engineering	0,88	Operations	0,87	Quality Assurance	0,86
Pakistan	0,85	Purchasing	0,93	Engineering	0,91	Operations	0,9
Panama	0,54	Engineering	0,79	Operations	0,72	Administrative	0,3
Papua New Guinea	0,64	Engineering	0,88	Operations	0,78	Military and Protective Services	0,76
Paraguay	0,56	Engineering	0,76	Information Technology	0,71	Military and Protective Services	0,7
Peru	0,58	Engineering	0,84	Military and Protective Services	0,79	Operations	0,78
Philippines	0,47	Engineering	0,75	Military and Protective Services	0,72	Administrative	0,31
Poland	0,5	Engineering	0,81	Human Resources	0,2	Accounting	0,23
Portugal	0,53	Information Technology	0,75	Military and Protective Services	0,75	Human Resources	0,29
Puerto Rico	0,48	Administrative	0,19	Engineering	0,8	Human Resources	0,32
Qatar	0,77	Engineering	0,9	Quality Assurance	0,88	Military and Protective Services	0,88
Republic of the Congo	0,7	Engineering	0,83	Information Technology	0,82	Operations	0,79
Reunion	0,52	Information Technology	0,81	Engineering	0,76	Administrative	0,25
Romania	0,46	Human Resources	0,19	Accounting	0,25	Military and Protective Services	0,72
Russia	0,47	Accounting	0,1	Human Resources	0,14	Engineering	0,78
Rwanda	0,65	Engineering	0,78	Information Technology	0,71	Business Development	0,71
Samoa	0,47	Operations	0,47	Accounting		Administrative	
Saudi Arabia	0,85	Engineering	0,94	Military and Protective Services	0,92	Operations	0,91
Senegal	0,66	Engineering	0,8	Information Technology	0,79	Operations	0,76
Serbia	0,49	Human Resources	0,24	Accounting	0,28	Engineering	0,71
Seychelles	0,56	Business Development	0,67	Operations	0,67	Education	0,4
Sierra Leone	0,76	Information Technology	0,84	Operations	0,83	Business Development	0,8
Singapore	0,52	Engineering	0,72	Military and Protective Services	0,71	Administrative	0,3
Slovakia	0,53	Engineering	0,86	Human Resources	0,23	Military and Protective Services	0,75
Slovenia	0,44	Human Resources	0,18	Engineering	0,79	Accounting	0,23
Solomon Islands		Accounting		Administrative		Business Development	
Somalia	0,82	Operations	0,9	Business Development	0,88	Education	0,84

South Africa	0,51	Engineering	0,8	Administrative	0,24	Military and Protective Services	0,68
South Korea	0,51	Military and Protective Services	0,7	Engineering	0,63	Business Development	0,63
South Sudan	0,75	Operations	0,81	Business Development	0,77	Community and Social Services	0,74
Spain	0,57	Information Technology	0,74	Engineering	0,73	Business Development	0,69
Sri Lanka	0,61	Sales	0,67	Operations	0,66	Entrepreneurship	0,66
St Kitts and Nevis	0,62	Business Development	0,63	Operations	0,61	Accounting	
St Vincent and the Grenadines	0,59	Education	0,35	Operations	0,62	Business Development	0,59
St, Lucie County, Florida, United States	0,45	Engineering	0,84	Administrative	0,16	Healthcare Services	0,22
Suriname	0,47	Administrative	0,27	Operations	0,7	Information Technology	0,68
Sweden	0,52	Engineering	0,81	Healthcare Services	0,25	Human Resources	0,29
Switzerland	0,59	Engineering	0,82	Information Technology	0,77	Military and Protective Services	0,74
Taiwan	0,51	Engineering	0,74	Human Resources	0,29	Business Development	0,68
Tajikistan	0,6	Education	0,6	Accounting		Administrative	
Tanzania	0,68	Engineering	0,85	Operations	0,77	Information Technology	0,77
Thailand	0,59	Engineering	0,72	Business Development	0,7	Consulting	0,69
The Bahamas	0,51	Administrative	0,17	Engineering	0,81	Support	0,22
The Gambia	0,7	Operations	0,81	Information Technology	0,8	Education	0,78
Timor-Leste	0,54	Education	0,62	Administrative	0,46	Accounting	
Togo	0,72	Engineering	0,86	Legal	0,82	Information Technology	0,81
Trinidad and Tobago	0,44	Engineering	0,83	Administrative	0,18	Human Resources	0,23
Tunisia	0,61	Military and Protective Services	0,78	Operations	0,78	Business Development	0,77
Turkey	0,62	Military and Protective Services	0,86	Engineering	0,83	Entrepreneurship	0,8
Turkmenistan	0,8	Operations	0,8	Accounting		Administrative	
Turks and Caicos Islands	0,68	Business Development	0,69	Operations	0,66	Administrative	
Uganda	0,64	Engineering	0,84	Military and Protective Services	0,74	Business Development	0,72
Ukraine	0,48	Accounting	0,12	Human Resources	0,16	Engineering	0,8
United Arab Emirates	0,71	Engineering	0,88	Military and Protective Services	0,84	Quality Assurance	0,81

