



ugr

Universidad
de Granada

TRABAJO FIN DE MÁSTER
INGENIERÍA INFORMÁTICA

Diseño y Desarrollo de una Plataforma ETL para Minería de Opiniones en Redes Sociales.

Autora

Lidia Sánchez Mérida

Director

Antonio Gabriel López Herrera



Escuela Técnica Superior de Ingenierías
Informática y de Telecomunicación

Granada, 1 de julio de 2021

Diseño y Desarrollo de una Plataforma ETL para Minería de Opiniones en Redes Sociales

Lidia Sánchez Mérida

Palabras clave: Red social, Análisis inteligente de datos, Análisis estadístico, Análisis de sentimientos, Comunidad, Usuario, Seguidores, Seguidos, Publicaciones, Preprocesamiento de datos, Fuentes de información.

Resumen

Las redes sociales están adquiriendo un papel cada vez más relevante en los diferentes mercados existentes, gracias al volumen de datos que generan. Es por ello por lo que la mayoría de las compañías invierten más recursos temporales, económicos y personales con el objetivo de aumentar y mejorar su presencia en las diversas comunidades virtuales, así como en la extracción de conocimiento útil que les ayude a maximizar sus beneficios. Como consecuencia, en este proyecto se plantea un estudio detallado acerca de las características comunes y propias de las redes sociales más populares a nivel global, con el fin de diseñar una estructura de información y un flujo de preprocesamiento que faciliten la aplicación de diversos tipos de análisis. Para demostrar su funcionamiento, se pretende desarrollar un prototipo *software* que integre el proceso completo de descarga, filtrado, preprocesamiento y almacenamiento de información procedente de una o varias redes sociales, además de diversos análisis estadísticos e inteligentes que sirvan como ejemplo del potencial que se esconde tras los datos recopilados previamente.

Design and Development of a ETL Platform for Opinion Mining in Social Networks

Lidia Sánchez Mérida

Keywords: Social networks, Social media, Intelligent data analysis, Statistical analysis, Sentiment analysis, Community, Users, Accounts, Followers, Followings, Posts, Media, Data processing, Data sources.

Abstract

Social networks are gaining importance on the different markets, thanks to the amount of data they create. That is why most business invest more economic and human resources to increase and improve their presence in the virtual communities, as well as the ability to get useful knowledge which helps them to maximize their benefits. As a consequence, this project presents a detailed study about the common and particular features of the most popular social networks in the world. The main goal is to design a data schema and preprocessing flow to then apply some types of analysis over the downloaded and preprocessed information. In order to prove how it works a software prototype will be developed by integrating the complete process of downloading, filtering, preprocessing and storing information from one or more social networks, as well as by integrating some statistical and intelligent analysis, which will be an example of the power that is hidden after the collected data.

Agradecimientos

En primer lugar, me gustaría agradecer el apoyo incondicional que me ha proporcionado mi madre en todo momento, puesto que ha sido una de las personas que más fuerza y sabiduría me ha transmitido durante el desarrollo de mi formación. Sin ella nada de lo que he conseguido hubiese sido posible.

A continuación, me gustaría recordar a mis compañeros del máster con los que he compartido momentos de alegría, tristeza, celebración, pero especialmente, por el apoyo que nos hemos brindado mutuamente en momentos tan complicados como los que hemos experimentado durante esta terrorífica pandemia.

Finalmente, debo agradecerle a mi tutor la proposición de este proyecto que tanto me ha inspirado, emocionado y alentado a continuar mi aprendizaje en el ámbito de la Ciencia de Datos, así como la confianza y la libertad para poder orientarlo según mis preferencias.

Índice de contenidos

Capítulo 1. Introducción.....	30
1.1. Motivación.....	31
1.2. Objetivos.....	32
1.3. Estructura de la memoria.....	32
Capítulo 2. Estado del Arte. Primera parte.....	34
2.1. Origen y evolución de las redes sociales virtuales.....	34
2.2. El auge de las redes sociales virtuales.....	37
2.3. Tipos de redes sociales virtuales.....	38
2.3.1 Correo electrónico.....	38
2.3.2 Aplicaciones y webs de mensajería.....	38
2.3.3 Redes sociales basadas en opiniones.....	38
2.3.4 Blogs.....	39
2.3.5 Redes sociales multimedia.....	39
2.4 Funcionalidades comunes y sus implicaciones.....	39
2.4.1 Identidad.....	40
2.4.2 Conversaciones.....	40
2.4.3 Compartir.....	41
2.4.4 Presencia.....	42
2.4.5 Relaciones sociales.....	42
2.4.6 Reputación.....	43
2.4.7 Comunidades.....	43
2.5. Aplicaciones en el ámbito laboral y económico.....	44
2.5.1 Mercado laboral.....	44
2.5.2 Mercado económico.....	46
2.6. Redes sociales como fuentes de datos.....	47
2.6.1 Twitter API.....	48
2.6.2 Facebook API.....	48
2.6.3 Instagram API.....	48
Capítulo 2. Estado del Arte. Segunda parte.....	50
3.1. Origen del análisis inteligente de datos.....	50
3.2. Disciplinas influyentes.....	51
3.3. Software de análisis de redes sociales.....	52
3.3.1 Herramientas internas.....	52
3.3.2 Herramientas de terceros.....	53
Capítulo 3. Objetivos generales y específicos.....	55
Capítulo 4. Metodología de desarrollo.....	59
Capítulo 5. Desarrollo del proyecto.....	62

5.1. Planificación del proyecto.....	62
5.3. Extracción de requisitos.....	66
Requisitos de datos.....	66
Requisitos funcionales.....	67
Requisitos no funcionales.....	67
5.4. Casos de uso.....	68
5.5. Diseño de la plataforma.....	71
5.5.1. Esquema de información.....	71
5.5.2. Flujo de procesamiento de datos.....	73
5.5.3. Lenguajes y librerías utilizadas.....	74
5.5.4. Tecnologías y despliegue.....	78
5.5.5. Arquitectura del sistema.....	79
5.5.6. Diseño de la interfaz.....	80
Capítulo 6. Ejemplo de aplicación.....	83
6.1. Evolución del perfil.....	83
6.2. Actividad del usuario.....	86
6.3. Interés de las publicaciones.....	88
6.4. Popularidad de las publicaciones.....	90
6.5. Análisis de sentimientos basados en texto.....	92
6.6. Análisis de patrones de conducta.....	95
Capítulo 7. Conclusiones y trabajo futuro.....	98
Bibliografía.....	100
Apéndice 1.....	104
Manual de usuario.....	104

Índice de ilustraciones

Ilustración 1: Página principal de la red social Six Degrees. Fuente: [4].....	21
Ilustración 2: Página principal de la red social Friendster. Fuente: [4].....	21
Ilustración 3: Funcionalidades comunes a las redes sociales y sus implicaciones. Fuente: [12].....	25
Ilustración 4: Los funciones comunes más destacadas de Youtube y Facebook. Fuente: [12].....	29
Ilustración 5: Uso de las redes sociales para lanzar campañas de publicidad. Fuente: [19].....	32
Ilustración 6: Diseño de la máquina tabuladora. Fuente: [33].....	36
Ilustración 7: Disciplinas que componen el análisis inteligente de datos. Fuente: [36]....	37
Ilustración 8: Esquema representativo de la metodología incremental de desarrollo. Fuente: [50].....	44
Ilustración 9: Diagrama de Gantt que se planteó al principio del proyecto. Fuente: realización propia.....	48
Ilustración 10: Diagrama de Gantt de la planificación real del proyecto. Fuente: realización propia.....	49
Ilustración 11: Diagrama de casos de uso asociado a la plataforma. Fuente: realización propia.....	53
Ilustración 12: Flujo de procesamiento aplicado a la información obtenida de las redes sociales. Fuente: realización propia.....	58
Ilustración 13: Esquema relacional para el almacén de datos SQL. Fuente: realización propia.....	63
Ilustración 14: Arquitectura de la herramienta. Fuente: realización propia.....	64
Ilustración 15: Diagrama HTA de la interfaz de la plataforma. Fuente: realización propia.	66
Ilustración 16: Diagrama de conceptos de la plataforma. Fuente: realización propia.....	66

Ilustración 17: Diagrama Wireflow de la interfaz de la aplicación. Fuente: realización propia.....	67
Ilustración 18: Recopilación de datos sobre la cuenta de Audi España en Instagram durante octubre de 2020.....	68
Ilustración 19: Recopilación de datos sobre la cuenta de Audi España en Instagram durante noviembre de 2020.....	68
Ilustración 20: Recopilación de datos sobre la cuenta de Audi España en Instagram durante diciembre de 2020.....	68
Ilustración 21: Evolución del perfil de la cuenta de Audi en Instagram durante siete días.	69
Ilustración 22: Evolución del perfil de la cuenta de Audi en Instagram durante siete días en mayor profundidad.....	69
Ilustración 23: Evolución del perfil de la cuenta de Audi en España durante cuatro semanas.....	70
Ilustración 24: Evolución del perfil de la cuenta de Audi en España durante cuatro semanas en mayor profundidad.....	71
Ilustración 25: Estudio de la actividad de la cuenta de Audi en España durante siete días.	72
Ilustración 26: Estudio de la actividad de la cuenta de Audi en España durante cuatro semanas.....	73
Ilustración 27: Estudio del interés de las publicaciones de la cuenta Audi en España durante siete días.....	74
Ilustración 28: Estudio del interés de las publicaciones de la cuenta Audi en España durante cuatro semanas.....	74
Ilustración 29: Estudio de las publicaciones mejor valoradas de la cuenta de Audi en España durante cuatro semanas.....	75
Ilustración 30: Captura de la publicación mejor valorada de la cuenta de Audi en España.	75
Ilustración 31: Estudio de las publicaciones peor valoradas de la cuenta de Audi en España durante cuatro semanas.....	76
Ilustración 32: Captura de la publicación peor valorada de la cuenta de Audi en España.	76
Ilustración 33: Captura de la segunda publicación peor valorada de la cuenta de Audi en España.....	77

Ilustración 34: Análisis de sentimientos de los títulos de las publicaciones de Audi en España durante cuatros semanas.....	77
Ilustración 35: Análisis de sentimientos de los comentarios de las publicaciones de Audi en España durante una semana del mes de noviembre.....	78
Ilustración 36: Análisis de sentimientos de los comentarios de las publicaciones de Audi en España durante una semana del mes de diciembre.....	78
Ilustración 37: Comentarios irónicos encontrados en las publicaciones de la cuenta de Audi en España.....	80
Ilustración 38: Análisis de los patrones de comportamiento de la cuenta de Audi en España durante noviembre.....	80
Ilustración 39: Análisis de los patrones de comportamiento de la cuenta de Audi en España durante diciembre.....	81
Ilustración 40: Sección de recopilación de datos sobre un usuario en particular y un conjunto de fuentes de información.....	89
Ilustración 41: Filtros a completar para realizar un análisis sobre la evolución del perfil de un usuario.....	90
Ilustración 42: Resultados gráficos del análisis sobre la evolución del perfil de un usuario.....	90
Ilustración 43: Filtros a completar para realizar un análisis sobre la popularidad de las publicaciones de un usuario.....	91
Ilustración 44: Filtros a completar para realizar un análisis de sentimientos sobre los títulos o los comentarios de las publicaciones de un usuario.....	91

Índice de tablas

Tabla 1. Resumen de los costes asociados al desarrollo del proyecto.....	52
Tabla 2. Entidades comunes a tres de las redes sociales más populares, destacando aquellas que se pretenden añadir en este proyecto.....	58
Tabla 3. Datos de los perfiles de usuario de las tres redes sociales más populares, destacando aquellos que se pretenden añadir al proyecto.....	58
Tabla 4. Datos de las publicaciones de las tres redes sociales más populares, destacando aquellos que se pretenden añadir al proyecto.....	59
Tabla 5. Comparación entre los sentimientos identificados en varios comentarios por diferentes analizadores de sentimientos.....	62
Tabla 6. Ejemplos de comentarios negativos recopilados en el mes de diciembre sobre la cuenta de Audi España en Instagram.....	80

Capítulo 1

Introducción

Desde que comenzó la revolución tecnológica, se han generado una gran cantidad de datos de todo tipo a partir del uso que realizan las personas de los recursos alojados en la red. La mayor parte de la información se encuentra vinculada a los propios usuarios, como son sus datos personales, pero también es posible analizar la interacción que realizan estos con las aplicaciones que poseen para descubrir información que no se aporta de forma directa, como los patrones de comportamiento. Si bien estos datos se llevan generando desde hace décadas, no ha sido hasta hace unos años cuando nos hemos percatado de la importancia que reside en analizar la información tanto proporcionada como generada por las aplicaciones. Gracias a ello, se han desarrollado nuevas disciplinas que facilitan esta tarea como la *Inteligencia de Negocio*, que combina la gestión de información con técnicas de *Minería de datos* para realizar análisis de diversos tipos, como aquellos basados en modelos de Aprendizaje Automático [1].

Debido a las numerosas ventajas de los estudios inteligentes de datos, la gran mayoría de compañías han desarrollado o adquirido herramientas específicas para la extracción, procesamiento y análisis de datos, con las que poder analizar la evolución de la sociedad actual y así poder adaptar sus procesos de contratación y estrategias de mercado. Sin embargo, para aplicar este procedimiento es necesario disponer de grupos sociales con suficientes usuarios como para poder obtener resultados representativos. Uno de los mejores ambientes disponibles actualmente son las conocidas *redes sociales*. Su principal característica es la interconexión de millones de personas de todo el mundo a las que se les proporciona un entorno en el que poder expresar libremente sus ideas, gustos, inquietudes, opiniones, etc. Gracias a su ámbito global, repercusión y la influencia que ejercen las redes sociales, muchas organizaciones están invirtiendo gran parte de sus recursos en monitorizarlas y extraer conocimiento inteligente a partir de la información que almacenan.

Asimismo, las redes sociales han posibilitado la creación de un nuevo mercado económico y laboral en el que tanto empresas como particulares desean aumentar su presencia para fines muy diversos como los que se presentan a continuación.

- Elaborar estrategias de inteligencia de negocio.
- Desarrollar y entrenar modelos de predicción que ayuden a identificar patrones de conducta de los usuarios.
- Analizar la repercusión de la comercialización de un nuevo producto antes de lanzarlo al mercado con el fin de minimizar riesgos.
- Elaboración de perfiles socio-psicológico de candidatos a un puesto de trabajo.
- Análisis de la actividad de ciertos usuarios para la detección de delitos cibernéticos.

Para ello, existen diferentes redes sociales que se presentan como fuentes de cantidades masivas de datos que se pueden analizar aplicando diversas técnicas de Minería de Datos así como utilizando modelos predictivos y algoritmos de Aprendizaje Automático. Con el fin de facilitar la interpretación de los resultados obtenidos tras los análisis de información, la mayoría de herramientas incorporan sistemas de visualización sencillos que permiten expresar las conclusiones extraídas en diversos formatos, como gráficas y tablas estadísticas.

Debido a la gran relevancia que están adquiriendo los análisis inteligentes de datos además de la influencia de las redes sociales que almacenan cantidades masivas de información, en este proyecto se desarrollará una plataforma capaz de obtener los datos de una cuenta de usuario procedente de cualquier medio social para procesarlos, analizarlos y visualizar los resultados. El objetivo principal es disponer de una herramienta capaz de analizar, de forma inteligente, el perfil y los patrones de comportamiento de los usuarios que interaccionan con el contenido de una cuenta en particular. Los estudios que se podrán llevar a cabo en esta aplicación se han establecido en función al tipo de información que se encuentra disponible en los diferentes medios sociales, desde medidas estadísticas en función de los datos personales de los usuarios, análisis de sus redes de contactos, hasta el estudio de su comportamiento, gustos y opiniones en función de las interacciones que realizan con las publicaciones que visualizan.

Como se ha mencionado anteriormente, las aplicaciones de esta herramienta son muy diversas puesto que gracias al estudio de un conjunto de usuarios dentro de un medio de comunicación, se puede extraer información muy valiosa tanto para el mercado laboral, económico así como para el desarrollo de distintas estrategias de *marketing* con las cuales alcanzar un rango más amplio de personas.

1.1. Motivación

La aparición de *Internet* ha provocado el desarrollo de nuevas estrategias para obtener datos, generar conocimiento útil con el fin de difundir información, productos y servicios a un mayor número de usuarios. Además, ha facilitado la creación de una gran cantidad de empresas, lo que se traduce en un aumento de la competitividad por obtener el liderazgo en uno o varios mercados. Hace unas décadas, los métodos tradicionales que se aplicaban para obtener información útil con la que ayudar en la toma de decisiones, se realizaban mediante encuestas y entrevistas físicas o telefónicas, asistencia a ferias o congresos de muestras así como anuncios en medios de comunicación como radio, periódicos y televisión. Si bien estas técnicas se presentaban como las más adecuadas debido a la poca cantidad de información que se generaba en aquella época, actualmente son prácticamente inviables debido al aumento tanto de la velocidad como de la cantidad de datos que se generan hoy en día .

Sin embargo, gracias al descubrimiento de nuevas tecnologías y medios de comunicación masivos, como las redes sociales, estos métodos son insuficientes hoy en día para consultar si un producto tendrá éxito en el mercado o si un candidato a un puesto de trabajo es el más idóneo. La principal razón reside en que la aparición de estos recursos han facilitado el desarrollo y evolución de una sociedad tecnológica, exigente, que demanda productos y servicios adaptados a los nuevos requisitos que surgen a una velocidad vertiginosa. Para cubrir estas necesidades, es indispensable la invención y aplicación de métodos inteligentes, capaces de realizar las mismas actividades que las técnicas tradicionales de una manera más eficaz, eficiente y rápida.

Así, surgieron diversas técnicas de Minería de Datos y Vigilancia Tecnológica, con las que es posible monitorizar la actividad de los usuarios en diferentes medios sociales. El objetivo principal consiste en obtener cantidades masivas de datos para analizarlos de forma rápida e inteligente y, de este modo, poder descubrir los nuevos hábitos y actitudes de los ciudadanos. Con estas técnicas, las empresas son capaces de reducir la incertidumbre y los riesgos asociados a la elaboración de sus productos, servicios así como a la contratación de personal.

Adicionalmente, las redes sociales juegan un papel de medios de comunicación que conectan a millones de usuarios, lo que posibilita generar y difundir nuevas tendencias de manera global, gracias a la influencia que ejercen sobre los ciudadanos de todo el mundo. No obstante, para ello es necesario estudiar la popularidad de los distintos usuarios de una red con el fin de seleccionar a los más influyentes y así poder extender la propaganda sobre un determinado producto o servicio. Para esta tarea, también se puede utilizar la herramienta que se presenta en este proyecto, de modo que se analicen algunos de los usuarios más influyentes de un medio social

con el fin de poder comparar los resultados y descubrir cuáles son los candidatos más prometedores.

Por último, también cabe la posibilidad de llevar a cabo auto-análisis con el objetivo de conocer qué opinión tienen los usuarios de diferentes medios sociales con respecto a la propia organización y a los productos y/o servicios que ofrece. De este modo, pueden conocer los aspectos positivos que destacan los usuarios así como aquellos en los que se debe de mejorar.

Por otro lado, es igualmente posible recopilar y analizar información de forma inteligente acerca de las opiniones de los usuarios con respecto a la competencia. El objetivo principal consiste en conocer su reputación asociada como entidad además de la de sus productos y/o servicios, así como las técnicas que utilizan y les son beneficiosas.

Cada una de las redes sociales dispone de una serie de datos particulares y es por ello por lo que, en un principio, se comenzaron a desarrollar herramientas de análisis específicas para cada una. No obstante, si analizamos la información que ofrecen, podemos comprobar que existe una multitud de datos comunes a todas ellas, como los que componen el perfil, la red de contactos, los *likes* así como los comentarios situados en las publicaciones. Es por ello por lo que se ha diseñado esta herramienta de modo que se pueda extraer y analizar información procedente de cualquier red social. De esta forma se proporciona una única plataforma capaz de obtener informes estadísticos independientemente de la fuente de datos, lo que proporciona una mayor robustez al sistema así como una gran flexibilidad para que, en un futuro, el cliente pueda analizar los mismos aspectos en diferentes entornos virtuales.

Otro aspecto a destacar reside en la facilidad de uso de la herramienta, puesto que el objetivo principal es que pueda ser utilizada por el mayor rango posible de usuarios, independientemente de su perfil. Uno de los recursos alojados en la red que resulta más familiar es la web, y es por ello por lo que en este proyecto se propone desarrollar esta herramienta como un sistema web. Desde él se podrá acceder a todas las operaciones comentadas anteriormente, como elegir la fuente de datos para recopilar información de una cuenta de usuario, realizar los diversos análisis disponibles o visualizar los resultados obtenidos, de forma clara y sencilla.

1.2. Objetivos

Este proyecto consiste en el desarrollo de una plataforma web capaz de obtener información de medios sociales para llevar a cabo tanto el preprocesamiento, almacenamiento y análisis inteligente de los datos recopilados, así como la visualización de los resultados obtenidos. En particular, los objetivos concretos que abordaremos en este proyecto son los siguientes:

1. Revisar el estado del arte para conocer las diferentes herramientas de análisis de redes sociales existentes en el mercado.
2. Desarrollar un sistema capaz para extraer datos de redes sociales para su posterior almacenamiento en distintas bases de datos, que faciliten los diferentes análisis que se puedan realizar así como la visualización de los resultados obtenidos en última instancia.
3. Implementar diversos análisis combinados que consideren tanto información personal como relacionada con productos o servicios de interés.
4. Analizar los comentarios de los usuarios para conocer el sentido positivo, negativo o neutral de estos con el objetivo de identificar los patrones de comportamiento que demuestran los usuarios a partir de sus comentarios.

1.3. Estructura de la memoria

La memoria de este proyecto se ha organizado en siete capítulos, comenzando por una breve introducción a las redes sociales y a los análisis inteligentes de datos, cuya información se detalla en la siguiente sección en la que se realiza un estudio

acerca del estado del arte de sendos ámbitos. En él se pretende explicar las definiciones formales de ambas tecnologías, así como ilustrar el origen de su aparición y sus componentes actuales. Además, también se pone de manifiesto la creciente necesidad de investigar sus fundamentos e idear nuevas mejoras con el objetivo de ampliar, aún más si cabe, la gama de aplicaciones en las que incorporar tanto las redes sociales como los análisis inteligentes de datos.

En el tercer capítulo se recopilan los fines generales y específicos que se persiguen con la realización de este proyecto, detallando el enfoque hacia el que se desea orientar el diseño de la plataforma en cuestión. Será en el cuarto capítulo en el que, además, se explique en detalle la metodología de desarrollo que se ha llevado a cabo durante el proyecto para diseñar e implementar la arquitectura de la herramienta.

A continuación, se ejemplificará el funcionamiento de la plataforma desarrollada con el objetivo de observar y comprobar las posibles aplicaciones que se pueden llevar a cabo, tanto desde el punto de vista de un usuario convencional como desde la visión de un desarrollador medianamente experimentado. Finalmente, se termina la memoria de este proyecto proporcionando unas pinceladas acerca del planteamiento y el desarrollo futuro en el que se podría continuar mejorando y ampliando las funcionalidades de la aplicación.

Capítulo 2

Estado del Arte

Primera parte. Redes Sociales.

Las personas, como seres sociales que somos, siempre nos encontramos en la búsqueda de nuevos contactos para ampliar nuestro círculo social. Desde hace unos años, también nos hemos enfocado en generar nuevas metodologías de interacción social con el fin de disponer de diversas técnicas que nos permitan conectar con un mayor número de usuarios de todo el mundo. Sin embargo, gracias a la enorme evolución, su extensión global y su gran impacto en la sociedad, las redes sociales virtuales se han convertido en sistemas tecnológicos con muchas más funcionalidades que la de estar en contacto con tus seres queridos.

El término **red social** hace referencia a las interconexiones existentes entre personas y a la posibilidad de generar nuevos enlaces con otros individuos que, a priori, se personan como desconocidos [2]. Basándose en este significado del término original, surgieron las **redes sociales virtuales**, que mediante plataformas web proporcionan un entorno social en el que cientos de millones de personas pueden conectar entre sí conversando, compartiendo gustos e intereses a la vez que formando nuevas comunidades con ideales similares [2] [3]. A diferencia de las redes sociales tradicionales, que suelen estar compuestas por unas diez personas de media, las redes sociales virtuales no conocen barreras físicas ni demográficas que les impidan conectar a gente de todo el mundo [3].

Esta forma de comunicación ha provocado una revolución tecnológica y social que ha cambiado muchas de nuestras costumbres y que ha añadido otras nuevas desconocidas hasta el momento. Debido a la relevancia tanto en el mundo real como en este proyecto, en la primera parte de este capítulo se explicarán los principales aspectos de las redes sociales para conocer, en detalle, cuáles son sus características y cuál es el papel que juegan en ámbitos tan relevantes como el comercio o el mercado laboral.

2.1. Origen y evolución de las redes sociales virtuales

Si bien el origen de las redes sociales virtuales es un tema muy controvertido debido a su origen desconocido, es innegable el hecho de que uno de sus precursores fue la aparición de Internet. La posibilidad de establecer una comunicación de forma casi instantánea sin ningún tipo de barreras permitió el planteamiento de un entorno virtual de contacto entre personas de todo el mundo [4].

La primera comunicación virtual se produjo a comienzos de la década de los setenta cuando el primer correo electrónico fue enviado entre dos ordenadores situados uno al lado del otro. Aunque el contenido del mensaje fuese la sucesión de caracteres de la primera línea de letras del teclado, se convirtió en una de las acciones más revolucionarias del siglo veinte [4] [5]. A partir de la misma, se desarrolló una primera plataforma online denominada **CompuServe**, a la

que cientos de usuarios podían acceder para conversar a través del paso de mensajes. Esta especie de foro se convirtió en el primer ejemplo de lo que podría ser una red social virtual [4].

Posteriormente, en 1978 se presentó una primera versión de un sistema de transferencia de archivos online denominado **Bulletin Board System (BBS)**, el cual estaba basado en redes de área local, que hacían uso de redes telefónicas y un módem para permitir el acceso a los usuarios. Asimismo, este sistema se encontraba instalado de forma particular en los ordenadores de los usuarios. Una vez se había identificado con sus credenciales personales, eran capaces de comunicarse entre ellos mediante el envío de mensajes así como el acceso y la descarga de todo tipo de contenido multimedia, como películas o videojuegos. La arquitectura de este primer sistema pudo ser el que estableció las bases para un modelos **peer-to-peer** [4].

Finalmente, años más tarde comenzaron a popularizarse los primeros **navegadores web** gracias a una plataforma online denominada **Usenet**, que jugaba un papel similar al de un tablón de anuncios. En él los usuarios publicaban noticias y artículos de libre acceso a nivel global. A diferencia de la BBS, esta plataforma disponía de una **arquitectura distribuida** en la que el contenido se agrupaba por temáticas. Los elementos que pertenecían a cada una de ellas se encontraban disponibles gracias a redes individuales compuestas por un conjunto de servidores en los que se almacenaban los artículos de una temática particular [5] [6].

Debido a la veloz evolución de las conexiones red y a la comercialización de los ordenadores personales, en 1988 se popularizó un servicio online denominado **Internet Relay Chat (IRC)**, el cual posibilita la comunicación instantánea entre usuarios a través de mensajes solo de texto, por lo que aún no era posible enviar contenido multimedia en tiempo real [5] [7]. Si bien se trataba de una plataforma muy primitiva, se convirtió en el primer servicio de chat remoto que se alojaba en la red y no en los ordenadores de los usuarios, como se había realizado hasta entonces [5]. Gracias a los continuos avances y nuevos desarrollos que se realizaron en la década de los noventa, surgió la primera red social virtual denominada **Geocities**, que permitía a los usuarios crear su propia página web de forma personalizada para luego clasificarlos en diferentes ciudades según el contenido del sitio web. Posteriormente, añadieron diversas funcionalidades como la de comunicarse entre ellos así como publicar su propio contenido [5].

Sin embargo, no fue hasta 1997 cuando apareció una red social virtual con características similares a las actuales. **Six Degrees** era una plataforma online que permitía a los usuarios elaborar su perfil a partir de los datos personales proporcionados, establecer relaciones de amistad con otros usuarios, buscar perfiles e invitar a otras personas para que se registrasen en este medio social [4] [5]. El origen de su nombre se basó en la teoría de los *Seis Grados de Separación*, propuesta en 1929 por Frigyes Karinthy, en la que se afirmaba que todos estamos conectados a través de seis individuos, como máximo. Si bien se considera la precursora de las redes sociales actuales, los usuarios comenzaron a abandonarla alrededor de 2002 debido a la falta de filtros de *spam* [4].

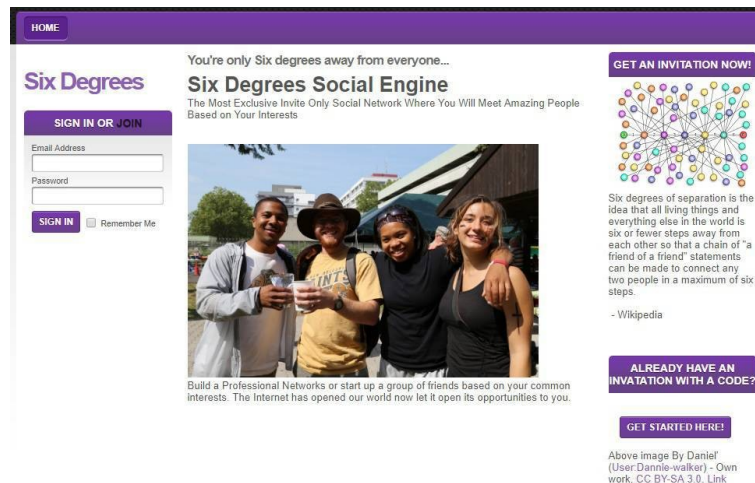


Ilustración 1: Página principal de la red social Six Degrees. Fuente: [4].

En esta época se produjo un punto de inflexión que provocó la creación de varias redes sociales virtuales en las que, con el paso del tiempo, se congregaba un mayor número de usuarios. Entre ellas, destacó la denominada **America Online Limited (AOL)**, por su doble papel como medio social y como medio de comunicación gracias a la inclusión de mecanismos que permitían difundir información entre sus usuarios. Debido a su diseño revolucionario y a su novedosa arquitectura, se convirtió en uno de los recursos online más utilizados por las generaciones más jóvenes. Esta popularidad impulsó la creación de los *blogs* personales y el comienzo de una nueva era tecnológica cuya principal característica es el **contenido viral** [4].

A partir del año 2002, en el que cada vez más personas disponían de un ordenador personal, comenzaron a surgir redes sociales virtuales más complejas como **Friendster**, que permitía la creación de comunidades de usuarios con gustos y opiniones similares con el fin de poder comunicarse y establecer discusiones sobre temas comunes, **MySpace** cuyo diseño fue prácticamente similar a la red anterior aunque con un abanico más amplio de funcionalidades, como la posibilidad de incluir contenido multimedia, **LinkedIn**, una red social virtual orientada a negocios y profesionales, además de los grandes sistemas sociales que conocemos actualmente como **Twitter** y **Facebook** [4] [5].



Ilustración 2: Página principal de la red social Friendster. Fuente: [4].

2.2. El auge de las redes sociales virtuales

Desde la evolución de la tecnología y el abaratamiento de sus costes, cada vez más personas disponen de uno o varios dispositivos digitales con los que pueden acceder a todos los servicios de la red. Esto favoreció el crecimiento de las redes sociales tanto en popularidad como en número de usuarios, forzando a sus compañías a introducir nuevas funcionalidades con las que mantener activos a sus usuarios, además de intentar captar a más personas para que formasen parte de estos medios sociales.

Sin embargo, el comportamiento humano es también un aspecto clave en la popularidad de las redes sociales, y por ello, sigue siendo estudiado por expertos de modo que las conclusiones extraídas puedan ser aplicadas también en este ámbito. Una de las más importantes consiste en **establecer relaciones personales** a partir de diferentes vías de comunicación con los que mantener el contacto. Al principio, cuando un usuario se registra en una red social por primera vez, es probable que busque las cuentas de sus seres queridos, como familiares y amigos. Una vez haya terminado de establecer vínculos sociales virtuales con sus más allegados, continuará con aquellas personas con las que tiene un menor grado de confianza. Así, seguirá expandiendo su círculo social virtual hasta aceptar y enviar peticiones de amistad a usuarios desconocidos debido a diferentes motivos, como por ejemplo, el hecho de compartir gustos y opiniones comunes, debido al interés que le suscita su contenido o al deseo de **hacer nuevas amistades** en la red [8]. De hecho, esta última actividad es la más realizada según un estudio realizado en Noruega en el que se le proponían encuestas online a los usuarios de diversas redes sociales muy populares en ese país, con el objetivo de medir qué actividades realizaban en los medios sociales de los que formaban parte [9].

Otra de las principales características del auge de las redes sociales es su capacidad de **generar contenido nuevo** y atractivo que animen a los usuarios a invertir más tiempo en ellas. Es por ello por lo que cada pocos minutos, las publicaciones se van actualizando, apareciendo las más recientes al comienzo de la lista. Por si no fuera suficiente, existe una segunda estrategia para llamar la atención al usuario: **las notificaciones**. Para diseñarlas, de nuevo aplican los estudios psicológicos acerca del comportamiento humano en los que se afirman que los elementos brillantes, de colores cálidos o sobre un fondo blanco en el que resalten su presencia, son el diseño perfecto para captar nuestra atención [8].

Además de establecer una red de contactos virtual y obtener contenido de diverso tipo actualizado periódicamente, una de las tareas que realizan la mayoría de usuarios es **indagar en los perfiles y publicaciones de otros**. Se trata de una ventana privada a la vida de otra persona, aunque depende del contenido y la información que publique y cada cuánto lo haga. Esta actividad, si bien conocemos que no es del todo correcta y que suele acarrear más desventajas que beneficios, es una de las más realizadas por todos los usuarios de casi cualquier red social. Se puede calificar de inevitable, considerando el hecho de que somos seres sumamente curiosos. Sin embargo, el problema suele originarse cuando comparamos nuestra realidad con la que otros muestran, nuestro estatus social, posesiones, entre otros, en cuyos casos aparecen sentimientos negativos a la par que peligrosos, como la envidia o la frustración [8]. Si bien las publicaciones no suelen representar la vida real de una persona, debido al incesante deseo por mostrar abiertamente muchas de las actividades que realizamos y de pensamientos que nos surgen, las organizaciones y empresas cada vez disponen de más recursos para analizar las redes sociales tanto de sus candidatos como de los miembros de la plantilla, con el objetivo de elaborar un perfil virtual que exponga más características personales de estas personas que de otro modo hubiese sido más costoso o casi imposible de conseguir.

Asociado a las dos últimas características, se encuentran dos de los aspectos fundamentales por los que las redes sociales pueden beneficiar a los usuarios. En primer lugar disponemos del **sistema de notificaciones**, que nos avisa de las fechas de cumpleaños de nuestros contactos, aniversarios, entre otros eventos. Mientras que, por otro lado, los usuarios también pueden obtener **beneficios** propios gracias a las publicaciones existentes, ya sean propias para

intentar llevar a cabo campañas de publicidad de sus productos y servicios, o sean propiedad de otros usuarios como, por ejemplo, ofertas de empleo. Sin embargo, existen entidades y organizaciones que se encuentran más interesadas en llevar a cabo **actividades con fines malintencionados**, como la manipulación de la información que se difunde o la divulgación de noticias, creencias y estudios falsos para alcanzar sus propios intereses [8].

Apelando a la parte socio-psicológica del ser humano, una de las necesidades más crecientes actualmente consiste en formar parte de aquellas comunidades o servicios en los que se encuentran las personas de nuestro círculo social. Este efecto ha sido estudiado y denominado como la **ley de Metcalfe**, la cual afirma que *“el valor de una red virtual es proporcional al cuadrado del número de usuarios”*. Es por ello por lo que todos los medios sociales disponen de una funcionalidad que les permite a los usuarios ya registrados, enviar peticiones a aquellas personas que aún no forman parte de la red social. Este efecto, unido a la presión social y a la necesidad de estar sintonía con los de tu alrededor, provocan como consecuencia que, cada vez, más personas dispongan de más cuentas en diferentes redes sociales solo por el hecho de que sus contactos se encuentran en ellas [8].

Si a las funcionalidades y características descritas anteriormente, añadimos el hecho de que el uso de las redes sociales, en principio, es **gratuito**, se convierte en la receta perfecta para incrementar, de forma exponencial, el número de usuarios que utilizan estos sistemas sociales y de información [8] [9].

2.3. Tipos de redes sociales virtuales

Tras conocer el concepto de red social virtual además de los motivos por los cuales se han convertido en uno de los servicios más populares de Internet, a continuación se presenta una clasificación general de los tipos de medios sociales existentes dependiendo de sus propósitos.

2.3.1 Correo electrónico

Como se ha explicado anteriormente, el correo electrónico se postula como una de las primeras aproximaciones a las redes sociales virtuales. Dispone de cientos de millones de usuarios que proporcionan sus credenciales para acceder a su cuenta y comenzar a establecer contacto con otros usuarios a través del envío de mensajes. Para ello, es necesario que conozcan la dirección de correo electrónico de la entidad con la que pretenden relacionarse. En resumen, se trata de un medio social virtual orientado a un tipo de **comunicación indirecta**, en la que se puede incluir contenido de todo tipo, tanto textual como multimedia. Sin embargo, a diferencia de la mayoría de redes sociales, está pensado para establecer **conversaciones privadas con individuos específicos** [10].

2.3.2 Aplicaciones y webs de mensajería

Tanto si se trata de una aplicación como de una sala de chat, en este tipo de redes sociales se establecen **comunicaciones directas**, ya que los usuarios pueden conversar entre sí en tiempo real. Si bien el contenido principal suele ser texto, con el paso del tiempo se ha añadido otro tipo de expresiones como los emoticonos, los *stickers* y los *gifs*, además de la posibilidad de compartir imágenes y vídeos tradicionales. Este tipo de medios sociales son los **más populares y extendidos** a nivel mundial, tanto por el uso por parte de la sociedad como por su integración en diferentes aplicaciones o plataformas, como los juegos online [10].

2.3.3 Redes sociales basadas en opiniones

En este ámbito existen diferentes tipos de plataformas sociales cuyo principal objetivo consiste en brindar la posibilidad a los usuarios de expresar sus opiniones con respecto a una

temática, producto o servicio. En primer lugar disponemos de los **foros de discusión**, como *Reddit* o *Quora*, que se caracterizan por ser **indirectos** y por la flexibilidad de que cualquier usuario pueda interaccionar con el resto de miembros sin la necesidad de conocerlos previamente, a diferencia de los dos casos anteriores. Además, generalmente suelen disponer de diferentes categorías para clasificar las discusiones existentes en función de su temática [11].

Por otro lado existen redes sociales especializadas en la crítica a productos, servicios o empresas como es el caso de *TripAdvisor* o *ElTenedor*. Este tipo de plataformas están adquiriendo, cada vez, una mayor relevancia e influencia en la sociedad puesto que proporcionan la posibilidad de **conocer la opinión de otros usuarios** con respecto a un producto, servicio o empresa que deseemos contratar. De este modo, se minimizan los posibles **riesgos y la incertidumbre** de si lo que deseamos comprar se adapta a nuestras necesidades [11].

2.3.4 Blogs

El término *blog* proviene de la palabra *weblog*, que designa una página web en la que una persona **escribe y comparte artículos de forma pública** para que otras personas puedan visualizarlos y realizar comentarios. Actualmente, existen una gran diversidad de usuarios que disponen de blogs, desde particulares que expresan sus gustos e ideas, creadores de contenido que utilizan este medio como una forma más de publicitar sus productos, hasta expertos y divulgadores de diversa naturaleza que publican diferentes tipos de artículos [10].

A raíz de su popularidad hace unos años, surgieron plataformas específicas en este tipo de medios de comunicación, como es el caso de *Medium* o *Tumblr*. Se trata de redes sociales que facilitan la creación, personalización, actualización y mantenimiento de uno o varios blogs, de modo que los usuarios solo tienen que concentrarse en generar contenido, puesto que para el resto de operaciones, las plataformas disponen de mecanismos automáticos que se encargan de llevarlas a cabo. Asimismo, al formar una comunidad extensa de usuarios, existe un mayor número de probabilidades de que los artículos publicados puedan ser visualizados por un amplio rango de personas de todo el mundo [11].

2.3.5 Redes sociales multimedia

Este tipo de medios sociales se caracterizan por la capacidad de publicar contenido de cualquier naturaleza. Plataformas como *Instagram* o *Snapchat*, son claros ejemplos de redes sociales cuyo contenido es únicamente visual, apelando al dicho popular que afirma que *una imagen, vale más que mil palabras*. Por otro lado existen medios sociales como *Youtube* que se centran **exclusivamente en contenido audiovisual**, ya sea estático o en tiempo real. Si bien se puede pensar que cuanto más restrictivo sea el tipo de contenido que se pueda publicar, menor interés puede suscitar, lo cierto es que este tipo de plataformas son las que disponen de un **mayor número de usuarios**. La dificultad de generar una imagen atractiva visualmente es mucho mayor que la de crear contenido audiovisual con un mínimo de interés, puesto que el mero hecho de disponer de movimiento suele ser suficiente como para captar nuestra atención aunque sea por poco tiempo. Es por ello por lo que este tipo de medios sociales también son los escogidos para difundir productos, servicios y noticias con el objetivo de alcanzar un rango más amplio de personas [11].

2.4 Funcionalidades comunes y sus implicaciones

Si bien en la sección anterior se han presentado los principales tipos de redes sociales existentes en la actualidad, todas ellas comparten un conjunto de funcionalidades comunes. Tanto para las compañías que poseen este tipo de plataformas, como para las que están interesadas en desarrollar nuevos medios sociales o aumentar su presencia en los ya

existentes, es necesario estudiar sus características e implicaciones asociadas. En particular, diversos estudios han definido siete grandes bloques que reflejan los aspectos más representativos de las redes sociales. Cabe destacar que no son mutuamente excluyentes y que no todas deben de integrar las siete funcionalidades. En la siguiente figura se resumen tanto las categorías como sus posibles implicaciones [12].

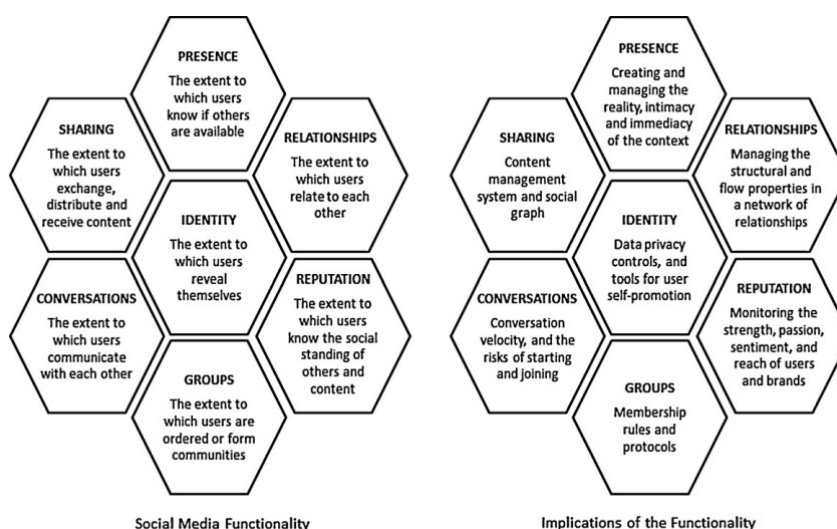


Ilustración 3: Funcionalidades comunes a las redes sociales y sus implicaciones.
Fuente: [12].

2.4.1 Identidad

Se trata del bloque principal que representa la posibilidad de identificar a un usuario a partir de la elaboración de su perfil, en el que se encuentran algunos de sus datos personales. Debido a su naturaleza sensible, la primera implicación procedente de obtener este tipo de información es la **privacidad**. Todas las entidades que dispongan de plataformas sociales en las que se recopilen datos de carácter personal, tienen como obligación legal integrar medidas de seguridad acorde al nivel de sensibilidad de la información para protegerla de accesos no autorizados. Asimismo, deben hacer lo propio con las operaciones que se les puede aplicar para diversos fines, ya sean estadísticos o relacionados con el *marketing* [12].

Por otro lado, las organizaciones cuyas intenciones residan en analizar los perfiles de los usuarios, deben considerar la estrategia, desarrollada por los individuos, de generar **diferentes identidades en función de la red social**. Por ejemplo, las publicaciones subidas a *Facebook* suelen caracterizarse por pertenecer a ámbitos relativos al ocio o las relaciones personales, mientras que el contenido en *LinkedIn* suele ser más profesional. Como segundo ejemplo destacan aquellas cuentas cuyos propietarios no aportan información personal veraz ya que su propósito, generalmente, es generar contenido humorístico. Este tipo de comportamientos son sumamente comunes en redes sociales como Twitter, en la que existen multitud de cuentas que representan identidades imaginarias como personajes de ficción, históricos, políticos, entre otros. Asimismo, existen plataformas cuya principal característica es el **anonimato**, por lo que, o bien no se recopilan datos con los que poder identificar a una persona total o parcialmente, o este tipo de información no se comparte con ninguna otra entidad [12].

Debido a la diversidad de identificaciones existentes en las redes sociales, se han desarrollado diferentes técnicas para el inicio de sesión por parte de un usuario, así como la visualización de ciertos campos personales por parte del resto de individuos de la comunidad [12].

2.4.2 Conversaciones

Existen diversos tipos de comunicación que pueden surgir en las diferentes redes sociales, tanto a nivel individual como en grupo, como son los *tweets* o los comentarios asociados a las publicaciones. La motivación que lleva a un usuario a expresar su idea u opinión acerca de una temática en concreto, suscita mucho interés a la mayoría de organizaciones. Es por ello por lo que todas ellas buscan herramientas con las que poder obtener las conversaciones que surgen en los medios sociales con el objetivo de analizarlas y así conseguir beneficios propios [12].

Dentro del ámbito del análisis inteligente de datos, existen dos conceptos que las compañías toman en consideración a la hora de elegir una u otra aplicación analítica. Por un lado se encuentra la **tasa de cambio**, que especifica el número de conversaciones que están surgiendo en un determinado período de tiempo. Mientras que también existe un segundo término referente a la **dirección del cambio** en una conversación, que mide la continuidad de la misma en relación a si el sentido de la discusión ha cambiado de positivo a negativo, o viceversa. De este modo, las diferentes discusiones que surgen en redes sociales en las que se proporcionan una gran cantidad de opiniones, como Twitter, se consideran piezas de un puzzle que es necesario unir para extraer conclusiones útiles sobre la opinión general de un producto, servicio u organización concreta. Este procedimiento se establece como requisito indispensable para las compañías al contratar herramientas analíticas profesionales para obtener conocimiento útil de los datos provenientes de redes sociales [12].

Otra de las actividades que más se está popularizando entre las marcas que disponen de redes sociales, consiste en iniciar discusiones específicas acerca de sus productos, servicios o eventos. De este modo consiguen que los usuarios continúen engrandeciendo el hilo con sus aportaciones con el objetivo de **difundir la información que les interesa**. Asimismo, las redes sociales se posicionan como un ambiente bastante propicio para realizar **encuestas** con las que determinar qué opción es la que más les interesa a los usuarios, con el fin de minimizar riesgos a la hora de diseñar un nuevo producto o servicio [12].

2.4.3 Compartir

Esta acción se presenta como una de las principales operaciones en las que, tanto redes sociales como las distintas entidades se encuentran enormemente interesadas. La posibilidad de compartir información genera una **red social de difusión**, en la que los usuarios se encuentran vinculados con aquellas personas a las que les han enviado el contenido. Es uno de los movimientos más poderosos para divulgar todo tipo de contenido, y es por ello, por lo que cada vez más se puede apreciar cómo las campañas de *marketing* instan a la población a compartir la propaganda que realizan acerca de sus productos o servicios. Tal es así, que independientemente del medio digital en el que se encuentre un artículo, la acción de compartir se puede llevar a cabo entre una gran variedad de plataformas virtuales [12].

Sin embargo, para analizar una red social basada en la divulgación de contenido, es necesario considerar dos aspectos fundamentales. En primer lugar, es necesario descubrir qué es lo que relaciona a los usuarios, es decir, **cuál es el contenido que están compartiendo** los unos con los otros. En el caso de plataformas de vídeo, como Youtube, es más sencillo puesto que todo el contenido que dispone es del mismo tipo. Pero en redes sociales de carácter más general, como *Facebook* o *Twitter*, esta identificación puede convertirse en una tarea de gran complejidad. La segunda implicación reside en conocer si el contenido que se está compartiendo puede causar **repercusiones negativas** a la compañía que lo almacena. Uno de los casos más populares, fueron las denuncias por derechos de autor que muchos usuarios le interpusieron a Youtube, por no disponer de filtros que confirmara si los autores de los vídeos publicados se correspondían con las personas que los subían a la plataforma. De igual modo, tampoco existían mecanismos con los que detectar si el contenido de los vídeos era inapropiado [12].

2.4.4 Presencia

Este bloque trata de identificar **dónde se encuentran los usuarios**, tanto en los entornos virtuales como en la vida real. Para el primer caso existen diversos mecanismos, en la mayoría de redes sociales, que informan a su red de contactos si un usuario se encuentra **en línea**. Ejemplos representativos de este concepto los podemos encontrar en *Whatsapp*, en el que se puede mostrar si un usuario está utilizando la aplicación en tiempo real, así como en el chat propio de *Facebook*. Mientras que en la segunda situación, generalmente es el usuario el que proporciona información acerca de dónde se encuentra, ya sea publicando una imagen o mediante contenido textual. Sin embargo, existen aplicaciones específicas que son capaces de mostrar los individuos que se encuentran **cerca de una determinada ubicación**. Entre los diversos propósitos que persiguen estas herramientas, se encuentra la posibilidad de conocer gente que comparte tus mismos intereses. No obstante, la geolocalización también es utilizada con otros fines no tan beneficiosos, como es el de **localizar los controles policiales** en los tramos de carreteras para advertir a los conductores de su presencia con cierta antelación.

Las implicaciones que conlleva el hecho de mostrar la disponibilidad y la ubicación actual de un usuario afectan directamente a los propietarios de los medios sociales, en tanto en cuanto les proporcionen a estos la posibilidad de **restringir la visualización de este tipo de datos** de forma personalizada. Así, cada individuo podría elegir el rango de personas a las que mostrar esta información o mantenerla privada de modo que nadie pueda acceder a ella [12].

2.4.5 Relaciones sociales

El término relaciones sociales en el ámbito virtual hace referencia a la capacidad tanto de conversar, compartir publicaciones como a la de aparecer en la lista de contactos de otro usuario. Las interacciones que surgen entre dos sujetos dependerán del tipo de conexión que los relacione.

Del mismo modo, dependiendo de los propósitos vinculados a las redes de contactos dentro de las diferentes plataformas sociales, las sugerencias de relación con otros usuarios se realizarán en base a diferentes criterios. En el caso de redes sociales integradas en un ámbito más profesional, como *LinkedIn*, llevan a cabo análisis con los cuales se estima el número de **grados de separación** entre una persona y un usuario al que esta desea conocer. Para ello se basa en las conexiones existentes entre su red de contactos y el individuo de interés. De esta forma, *LinkedIn* sugiere como nuevos contactos a aquellos con los que se pueda disminuir la distancia hasta conseguir establecer una relación con el usuario deseado. Por otro lado, existen redes sociales cuyo objetivo reside en motivar al usuario a **expandir su red de contactos** a través de sugerencias basadas en criterios más sencillos, como los amigos de tus amigos.