United Kingdom	0,57	Engineering	0,86	Administrative	0,26	Military and Protective Services	0,71
United States	0,51	Engineering	0,8	Administrative	0,22	Military and Protective Services	0,75
Uruguay	0,54	Military and Protective Services	0,77	Engineering	0,75	Healthcare Services	0,29
Uzbekistan	0,67	Engineering	0,87	Information Technology	0,81	Business Development	0,8
Vanuatu	0,6	Operations	0,66	Education	0,53	Accounting	
Venezuela	0,57	Military and Protective Services	0,8	Engineering	0,78	Operations	0,74
Vietnam	0,49	Accounting	0,33	Human Resources	0,33	Engineering	0,63
Western Sahara, Morocco		Accounting		Administrative		Business Development	0
Yemen	0,83	Engineering	0,94	Sales	0,92	Operations	0,92
Zambia	0,67	Engineering	0,86	Operations	0,8	Military and Protective Services	0,77
Zimbabwe	0,58	Engineering	0,8	Military and Protective Services	0,72	Operations	0,72

BIBLIOGRAFÍA

[1] PIMCity, «PIMCity». [En línea]. Disponible en: https://www.pimcity.eu/#.

[2] «PIMCity Deliverable 3.2: Design of data valuation tools and trading engine», 2020.

[3] LinkedIn, «LinkedIn Campaign Manager». [En línea]. Disponible en: linkedin.com/campaignmanager.

[4] LinkedIn, «LinkedIn Ads Agreement». [En línea]. Disponible en: https://www.linkedin.com/legal/sas-terms.

[5] LinkedIn, «Campaign Groups - Overview». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a424035.

[6] LinkedIn , «Forecasted Results in Campaign Manager». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a427257. [Último acceso: 27 mayo 2021].

[7] LinkedIn, «Buscar Más Acerca de LinkedIn audiencias con los resultados previstos: preguntas frecuentes». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a422276. [Último acceso: 27 mayo 2021].

[8] LinkedIn, «Presupuestos de campaña – Descripción general». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a422101. [Último acceso: 27 mayo 2021].

[9] LinkedIn, «Indicadores de vídeo en el administrador de campañas - Definiciones». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a426666.

[10] LinkedIn, «Campaign Bidding Strategies – Overview». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a421112.

[11] LinkedIn, «Maximum Delivery Bidding Strategy – Overview». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a423490.

[12] LinkedIn, «Target Cost Bidding Strategy – Overview». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a424652.

[13] LinkedIn, «Puja manual - Resumen». [En línea]. Disponible en: https://www.linkedin.com/help/lms/answer/a423484.

[14] A. Cuevas, J. Gonzalez Cabañas, A. Arrate y R. Cuevas, «Does Facebook Use Sensitive Data for Advertising Purposes?», *Communications of the ACM*, vol. 64, nº 1, pp. 62-69, 2019.

[15] J. Gonzalez Cabañas, A. Cuevas y R. Cuevas, «FDVT: Data Valuation Tool for Facebook Users», *CHI Conference*, 2017.

[16] I. Basileal, A. Korolova y J. Heidemann, «Auditing for Discrimination in Algorithms Delivering Job Ads», *Proceedings of the Web Conference*, 2021.

[17] M. Fatehkia, B. Coles, F. Ofli y I. Weber, «The Relative Value of Facebook Advertising Data for Poverty Mapping», *ICWSM*, vol. 14, nº 1, pp. 934-938, 2020.

[18] J. Palotti, N. Adler, A. Morales-Guzman, J. Villaveces, S. Vedran, M. Garcia Herranz, M. Al-Asad y I. Weber, «Monitoring of the Venezuelan exodus through Facebook's advertising platform», *PLoS ONE*, vol. 2, nº 15, 2020.

[19] D. García, Y. M. Kassa, A. Cuevas, M. Cebrian, E. Moro, I. Rahwan y R. Cuevas, «Analyzing gender inequality through large-scale Facebook advertising data», *PNAS*, vol. 115, nº 27, pp. 6958–6963, 2018.

[20] D. Garcia, «FacebookGenderDivide», 2018. [En línea]. Disponible en: https://github.com/dgarcia-eu/FacebookGenderDivide.

[21] N. Obradovich, Ö. Özak, I. Martin, I. Ortuño-Ortin, E. Awad, M. Cebrian, R. Cuevas, K. Desmet, I. Rahwan y A. Cuevas, «Expanding the Measurement of Culture with a Sample of Two Billion Humans» *IZA DP*, nº 13735, 2020.

[22] K. Haranko, E. Zagheni, K. Garimella y I. Weber, «Professional Gender Gaps Across US Cities», *ICWSM*, pp. 604-607, 2018.

[23] F. C. Verkroost, R. Kashyap, K. Garimella, I. Weber y E. Zagheni, «Tracking global gender gaps in information technology using online data», *Digital Skills Inights*, pp. 81-93, 2020.

[24] World Bank, «LinkedIn Data for Development». [En línea]. Disponible en: https://datacatalog.worldbank.org/.

[25] Microsoft, «Getting Started with LinkedIn's Marketing APIs». [En línea]. Disponible en: https://docs.microsoft.com/en-gb/linkedin/marketing/getting-started.

[26] Mozilla and individual contributors, «JavaScript reference». [En línea]. Disponible en: https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference.

[27] K. Simpson, *Up & Going*, Sebastopol, CA: O'Relly Media, 2015.

[28] «Chrome DevTools Protocol». [En línea]. Disponible en: https://chromedevtools.github.io/devtools-protocol/.

[29] Google, «Web Page Replay». [En línea]. Disponible en: https://chromium.googlesource.com/catapult/+/HEAD/web_page_replay_go/.

[30] IETF, «RFC 33986», 2005.

[31] Ecma International, «The JSON Data Interchange Syntax». Ginebra, 2017.

[32] Mozilla, «Percent-encoding». [En línea]. Disponible en: https://developer.mozilla.org/en-US/docs/Glossary/percent-encoding.

[33] Mozilla, «encodeURIComponent». [En línea]. Disponible en: https://developer.mozilla.org/es/docs/Web/JavaScript/Reference/Global_Objects/encodeURIComponent.

[34] Berners-Lee et al., « Uniform Resource Identifier (URI): Generic Syntax».

[35] S. Abuse, «Cómo crear un módulo Node.js». 15 abril 2020. [En línea]. Disponible en: https://www.digitalocean.com/community/tutorials/how-to-create-a-node-js-module-es#:~:text=En%20Node.,hace%3B%20cualquier%20archivo%20de%20Node..

[36] «Tmux». [En línea]. Disponible en: https://github.com/tmux/tmux/wiki.

[37] Instituto Nacional de Estadística, *Clasificación Nacional de Ocupaciones*, Madrid, 2010.

[38] Organización Internacional del Trabajo , *Clasificación Internacional Uniforme de Ocupaciones*, 2008.

[39] «onetonline.org». [En línea]. Disponible en: https://www.onetonline.org/.

[40] M. R. Frank, D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. Fledman, M. Groh, J. Lobo, E. Moro, D. Wang, H. Youn y I. Rahwan, «Toward understanding the impact of artificial intelligence on labor», *PNAS*, vol. 116, nº 14, pp. 6531-6539, 2019.

[41] country-list, «github.com/umpirsky/country-list», 6 mayo 2020. [En línea]. Disponible en: https://github.com/umpirsky/country-list/blob/master/data/en/country.csv. [Último acceso: 22 marzo 2021].

[42] scikit-learn developers, «Agglomerative Clustering». [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html.

[43] Unión Europea, *Reglamento (UE) 2016/679 del Paralamento Europeo y del Consejo*, 2016.

[44] Unión Europea, *Directiva 95/46 del Consejo*, 1995.

[45] Agencia Española de Protección de Datos, «Proteccion de datos: Guía para el Ciudadano».

[46] LinkedIn, «Condiciones de uso». [En línea]. Disponible en: https://es.linkedin.com/legal/user-agreement?.

[47] LinkedIn, «Política de privacidad». [En línea]. Disponible en: https://es.linkedin.com/legal/privacy-policy?.