Este quinto bloque se encuentra estrechamente ligado al primero que se corresponde con la **identidad de los usuarios**. La razón principal reside en que, generalmente, aquellas plataformas que no realizan demasiado énfasis en la identificación y la elaboración de un perfil medianamente complejo, no le otorgan apenas importancia a las relaciones sociales. Ejemplos representativos de este tipo de medios son *Youtube* y *Twitter*, en las que el contenido es el verdadero protagonista.

Existen dos implicaciones principales para los interesados en analizar las propiedades de la red de contactos de un determinado usuario. Por un lado, es necesario conocer el grado de influencia que posee. Generalmente, aquellas cuentas que se encuentran vinculadas con un mayor número de usuarios y que realizan una gran cantidad de interacciones, se suelen corresponder con los denominados **influencers**. Sin embargo, para realizar un estudio acerca de las propiedades relativas a las relaciones sociales de un usuario, es necesario conocer el **tipo de conexiones** que establece un individuo con sus contactos, ya que es posible que con una sola persona haya desarrollado tanto una relación de amistad como otra laboral. Esta cualidad también es considerada por los desarrolladores de las redes sociales, con el fin de

establecer los mecanismos con los que un usuario puede establecer un vínculo con otro. Así, aquellas plataformas que se caracterizan por disponer de relaciones más complejas, como es el caso de *LinkedIn* cuyo propósito es profesional, deberán integrar dispositivos de identificación que autentiquen al usuario antes de enviar la petición. Mientras, que en otros medios sociales en los que sus conexiones son solo amistosas, basta con incluir técnicas identificativas más sencillas [12].

2.4.6 Reputación

Esta característica representa el **nivel de confianza asociado a un usuario o a su contenido**, dependiendo de cuál sea la entidad predominante en una red social. Asimismo, en función de lo anterior, el mecanismo para determinar su nivel de confianza también será diferente. Por ejemplo, en el caso de *YouTube* lo más importante es el contenido y para calcularla puede utilizar tanto el número de visualizaciones como la puntuación otorgada por otros usuarios. Mientras que en el caso de *LinkedIn*, cuya entidad principal es el usuario, su reputación se calcula en función del apoyo que le proporcionan sus contactos.

Su principal implicación consiste en elegir una **métrica adecuada para calcular la reputación** de un usuario. La razón fundamental reside en que si escogemos la información incorrecta con la que estimar la confianza de un individuo, las interpretaciones obtenidas podrían no ser precisas. Un ejemplo representativo reside en analizar el número de seguidores de un usuario. Este dato no es válido para calcular su confianza, puesto que puede disponer de muchos seguidores y que en realidad ninguno visualice sus publicaciones.

Una vez escogido el tipo de información que mejor represente la reputación del usuario, a continuación es necesario escoger el **método de evaluación** que nos permita conocer cuán popular es un usuario en una determinada red social. Para ello se puede hacer uso de datos objetivos, como el número de visualizaciones de sus publicaciones, o técnicas inteligentes con las que analizar un conjunto de datos con el fin de obtener una puntuación que represente el grado de su reputación. Entre los diferentes campos que se pueden estudiar para este propósito, se encuentra el sentimiento que existe tras los comentarios en una publicación, el número de usuarios que interaccionan con su contenido, entre otros [12].

2.4.7 Comunidades

Se trata de la posibilidad que ofrecen las distintas redes sociales a los usuarios de **conformar grupos** entre ellos. Si bien una de las teorías más populares acerca del número de miembros como máximo que debe poseer una comunidad para encontrarse bien equilibrada ronda las 150 personas, en la realidad se ha demostrado la existencia de grupos muy prolíferos con un mayor número de usuarios. Dependiendo de los mecanismos integrados en las distintas plataformas sociales, existen diversas **categorías de comunidades**: públicas, privadas o por invitación. Dentro de ellas, los usuarios pueden disponer de la oportunidad de **clasificar a sus miembros** con diferentes etiquetas de modo que puedan diferenciar, por ejemplo, cuáles son amigos, conocidos, *fans*, etc.

Como implicación directa reside la posibilidad de implementar mecanismos con los que poder **ordenar y categorizar a los miembros de una comunidad**, permitiendo elegir si sus etiquetas son públicas o no. Este tipo de acciones pueden ser muy útiles al **administrar los permisos** para cada uno de los usuarios, de manera que se pueda restringir el acceso y la difusión de cada tipo de información disponible [12].

Por último, tal y como se ha destacado al comienzo de esta sección, no todas las redes sociales apuestan por incluir o potenciar los siete bloques existentes. En la siguiente imagen se pueden observar las funcionalidades más destacadas de dos de las redes sociales más populares actualmente. En el caso de *Youtube* podemos apreciar que la **más relevante es la función de compartir**. Tiene sentido que sea el bloque más distinguido puesto que sus vídeos

más populares son los que más veces han sido compartidos por los usuarios. A continuación, predominan las funciones asociadas a las **conversaciones, comunidades y reputación**. La primera funcionalidad se encuentra relacionada con la lista de comentarios existentes en los vídeos publicados. Los más populares llegan a almacenar hasta cientos de millones de opiniones de usuarios. Por otro lado, al igual que ocurre en *Twitter*, existen comunidades con numerosos miembros que comparten gustos similares y, por lo tanto, también suelen visualizar vídeos de los mismos creadores de contenido. De este modo, todos conforman un grupo asociado a la temática a la que pertenezcan las publicaciones que visualizan. Finalmente, existen diferentes sistemas integrados por la plataforma para determinar el nivel de confianza asociado al contenido de un determinado usuario, ya sea por el número de visualizaciones o por las veces que haya incumplido alguna de las normas relacionadas con la publicación de contenido audiovisual [12] [13].

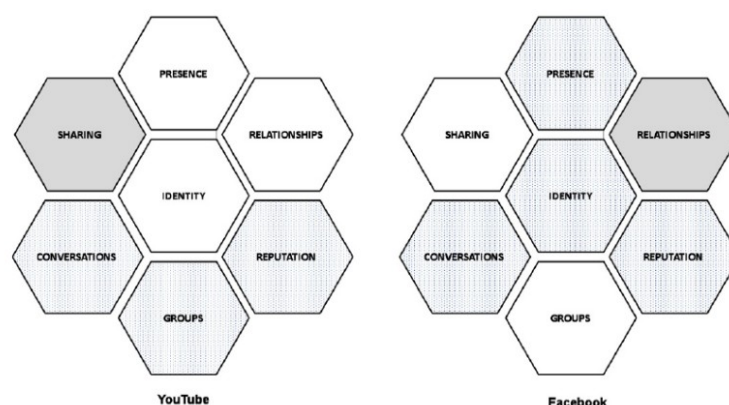


Ilustración 4: Las funciones comunes más destacadas de Youtube y Facebook. Fuente: [12].

Al contrario que *Youtube*, **Facebook** apuesta por fomentar la funcionalidad asociada a las **relaciones entre usuarios**. Esta es una de las principales características de *Facebook* y se puede apreciar en la enorme lista de contactos que le sugiere a cada usuario miembro de esta plataforma para entablar una nueva conexión. Con menor fuerza también destacan otros cuatro pilares como **la identidad, presencia, reputación y las conversaciones**. La primera funcionalidad se corresponde con los numerosos datos personales que un usuario puede aportar para elaborar un perfil más completo, con el que sea más fácilmente identificable. Mientras que en la segunda destaca por la posibilidad que les brinda a sus usuarios de publicar contenido en el que se muestre su localización real, así como mecanismos para conocer si un usuario se encuentra activo en el chat que integra la plataforma. Este último junto a la posibilidad de escribir comentarios en las publicaciones, son las dos operaciones principales por las que fomenta las conversaciones entre sus miembros. Finalmente, al igual que otras redes sociales, también dispone de sistemas específicos para evaluar la reputación de un usuario en base a las interacciones relativas a sus publicaciones [12] [13].

2.5. Aplicaciones en el ámbito laboral y económico

Dependiendo de los propósitos de los usuarios y del diseño de las redes sociales, estas plataformas pueden ser utilizadas para diversas finalidades. La mayoría de los objetivos que persiguen los usuarios en general ya han sido mencionados en anteriores secciones, como son las relaciones sociales, el entretenimiento, la búsqueda de noticias e información, etc. No obstante, algunas entidades han integrado los medios sociales como herramientas adicionales con las que realizar tanto tareas existentes como nuevas actividades. En particular, cabe destacar tanto el mercado laboral como el económico, puesto que son los que se han visto más afectados debido a la aparición de las comunidades sociales virtuales.

2.5.1 Mercado laboral

En este primer ámbito se pretende estudiar el **proceso de selección** que realizan las compañías con el fin de incorporar nuevos miembros a su plantilla. Para ello, es necesario distinguir las situaciones en las que se encuentran las dos entidades protagonistas de esta actividad: **los reclutadores y los candidatos**.

Los primeros individuos son los encargados de determinar los medios a utilizar en los procesos de reclutamiento y selección de personal. Una de las técnicas tradicionales más populares hasta hace unos años, consistía en la publicación de ofertas de empleo en los principales medios de comunicación y entretenimiento, como **periódicos y revistas**. Sin embargo, actualmente estos métodos han perdido una gran cantidad de usuarios, por lo que resulta bastante complicado que los más jóvenes visualicen las ofertas de empleo en los medios de comunicación convencionales. Otro de los recursos que solían utilizar son las **agencias de contratación**, las cuales realizan este costoso procedimiento por las empresas que contratan sus servicios con el objetivo de proporcionarles candidatos para completar sus vacantes. Sin embargo, además de las elevadas tasas que requieren por realizar la contratación de personal, en la mayoría de ocasiones no se realiza un estudio en detalle para conocer si el candidato seleccionado es el más adecuado al puesto.

Por otro lado, las denominadas **ferias de empleo** proporcionan un ambiente muy prolifero para que los reclutadores puedan conocer, en persona, a diferentes candidatos para un empleo. No obstante, este proceso es sumamente costoso debido a los recursos materiales y económicos que debe suministrar la empresa para asistir a este tipo de eventos [14]. Además, es necesario considerar que durante el **proceso de selección**, los aspirantes proporcionan datos tanto personales como profesionales. Su verificación puede ser un procedimiento bastante complejo y complicado debido al acceso restringido o a la inexistencia de fuentes de datos que contienen información real [14].

Gracias a la aparición de las redes sociales, las compañías han conseguido reducir la mayor parte del **coste económico**, puesto que estas plataformas son de uso gratuito y las pueden utilizar como medio de promoción y publicación de ofertas laborales. Del mismo modo, también se reduce enormemente el **coste temporal** puesto que los reclutadores invierten hasta un 60% menos de tiempo en el proceso de selección y contratación [14]. Además de estas dos ventajas principales, las redes sociales también pueden presentarse como fuentes de información que permiten extraer algunas conclusiones de la vida personal de los candidatos. En primera instancia, los reclutadores suelen investigar el contenido de las cuentas de los candidatos con el objetivo de elaborar una **primera impresión** acerca de su posible grado de adaptación a la compañía. Sin embargo, también pueden llevar a cabo estudios minuciosos de sus perfiles para contrastar la información que han aportado, e incluso, elaborar un **perfil psicológico**. De este modo, pueden deducir si poseen dotes de comunicación, si disponen de capacidad creativa, las personas que se encuentran en su círculo social, los lugares que suelen frecuentar, sus hábitos, así como su personalidad y comportamiento basado en las interacciones de sus publicaciones y las de otros usuarios.

Si bien, por norma general, los reclutadores buscan más información acerca de las personas que han solicitado un puesto de trabajo, también están comenzando a utilizar las redes sociales para **buscar posibles candidatos** con el fin de contactar con ellos y ofrecerles la posibilidad de participar en el proceso de selección [15].

Por otra parte, los **candidatos** también disponen de diversos beneficios a la hora de utilizar las redes sociales como portales de búsqueda de empleo. En primer lugar, son consideradas como las comunidades sociales más numerosas hasta el momento, con la capacidad de conectar usuarios de todo el mundo. Además, la especialización de algunos medios sociales en el ámbito laboral, como es el caso de *LinkedIn*, facilita la **elaboración y publicación del perfil laboral** a todos los miembros de la red. Para los candidatos, se trata de un medio virtual más

rápido y cómodo en el que poder visualizar un mayor número de ofertas de empleo además de la posibilidad de contactar directamente con las empresas de su interés [16].

Debido a la cantidad de información que pueden aportar nuestras redes sociales acerca de nuestra identidad, carrera profesional, gustos y opiniones, es sumamente importante estudiar el tipo de contenido que publicamos en ellas. Para conocer en más detalle los **criterios de evaluación** de los reclutadores, se han realizado diversos estudios como el que se presenta a continuación, en el que ambas entidades han participado. La primera conclusión que se obtiene es que un alto porcentaje de compañías utilizan redes sociales, como *Facebook* y *Twitter* para publicar ofertas de empleo y buscar información adicional acerca de los candidatos que dispongan de cuentas en estas plataformas virtuales. A continuación, se ha realizado un análisis comparativo acerca de la opinión que tienen los dos protagonistas en función de ciertos comportamientos. Los reclutadores penalizan más las publicaciones que muestren el uso de **alcohol, drogas o faltas de ortografía** que los candidatos, mientras que el contenido asociado a **voluntariados, actividades religiosas o profesionales**, son mejor valorados por las compañías que por los usuarios. Finalmente, los candidatos se muestran más contrariados a que las compañías puedan acceder y analizar el contenido de sus redes sociales, así como que las conclusiones extraídas se tengan en consideración durante el proceso de selección [16].

2.5.2 Mercado económico

Gracias al poder de **difusión a nivel mundial** tan característico de las redes sociales, estos medios han supuesto un impacto extraordinario en el ámbito económico. Es tal su influencia, que ha favorecido la aparición de una nueva estrategia de *marketing* orientada a estas comunidades virtuales. Además de su gran alcance, también ha provocado una gran revolución en las principales funciones de la publicidad. En primer lugar, se ha integrado la posibilidad de conocer las **opiniones** de los usuarios acerca de productos y servicios, mientras que por otro lado son los propios miembros los que, incluso, pueden realizar campañas de **promoción** en colaboración con la empresa o de manera totalmente altruista [17].

Otra de las actividades que favorecen la transmisión de información reside en brindar la posibilidad a los usuarios, de cualquier medio social, de **compartir** todo tipo de contenido entre sus contactos. Se trata del método de propagación más efectivo en comparación con los más tradicionales, como los anuncios de televisión y periódicos. Sin embargo, el diseño de las plataformas virtuales también favorece la difusión de contenido, a través de la visualización de las **interacciones** que realizan nuestros contactos, como los *likes* que otorgan a las publicaciones o los comentarios que les añaden. Así, las redes sociales pueden suscitar nuestro interés al comprobar que algún producto o servicio le parece atractivo a algunos miembros de nuestro círculo social [17]. Adicionalmente, existe una segunda estrategia ligada a las interacciones que realizamos con las publicaciones de otros usuarios, consistente en analizar su contenido con el objetivo de identificar nuestros posibles intereses. Las redes sociales la aplican a la hora de **personalizar la publicidad** para cada usuario, de modo que maximizan las posibilidades de promocionar un producto o servicio de interés para cada miembro de una comunidad virtual [18].

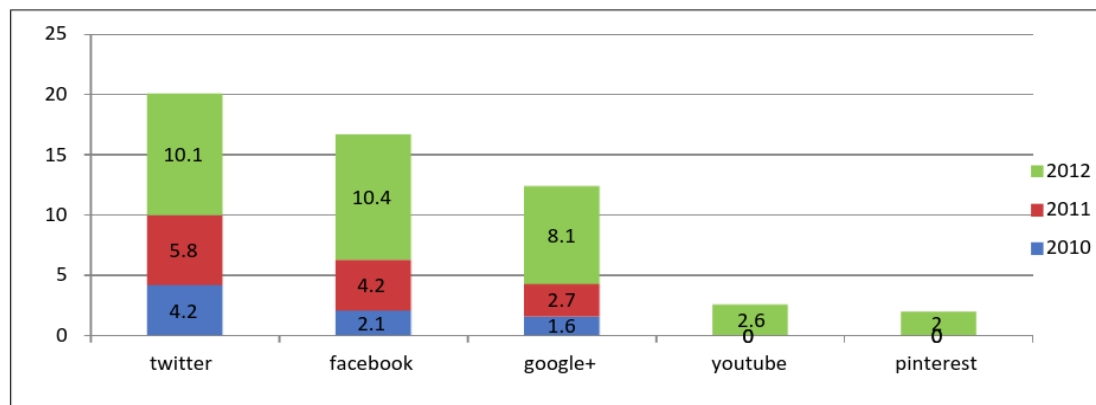


Ilustración 5: Uso de las redes sociales para lanzar campañas de publicidad. Fuente: [19].

Debido a su amplio abanico de diversas ventajas, con el paso del tiempo las redes sociales se están convirtiendo en el centro de atención para la promoción de todo tipo de productos y servicios. En la siguiente gráfica podemos observar el crecimiento exponencial entre 2010 y 2012 del uso de las plataformas sociales para integrar campañas de publicidad. No obstante, uno de los principales motivos de esta expansión reside en la **reducción de costes**, que experimentan las empresas, a la hora de anunciar sus productos en redes sociales. En primera instancia, puede ser, incluso, **gratuito**, en el caso en el que la entidad disponga de varias cuentas en diversas redes sociales y las administre por sí misma. De este modo, basta con publicar contenido llamativo e intentar expandir la red de contactos de modo que se amplíe el rango de posibles consumidores [18] [19].

Existe una segunda técnica que consiste en producir tus propios anuncios para posteriormente publicarlos en diferentes redes sociales por un precio mucho menor, comparado con los medios de comunicación convencionales. Además, las propias compañías de las plataformas virtuales proporcionan **herramientas y APIs que facilitan el desarrollo y despliegue de las campañas de publicidad**, permitiendo especificar el lugar que ocupará el anuncio, la audiencia objetivo e incluso, la geolocalización en la que se emitirá [20].

Igualmente, existen multitud de recursos online en los que se recopilan diversas guías que contienen, tanto consejos como actividades a realizar, con el fin de poder orientar todo tipo de promociones a las redes sociales de forma personalizada. En primer lugar es necesario realizar un análisis de la situación tanto de la empresa, de su posición con respecto a sus competidores así como el público actual y el que se desea alcanzar. De este modo, se determinan las **plataformas sociales más adecuadas** a los objetivos de la organización de forma que se inviertan sus recursos con las máximas garantías de éxito [20] [21]. Otro de los aspectos fundamentales reside en establecer **métricas** con las que evaluar el éxito de la estrategia implantada en función de las metas propuestas. Entre las más populares se encuentran la evolución de los seguidores, el número de clics en los enlaces, el número de ocasiones en las que el contenido se comparte entre usuarios, así como el volumen de visualizaciones. Para ello, existen herramientas que facilitan la monitorización, análisis y extracción de conclusiones en tiempo real, tanto gratuitas como de pago, que ayudan a conocer qué aspectos pueden mejorarse y que aportan sugerencias útiles para aumentar los beneficios de las campañas de publicidad en medios virtuales [21].

2.6. Redes sociales como fuentes de datos

La gran mayoría de redes sociales disponen de una **API** (*Application Programming Interfaces*) de desarrollo propio [22]. Se trata de un conjunto de protocolos de comunicación que permiten la interacción entre un cliente, como por ejemplo una aplicación, y un servidor, que es el que dispone de una serie de operaciones. A efectos prácticos, una API es una entidad intermediaria que posibilita el envío de peticiones por parte de clientes, para posteriormente ser

redireccionadas al servidor, que es el encargado de efectuarlas y devolver una respuesta. El hecho de que las compañías propietarias de los diferentes medios sociales desarrollen APIs propias facilita a los desarrolladores su **integración con las aplicaciones** [24] [25]. Generalmente, una API de cualquier red social posibilita, entre otras operaciones, la descarga de datos y la realización de las funcionalidades más básicas. A continuación se presentan las diferentes fuentes de datos que permiten la descarga de información de las redes sociales más populares actualmente [22].

2.6.1 Twitter API

La API de Twitter es una de las más utilizadas para realizar análisis inteligentes de datos sobre la información generada por sus usuarios. Se trata de una API REST que fue publicada en 2006 y cuyo acceso es posible a través de una autenticación particular para la misma. Entre sus diversas funcionalidades, permite la descarga de datos de usuarios tales como los *tweets* asociados a un *hashtag*, el histórico de *tweets* de un usuario en concreto, así como la obtención de la red de contactos de una determinada persona. Si bien en principio su uso es gratuito, se encuentra limitado a un máximo de 900 peticiones por cada 15 minutos. No obstante, se puede ampliar este límite contratando alguna de las versiones de pago disponibles.

Adicionalmente, cabe destacar la existencia de una gran cantidad de ejemplos y documentación, además de *plugins* específicos para aplicaciones orientadas al estudio de información de usuarios, como *Gephi*. Se trata de una herramienta que he estudiado y utilizado en la asignatura optativa de **Gestión de Información en la Web** de este máster profesional para analizar las propiedades de una red social [22] [23].

2.6.2 Facebook API

La gran compañía de Mark Zuckerberg publicó su primera versión de la **API Graph** en 2010. De nuevo, al igual que en el caso anterior, se trata de una API REST basada en protocolos HTTP para enviar peticiones relacionadas tanto con la descarga de datos de usuarios, como para otros fines, como subir nuevo contenido o crear campañas de publicidad automáticas. En este caso, el uso de la API es gratuito aunque para evitar la saturación de los servidores, existe un límite de 200 peticiones por hora, el cual es bastante restrictivo si lo comparamos con la API de Twitter. De forma similar al caso anterior, existe una amplia gama de documentación y ejemplos de las acciones que se pueden realizar con esta API, así como una extensa comunidad de desarrolladores que continúan actualizando, manteniendo e incorporando nuevas funciones [22] [23].

2.6.3 Instagram API

La API oficial de Instagram fue publicada en 2014 aunque actualmente no se encuentra en mantenimiento, por lo que utilizarla supondría una serie de riesgos para desarrollar la herramienta planteada en este proyecto [22] [23]. En su lugar, *Facebook* implementó la API mencionada en la sección anterior, que facilita la integración de las aplicaciones y plataformas desarrolladas para darse a conocer mediante todas sus redes sociales [26] [27]. Sin embargo, esta API se encuentra enfocada, principalmente, a creadores de contenido y empresas. Como alternativa proponen una **API básica de visualización** con la que se puede obtener información acerca de las cuentas de los usuarios, aunque los datos disponibles son sumamente limitados [27].

A diferencia de Twitter, Instagram es una red social que no ha sido tan estudiada hasta el momento, por lo que si bien la herramienta que se presenta en este proyecto se ha diseñado para recopilar datos de cualquier medio social, nos pareció interesante elaborar las pruebas para su desarrollo y testeo utilizando datos de Instagram. Sin embargo, no contamos con una API oficial y pública con una diversidad de datos suficiente como para realizar diferentes análisis con los que estudiar a sus usuarios. Por lo tanto, comenzamos a investigar y a testear varias APIs de

Instagram implementadas por terceros, las cuales actúan como intermediarias entre el cliente y la **API privada que Instagram posee**, facilitando la extracción de información y la realización de algunas operaciones.

Después de probar varias, el principal inconveniente que tienen en común es la **gestión del inicio de sesión**, ya que para ejecutar las funciones disponibles es necesario poseer una cuenta de usuario en Instagram y proporcionar las credenciales. La API privada de esta red social dispone de unas medidas de seguridad extremadamente altas, entre las cuales se incluye el bloqueo de un usuario que realice demasiados intentos de identificación en un corto período de tiempo. Si esto sucede, Instagram es capaz de impedir que el usuario bloqueado utilice su cuenta para enviar peticiones desde varias horas hasta meses, como he podido experimentar durante el desarrollo de este proyecto.

La primera solución que se podría aplicar es la de utilizar **diferentes cuentas** de Instagram para conectar con alguno de los usuarios que aún no se encuentren bloqueados. Sin embargo, en menos de una semana Instagram me bloqueó hasta dos cuentas diferentes. Por lo que realmente, esta primera propuesta no es viable por la velocidad a la que la compañía impide que una cuenta de usuario realice peticiones a su API a través de una de terceros.

Tras investigar y probar el funcionamiento de más APIs de Instagram publicadas en repositorios de *GitHub*, encontré una denominada *Instaloder* [28]. En ella se incorpora la posibilidad de **almacenar la sesión** iniciada como un objeto para reutilizarlo tantas veces como sea posible hasta que expire la misma y se deba realizar este proceso de nuevo. No obstante, esta API presenta un inconveniente aún mayor que el anterior al incorporar un manejador de excepciones para administrar los errores devueltos por la API privada de Instagram si se supera el **número máximo de peticiones** permitidas en un determinado período de tiempo. Su implementación consiste en calcular una estimación de este valor, puesto que es desconocido, para cualquier tipo de función, y en caso de haber realizado un número de peticiones cercano a la aproximación hallada, añade un determinado tiempo de espera antes de enviar otra petición. El problema reside en que puede encadenar tiempos de espera de forma consecutiva hasta el límite de, por ejemplo, obtener veinte *posts* y esperar una hora y media para descargar los veinte siguientes. Si bien es plausible que los procesos de descarga de datos sean costosos, en este caso nos pareció exagerada la solución encontrada por el desarrollador de esta API al segundo inconveniente común a todas las implementaciones: colapsar los servidores mediante el envío de peticiones masivas.

Continuando con la búsqueda de una API que pudiese obtener datos de usuarios de Instagram algo más veloz, encontré una implementada por un usuario en *GitHub* llamado *LevPasha* [29]. En este caso, no contaba con un manejador de excepciones como en la API anterior pero la capacidad de obtener una mayor cantidad de datos en menor tiempo y sin recibir tantos bloqueos por parte de Instagram, fue el detonante para **escogerla como una de las fuentes de datos** definidas para demostrar las capacidades analíticas de esta herramienta. Además, esta API dispone de una gran diversidad de funciones que permiten el acceso a un amplio abanico de datos, tanto personales asociados al perfil de los usuarios, como a las publicaciones y las interacciones que estas reciben. Sin embargo, no contaba con ningún método capaz de almacenar la sesión iniciada para su uso posterior, como en el caso de *Instaloder*, por lo que para evitar los sucesivos bloqueos de la cuenta que utilizaba lo implementé yo misma sin gran dificultad. Esto supuso una enorme reducción de peticiones para iniciar sesión además de evitar la creación de numerosas cuentas de Instagram para descargar información de los usuarios de este medio social.

Segunda parte. Análisis Inteligentes de Datos.

Gracias a la revolución tanto social como digital, los datos se han convertido en una de las entidades más relevantes e influyentes para casi cualquier tipo de actividad. Con el paso del tiempo, su diversidad ha aumentado exponencialmente, provocando la aparición de información de distinta naturaleza. Es por ello por lo que, desde hace unos años, se están desarrollando y utilizando técnicas para estudiar los datos procedentes de casi cualquier fuente, con el fin de extraer información útil con la que poder plantear nuevas estrategias y así maximizar los beneficios.

Una de las disciplinas que surgen es el **análisis inteligente de datos**. Este se define como el uso de medidas estadísticas, en combinación con técnicas de reconocimiento de patrones y de algoritmos de aprendizaje automático, que se encuentran integrados en herramientas de abstracción de datos y extracción de conocimiento útil [30]. Estos métodos disponen de la capacidad de complementarse entre sí. La razón principal reside en que, si bien existen cálculos estadísticos que requieren de una alta capacidad de computación, estos recursos no son capaces de sustituir las bases estadísticas en las que se fundamentan los análisis inteligentes de datos [31].

Debido a que el planteamiento de este proyecto tiene como principal objetivo la realización de diversos análisis inteligentes aplicados a datos procedentes de cualquier medio social, en la segunda parte de este capítulo se detallan los aspectos más relevantes de una de las técnicas más populares en el ámbito asociado a la Ciencia de Datos.

3.1. Origen del análisis inteligente de datos

Como se ha mencionado anteriormente, los análisis inteligentes de datos se encuentran basados en multitud de teorías estadísticas, las cuales comenzaron a desarrollarse en la época del antiguo Egipto para construir las pirámides [32]. Gracias a la constante evolución de la sociedad, cada vez se generaban una mayor cantidad de datos. Una de las fuentes de información de mayor tamaño fue registrada en el siglo diecinueve, cuando apareció uno de los sistemas más populares actualmente: el censo de la población. Sin embargo, conforme aumentaba el número de personas, la tarea de recopilar y registrar a los nuevos habitantes se convirtió en misión imposible. De este modo, Herman Hollerith, el que posteriormente fue fundador de IBM, diseñó una máquina capaz de agilizar considerablemente este proceso, a la que denominó **la máquina tabuladora** [32]. Este dispositivo era capaz de registrar a un individuo a través de tarjetas agujereadas en las que cada punto representaba un dato distinto. Asimismo, también contaba con un clasificador con el que se podía filtrar la población, por ejemplo utilizando la ciudad de nacimiento. En la siguiente captura se puede visualizar el aspecto de esta máquina tan revolucionaria [33].



Ilustración 6: Diseño de la máquina tabuladora.
Fuente: [33].

La revolución tecnológica y social continuó su avance hasta que, a comienzos del siglo veinte, las cantidades de información generadas eran prácticamente ingestionables. Así, surgieron los dispositivos de memoria más popularmente utilizados en la actualidad: las memorias RAM y los discos duros [34]. Gracias al diseño de una de las arquitecturas de ordenador más básicas como la que ideó John von Neumann, aparecieron los denominados **gestores de bases de datos**, como el relacional. Con este planteamiento se brindó la posibilidad de almacenar, gestionar y obtener datos bajo demanda de una forma rápida y eficaz [34] [35]. No obstante, debido a la aparición de las nuevas tecnologías y al incremento de la cantidad y de la diversidad de datos, se implementaron nuevos sistemas de almacenamiento específicamente optimizados para la recuperación de información en menor tiempo. Así, surgieron los denominados **data warehouse**, que además permitieron realizar las primeras operaciones estadísticas con datos, como el filtrado y la comparación entre ellos. Una vez se desveló el enorme potencial del análisis de información, se originó una de las disciplinas más populares hoy en día: **la Inteligencia de Negocio**. En ella se recopilan las mejores estrategias a aplicar para maximizar los beneficios de una compañía a través de la recopilación y el estudio de datos [35].

En los noventa se produjo un punto de inflexión con la aparición de Internet, que aceleró las distintas revoluciones que se desarrollaban en aquella época. Gracias a la interconexión de millones de personas y aplicaciones a nivel global, se desarrollaron nuevos sistemas de bases de datos, como el **NoSQL**, que permite el almacenamiento y la gestión de información desestructurada y de diversa naturaleza. Es por ello por lo que, junto al considerable aumento de la capacidad computacional de los ordenadores, comienza a desarrollarse uno de los nuevos paradigmas más utilizados hoy en día: **la Minería de Datos**. En ella se empezaron a usar técnicas de **análisis predictivos**, procedentes de diferentes ámbitos como la Inteligencia Artificial o el Aprendizaje Automático, para diversas finalidades, como descubrir las futuras necesidades de la sociedad, obtener los patrones de comportamiento de los usuarios, además de analizar tanto los riesgos y beneficios de distintas acciones comerciales [33].

3.2. Disciplinas influyentes

Tal y como se ha explicado en la definición proporcionada al comienzo de este capítulo, el análisis de datos se encuentra compuesto por un conjunto de diferentes disciplinas, pertenecientes tanto al ámbito de las matemáticas como al de la informática. En la siguiente imagen se puede apreciar un esquema en el que se representan las principales materias, de las que se extraen una gran cantidad de principios y técnicas, para su integración en el análisis inteligente de datos [36].

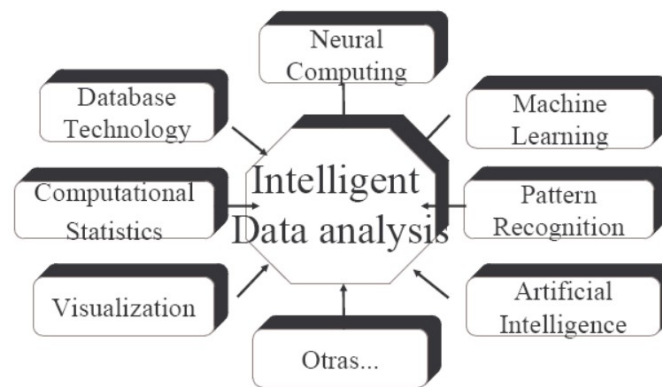


Ilustración 7: Disciplinas que componen el análisis inteligente de datos. Fuente: [36].

Una de las bases más antiguas sobre la que se sustentan las técnicas analíticas de información es la **Estadística**. Esta ciencia es ampliamente utilizada para diversas finalidades, como la experimentación apoyada en teorías ampliamente refutadas con el objetivo de que los resultados sean fiables, visualizar las conclusiones extraídas haciendo uso de multitud de gráficos estadísticos, recopilar, generar e inferir nuevas muestras con el fin de componer una población razonablemente representativa, u obtener nuevo conocimiento a partir de los datos disponibles [36] [37].

Gracias a la constante evolución de la tecnología y sus componentes fundamentales, como los dispositivos de almacenamiento, otra de las áreas que ha facilitado el desarrollo de los análisis masivos de datos es el **Almacenamiento de Información**. Existen una gran variedad de **Sistemas Gestores de Bases de Datos**, caracterizados por diferentes diseños en función de los propósitos para los que fueron desarrollados. Dependiendo de la estructura en la que se representa la información, pueden facilitar la realización de diversos estudios de datos. Uno de los ejemplos más conocidos son los sistemas de almacenamiento **relacionales**, en los cuales se definen estructuras de datos determinadas que permiten interconectarlos y facilitar la ejecución de consultas complejas de forma eficiente. Mientras que, por otro lado, existen almacenes de datos basados en **grafos** que facilitan el estudio de las propiedades relativas a la información estructurada como una red de datos. Ambos sistemas serán utilizados en este proyecto y detallados en capítulos posteriores. Adicionalmente, debido al aumento de la capacidad de computación que se encuentra presente en la gran mayoría de ordenadores, se han podido desarrollar disciplinas como la **Inteligencia Artificial** o el **Aprendizaje Automático**, en las que se definen diferentes metodologías con las que desarrollar nuevos algoritmos y entrenar modelos de predicción para optimizar todo tipo de operaciones, como el procesamiento de datos, la detección de patrones de información, la búsqueda de relaciones entre los propios datos, así como la implementación de nuevas representaciones con las que agilizar diversos cálculos [36] [38] [39].

3.3. Software de análisis de redes sociales

Una vez conocemos el potencial que se esconde tras la información almacenada en las redes sociales, así como los principios y operaciones que conforman los análisis inteligentes de datos, a continuación se ha realizado una investigación acerca de las herramientas actuales que se encuentran orientadas a este tipo de procedimientos. Algunas de ellas han sido desarrolladas por los propietarios de las redes sociales más populares, y por ello se las conoce como **internas** [40]. Sin embargo, también existe multitud de software diseñado con los mismos propósitos por terceras partes.

3.3.1 Herramientas internas

En el caso de **Facebook**, dispone de diversas herramientas en función del objetivo que se persiga. Una de las más populares es **Audience Insights**, que ayuda a captar **más consumidores** comparando la actividad de su público actual y la del resto de usuarios de esta plataforma. Para ello recopila información de diversa naturaleza, como los datos demográficos de los individuos, su frecuencia de uso o las páginas mejor valoradas para diferentes categorías. Según la documentación de esta herramienta, la información que se recoge es en términos generales, por lo que realmente este software **no es capaz de recopilar y analizar los datos de usuarios concretos** [41].

Otra de las grandes corporaciones que ha desarrollado su propia herramienta analítica es **Twitter**. Su principal finalidad consiste en proporcionar al usuario un resumen acerca de la **actividad que ha recibido su perfil mes a mes**. Para ello es capaz de recabar datos que se encuentran accesibles en la propia plataforma, como el número de *likes* y *retweets*, e información no disponible públicamente, como el número de visitas que ha recibido la cuenta o los *tweets* más populares del mes. No obstante, para hacer uso de esta herramienta estadística es necesario utilizar las **tarjetas de Twitter**, una funcionalidad que permite representar el contenido del perfil [40] [42].

Un último ejemplo de las herramientas internas podemos encontrarlo en **Instagram Insights**, a la cual únicamente tienen acceso las **cuentas de negocios o los grandes influencers**. Este software solo es capaz de recopilar un rango muy limitado de datos, como el número de *likes*, las visitas que recibe el perfil, información demográfica básica de los seguidores, etc [40] [43].

Como se ha podido apreciar hasta el momento, las compañías propietarias de las redes sociales más populares han proporcionado a sus usuarios herramientas analíticas con las que estudiar la evolución de sus perfiles. Sin embargo, todas ellas presentan diversos **inconvenientes** comunes. El primero de ellos reside en que cada herramienta se encuentra orientada exclusivamente al análisis de los **datos asociados a la plataforma en cuestión**. Por lo que un usuario que desee examinar el rendimiento de su negocio en las diferentes plataformas virtuales, tendría que hacer uso de cada una de ellas, lo cual le resultaría bastante laborioso. Por otro lado, cada software dispone de unas **métricas diferentes** para obtener las distintas medidas estadísticas que proporciona. Un ejemplo representativo de esta desigualdad reside en la forma de establecer el número de visualizaciones de un vídeo. Mientras que en **Facebook** es necesario que un usuario vea el vídeo durante tres segundos para contar como una visualización, en **Youtube** este período se amplía hasta los treinta segundos. Por lo que los resultados obtenidos en las diferentes herramientas para una misma medida no serían comparables. Asimismo, la **variedad de datos** a los que permiten acceder de forma gratuita son muy limitados, en la mayoría de ocasiones, lo cual también limita los tipos de análisis que se pueden realizar [44]. Por último, en algunos casos como sucede en Twitter, es necesario hacer uso de ciertas **funcionalidades propias** de la plataforma para beneficiarse del acceso a sus herramientas analíticas. Esto puede suponer un inconveniente a la hora de planificar la estructura del contenido que se pretende publicar.

3.3.2 Herramientas de terceros

Por otro lado, al igual que las compañías propietarias de redes sociales han desarrollado sus softwares analíticos, ciertas organizaciones externas han contribuido con sus propias herramientas. La principal diferencia con las anteriores reside en que la mayoría de los programas de terceros son **de pago**, al contrario que las herramientas internas en las que su uso era gratuito. Sin embargo, esto explica la adición de un mayor número de **análisis más complejos** que permiten a los usuarios realizar una gran diversidad de estudios de sus redes sociales. Además, la gran mayoría de las herramientas de terceros son capaces de obtener y examinar información procedente de **diversas plataformas** virtuales, por lo que con un solo programa un individuo podría gestionar y analizar todas sus cuentas.

No obstante, también existen herramientas enfocadas específicamente en un tipo de red social o contenido, como es **Snaplytics**, que se basa en el análisis de las publicaciones temporales, como las historias de Instagram o el contenido de *Snapchat*, **SHIELD App** que se centra en el estudio de la evolución de perfiles profesionales alojados en redes sociales vinculadas al mercado laboral como *LinkedIn*, o **TapInfluence**, cuyo objetivo principal reside en el estudio de los denominados *influencers* a través de diversos factores como su nivel de compromiso, el alcance mediático, así como una estimación de la recaudación que podrían conseguir [45].

Las herramientas de terceros también disponen de multitud de características comunes, independientemente de si son generales o especializadas. Entre ellas se encuentra la posibilidad de realizar **análisis de información básica**, como el número de clics que ha recibido una publicación, la evolución del número de seguidores, etc. Asimismo, también permiten la **personalización de los informes** en los que se visualizan las conclusiones obtenidas, mayormente, mediante representaciones gráficas. Por otro lado, solo algunas de ellas incluyen funcionalidades de estudio y comparación de datos con las organizaciones de la **competencia**, para analizar la situación de una empresa con respecto a las demás que se encuentran en su mismo sector [45] [46].

Si bien estos programas externos integran un mayor número de funcionalidades, como principal inconveniente destaca el enorme **incremento del precio** conforme deseamos disponer de una mayor flexibilidad y variedad de análisis. De igual modo, también se encarece la tarifa si deseamos disponer de la herramienta en diferentes dispositivos o si estudiar varias cuentas procedentes de diversos medios sociales. Una segunda limitación que he observado en todas ellas es que el usuario dispondrá de las **fuentes de información** que se hayan incluido en la herramienta. Por lo que siempre va a estar sujeto a los acuerdos que pacten la compañía propietaria del programa con las empresas poseedoras de los diferentes medios sociales. Asimismo, como hemos podido observar, la mayoría de programas se encuentran orientados al análisis de **información de mercado**, por lo que no existen demasiadas herramientas capaces de estudiar perfiles de usuarios concretos para diversos fines.

Por otro lado, estos programas presentan ciertas limitaciones que se encuentran dentro de un ámbito más técnico. Una de las más comunes consiste en **almacenar todos los datos** proporcionados por las distintas fuentes de información, en lugar de filtrarlos para obtener únicamente aquellos que son realmente útiles. Esto supone un sobre-esfuerzo tanto en tareas de almacenamiento como de análisis, lo que conlleva un coste adicional en materia económica y temporal. Otra de las principales dificultades que encuentran los usuarios de estas herramientas reside en la capacidad de traducir sus necesidades analíticas a las **métricas** integradas en el programa. Asimismo, cabe destacar que la mayoría de herramientas disponen de análisis estadísticos básicos y que son pocas las que integran técnicas relativas al **Aprendizaje Automático o Inteligencia Artificial**, con el objetivo de inferir un conocimiento de mayor interés y complejidad. Por ejemplo, obtener el número de comentarios de las publicaciones de una determinada cuenta, puede no ser tan útil como **analizar sus sentimientos** y presentar un informe acerca de las publicaciones mejor o peor valoradas por los miembros de una comunidad virtual [47].

Capítulo 3

Objetivos generales y específicos

En este tercer capítulo se detallan los fines que se persiguen con la realización de este proyecto. Si bien los objetivos han sido expuestos de forma abstracta al comienzo de la memoria, a continuación se especifica el contexto en el que se han enmarcado para definir el planteamiento hacia el que se ha orientado y así poder describir las funcionalidades que han surgido durante el diseño.

El **primer objetivo** consiste en realizar una revisión sobre el estado del arte acerca de los software actuales especializados en el análisis de redes sociales. Durante la primera parte del capítulo anterior, se han detallado los conceptos y los elementos propios de estas plataformas, así como su historia, evolución y tipología. Esta lectura sirve como preámbulo inicial para comprender la segunda parte, en la que se detallan los entresijos de los actuales programas especializados en el análisis de información procedente de diferentes medios sociales. En particular, en el último epígrafe, se proporciona una clasificación en la que se encuentran los diferentes tipos de herramientas según pertenecen a las propias compañías propietarias de las redes sociales, o si han sido desarrolladas por terceros. Aquellas que pertenecen a la primera categoría se caracterizan por analizar únicamente los datos de redes sociales particulares, siendo bastante restrictivas en la diversidad de información a la que permiten acceder, además de la desigualdad de las métricas utilizadas para calcular los mismos conceptos, lo cual no permite realizar una comparación de los resultados obtenidos. Mientras que las plataformas desarrolladas por terceros disponen de planes económicos muy elevados y de análisis demasiado sencillos y poco optimizados para un gran volumen de datos.

En el **segundo objetivo** se desea plantear un sistema capaz de obtener información sobre una o varias redes sociales, con la intención de poder ser procesada y guardada en uno o varios sistemas de almacenamiento, y así poder utilizarla para realizar diversos análisis. La extracción de datos de diversos medios sociales se fundamenta en una estructura de información común y de diferentes esquemas particulares a cada una de las plataformas. En el primer caso, se aglutinan aquellos datos que se encuentran presentes en la mayoría de los medios sociales existentes en la actualidad, como por ejemplo la información personal asociada a los perfiles de los usuarios. Mientras que, por otro lado, el sistema dispondrá de estructuras personalizadas para las diferentes redes sociales, con el objetivo de unificar aquellos datos que son exclusivos de cada una de ellas. Así, este diseño permite que la **extracción de datos sea independiente de las fuentes** de información, puesto que es posible integrar varias de ellas con el fin de obtener datos de una misma plataforma a través de la aplicación de una misma estructura que especifica un conjunto de campos uniforme.

De igual modo, cada esquema de información tiene asociado un conjunto de métodos de **procesamiento**, para verificar que los datos obtenidos para cada campo son de la tipología esperada y se encuentran dentro de un rango de valores válido. Esta actividad supone un paso previo al almacenamiento en una **primera base de datos orientada a documentos**, cuyo principal cometido consiste en mantener la información obtenida para la posterior realización de análisis. Generalmente, los datos procedentes de las redes sociales pueden combinarse en diferentes agrupaciones dependiendo del ámbito al que pertenecen. Es por ello, por lo que como segundo método de almacenamiento se ha escogido un sistema **relacional** con el que

establecer un diagrama en el que se interconectan las diferentes entidades identificadas. En esta misma base de datos también serán almacenados los resultados de los estudios realizados para ser visualizados directamente, así como para facilitar el cálculo de otros similares, de modo que se reutilice la información generada previamente con el objetivo de disminuir la inversión temporal y computacional de los análisis futuros.

Para completar el **tercer objetivo** se detallan a continuación los diferentes análisis que se han diseñado para este proyecto.

- **Evolución del perfil del usuario.** En este estudio se plantea una comparación entre los diferentes datos que se encuentran asociados directamente al perfil de cualquier cuenta, independientemente del medio social en el que se encuentre. En particular, se pretende mostrar el **número de seguidores, seguidos y publicaciones** de un usuario durante un período de tiempo determinado. Si este excede los siete días, se procede a calcular la media acumulada por semanas de cada uno de los campos mencionados anteriormente. En caso contrario, la información será visualizada directamente. Los diversos objetivos que se persiguen con la realización de este análisis se detallan a continuación.
 - En primer lugar se desea demostrar si existe una **relación entre el número de publicaciones y el número de seguidores**. Hasta cierto punto es lógico pensar que a mayor contenido, mayor interés puede suscitar al resto de miembros de la comunidad, y por lo tanto, mayor probabilidad hay de que se suscriban para seguir recibiendo las nuevas publicaciones. Sin embargo, existen otros motivos por los que esta relación puede no ocurrir, como por ejemplo, el tipo o la calidad del contenido. También se plantean otros factores externos a las redes sociales como por ejemplo la fama o popularidad que ya estuviese asociada a una entidad particular.
 - El segundo ámbito a analizar reside en determinar la **consolidación del usuario** dentro de un determinado medio social. Para ello, se puede observar la variabilidad de los tres campos implicados durante el período de tiempo seleccionado. Por norma general, las cuentas recién creadas tienden a sufrir graves alteraciones en el número de publicaciones, seguidores y seguidos. Mientras que los usuarios que ya se encuentran integrados en la comunidad suelen gozar de cierta estabilidad, la cual se refleja en unos valores más equilibrados de los tres campos mencionados.
- **Actividad del usuario.** Con este estudio se pretende poner de manifiesto la periodicidad con la que una determinada cuenta genera nuevo contenido. El cálculo de la media acumulada por semanas explicado en el anterior análisis, también es aplicable en este segundo estudio si el período de tiempo seleccionado en el análisis es superior a los siete días. El objetivo que persigue es doble.
 - Por un lado, a través del análisis del número de publicaciones que realiza durante un determinado tiempo, se puede determinar si existe algún patrón que indique que el responsable de la cuenta sigue una **planificación** concreta para subir nuevo contenido. Esta característica demostraría que el usuario tiene un importante interés en mantener su cuenta activa y lo más actualizada posible. Uno de los motivos más probables que explicaría esta situación consistiría en utilizar las redes sociales como un **medio publicitario** para dar a conocer sus productos y/o servicios.
 - Por otro lado, mediante el estudio de la actividad de la cuenta también se puede determinar el **grado de influencia** que el responsable ejerce sobre la comunidad que se encuentra a su alrededor. A mayor número de seguidores, mayor capacidad de difusión tiene su contenido. Las redes sociales son herramientas que, bien

utilizadas, pueden jugar un papel de divulgación muy beneficioso, como la capacidad de proporcionar noticias mundiales, apoyar campañas humanitarias o realizar denuncias sociales colectivas contra una causa común. Sin embargo, también pueden provocar consecuencias negativas procedentes de la difusión de información falsa y la manipulación, entre otros actos.

- **Evolución del interés de las publicaciones.** Este tercer análisis intenta representar el interés que suscitan las publicaciones de una determinada cuenta, basándose en el número de **interacciones recibidas** dentro de una duración temporal concreta, que si excede los siete días se le aplicará el cálculo de la media acumulada por semanas explicado en los análisis anteriores. En este estudio se permite la combinación de las diferentes acciones comunes a todas las redes sociales, como los me gusta y los comentarios, así como los particulares a las diferentes plataformas. Por lo que en este caso, el estudio se encuentra enfocado en la comunidad que se encuentra a su alrededor, en particular, al análisis de los diferentes modos de expresión que utilizan para dar su opinión acerca del contenido publicado.
 - Así, en primer lugar puede conocer cuál es el tipo de interacción que más reciben sus publicaciones durante un período determinado de tiempo, con el objetivo de poder idear una estrategia con la que **incentivar otro tipo de participación** que le sea más beneficiosa.
 - Además, también puede visualizar cuáles son los intervalos temporales en los que existe una mayor y una menor interacción por parte de los miembros de la comunidad. El principal objetivo reside en preparar un calendario adecuado que potencie la generación de nuevo contenido en aquellas épocas en las que los usuarios están más activos con el fin de poder **maximizar su difusión**.
- **Popularidad de las publicaciones.** El objetivo principal de este estudio consiste en presentar un *ranking* de las **diez publicaciones mejor o peor valoradas** en función de un conjunto de interacciones. Como en el caso anterior, se puede realizar una combinación entre las comunes y las particulares a la red social en la que estemos realizando el análisis. De este modo, permite conocer las características y el tipo de contenido que ha recibido un mayor número de interacciones, y por lo tanto, ha suscitado un mayor interés en la comunidad. Los resultados de este análisis pueden ayudar al responsable de la cuenta, a considerar las cualidades de las publicaciones más populares para integrarlas en el nuevo contenido que genere y así intentar **mejorar la opinión y difusión** del mismo. Sin embargo, también pienso que es importante conocer los motivos por los cuales el contenido publicado no ha recibido tanto apoyo. Es por ello por lo que este cuarto análisis también se puede orientar hacia el sentido opuesto, para detectar cuál ha sido el contenido que menor aceptación ha tenido. Así, el usuario también podrá analizar los motivos por los cuales ciertas publicaciones han caído en el olvido.
- **Análisis de sentimientos basados en texto.** En este quinto estudio se integran técnicas de análisis inteligente de datos para determinar la bondad de los títulos y los comentarios de las publicaciones. Así se cumple la primera parte del **cuarto objetivo** de este proyecto, cuyo principal fin es el de determinar el número de textos clasificados como positivos, neutrales o negativos en un período específico de tiempo. Por un lado, se trata de uno de los métodos más utilizados para estudiar el denominado **termómetro social**, con el que se puede determinar el grado de satisfacción o descontento que expresan los miembros de una red social. De este modo, el responsable de la cuenta puede conocer el efecto que produce su contenido sobre la comunidad que le rodea, determinando qué temáticas, productos o servicios son los que mejor o peor han sido valorados mediante los comentarios publicados por el resto

de usuarios. Sin embargo, al aplicar un análisis de sentimientos sobre los títulos de las publicaciones que escribe el propio usuario en estudio, también se puede entrever la **intencionalidad y el tipo de contenido que publica** en las redes sociales, a través de la descripción que presenta su contenido al resto de los miembros de la comunidad. De esta manera, también se establece una especie de termómetro particular a una determinada cuenta con el objetivo de conocer el tipo de conducta que demuestra durante un período de tiempo concreto.

- **Análisis de patrones de comportamiento.** Finalmente, este último análisis se encuentra ligado al anterior, en tanto en cuanto es capaz de identificar el **tipo de usuarios** existentes en una comunidad, a partir del comportamiento que se refleja en sus comentarios. Debido a la variedad de sentimientos que se pueden considerar para clasificar los textos de las publicaciones, en este estudio se realiza un conteo del número de usuarios que mayoritariamente redactan comentarios positivos, y de aquellos miembros que, por el contrario, se dedican a realizar una mayor cantidad de críticas en las publicaciones de una determinada cuenta. Así, el responsable puede conocer y analizar los motivos por los que en ciertos periodos su contenido ha provocado más reacciones positivas, mientras que en otras épocas sus publicaciones han suscitado un mayor malestar entre sus seguidores.

Capítulo 4

Metodología de desarrollo

Uno de los pilares fundamentales en el desarrollo de un producto o servicio de cualquier tipo es la planificación. En casi cualquier ámbito existe un conjunto de metodologías que proporcionan los pasos a seguir para conseguir un resultado de calidad y con la mínima inversión monetaria y temporal. En particular, en el mundo de la informática se han ideado una gran variedad de metodologías de desarrollo, debido a la diversidad y la complejidad de los productos y servicios que se pueden generar. Las **metodologías tradicionales** son las que apuestan por un desarrollo basado en etapas cíclicas, que comienzan con la descripción del objetivo, la planificación, la implementación y testeo, la monitorización y finalizan con el despliegue o entrega al cliente. Generalmente, se caracterizan por su claridad en la definición de los requisitos y por un mayor control en el proceso de desarrollo y documentación. Por otro lado, las **metodologías ágiles** se centran en el factor humano, fomentando la comunicación y colaboración entre los miembros del equipo de desarrollo, además de considerar al cliente como una entidad fundamental para la realización del proyecto. Gracias al continuo flujo de información entre ambas partes, esta metodología proporciona la capacidad de responder rápidamente a los cambios de requisitos con el mínimo gasto de recursos [48] [49].

Tras conocer los detalles de las diferentes alternativas disponibles, la que más se ha ajustado a mi planificación es la **incremental**. Se trata de una metodología tradicional basada en un conjunto de etapas que se repiten durante varias iteraciones hasta la finalización del sistema. La siguiente imagen define las cuatro fases que propone esta metodología así como el orden en el que se llevan a cabo dentro de cada uno de los ciclos de desarrollo. En la primera etapa de análisis se identifican los nuevos requisitos que se desean integrar en el sistema, para posteriormente implementar su diseño y testear su comportamiento final [50].



Ilustración 8: Esquema representativo de la metodología incremental de desarrollo. Fuente: [50].

Uno de los principales motivos por los que elegí esta metodología es por la posibilidad de comenzar a **utilizar el producto sin la necesidad de estar finalizado por completo**. De esta forma, tras integrar la descarga y el almacenamiento de información procedente de redes sociales, la plataforma comenzó a explotar esta funcionalidad para obtener los volúmenes de

datos que posteriormente serían estudiados, mientras que al mismo tiempo, estaba planteando y diseñando los análisis que estarían disponibles así como su almacenamiento en una base de datos relacional. Esta metodología facilita la división del sistema en varios módulos de modo que, en cada iteración, solo debemos de ocuparnos de uno en concreto [50].

Tras detallar la metodología de software que se ha aplicado para planificar el proyecto, a continuación se explica la metodología propia que se ha seguido para desarrollar este trabajo. Para ello me he basado en las diferentes fases que se llevan a cabo en los análisis de información [36].

1. **Revisión del estado del arte.** Como se ha podido apreciar en la segunda parte del segundo capítulo, se ha realizado un estudio acerca de las características y funcionalidades que ofrecen las diferentes herramientas de análisis de redes sociales existentes en el mercado. El principal objetivo de esta revisión consiste en verificar sus fortalezas y debilidades para intentar plantear un proyecto que aporte un valor añadido a los programas analíticos que se encuentran disponibles actualmente.
2. **Planificación de objetivos.** A continuación se establecen los objetivos generales y específicos que se pretenden cumplir para generar un producto acorde con la idea inicial que se planteaba en este proyecto.
3. **Estudio de recursos.** En esta tercera fase se ha realizado, en primer lugar, una revisión acerca de las fuentes de información disponibles para obtener datos de redes sociales. Este paso es fundamental para plantear aquellos análisis de datos que mayor conocimiento útil puedan proporcionar en función de la información que se pueda extraer. Otro de los aspectos que ha influido en la elección de las fuentes de datos ha sido el nivel analítico que han sufrido las redes sociales más populares en los últimos años. Mientras que Twitter ha protagonizado la mayoría de estudios de comunidades virtuales, otras plataformas tan influyentes como Instagram han sido menos estudiadas. Del mismo modo, se ha efectuado una revisión acerca de los diferentes algoritmos existentes para realizar análisis de sentimientos basados en texto. Así, he podido conocer qué tipo de librerías y herramientas tengo a mi alcance, cuál es su funcionamiento y sus limitaciones con el fin de poder realizar diversos experimentos para decidir cuáles son las que mejor se adaptan a este proyecto.
4. **Diseño y desarrollo de la aplicación.**
 - a) Tras conocer la disponibilidad de las diferentes fuentes de información y la diversidad de datos a la que se puede acceder, se lleva a cabo el diseño de los análisis estadísticos e inteligentes que se encontrarán disponibles en la aplicación.
 - b) Una vez han sido planteados, se implementa la extracción, filtrado y almacenamiento de la información necesaria para cada uno de ellos con el objetivo de disminuir la inversión temporal y computacional asociada a los futuros procesos analíticos [37]. Como se mencionó anteriormente en la metodología de software elegida, este proceso de descarga es el primero que comienza a funcionar para recopilar información sobre varias cuentas situadas en la red social seleccionada como fuente de datos: Instagram.
 - c) El tercer paso consiste en definir el formato en el que los datos deben ser proporcionados como entrada para los diferentes análisis existentes. En particular, destaca el preprocesamiento asociado a los textos a los que se pretende aplicar un análisis de sentimientos. Las técnicas de limpieza y preparación de los textos dependerán del tipo de algoritmos que se vayan a utilizar. Este proceso se encuentra detallado en el apartado 5.5.3 de esta memoria.

- d) A continuación se plantea el flujo de operaciones y datos propio de cada uno de los análisis que se pretenden implementar. De este modo, se reflexiona acerca de los motores de bases de datos que mejor se pueden adaptar al almacenamiento de la información preprocesada y de los resultados analíticos. Como se ha podido apreciar en la descripción de cada uno de los análisis disponibles, en la mayoría se pretende establecer una relación entre diferentes tipos de datos. Es por ello por lo que se diseña un esquema relacional, identificando las diferentes entidades y sus interconexiones en base a la información que se pretende analizar y a los resultados obtenidos. El objetivo consiste en maximizar la reutilización tanto de los datos preprocesados como los resultados de análisis anteriores para facilitar la ejecución de los futuros, disminuyendo el coste temporal y computacional.
 - e) Finalmente se reflexiona acerca de los diferentes tipos de visualizaciones que se suelen utilizar para mostrar los resultados obtenidos a partir de estudios analíticos. Por norma general, la mayoría de herramientas suelen optar por gráficas estadísticas que facilitan tanto la representación como la interpretación de los datos [37] [38]. Sin embargo, para este proyecto se ha intentado aportar cierta singularidad incluyendo gráficas poco corrientes con el objetivo de conseguir encontrar un equilibrio entre la originalidad de las visualizaciones y la facilidad para su interpretación y comprensión por parte de los usuarios finales.
5. **Análisis de un caso práctico.** Una vez la herramienta había sido desarrollada completamente, se aplicaron los diferentes análisis disponibles sobre el conjunto de datos recopilados sobre una cuenta de Instagram con el objetivo de poder estudiar los resultados obtenidos y extraer las posibles conclusiones e interpretaciones posibles para cada uno de ellos.
6. **Redacción de la memoria.** A pesar de ser la última fase, en realidad se ha ido desarrollando en paralelo con las anteriores de modo que se quedase constancia de los avances que se realizaban en los ámbitos de investigación, desarrollo y análisis de información sobre un usuario concreto.

Capítulo 5

Desarrollo del proyecto

Este capítulo está dotado de un enfoque más técnico puesto que su principal objetivo consiste en proporcionar los detalles relativos al planteamiento del proyecto, al desarrollo de la herramienta asociada así como a la redacción de la documentación necesaria. Para explicar las diferentes etapas que han permitido la realización de este proyecto, y en función de la metodología seleccionada anteriormente, se ha organizado el contenido de este capítulo en las siguientes secciones:

1. Planificación y presupuesto sobre el desarrollo del proyecto.
2. Identificación de todo tipo de requisitos.
3. Descripción de los casos de uso.
4. Diseño del sistema.
5. Implementación de la plataforma.
6. Manual de usuario de la herramienta.

5.1. Planificación del proyecto

Uno de los componentes fundamentales a la hora de desarrollar cualquier tipo de proyecto es la planificación previa. Si la programación definida es viable y realista, puede ayudar a reducir considerablemente los recursos temporales, humanos y monetarios de modo que se minimicen los riesgos que puedan surgir. Existen diversas herramientas que se pueden utilizar para definir la planificación de un proyecto. Una de ellas es el **diagrama de Gantt**, que permite representar el conjunto de actividades que se deben realizar así como el tiempo asignado a cada una de ellas para su finalización. En la ilustración 9 se presenta la programación temporal que se ideó al comienzo del desarrollo de este proyecto. Como se puede observar, las primeras tareas se encuentran relacionadas con la búsqueda de **documentación** sobre el funcionamiento de las **redes sociales**, sus componentes y la información que contienen cada una de ellas. Asimismo, también se ha realizado una investigación acerca de las **herramientas** actuales de análisis de medios sociales con el objetivo de conocer qué funcionalidades proporcionan.

Tras estudiar las diferentes entidades involucradas en este proyecto y las herramientas existentes en el mercado, a continuación se definen los diversos **casos de uso** asociados, con el objetivo de especificar las acciones que podría realizar el usuario con la plataforma. Una vez conocemos las diversas operaciones que se encontrarán disponibles, es necesario plantear la arquitectura general del sistema. En mi caso, he optado por un **diseño modular** que se encuentra constituido por diferentes componentes con el principal objetivo de facilitar el desarrollo por separado de cada uno, para posteriormente, conectarlos entre sí de modo que funcionen como si fuesen uno. De acuerdo a la metodología de desarrollo que he seleccionado y explicado en el capítulo anterior, en primer lugar he comenzado con el diseño, implementación y testeo del **módulo de datos**. Este primer componente se centra en la **descarga de información** procedente de las diferentes APIs que se pueden integrar para obtener datos sobre usuarios de redes sociales. Asimismo, también comprende las técnicas de **procesamiento** requeridas para filtrar los campos obtenidos, de modo que solo almacenemos los necesarios para realizar los análisis, además de verificar su tipología y su rango de valores.

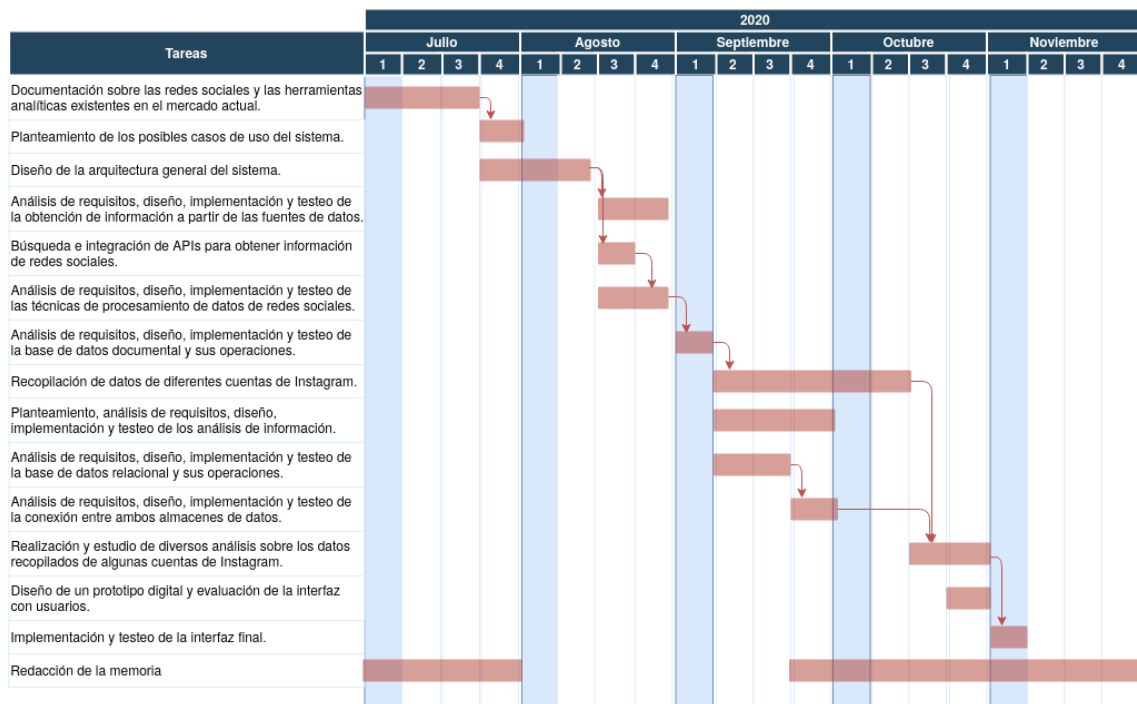


Ilustración 9: Diagrama de Gantt que se planteó al principio del proyecto. Fuente: realización propia.

Una vez integrado el proceso de descarga y procesamiento, continuamos con el desarrollo del **segundo módulo orientado al almacenamiento**. En primer lugar, comenzamos con el diseño de la base de datos documental y sus operaciones con el fin de empezar la recopilación y el almacenamiento de datos de usuarios. Así, podremos acumular un volumen de información suficiente como para, posteriormente, realizar diversos análisis que nos permitan extraer conocimiento útil y nos ayude a obtener conclusiones valiosas.

De nuevo, retomamos el diseño del módulo de datos para incluir el último componente en el que se plantean e integran los diversos **análisis de datos** que se encontrarán disponibles en la plataforma. Todos ellos han sido detallados en el capítulo tercero de este documento. Al mismo tiempo, finalizamos el módulo de almacenaje diseñando, implementando y testeando las operaciones necesarias para trabajar con una **base de datos relacional** en la que almacenar la información a analizar, así como los resultados obtenidos de los estudios realizados. El objetivo de esta última acción consiste en reutilizarlos parcialmente para análisis similares, así como representarlos gráficamente de forma directa, lo que provocaría una disminución de la inversión temporal y computacional. Tras incorporar este último almacén de datos, se integra una conexión entre él y la base de datos documental para transferir la información recopilada de las diferentes fuentes de información, prepararla y jerarquizarla con el fin de facilitar los análisis que se le vayan a realizar posteriormente.

Tras completar el desarrollo de la herramienta y recopilar una cantidad de datos suficientes, comenzamos a **realizar diversos análisis** de todo tipo para estudiar sus resultados y extraer las conclusiones dependiendo de los objetivos de cada uno. Asimismo, comienza el **diseño de la interfaz del sistema y el flujo de operaciones** tal y como he cursado en la asignatura del máster **Desarrollo y Evaluación de Sistemas Software Interactivos**. Una vez disponemos de los bocetos finales relativos a la interfaz de la plataforma, así como del orden de las acciones para cada funcionalidad que esté disponible en la herramienta, se genera un **prototipo digital** no funcional para evaluar su calidad e interacción. Para ello escogí a tres usuarios, uno experto en el ámbito de la informática, otro usuario de nivel medio pero gran conocedor de las nuevas tecnologías y una tercera persona sin apenas habilidades informáticas. Cuando ya ha sido verificado, se modifican los elementos necesarios para corregir las posibles deficiencias que presente el prototipo para desarrollar y probar la interfaz del sistema.

Sin embargo, la planificación anterior no se pudo llevar a cabo por diversas dificultades que surgieron a lo largo del desarrollo del proyecto. En primer lugar, si bien esta plataforma se encuentra orientada hacia el análisis de datos procedentes de cualquier red social, hemos elegido **Instagram** para realizar los ejemplos que manifiestan el funcionamiento de esta herramienta, puesto que es una de las redes sociales que más importancia está cobrando en la actualidad y que, sin embargo, **no ha sido analizada tan en profundidad** como es el caso de Twitter. Como se comentó anteriormente, este medio social solo dispone de una API que proporciona información muy limitada. Por lo tanto, el objetivo consistía en encontrar una **librería de terceros** conectada a la API privada de Instagram para poder obtener una gran cantidad de datos de diversa naturaleza. Si bien existen una multitud de proyectos abiertos orientados a esta temática, la gran mayoría de los que he probado no disponen de un proceso de identificación correcto. O bien produce un error, o es necesario iniciar sesión en cada petición que se realiza, lo cual provoca el bloqueo temporal de la cuenta que se esté usando. Por lo tanto, como se puede observar en la ilustración 10, la búsqueda de una API de Instagram pública y con suficiente diversidad de información, **se demoró bastante** en el tiempo comparado con la planificación inicial.

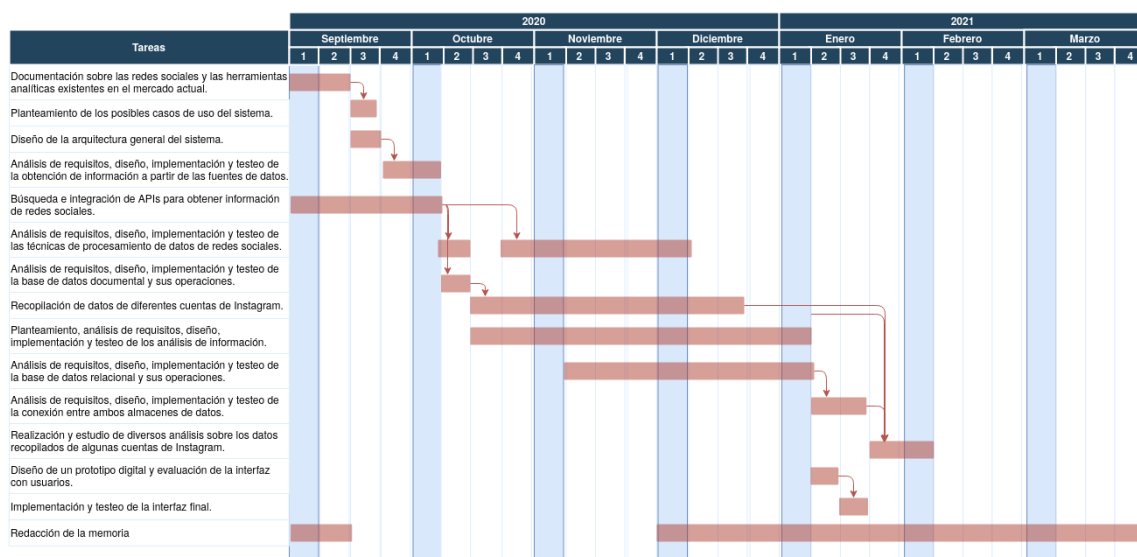


Ilustración 10: Diagrama de Gantt de la planificación real del proyecto. Fuente: realización propia.

Asimismo, otro de los inconvenientes que originó cierto retraso en el desarrollo del proyecto fue el diseño de los **análisis**, su procesamiento y almacenamiento. Se han planteado diferentes enfoques cuyos estudios han tomado más tiempo del que pensaba en un principio. La razón de ser es que no podía comprobar su viabilidad hasta que no estuviesen implementados y listos para su testeo. De igual modo, el procesamiento de la información generada y su almacenamiento se han tenido que ir adaptando a los diversos cambios que realizaba en la estructura de los análisis. Una vez había escogido el diseño que me pareció más adecuado, también he tenido que aplicar diversas **optimizaciones al esquema relacional** que almacenaba la información a analizar y los resultados de los estudios. En este paso, de nuevo, tuve que invertir más tiempo de lo que pensaba en realizar varios estudios para determinar aquel que se caracterizase por una mejor relación entre el conjunto de requisitos que se debían de cumplir, y la calidad y eficiencia del diseño relacional.

5.2. Presupuesto del proyecto

A partir de esta última planificación, se pueden estimar los costes económicos asociados al desarrollo del proyecto, considerando los aspectos relativos al personal, al esfuerzo y a los recursos utilizados. Con respecto al primer ámbito, procedemos a estimar el coste económico que supondría el desarrollo de este trabajo si dispusiese de un contrato indefinido. Asumiendo que el

salario bruto es de 1.500€/mes, el coste desglosado de la Seguridad Social sería el que se muestra a continuación.

- Considerando un 23,60% de **contingencias comunes** asociadas a la asistencia sanitaria del trabajador cuando este se encuentre enfermo, el coste calculado sería de 354€/mes [51] [52].
- Con una cuota de **desempleo** de un 5,5%, el gasto estipulado sería de 82,5€/mes [51] [52].
- Sumando un 0,2% del **FOGASA** o Fondo de Garantía Salarial, que se encarga de gestionar tanto la percepción monetaria durante el contrato así como durante el despido, el gasto asciende a 3€/mes [51] [53].
- Por último, es necesario añadir el 0,7% del salario bruto asociado a la **formación profesional** a la que tiene derecho el trabajador por un importe de 10,5€/mes [51] [54].

Todo ello desemboca en un coste de **1.950€ al mes** por una única persona contratada, lo que en total alcanzaría los **13.650€ por los siete meses** de duración del proyecto.

A continuación, procedemos a calcular los gastos asociados a los materiales y servicios que se han utilizado durante el período de desarrollo. Existen dos categorías principales, según pude estudiar en la asignatura del máster **Planificación y Gestión de Proyectos Informáticos**. Por una parte, disponemos de los siguientes **materiales inventariables**, de los cuales sus gastos asociados vendrán computados por la amortización que se pueda llevar a cabo a partir de su precio total, su duración estimada y su uso durante los seis meses que dura el desarrollo del proyecto:

- Un **portátil** valorado en 1.000€ con una vida útil estimada de 5 años supondría un coste de 200€/ año o 16,67€/mes, lo que significa que para este proyecto su gasto total asociado sería de **116,69€**.
- Una **mesa de escritorio** estimada en 80€ con una vida útil aproximada de 10 años tendría asociada un gasto de 8€/año o 0,67€/mes, añadiría un coste de **4,69€** al total del proyecto.
- Una **silla** tasada en 100€ con una vida útil de alrededor de 7 años constituye un gasto de 14,28€/año o 1,19€/mes, por lo que suma una cantidad de **8,33€**.

Por otro lado, existe una segunda clasificación referente a los **materiales fungibles**, cuyos gastos sí se encuentran asociados al coste de los mismos, puesto que se trata de bienes que, a diferencia de los anteriores, no disponen de una vida útil si no de un uso constante hasta finalizar el producto. A continuación se detalla la lista de los materiales fungibles implicados en el desarrollo de este proyecto:

- Material genérico de **papelería**, como bolígrafos y folios entre otros, supone un gasto estimado de **50€** en total.
- Un **ratón** inalámbrico para el ordenador supone un gasto de **11€**.
- Una **alfombrilla** para facilitar el uso del ratón tiene un coste asociado de **2€**.

Asimismo, es necesario incluir el importe asociado a los **servicios** básicos que se han estado utilizando durante el período en el que se ha trabajado sobre el proyecto. Entre otros se incluyen el suministro eléctrico o la tarifa de internet, lo cual se estima en una cantidad de 80€/mes por lo que supondría un coste total de **560€**.

Sumando los gastos calculados para cada uno de los productos y servicios anteriormente descritos, el **importe total de los materiales utilizados durante el proyecto es de 752,71€**.

En la tabla 1 se resumen todos los costes explicados anteriormente con el objetivo de clarificar el importe total del desarrollo del proyecto durante los siete meses que ha durado.

Total de gastos de personal.	13.650€
Gastos del personal de la trabajadora al mes.	1.950€
Total de gastos de ejecución.	752,71€
Costes de adquisición de material inventariable.	129,71€
Un portátil.	116,69€
Una mesa de escritorio.	4,69€
Una silla ergonómica.	8,33€
Costes de adquisición de material fungible.	63€
Material de papelería.	50€
Un ratón inalámbrico para el ordenador.	11€
Una alfombrilla para el ratón.	2€
Gastos operativos	560€
Servicios básicos de suministro eléctrico, tarifas de internet, etc.	560€
TOTAL	14.935,71€

Tabla 1. Resumen de los costes asociados al desarrollo del proyecto.

Finalmente, podría ser interesante estimar el coste mensual que supondría el despliegue de la plataforma implementada en un proveedor en la nube, como por ejemplo **Azure**.

Si seleccionamos una instancia perteneciente a la serie Bs, que se corresponde con la gama de máquinas virtuales económicas y escalable, de dos núcleos de CPU, 4 GB de RAM, 8 GB de almacenamiento con sede en la zona oeste de EEUU y con un sistema operativo Linux, supondría un gasto de **29,92€/mes** [55].

5.3. Extracción de requisitos

En la tercera sección de este capítulo, presentamos los diferentes tipos de requisitos que debe satisfacer el sistema que se implementa en este trabajo, de acuerdo con los objetivos anteriormente planteados.

Requisitos de datos

1. Un esquema de información general en base a los datos comunes que se encuentran en todas las redes sociales.
2. Un sistema de filtros para capturar únicamente los datos que aparezcan en el esquema planteado anteriormente.
3. Un esquema de procesamiento que sea capaz de verificar los datos obtenidos de las diferentes fuentes de información, con el objetivo de comprobar la validez de los tipos y los rangos de sus valores.
4. Una primera base de datos que permita almacenar la información recopilada de las diferentes redes sociales sin la obligación de especificar una estructura previa.
5. Las técnicas de procesamiento necesarias para procesar la información textual que vaya a ser utilizada para realizar los análisis de sentimientos.

6. Un esquema relacional en el que aparezcan las distintas entidades, su información asociada y las conexiones existentes entre ellas. Esta tarea debe considerar los datos que van a ser analizados así como los que serán generados tras realizar los estudios.
7. Un segundo almacén de información para guardar los datos procesados que se pretenden analizar, así como los resultados generados para su posterior uso y visualización.
8. Una conexión entre los dos almacenes de datos para migrar la información indispensable para realizar el análisis solicitado bajo demanda.

Requisitos funcionales

1. El sistema permitirá al usuario especificar la cuenta de la que desea obtener su información disponible procedente de las fuentes de datos disponibles, que en este proyecto se particulariza a Instagram.
2. El usuario podrá seleccionar el análisis que desea efectuar y configurar los parámetros necesarios para realizarlo.
3. El usuario podrá interactuar con las gráficas estadísticas que representen los resultados de los análisis que se han realizado.
4. El usuario podrá descargar en una imagen la representación gráfica que muestra los resultados del análisis que se ha efectuado.

Requisitos no funcionales

1. El sistema deberá realizar el proceso de extracción, filtrado, procesamiento y almacenamiento de información en segundo plano, permitiendo al usuario utilizar la plataforma al mismo tiempo.
2. El sistema deberá ser capaz de procesar y jerarquizar únicamente la información necesaria para cada análisis bajo demanda, reduciendo al máximo la redundancia de información y haciendo un buen uso del espacio de almacenamiento.
3. El sistema debe permitir la integración de diferentes fuentes de datos, siempre y cuando se respete el esquema de información general, o se defina uno nuevo en caso de que sean datos específicos de una red social particular.
4. La interfaz del sistema debe ser lo más sencilla posible con el fin de que cualquier tipo de usuario pueda hacer uso de ella sin la necesidad de realizar un sobre-esfuerzo.
5. La plataforma deberá de contener una descripción acerca del cometido de cada sección con el objetivo de informar al usuario sobre cada una de las funcionalidades disponibles.
6. La arquitectura del sistema se organizará en diversos módulos que deben funcionar tanto independientemente como de forma conjunta, de modo que se facilite el posterior mantenimiento e incorporación de nuevas funcionalidades.
7. El despliegue del sistema deberá ser un proceso automático en el que se contemplen los siguientes requerimientos.
 - a) La construcción y configuración de los contenedores necesarios, para disponer de los almacenes de datos y de la lógica de negocio de la aplicación.

- b) La instalación de las dependencias y librerías imprescindibles para su correcto funcionamiento.
- c) La configuración de la persistencia de datos para almacenar la información de manera independiente a los contenedores.

5.4. Casos de uso

El objetivo de este apartado consiste en especificar el conjunto de funcionalidades que se encontrarán disponibles en la plataforma. En la ilustración 11 se pueden observar las diversas operaciones que se podrán llevar a cabo, desde la extracción automática de información asociada a una cuenta, la realización de cualquiera de los análisis disponibles, hasta la descarga de los resultados gráficos obtenidos del último estudio realizado.

Aquellas personas que deseen utilizar esta herramienta no tendrán que poseer conocimientos específicos, ni experiencia previa con sistemas informáticos para poder desenvolverse correctamente. Asimismo, tampoco se requiere su registro ni introducción de datos personales previo al uso de la plataforma.

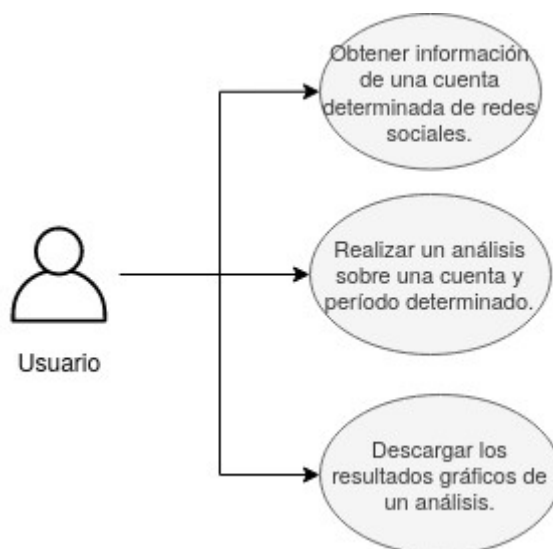


Ilustración 11: Diagrama de casos de uso asociado a la plataforma. Fuente: realización propia.

A continuación, se pretenden detallar todos los casos de uso expuestos en el diagrama de la ilustración 11 con el objetivo de conocer, en profundidad, su información asociada.

Caso de uso	Extracción de datos de una cuenta determinada.	
N.º de caso de uso	CU-1.	
Actores	Usuario.	
Tipo	Primario, esencial.	
Referencias	RD-1, RD-2, RD-3, RD-4, RF-1, RNF-1, RNF-4, RNF-5.	
Descripción	El usuario podrá especificar la cuenta de la que obtener información, así como las diferentes redes sociales en las que el sistema deberá buscarla.	
Precondición		
Secuencia Normal	Paso	Acción
	1	El usuario se dirige hacia la primera sección del menú lateral izquierdo.
	2	El sistema muestra la pantalla asociada a la recolección de datos procedentes de redes sociales.
	3	El usuario escribe el nombre de la cuenta de la que desea descargar sus datos y selecciona las redes sociales en las que desea buscar dicha información.
	4	El usuario pulsa sobre el botón apropiado para guardar la nueva configuración.
	5	El sistema consulta la cuenta de la que comenzar la extracción de información y las fuentes en las que buscarla para iniciar el proceso.
	6	Una vez obtiene los datos de la cuenta, el sistema filtra y preprocesa la información de acuerdo al esquema de datos definido.
	7	Finalmente, añade un identificador único además de la fecha de extracción y los almacena en la base de datos documental.
Postcondición	El número de datos almacenados aumentará proporcionalmente a la nueva información recopilada.	
Excepciones	Paso	Acción
	6a	El sistema no encuentra información sobre la cuenta proporcionada y por tanto continúa con la siguiente fuente seleccionada.
	6b	El sistema no encuentra información sobre la cuenta proporcionada y no hay más fuentes seleccionadas, por lo que termina el proceso.
Rendimiento	No se le exige un tiempo determinado puesto que es una operación costosa y depende de la disponibilidad y de la latencia de las fuentes de información que se consulten.	
Frecuencia esperada	Alta.	
Importancia	Vital.	
Urgencia	Baja.	
Comentarios	Este proceso se realiza en segundo plano con el objetivo de que el usuario pueda continuar utilizando la plataforma, mientras la extracción de la nueva información se lleva a cabo.	

Caso de uso	Realizar un análisis determinado.	
N.º de caso de uso	CU-2.	
Actores	Usuario.	
Tipo	Primario, esencial.	
Referencias	RD-6, RD-7, RD-8, RF-2, RF-3, RNF-1, RNF-2, RNF-4, RNF-5.	
Descripción	El usuario podrá seleccionar uno de los análisis disponibles para aplicarlo sobre una cuenta determinada y dentro de un período de tiempo concreto.	
Precondición	Que existan datos de la cuenta especificada en el rango de tiempo introducido.	
Secuencia Normal	Paso	Acción
	1	El usuario se dirige hacia la opción del análisis deseado que se encuentra en el menú lateral izquierdo.
	2	El sistema muestra la pantalla asociada al análisis escogido.
	3	El usuario selecciona la cuenta a estudiar y el período de tiempo en el que realizar el estudio.
	4	Una vez se ha especificado la configuración necesaria, el usuario pulsa sobre el botón asociado al comienzo del análisis.
	5	El sistema comprueba si el tipo de análisis requerido ya se ha realizado anteriormente.
	6	El sistema recupera los resultados del análisis y los muestra directamente de manera gráfica e interactiva.
Postcondición	Una nueva gráfica interactiva aparece en la pantalla actual mostrando los resultados del análisis realizado.	
Excepciones	Paso	Acción
	6a	El sistema detecta que el análisis no se ha realizado anteriormente e inicia el proceso desde el principio.
	6b	El sistema detecta que no dispone de la información requerida para realizar el análisis y por tanto, avisa al usuario del suceso y comienza con la extracción de los datos faltantes.
Rendimiento	En caso de que el análisis se haya realizado previamente, la representación de su resultado es inmediata. Si no existe, se demorará algo más, dependiendo de la complejidad de la información a estudiar.	
Frecuencia esperada	Alta.	
Importancia	Vital.	
Urgencia	Muy alta.	
Comentarios		

Caso de uso	Descarga de los resultados gráficos de un análisis.	
N.º de caso de uso	CU-3.	
Actores	Usuario.	
Tipo	Primario, esencial.	
Referencias	RF-4, RNF-4, RNF-5.	
Descripción	El usuario será capaz de descargar los resultados del análisis actual en una imagen.	
Precondición	Que se haya realizado un análisis previamente.	
Secuencia Normal	Paso	Acción
	1	El usuario pulsa sobre el primer icono de la gráfica que representa los resultados del último análisis realizado.
	2	El sistema escribe los resultados en una imagen y envía el fichero al navegador para posibilitar su descarga.
	3	El usuario recibe el archivo creado para descargarlo en su sistema.
Postcondición	Se ha creado una nueva imagen que contiene los resultados gráficos de un análisis determinado.	
Excepciones	Paso	Acción
Rendimiento	La imagen se crea y se envía para su descarga en cuestión de segundos.	
Frecuencia esperada	Baja.	
Importancia	Baja.	
Urgencia	Media.	
Comentarios		

5.5. Diseño de la plataforma

5.5.1. Esquema de información

En este quinto apartado comenzaremos describiendo el esquema de información que se ha planteado para la extracción de datos procedentes de redes sociales. El objetivo consiste en adaptar la **información común** de cualquier medio social a una misma estructura, de manera que las técnicas de procesamiento, almacenamiento y análisis sean **independientes de las fuentes de datos**. Una vez se han fijado los objetivos y los análisis que se pretenden implementar para ejemplificar el funcionamiento de este sistema, es necesario realizar un estudio acerca de la información que se puede encontrar en las redes sociales. En la tabla 2 se pueden observar las entidades pertenecientes a las tres plataformas virtuales más utilizadas a nivel global. Todas aquellas que se encuentran resaltadas serán las que se contemplen en la plataforma asociada a este proyecto.

Facebook	Instagram	Twitter
Perfil personal	Perfil personal	Perfil personal
Publicaciones	Publicaciones	<i>Tweets</i>
Seguidos	Seguidos	Seguidos
Seguidores	Seguidores	Seguidores
Historias	Historias	Historias
	<i>Hashtags</i>	

Tabla 2. Entidades comunes a tres de las redes sociales más populares, destacando aquellas que se pretenden añadir en este proyecto.

En primer lugar, los tres medios sociales tomados como ejemplo disponen de un **perfil de usuario** que se genera a partir de los datos personales que se proporcionan al crear una cuenta. Asimismo, esta entidad también se compone de otro tipo de información que, generalmente, se produce de forma automática como el identificador del usuario o la fecha de registro. De entre todos los posibles datos personales asociados a un perfil, a continuación en la tabla 3 se muestran aquellos que son comunes a las tres redes sociales de ejemplo, resaltando los que son considerados de interés para la plataforma.

Facebook	Instagram	Twitter
Nombre y apellidos	Nombre y apellidos	Nombre y apellidos
Foto de perfil	Foto de perfil	Foto de perfil
Nombre de usuario	Nombre de usuario	Nombre de usuario
	Sitio web	Sitio web
Biografía	Biografía	Biografía
Ubicaciones		Ubicaciones
Fecha de nacimiento	Fecha de nacimiento	Fecha de nacimiento
Fecha de registro	Fecha de registro	Fecha de registro
Género		
Número de publicaciones	Número de publicaciones	Número de publicaciones
Número de seguidores	Número de seguidores	Número de seguidores
Número de seguidos	Número de seguidos	Número de seguidos

Tabla 3. Datos de los perfiles de usuario de las tres redes sociales más populares, destacando aquellos que se pretenden añadir al proyecto.

En primer lugar, el **nombre de usuario** es extraído con el objetivo de identificar unívocamente a una cuenta dentro de cada una de las comunidades virtuales existentes para extraer su información asociada. Mientras que por otro lado, también almacenaremos directamente el **número total de publicaciones subidas, seguidores y seguidos**. Este tipo de información puede ayudar a la extracción de diversas conclusiones, como si existe una estrategia particular para subir nuevo contenido, la capacidad de influencia que puede ejercer en función de los usuarios que se han suscrito para visualizar su contenido, o el interés que tiene en otras cuentas a partir del conjunto de seguidos que ha ido acumulando.

La segunda entidad disponible son las **publicaciones**. Dependiendo del tipo de red social, el contenido de las publicaciones puede ser de diversa naturaleza en función de los propósitos que persigan las distintas plataformas virtuales. Sin embargo, en la gran mayoría de ellas, así como en los tres medios tomados como ejemplo, su contenido se caracteriza por ser multimedia, es decir, puede albergar tanto imágenes, vídeo como texto. A su vez, esta entidad permite la existencia de un conjunto de subentidades que representan las distintas **interacciones** que los usuarios pueden realizar sobre las publicaciones. Este tipo de información es muy valiosa, puesto que es la que permite estudiar y caracterizar una cuenta determinada a partir de las opiniones del resto de miembros de una comunidad. A continuación

se presentan en la tabla 4 las diversas interacciones disponibles en las tres redes sociales de ejemplo, resaltando aquellas que van a ser incluidas en la herramienta desarrollada.

Facebook	Instagram	Twitter
Me gusta	Me gusta	Favoritos
Comentarios	Comentarios	Comentarios
Mencionar usuarios	Mencionar usuarios	Mencionar usuarios
Etiquetar usuarios	Etiquetar usuarios	Etiquetar usuarios
Re-post		Retweets
Citar publicaciones		Citar tweets

Tabla 4. Datos de las publicaciones de las tres redes sociales más populares, destacando aquellos que se pretenden añadir al proyecto.

Una de las interacciones más comunes desde la aparición de este tipo de plataformas virtuales son los **me gusta**. Este tipo de dato permite conocer el nivel de agrado que provoca una publicación en los miembros que se encuentran alrededor de la cuenta que la ha publicado. Se trata de una métrica considerablemente sencilla que se puede utilizar para conocer rápidamente la opinión del resto de usuarios sobre un determinado contenido. Por otro lado, la posibilidad de que los usuarios puedan redactar sus pensamientos y opiniones acerca de una publicación permite tanto, profundizar en las emociones que provoca el contenido de una determinada cuenta, así como estudiar el tipo de comportamiento que tienen sus seguidores con respecto a las publicaciones. Es por ello por lo que los **comentarios** se posicionan como una subentidad fundamental para el desarrollo de este proyecto.

Finalmente, para completar el esquema de información común a todos los medios sociales, obtendremos la red social de **seguidores** de las cuentas a analizar, con el objetivo de estudiar e identificar los patrones de comportamiento que manifiestan las comunidades que se encuentran a su alrededor, en base a los dos tipos de interacciones comunes mencionadas anteriormente. De este modo, podremos conocer el tipo de usuarios que siguen a una cuenta determinada, además de sus gustos e intereses.

5.5.2. Flujo de procesamiento de datos

Una vez conocemos la estructura de la información común a las diferentes fuentes de datos, a continuación detallaremos las técnicas de procesamiento que se aplican sobre la información recopilada de cada fuente. Este flujo de preprocesamiento se encuentra representado en la ilustración 12.

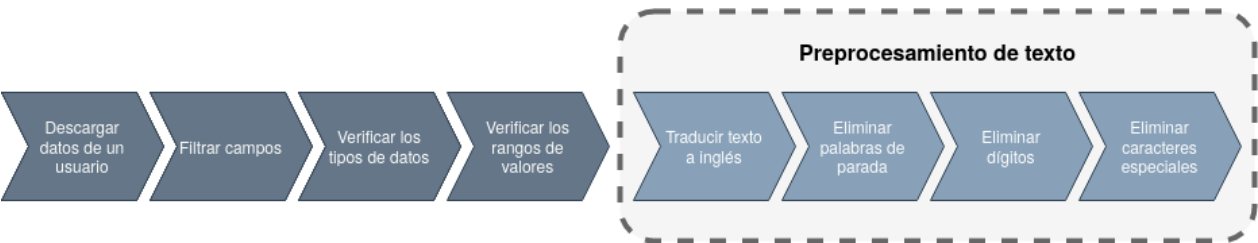


Ilustración 12: Flujo de procesamiento aplicado a la información obtenida de las redes sociales. Fuente: realización propia.

En primer lugar, construimos el esquema de datos general mencionado anteriormente. Para ello **filtramos la información descargada** obteniendo únicamente los campos de interés que nos ayudarán a realizar los análisis. En caso de que la fuente de datos no provea ningún valor para alguno de los campos definidos, se le asignará un valor por defecto. De este modo, solo

guardamos la información necesaria y gestionamos, de forma eficiente, el espacio de almacenamiento.

A continuación procedemos a comprobar si el **tipo de dato** recibido, para cada campo, es el esperado así como su **rango de valores**. Para ello, previamente, es necesario realizar un estudio sobre cada una de las fuentes con el fin de analizar qué tipo de información proporciona y en qué formato. El objetivo de esta fase de preprocesamiento consiste en evitar almacenar valores incorrectos que posteriormente dificulten o impidan obtener resultados válidos y fiables.

En última instancia, en caso de que el campo sea de naturaleza textual, deberá de sufrir un procesamiento específico con el objetivo de ser proporcionado como entrada al analizador de sentimientos. Por lo tanto, este procedimiento ha sido construido sobre los requisitos que exigen las librerías utilizadas, que serán detalladas más adelante. No obstante, tras el estudio de sus documentaciones y la realización de diversos experimentos para comprobar sus comportamientos frente a diferentes formas de procesamiento de textos, podemos concluir que las técnicas óptimas a aplicar son las siguientes por este orden:

- **Traducción del texto a inglés.** Por norma general, los diferentes tipos de analizadores suelen proporcionar mejores resultados cuando los textos se encuentran en inglés que en cualquier otro idioma. En el caso de las librerías que he utilizado, no han sido una excepción por lo que merece la pena invertir cierto tiempo en traducir los textos para mejorar la calidad de los resultados.
- **Eliminación de palabras de parada o *stopwords*.** Esta técnica es una de las más comunes en el procesamiento de textos, puesto que su fin consiste en eliminar cualquier palabra que no sea de utilidad a la hora de analizar un texto. Persigue un doble objetivo, por un lado evita desperdiciar tiempo en analizar palabras que no son relevantes para la identificación del sentimiento del texto, mientras que, a su vez, evita que la bondad de este tipo de palabras contaminen el resultado final.
- **Eliminación de dígitos.** Esta tercera técnica se encuentra en consonancia con la anterior puesto que, en este caso, los números no tienen asociado ningún sentimiento por lo que su eliminación puede ahorrar cierto tiempo.
- **Eliminación de caracteres especiales.** Para las librerías de análisis de sentimientos tradicionales suele ser conveniente la eliminación de todo tipo de signos de puntuación, exclamación, mayúsculas y caracteres especiales. Sin embargo, una de las que he utilizado sí que es capaz de analizar tanto los signos de exclamación como las mayúsculas en su beneficio para aportar un resultado más exacto. Ambos tipos de caracteres aumentan la intensidad del sentimiento identificado elevando, a su vez, el grado de confianza del analizador. Por ejemplo, si un determinado texto es considerado positivo, el hecho de que se encuentre acompañado de exclamaciones o esté escrito en mayúsculas, le otorga un plus de positividad.

5.5.3. Lenguajes y librerías utilizadas

El objetivo de esta sección consiste en detallar los lenguajes de programación y tecnologías que se han utilizado para posibilitar el desarrollo de la herramienta. En primer lugar, destacamos que tanto el *backend* como la interfaz han sido implementadas con **Python** comprobando su compatibilidad con las versiones 3.6 y 3.7 a través de la automatización de la ejecución de tests para cada una de las clases mediante herramientas de integración continua como **Travis**. La elección de este lenguaje se fundamenta en que se trata de uno de los más utilizados en el ámbito de la Ciencia de Datos gracias a la multitud de librerías existentes para la gestión, procesamiento y generación de información. Asimismo, dispone de

estructuras de datos altamente flexibles que permiten almacenar diferentes tipos de valores simultáneamente.

Para la **extracción de datos** procedentes de redes sociales, en particular de Instagram, se ha utilizado la librería **LevPasha** [29] como se comentó en el apartado 2.8.3 de este documento, a la que se le han añadido algunas modificaciones con el objetivo de mejorar la gestión de los inicios de sesión necesarios para poder identificarte como usuario antes de proceder con la descarga de información. En relación al modo en el que se realiza la extracción de datos, como se comentó anteriormente, el objetivo es llevar a cabo esta tarea en segundo plano de modo que el usuario pueda continuar utilizando la plataforma. Para ello he utilizado una librería denominada **Huey**, que proporciona un **servidor de tareas**, el cual dispone de un planificador con el que se puede programar la ejecución de tareas de forma automática, una cola de tareas en las que ir acumulando aquellas que no se han llevado a cabo aún y diversos mecanismos de memoria temporal en la que almacenar los datos recibidos [56].

Tal y como se ha explicado en el apartado anterior, la primera técnica de procesamiento aplicada a textos para posteriormente ser sometidos a análisis de sentimientos consiste en **traducir su contenido a inglés**. Para ello, he realizado numerosos experimentos con diversas librerías intentando medir la calidad de la traducción con diferentes tipos de textos, su robustez y tolerancia a errores ortográficos, así como la inversión temporal. Sin embargo, la gran mayoría de ellas se caracterizan por una **latencia** considerablemente alta, por lo que casi independientemente de la librería seleccionada, este paso se posiciona como un posible **cuello de botella** durante el procesamiento de la información textual. En primera instancia, una de las librerías que mejor relación calidad-tiempo he podido encontrar es **googletrans** [57], la cual utiliza internamente la API oficial de Google. Sin embargo, en una de sus últimas actualizaciones han realizado algunas modificaciones en el flujo de peticiones que no han sido adaptadas en esta librería por lo que generaba una gran cantidad de excepciones. No obstante, continuando con la búsqueda encontré una librería denominada **google-trans-new** [58], que sí contemplaba este nuevo sistema de peticiones para utilizar la API oficial de Google como recurso para la traducción. Por ello, ha sido la que he utilizado finalmente, además de por su flexibilidad a la hora de detectar automáticamente el idioma del texto a traducir.

En relación a los **analizadores de sentimientos** basados en texto he podido comprobar que existen dos tipos principales. Por un lado se encuentran los basados en **diccionarios**, que contienen un conjunto de términos asociados a sus correspondientes sentimientos e intensidades de los mismos. La principal ventaja reside en su gran velocidad de análisis pese a que, en ocasiones, la calidad de los resultados puede ser inferior. Una de las librerías más populares que aplica esta técnica analítica es **VADER** [59], cuyo funcionamiento se basa en la sumatoria de los sentimientos relativos a los términos del texto que se está analizando para, posteriormente, normalizar el resultado final entre -1, identificando el texto como negativo, y 1 que se consideraría positivo [60].

El segundo tipo de analizadores que he podido conocer son los **modelos predictivos** que, en su mayoría, se encuentran basados en técnicas de Aprendizaje Automático. Entre sus diversas ventajas podemos encontrar la posibilidad de generar un modelo particular para el problema que intentamos resolver, lo que proporciona un control total sobre el análisis de sentimientos. No obstante, en mi opinión la principal desventaja reside en crear un conjunto de entrenamiento con un número de textos etiquetados suficientemente amplio y representativo como para optimizar al máximo la precisión y la capacidad de generalización del modelo. Sin embargo, gracias al aumento de la importancia de la Ciencia de Datos, se han desarrollado analizadores inteligentes de sentimientos que se encuentran disponibles de manera pública. Una de las compañías especializadas en este ámbito es **MonkeyLearn**, que dispone de diversos analizadores de sentimientos tanto generales como específicos de diversas temáticas, como las redes sociales [61]. En principio proporcionan licencias gratuitas pero limitadas a un número escaso de peticiones al mes, por lo que debido al volumen de textos que puede analizar la plataforma no ha sido viable su incorporación.

Por otro lado, existen librerías que contienen modelos pre-entrenados listos para utilizarse y sin ningún tipo de limitación. Una de ellas es **TextBlob** [62], que dispone de un clasificador de textos basado en el algoritmo **Naive Bayes** entrenado mediante un conjunto de opiniones cinematográficas y sus correspondientes sentimientos identificativos [63]. Sin embargo, esta biblioteca también dispone de un analizador basado en **diccionario** como es el caso de **VADER**. La diferencia principal con esta librería es que **TextBlob** también contiene una tercera métrica consistente en determinar cuán **subjetivo** es el texto que se está analizando. Así, el algoritmo es capaz de identificar si se trata de una opinión personal o si, por el contrario, es un texto más objetivo [64]. Tras realizar diversos experimentos con esta biblioteca utilizando diferentes tipos de textos, puedo afirmar que los resultados obtenidos han sido inferiores a los proporcionados por la librería **VADER**, tanto en la calidad de la identificación de los sentimientos, como en la inversión temporal, especialmente en el caso del analizador basado en **Naive Bayes**, puesto que se demora enormemente en el tiempo.

Finalmente encontré una tercera librería denominada **flair** [65], la cual también proporciona clasificadores pre-entrenados especializados en la identificación de sentimientos. Sin embargo, a diferencia de las bibliotecas anteriores, estos modelos utilizan técnicas inteligentes más complejas. En particular, se trata de **Redes Neuronales Recurrentes**, que posibilitan la retención de la información generada en iteraciones previas para utilizarla como contexto en las futuras y así mejorar la calidad del modelo durante el entrenamiento. Este tipo de técnicas inteligentes suelen aplicarse al procesamiento del lenguaje y al análisis de sentimientos con el objetivo de aprender la distribución de probabilidad de los diferentes tipos de caracteres con la que, posteriormente, poder **predecir el sentimiento asociado de términos desconocidos** [60] [66]. Cabe destacar que los resultados proporcionados por esta librería han sido los que mejor calidad-tiempo han proporcionado de entre todas las experimentadas para los diferentes tipos de textos que se han analizado. En la tabla 5, para demostrar las diferentes características que se han detallado sobre cada librería, se presentan un conjunto de ejemplos en los que se utilizan diferentes tipos de textos procedentes de comentarios reales que se han recopilado de la red social Instagram.

	VADER	TextBlob diccionario	TextBlob Naive Bayes	Flair
Acho regalame uno 😘😘😘	Neutral (100%)	Neutral (100%)	Positivo (79%)	Positivo (97%)
Menuda pasada de interior....!!!!	Neutral (100%)	Neutral (100%)	Positivo (79%)	Positivo (99%)
@audispain atención al cliente no se han molestado en decirnos nada desde hace 15 días, pero nosotros sin coche y sin una alternativa MUY MUY DESCONTENTA DE AUDI Y SU ATENCIÓN	Neutral (100%)	Neutral (73%)	Positivo (79%)	Negativo (99%)
@audispain insisto que lo que está pasando en Audi y la comunicación que tiene taller- cliente , taller- atención al cliente , atención al cliente - cliente , fatal Un mes vamos a llevar sin coche y no han tenido ni el detalle de ponernos uno de sustitución o de darnos más información solo nos han dado cuando nosotros hemos insistido por wassap o redes sociales.	Neutral (100%)	Neutral (85%)	Positivo (79%)	Negativo (99%)

Tabla 5. Comparación entre los sentimientos identificados en varios comentarios por diferentes analizadores de sentimientos.

Como podemos observar, el **primer texto** está compuesto por expresiones coloquiales, faltas de ortografía y emoticonos que representan una cara de tristeza, pese a que en mi opinión están siendo utilizados de manera irónica, puesto que manifiesta un interés por probar un producto, que en particular se trata de un coche de la marca Audi. Mientras que los dos analizadores basados en diccionarios coinciden en que este texto demuestra un sentimiento neutral con un 100% de seguridad, los otros dos basados en técnicas inteligentes lo clasifican como positivo con un alto porcentaje. Por lo que podemos concluir que incorporar **algoritmos de Aprendizaje Automático** a este tipo de tareas no solo resulta beneficioso, sino que proporciona resultados más reales.

El **segundo texto** se caracteriza por varios **signos de exclamación**, lo que al criterio de los modelos de la librería *flair* son indicadores de una **mayor intensidad del sentimiento** identificado, que en este caso es positivo con una intensidad de un 99%. De nuevo, los dos analizadores basados en algoritmos inteligentes coinciden en el sentimiento identificado, aunque el de *Naïve Bayes* proporciona una menor confianza en sus resultados.

Finalmente, en los **dos últimos textos** podemos apreciar que, mientras los analizadores que utilizan diccionarios siguen decantándose por sentimientos neutrales, aunque con un mayor grado de duda en el caso del perteneciente a la librería *TextBlob*, los dos modelos inteligentes discrepan en sus resultados. El basado en *Naïve Bayes* los identifica como dos textos positivos aportando un porcentaje de confianza bastante alto, mientras que el modelo de la librería *flair* considera que son negativos con un porcentaje casi del 100%. Cabe destacar que, en el **tercer texto**, algunas de las palabras se encuentran en **mayúsculas**, por lo que aumenta la intensidad del sentimiento encontrado. Sin embargo, pese a tomarlo en consideración como hemos comentado anteriormente, el modelo perteneciente a la librería *flair* proporciona el mismo porcentaje de confianza en sendos textos, pese a que el cuarto no dispone de elementos adicionales que aumenten la intensidad del sentimiento. Es por ello por lo que se puede concluir que esta librería dispone de analizadores de sentimientos **mejores entrenados y más robustos** que utilizan principalmente lo que han aprendido para identificar los sentimientos de los términos de un texto, además de apoyar su resultado en otro tipo de caracteres como los signos de exclamación o mayúsculas.

Si bien tras los diversos experimentos realizados con las librerías consideradas se puede concluir que los modelos de *flair* son los más eficaces, cabe destacar que puede ocurrir una situación en la que para determinados textos no sea capaz de predecir el sentimiento que representan. Es por ello por lo que he planteado una **técnica híbrida** en la que, en primer lugar, el texto es analizado mediante el modelo proporcionado en la librería *flair*. En caso de que no pueda proporcionar un resultado o que la confianza del sentimiento identificado no sea superior al 60%, el texto será estudiado por la librería **VADER** con el objetivo de asignarle un sentimiento para el primer caso, o comparar los resultados de ambas bibliotecas para determinar en base al porcentaje de confianza cuál es el sentimiento que se encuentra más acorde al texto proporcionado.

Para finalizar esta sección cabe destacar que para implementar la **interfaz** con la que interactuar de forma gráfica con la plataforma he utilizado una librería denominada **Dash** [67]. Este *framework* está orientado a facilitar la construcción de aplicaciones web, especialmente si disponen de numerosos elementos visuales, como es el caso de esta herramienta. Se basa en las tres siguientes librerías:

- **Flask**: se trata de otro *framework* especializado en la implementación de sistemas web o de microservicios, como es el caso del proyecto que llevé a cabo en la asignatura **Cloud Computing** de este máster [68].
- **Plotly**: esta librería permite la visualización de una gran variedad de gráficos interactivos [69].
- **ReactJS**: es una biblioteca de JavaScript que facilita el desarrollo de interfaces web interactivas [70], la cual también ha sido estudiada en el máster durante la asignatura **Sistemas Software Basados en Web**.

5.5.4. Tecnologías y despliegue

En este cuarto apartado se detallan los tipos de motores de bases de datos que se han integrado en la plataforma con el objetivo de almacenar tanto la información extraída de fuentes externas, como los datos generados a partir de su procesamiento y análisis. Para el primer caso he utilizado un almacén NoSQL basado en documentos, como es **MongoDB**, por su facilidad de uso, su eficiencia, su flexibilidad al no tener que establecer un esquema previo, así como por el conocimiento que he adquirido trabajando con él durante la asignatura **Cloud Computing** cursada en este máster.

Su integración en este proyecto se basa en la construcción y despliegue de un **contenedor** con la última imagen disponible de esta base de datos, además de con una **persistencia de datos** configurada para evitar perder la información almacenada en caso de que surja algún problema con el contenedor. Para ello, basta con proporcionar el nombre del volumen que se creará como una carpeta más en el sistema en el que se despliegue la aplicación. Además, se han desarrollado los ficheros necesarios para crear automáticamente una base de datos en la que almacenar la información recopilada procedente de las diversas fuentes de datos disponibles, así como un usuario administrador con el que poder realizar las distintas operaciones necesarias. Dentro de ella, se encuentran las **colecciones** que contienen los diferentes tipos de entidades que se han explicado anteriormente: los perfiles de usuarios, la información asociada a las publicaciones y la lista de seguidores de las cuentas que se desean estudiar.

En el segundo caso en el que se pretende jerarquizar y relacionar la información procesada a analizar, así como almacenar los resultados obtenidos de los estudios realizados, he seleccionado una base de datos SQL como es **PostgreSQL**. En esta ocasión no disponía de experiencia previa con este motor de almacenamiento, pero lo escogí principalmente por su popular buen rendimiento tanto en operaciones de lectura como de escritura, además de por ser una **base de datos relacional basada en objetos**, lo que permite diseñar un esquema con herencia entre tablas [71], lo cual resulta beneficioso para facilitar la relación entre las diferentes entidades disponibles en este proyecto. El diagrama relacional que representa la estructura de este almacén de datos se puede visualizar en la ilustración 13.

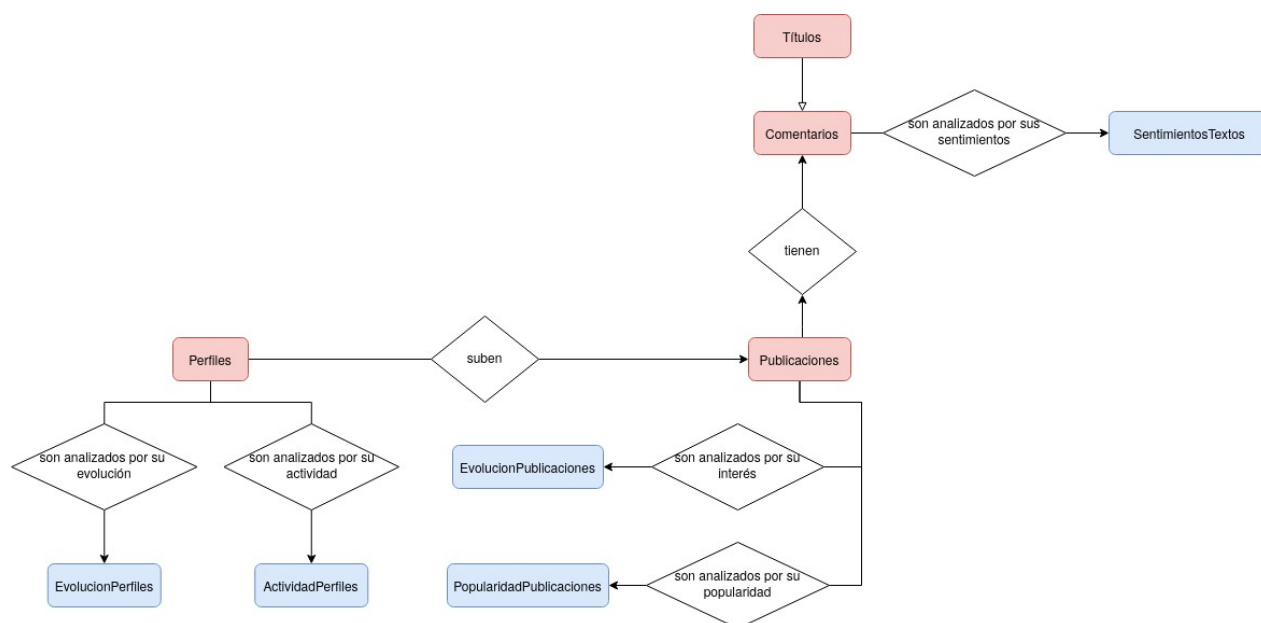


Ilustración 13: Esquema relacional para el almacén de datos SQL. Fuente: realización propia.

Como podemos observar, en un color cálido se representan las principales entidades identificadas en capítulos anteriores, mientras que en un tono más frío se reflejan las tablas encargadas de almacenar los

resultados de los diferentes análisis expuestos en secciones previas. En el primer caso, la tabla **Perfiles** almacenará la información procesada de los distintos perfiles de usuarios a estudiar, como es el identificador del usuario, la fecha de recopilación de la información, el nombre completo, así como el número de seguidores, seguidos y las publicaciones subidas. Esta entidad, por lo tanto, estará relacionada con los dos análisis de perfiles implementados para analizar su evolución a lo largo del tiempo basada en el interés que ha suscitado en el resto de miembros de la comunidad y en su actividad periódica.

Por otro lado, se encuentra la segunda entidad principal denominada **Publicaciones**, que almacenará el identificador de cada publicación proporcionado por la fuente de datos de la que provenga, la fecha de recopilación, además de cierta información básica como el número de *me gusta* y la fecha en la que se publicó. Como se detalló en capítulos previos, todas las redes sociales permiten realizar **comentarios**, por lo que se representa como la tercera entidad que se encuentra unida con la tabla **Publicaciones**, de modo que se puedan identificar fácilmente los comentarios asociados a cada una. Sin embargo, cada plataforma dispone de un tipo específico de contenido, como es el caso de Instagram puesto que se basa en imágenes aunque permite incluir un título para cada una. Debido a que la ejemplificación del funcionamiento de la herramienta se centra en esta red social, se ha integrado una cuarta entidad denominada **Títulos, que hereda de la entidad Comentarios** ya que contiene la misma estructura, aunque su objetivo consiste en almacenar los títulos originales y procesados de las publicaciones para luego proporcionarlos a los analizadores de sentimientos. Gracias a esta propiedad, basta con conectar la tabla **Comentarios** con una nueva tabla denominada **SentimientosTextos**, con el fin de almacenar los sentimientos e intensidades asociados de cada uno de los comentarios y títulos de las publicaciones.

Finalmente, cabe destacar que, al igual que en el caso de la base de datos no relacional, para este almacén también se ha planteado la construcción, configuración y despliegue mediante un **contenedor con persistencia de información**, así como la creación automática del esquema relacional explicado y de un nuevo usuario con los suficientes privilegios para operar con él.

5.5.5. Arquitectura del sistema

El objetivo de este penúltimo apartado consiste en detallar la arquitectura que se ha diseñado para implementar la herramienta de este proyecto. A continuación, en la ilustración 14 se muestra un diagrama en el que se representan las dos tareas principales que se pueden llevar a cabo dentro de la herramienta desarrollada.

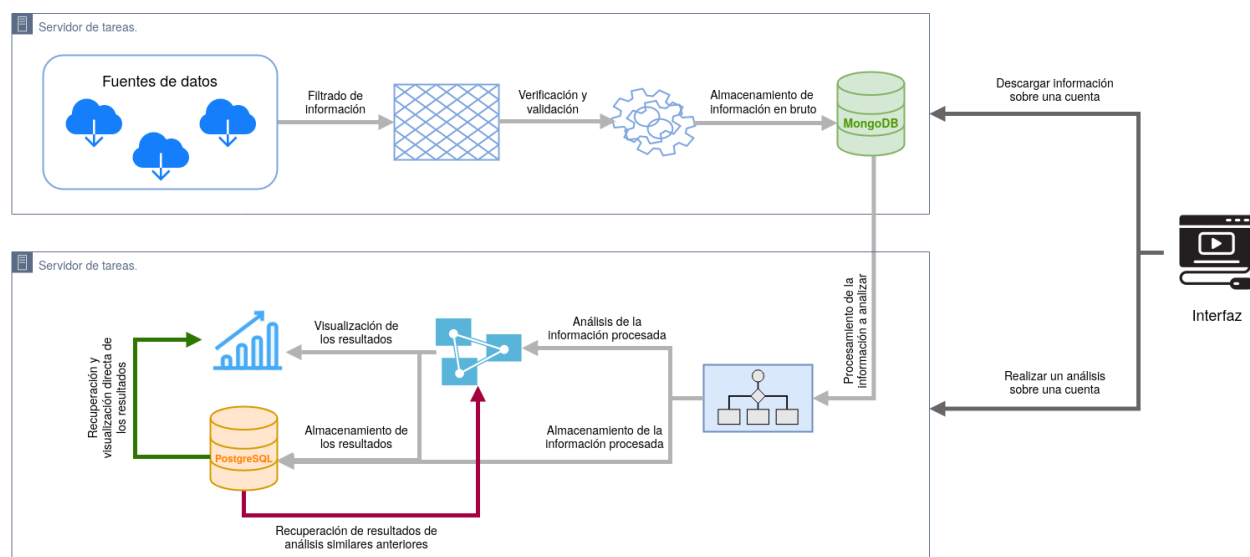


Ilustración 14: Arquitectura de la herramienta. Fuente: realización propia.

La **primera actividad** que se puede realizar consiste en especificar el nombre de la cuenta de la que deseamos **obtener información**, así como las fuentes en las que la herramienta realizará la búsqueda. Una vez se confirme la operación, a través de un servidor de tareas se recopilará la información disponible asociada a cada entidad descrita anteriormente, es decir, al perfil, las publicaciones y los seguidores de forma periódica y en segundo plano. A continuación se filtran y verifican los datos recopilados para únicamente almacenar aquellos que sean útiles y válidos para realizar los diversos análisis existentes.

Mientras que la **segunda acción** se enfoca hacia la **aplicación de un determinado análisis** sobre una cuenta y período de tiempo específicos también bajo una tarea en segundo plano. Para esta operación existen tres flujos posibles.

- Con un color grisáceo se representa el primer caso en el que **no existen resultados previos** similares al tipo de análisis demandado, por lo que en primer lugar se deberá de obtener la información en bruto asociada a la cuenta y al tiempo concretados. En caso de ser un análisis basado en texto, se realiza un preprocesamiento previo como el que se detalló en la sección 5.2.2. A continuación, por un lado se almacenan los datos procesados en la base de datos relacional con el fin de poder ser reutilizados en posteriores análisis, mientras que por otro lado comienza el análisis seleccionado. Una vez ha finalizado, se guardan los resultados obtenidos de nuevo en el almacén relacional y se visualizan en la interfaz de forma gráfica e interactiva.
- En el segundo caso, visualizado en color magenta, se contempla la posibilidad de **reutilizar los resultados previos de análisis similares** que sean compatibles con el período de tiempo especificado. De este modo, el nuevo estudio consiste en realizar los cálculos estrictamente necesarios para incorporar los nuevos datos a los resultados recuperados. Uno de los casos prácticos que más se beneficia de este método sucede en los **análisis de sentimientos** tanto de los comentarios como de los títulos de las publicaciones. El objetivo consiste en recuperar el sentimiento identificado de los textos estudiados anteriormente para disminuir la considerable inversión temporal y computacional que supone el procesamiento y la aplicación de los analizadores de sentimientos.
- Finalmente, en el tercer caso representado con un color verdoso se contempla la mejor situación posible en la que el **análisis demandado se ha realizado con anterioridad** sobre la misma cuenta y período de tiempo. Por lo que basta con recuperar los resultados asociados y representarlos directamente de forma gráfica.

5.5.6. Diseño de la interfaz

En la última sección de este quinto capítulo se pretende explicar el proceso de diseño que se ha realizado para implementar la interfaz con la que el usuario podrá interactuar con la herramienta. Para ello aplicaremos los conocimientos adquiridos en la asignatura del máster **Desarrollo y Evaluación de Sistemas Software Interactivos**. En primer lugar, en la ilustración 15 se puede observar el **diagrama HTA** elaborado a partir del análisis jerárquico de tareas que se ha efectuado con el objetivo de establecer un determinado orden para llevar a cabo las operaciones disponibles en la plataforma.

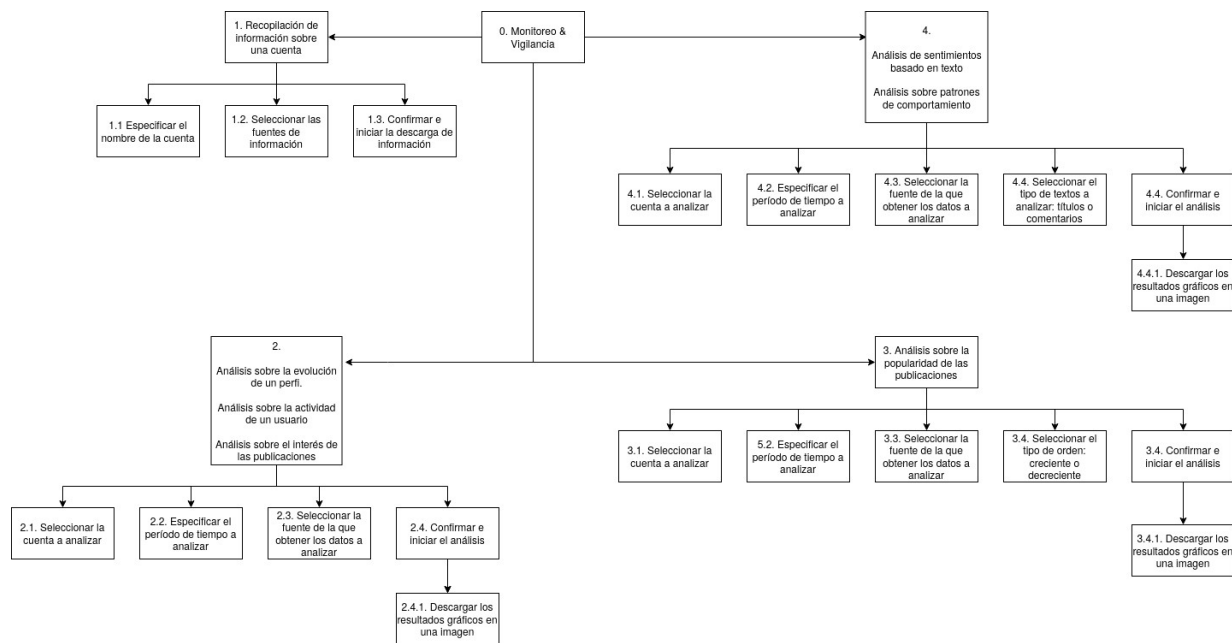


Ilustración 15: Diagrama HTA de la interfaz de la plataforma. Fuente: realización propia.

Como se puede apreciar dentro de la ilustración 15, en la primera actividad asociada a la recopilación de información sobre una cuenta determinada se deberá de especificar el nombre de usuario así como las fuentes de información en las que buscar sus respectivos datos. Mientras que en el resto de acciones encargadas de efectuar los diversos análisis existentes, se deberá concretar la cuenta sobre la que se desea realizar el estudio seleccionado, además del período de tiempo a considerar. Sin embargo existen dos excepciones en el flujo de las tareas necesarias para realizar un análisis. En el estudio de la popularidad de las publicaciones también se deberá de especificar si se deseamos visualizar las mejores o peores publicaciones. Mientras que en el análisis de sentimientos se deberá seleccionar el tipo de texto a estudiar entre los dos disponibles: los títulos o los comentarios de las publicaciones.

A continuación, en la ilustración 16 se presenta el **diagrama de conceptos** que representa las diferentes clases que se han implementado durante el desarrollo de la plataforma así como sus respectivas conexiones.

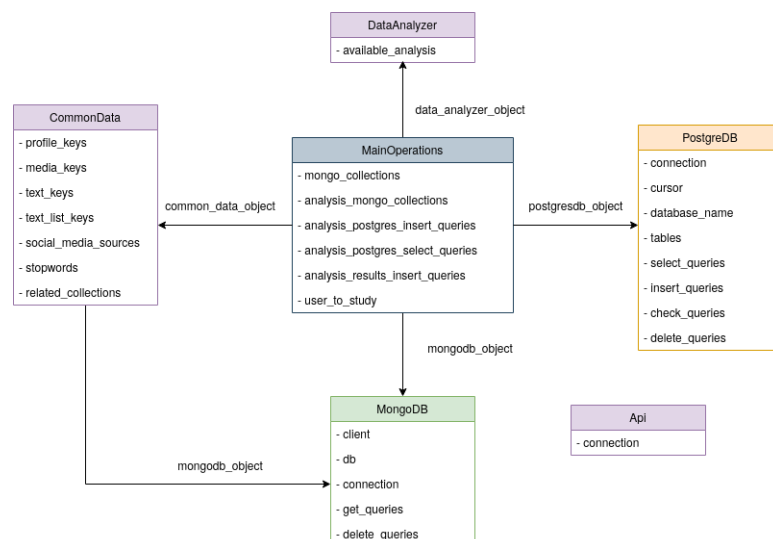


Ilustración 16: Diagrama de conceptos de la plataforma. Fuente: realización propia.

Las entidades coloreadas en un tono violáceo se corresponden con las tres clases que conforman el módulo de datos. En primer lugar disponemos de la clase **Api**, que contiene los métodos necesarios para establecer la conexión con las distintas fuentes de datos y extraer la información de una determinada cuenta. Por otro lado, se encuentra la clase **CommonData** cuyo principal cometido se centra en el preprocesamiento de los datos recopilados con respecto al esquema de información común que se explicó en la primera sección de este quinto capítulo. Finalmente, la clase **DataAnalyzer** contiene las implementaciones de los diferentes análisis disponibles en la plataforma.

Con el objetivo de interactuar con las dos bases de datos integradas en este proyecto, se ha diseñado una clase para cada una de ellas que juegan el papel de **Single Source of Truth**. Este término ha sido estudiado en la asignatura **Cloud Computing** de este máster y hace referencia a la capacidad de aglutinar las operaciones imprescindibles para trabajar con una base de datos en una única clase. Esta será utilizada por aquellas que necesiten interactuar con alguno de los almacenes a través de la generación de un único objeto. Así, se puede ejercer un mayor control sobre las acciones que se efectúan sobre los datos almacenados.

Por último, se ha diseñado la clase **MainOperations** como la única interconectada con todas las clases anteriores con el objetivo de proporcionar acceso a todas las actividades que se pueden realizar en la plataforma, desde establecer la cuenta de la que recopilar su información, hasta efectuar cualquiera de los análisis disponibles.

Para finalizar este capítulo, a continuación en la ilustración 17 se puede observar el **diagrama Wireflow** que muestra las diferentes pantallas que componen la interfaz así como su navegación. Como se puede apreciar, se trata de una estructura basada en secciones situadas sobre un único nivel. La primera que aparece se encuentra vinculada a la especificación del nombre de usuario relativo a la cuenta de la que se desea obtener información a través de las fuentes seleccionadas. Mientras que las seis restantes proporcionan acceso a cada uno de los diferentes estudios existentes en la herramienta. En cada una de las secciones analíticas se visualizarán los distintos parámetros que se deberán configurar antes de la realización del análisis. En estas mismas pantallas será donde se visualicen también los resultados gráficos una vez haya finalizado el proceso.



Ilustración 17: Diagrama Wireflow de la interfaz de la aplicación. Fuente: realización propia.

Capítulo 6

Ejemplo de aplicación

En este penúltimo capítulo se pretende mostrar el comportamiento de la herramienta desarrollada mediante el estudio de un perfil real. Si bien se ha recopilado y analizado información procedente de diversas cuentas de **Instagram**, a continuación se detalla un caso concreto que permite demostrar el verdadero potencial de este proyecto. En particular, se trata del usuario *audispain*, la cuenta oficial de la marca de automóviles **Audi** en España. A continuación, en las ilustraciones 18, 19 y 20 se puede visualizar el calendario en el que se realizó la descarga de su respectiva información. Como se puede apreciar, en total disponemos de un volumen de datos de hasta un período de cuatro semanas alternas.

OCTUBRE 2020						
			01	02	03	04
05	06	07	08	09	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Ilustración 18: Recopilación de datos sobre la cuenta de Audi España en Instagram durante octubre de 2020.

NOVIEMBRE 2020						
						01
02	03	04	05	06	07	08
09	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Ilustración 19: Recopilación de datos sobre la cuenta de Audi España en Instagram durante noviembre de 2020.

DICIEMBRE 2020						
	01	02	03	04	05	06
07	08	09	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Ilustración 20: Recopilación de datos sobre la cuenta de Audi España en Instagram durante diciembre de 2020.

Cabe destacar que el volumen de datos analizado para ejemplificar el funcionamiento de la aplicación **no contiene toda la información disponible** en Instagram sobre la cuenta mencionada anteriormente. Si bien la aplicación permite la descarga de todos los datos existentes de un usuario determinado, es necesario considerar el elevado **coste computacional y temporal** que conlleva tanto esta acción, como los posteriores análisis que se apliquen. Dependiendo de los objetivos que deseemos cumplir y del tiempo del que dispongamos para tomar ciertas decisiones, se podrá optar por recopilar el conjunto completo de información disponible, o solamente un volumen de datos con suficientes muestras como para obtener resultados analíticos fiables con una inversión de recursos moderada.

La librería de Instagram utilizada para recopilar información sobre usuarios de esta plataforma consiste en la **descarga de la información más reciente**. En el caso de las publicaciones, por cada petición se obtienen las interacciones recibidas de las cien últimas. De igual modo se recopilan los comentarios escritos en cada publicación, descargando, así, los ciento cuarenta más recientes.

6.1. Evolución del perfil

Comenzamos con un estudio general de las principales características de la cuenta seleccionada a pequeña escala. En la ilustración 21 podemos apreciar la evolución del **número de seguidores, seguidos y publicaciones** subidas durante un período de **siete días** comprendido entre el 1 y el 7 de diciembre de 2020. A simple vista podemos comprobar que el volumen de **seguidores es el valor predominante**, puesto que el color morado, que representa esta propiedad, es el que prevalece casi

en la totalidad del gráfico. De hecho, las dos métricas restantes apenas son apreciables en comparación con la comunidad de seguidores.

Evolución del número de seguidores, seguidos y publicaciones.

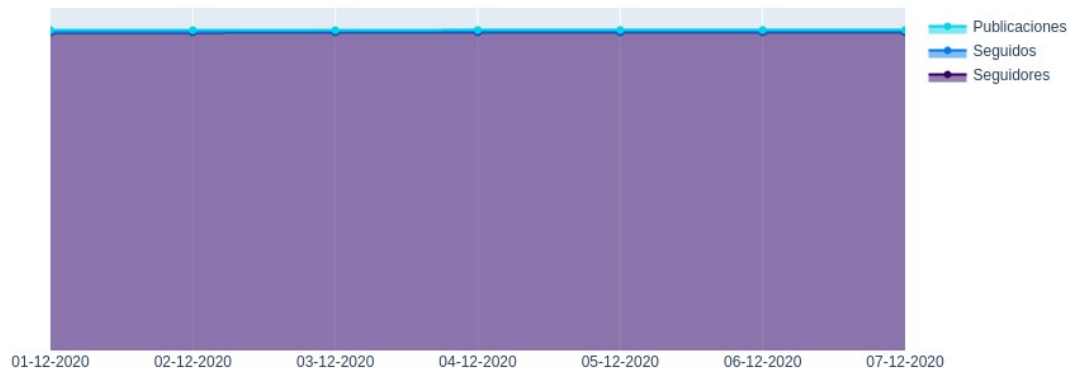


Ilustración 21: Evolución del perfil de la cuenta de Audi en Instagram durante siete días.

Sin embargo, gracias a la interactividad que ofrece la librería, podemos aplicar un **zoom** personalizado en cualquier área del gráfico. Así, nos permite visualizar en detalle el comportamiento de las tres medidas incluidas en este análisis, durante el período de tiempo especificado con el objetivo de conocer si existe algún tipo de relación que las vincule entre sí. Y, tal y como podemos observar en la ilustración 22, así es puesto que sus valores sufren las **mismas variaciones simultáneamente**. En particular, podemos apreciar que durante la semana estudiada el número de seguidores, seguidos y publicaciones subidas se ha mantenido prácticamente igual, lo que nos indica que en este período de tiempo la cuenta de Audi en Instagram no ha experimentado apenas ningún cambio en su contenido o sus comunidades.

Evolución del número de seguidores, seguidos y publicaciones.

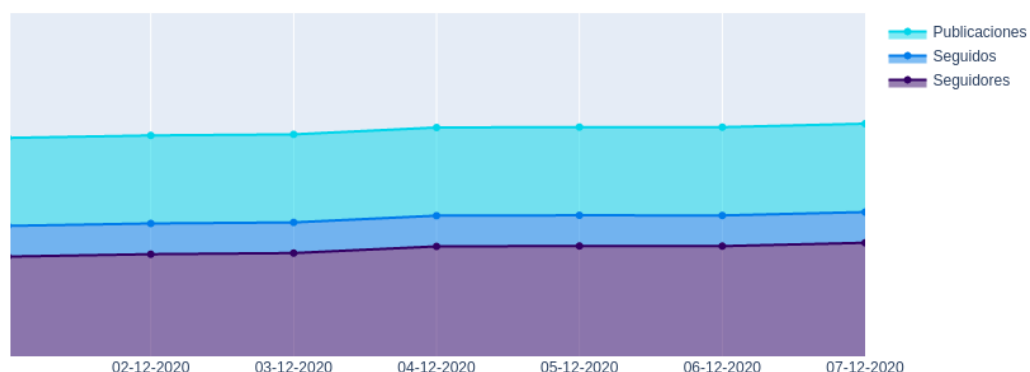


Ilustración 22: Evolución del perfil de la cuenta de Audi en Instagram durante siete días en mayor profundidad.

A continuación se repite este mismo análisis pero tomando el conjunto de datos al completo, que dispone de una duración de cuatro semanas, como se pudo apreciar en el calendario explicado anteriormente. En este caso, en lugar de representar los

valores reales obtenidos de Instagram, se calcula **una media por semana** para el número de seguidores, seguidos y publicaciones.

En la ilustración 23 se puede observar que la perspectiva general es **similar al del caso anterior**, puesto que el número de seguidores sigue siendo más visible que el número de seguidos y de publicaciones subidas. Por lo que la primera conclusión que podemos extraer es que esta cuenta puede ejercer una **gran influencia** sobre la red social en general, ya que dispone de una capacidad de difusión masiva gracias al poder que tiene su voluminosa comunidad de seguidores para compartir su contenido con otros usuarios. Asimismo, resalta la gran diferencia entre el número de seguidos con respecto a los usuarios suscritos. Este aspecto nos indica que la marca automovilística no parece tener interés en visualizar las publicaciones de otros usuarios de Instagram, por lo que su principal objetivo consiste en **dar a conocer sus productos**, lo que significaría que está utilizando esta red social como un medio más de publicidad con fines principalmente comerciales.

Evolución del número de seguidores, seguidos y publicaciones.



Ilustración 23: Evolución del perfil de la cuenta de Audi en España durante cuatro semanas.

De nuevo, tras aplicar cierto *zoom* a la gráfica anterior para visualizar el comportamiento de las métricas en conjunto, en la ilustración 24 podemos apreciar que en este lote de datos más voluminoso presentan un comportamiento similar. Por un lado, la variabilidad del número de seguidores y seguidos no es considerablemente alta, por lo que nos indica que la cuenta de Audi en Instagram se caracteriza por una **cierta estabilidad**. De esta forma, la gráfica no muestra un crecimiento o pérdida importante del número de usuarios en cada comunidad, algo que podría ser más propio de una cuenta recién creada.

Evolución del número de seguidores, seguidos y publicaciones.

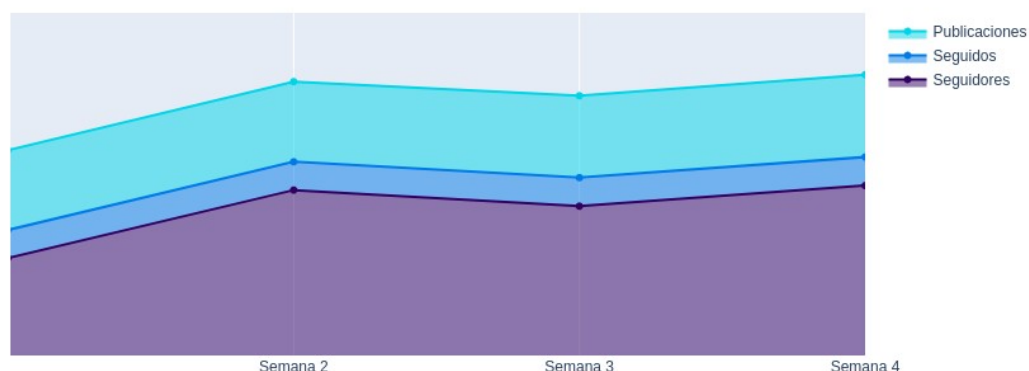


Ilustración 24: Evolución del perfil de la cuenta de Audi en España durante cuatro semanas en mayor profundidad.

Asimismo, también destaca el aumento y decremento acompasado de las tres métricas de forma simultánea. En particular, el pico de actividad se sitúa sobre la segunda semana en la que la cuenta publica más contenido y se incorporan un mayor número de usuarios a las comunidades de seguidores y seguidos. Este comportamiento confirma la teoría anterior que pone de manifiesto que, para esta cuenta en particular, **a mayor número de publicaciones, mayor número de seguidores y seguidos**. La explicación para el primer caso es relativamente sencilla, y es que cuanto más contenido se publique en una cuenta, mayor interés puede generar en el resto de miembros de la red social y mayor probabilidad existe de que más usuarios se suscriban a su contenido.

Sin embargo, la razón del aumento simultáneo del número de seguidos no es tan evidente. Una posible teoría se fundamenta en la fecha en la que se tomaron los datos, que en el caso de la segunda semana se corresponde con la segunda semana de noviembre de 2020, en la cual las empresas suelen plantear la **campaña de Navidad**. Por lo que la marca Audi ha podido utilizar Instagram para seguir cuentas de otras compañías con el objetivo de realizar un estudio sobre sus tácticas publicitarias, así como de los productos que pretenden promover en esta red social para orientar su propia estrategia comercial.

6.2. Actividad del usuario

El segundo análisis que he aplicado a los datos recopilados de la cuenta de Audi en España procedentes de Instagram consiste en visualizar concretamente, el **número de publicaciones** que ha subido durante un determinado período de tiempo. Como en el caso anterior, comenzamos con un primer experimento de menor volumen situando el rango temporal en **siete días** entre el 1 y el 7 de diciembre de 2020. Se ha seleccionado este período en particular para comprobar el efecto que puede tener una época tan especial para el comercio como es la Navidad.

Como podemos observar en la ilustración 25, al inicio del estudio cuenta con 1.242 publicaciones que se van incrementando conforme avanza la semana. El número de contenido que publica cada día se puede visualizar de forma numérica en el interior de las cajas del gráfico. Así, podemos apreciar que casi en la totalidad de los días se ha estado subiendo una publicación, excepto el 4 de diciembre que se publicaron dos. Cabe puntualizar que el eje de ordenadas es calculado automáticamente por la librería *Dash* a partir de la sumatoria del número de publicaciones de cada día.

Como primera conclusión para este estudio podemos confirmar que la cuenta de Audi en Instagram es **bastante activa** puesto que publica contenido regularmente, casi a diario, lo que generalmente ayuda a conformar una comunidad de seguidores tan voluminosa como la que tiene actualmente.

Evolución del número de publicaciones

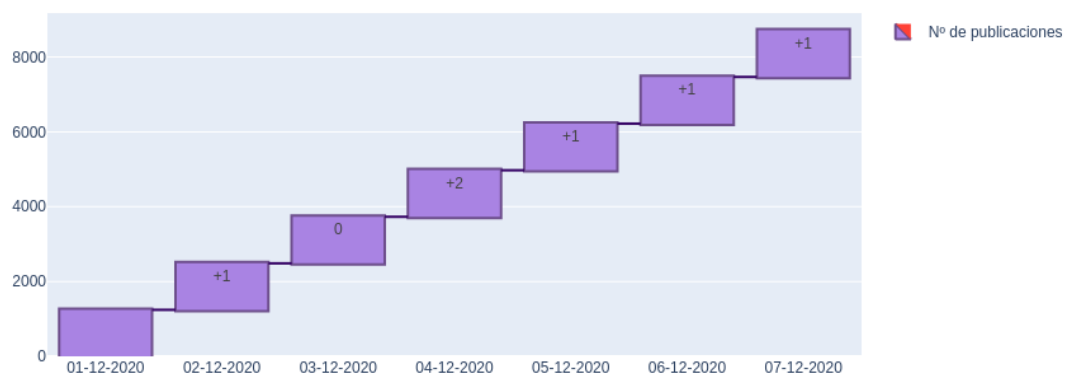


Ilustración 25: Estudio de la actividad de la cuenta de Audi en España durante siete días.

A continuación, repetimos el mismo tipo de análisis aunque sobre el conjunto de datos completo de cuatro semanas de duración. El objetivo consiste en comprobar si a una mayor escala el comportamiento visualizado anteriormente se mantiene. Como en el análisis del apartado anterior, en este caso se visualiza la **media del número de publicaciones** por semana. Así, podemos observar en la ilustración 26 que en la primera dispone de una media de 1.212 publicaciones, y que como en el estudio a menor escala, cada semana incrementa el número de contenido publicado. Por lo tanto, en este análisis también podemos visualizar cierta **regularidad en la actividad** de esta cuenta. Sin embargo, se puede apreciar una notable diferencia en el número de publicaciones subidas en la tercera semana que asciende a 26. La razón de este incremento puede residir en que los datos recopilados en este período de tiempo se sitúan en la primera semana de diciembre de 2020, por lo que nos encontrábamos en plena **campaña de Navidad**. En esta época es en la que las marcas realizan sus mayores esfuerzos por publicitar sus productos a la sociedad con fines principalmente comerciales. Por lo tanto, este análisis confirma la teoría de que la marca automovilística utiliza **Instagram como un medio más de publicidad** en el que invierte recursos monetarios, temporales y humanos para dar a conocer sus vehículos a una enorme comunidad virtual e internacional.

Evolución del número de publicaciones

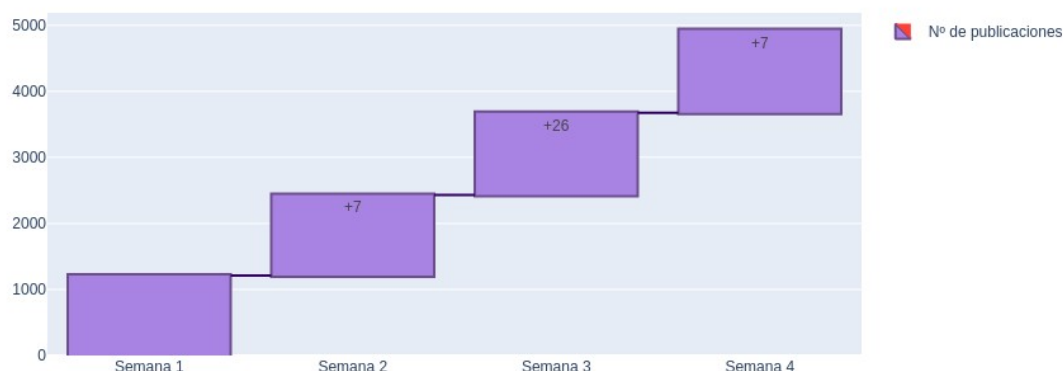


Ilustración 26: Estudio de la actividad de la cuenta de Audi en España durante cuatro semanas.

6.3. Interés de las publicaciones

En este tercer análisis se pretende comprobar el grado de interés que suscitan las publicaciones de la cuenta de Audi España en Instagram. Para ello estudiaremos el número de interacciones que realizan los usuarios de esta comunidad virtual en forma de *me gusta* y comentarios. De nuevo, comenzamos aplicando este análisis a un conjunto de datos más reducido de una duración de **siete días**, entre el 1 de diciembre y el 7 de diciembre de 2020.

Tal y como se puede observar en la ilustración 27, la interacción que más reciben las publicaciones de este usuario son los *me gusta*, puesto que su representación en color azul acapara casi la totalidad del gráfico. Mientras que, durante este período, el número de comentarios por día es bastante más reducido en comparación. Por lo que, de momento, parece ser que la comunidad de seguidores de esta cuenta **utiliza mayoritariamente el botón de *me gusta*** para opinar sobre el contenido que publica. Si bien se trata de una interacción bastante común en el resto de plataformas virtuales, no proporciona demasiada información acerca de las sensaciones que provoca el contenido sobre los miembros de la red social. Es por ello por lo que, en análisis posteriores estudiaremos los sentimientos que se encuentran detrás de los comentarios recopilados durante este período.

Otro aspecto a destacar es el progresivo **decremento** del número de sendas interacciones conforme avanza la semana. Esto nos indica que en este período de tiempo en el que se ha realizado el análisis, los usuarios de esta red social demuestran una menor actividad en base al número de *me gusta* y comentarios que proporcionan al contenido de la cuenta de Audi en particular.

Evolución del interés de las publicaciones

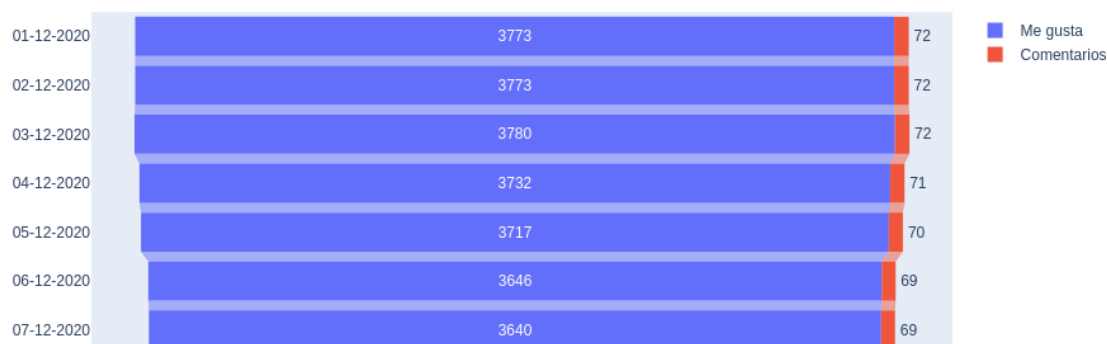


Ilustración 27: Estudio del interés de las publicaciones de la cuenta Audi en España durante siete días.

A continuación realizamos el mismo estudio aunque utilizando el conjunto de datos completo. Como podemos visualizar en la ilustración 28, la interacción que más efectúan los usuarios de esta red social sigue siendo los **me gusta**. La explicación de este hecho puede consistir en la **facilidad e inmediatez** de esta acción, puesto que basta con pulsar un botón o realizar un doble toque sobre la publicación para opinar de manera favorable sobre ella, por lo que conlleva un mínimo esfuerzo y tiempo.

Asimismo, en un volumen más amplio de información se puede apreciar la misma tendencia a la baja del número de interacciones conforme avanzan las semanas. En particular, las dos últimas se sitúan en el mes de diciembre de 2020, como se pudo observar en el calendario que se detalló al comienzo del capítulo. Este hecho indica que, en base a los datos recopilados de esta cuenta y las interacciones obtenidas, se aprecia una **menor actividad** por parte de los miembros de Instagram durante las **semanas anteriores a la Navidad**, lo que contrasta con el incremento de recursos que invierte la compañía en publicar más contenido durante estas fechas, tal y como hemos podido observar en el análisis de la sección anterior.

Evolución del interés de las publicaciones

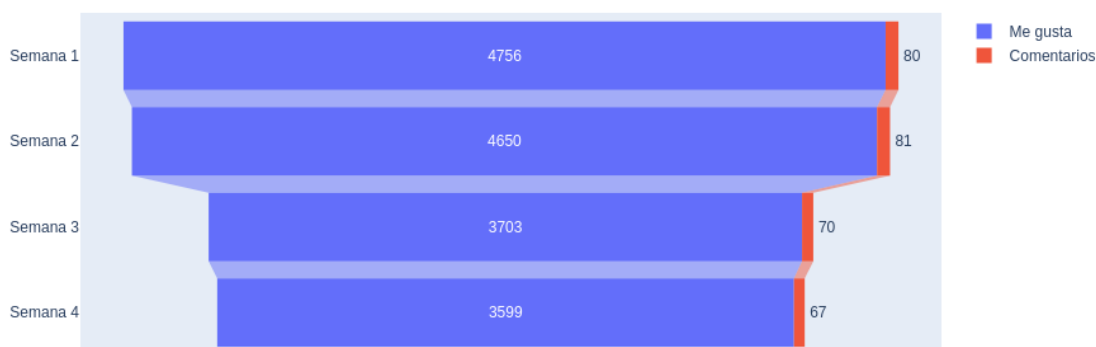


Ilustración 28: Estudio del interés de las publicaciones de la cuenta Audi en España durante cuatro semanas.

6.4. Popularidad de las publicaciones

El cuarto análisis se aplica sobre el conjunto de datos completo de la cuenta de Audi España en Instagram cuya duración es de **cuatro semanas**. El principal objetivo de este estudio consiste en mostrar un *ranking* de las **diez publicaciones mejor y peor valoradas** en base a las dos interacciones consideradas en el estudio anterior: el número de *me gusta* y de comentarios. Así, podremos analizar las características de las publicaciones que más sensación han causado entre los seguidores de la cuenta, con el fin de conocer si existe algún tipo de publicación en particular que tiene una mejor acogida. De igual modo, se pretenden estudiar los posibles motivos por los que algunas publicaciones no han sido tan bien recibidas por los miembros de esta comunidad virtual.

Comenzamos visualizando en la ilustración 29 las diez publicaciones mejor valoradas en base al número de *me gusta* y comentarios recibidos. Tal y como se puede apreciar, la publicación que ocupa el **primer puesto destaca notablemente** puesto que las interacciones que ha recibido son considerablemente mayores que las del resto. Para visualizar las publicaciones situadas en el *ranking*, se ha planteado la recopilación y almacenamiento de sus **enlaces** proporcionados por la librería *LevPasha*. Así, podemos observar en la ilustración 30 las características asociadas a la publicación mejor valorada. Como se puede apreciar, se trata de una fotografía singular de uno de los automóviles deportivos de la marca situado sobre un paisaje natural.

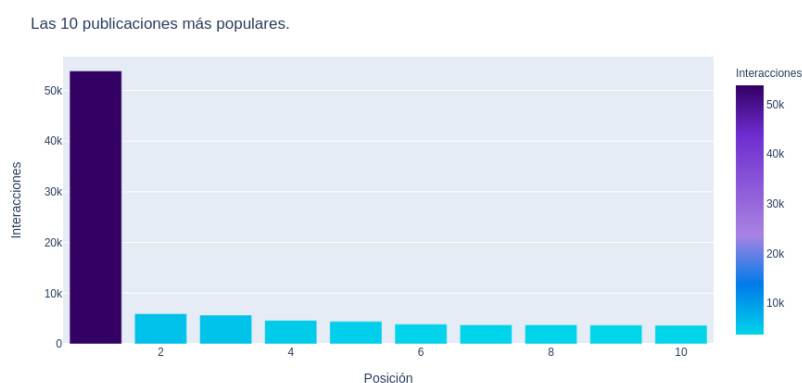


Ilustración 29: Estudio de las publicaciones mejor valoradas de la cuenta de Audi en España durante cuatro semanas.

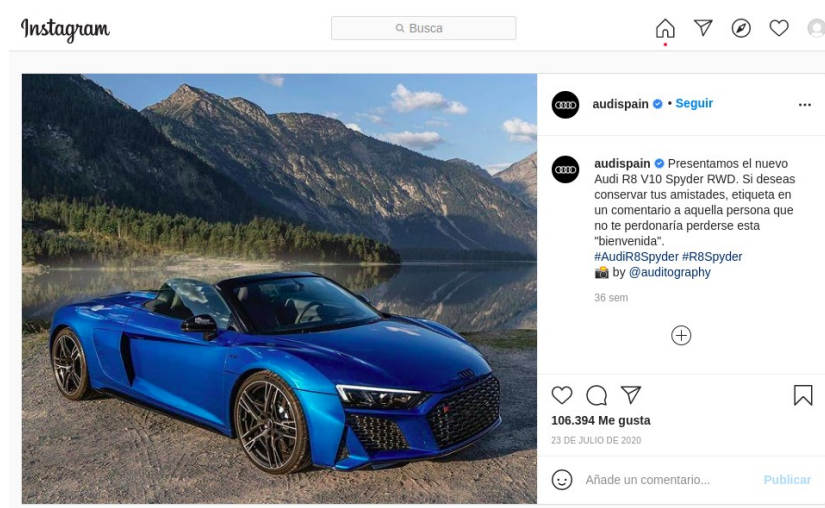


Ilustración 30: Captura de la publicación mejor valorada de la cuenta de Audi en España.

A continuación procedo a replicar este mismo análisis pero orientado a las **publicaciones peor valoradas** en función del número de *me gusta* y de comentarios. A diferencia del caso anterior, no existe una publicación en particular que destaque en el *ranking*, puesto que la mayoría de los valores representados en la ilustración 31 son considerablemente similares. De hecho, las dos publicaciones peor valoradas según las métricas escogidas para este análisis, han recibido prácticamente el mismo número de interacciones.

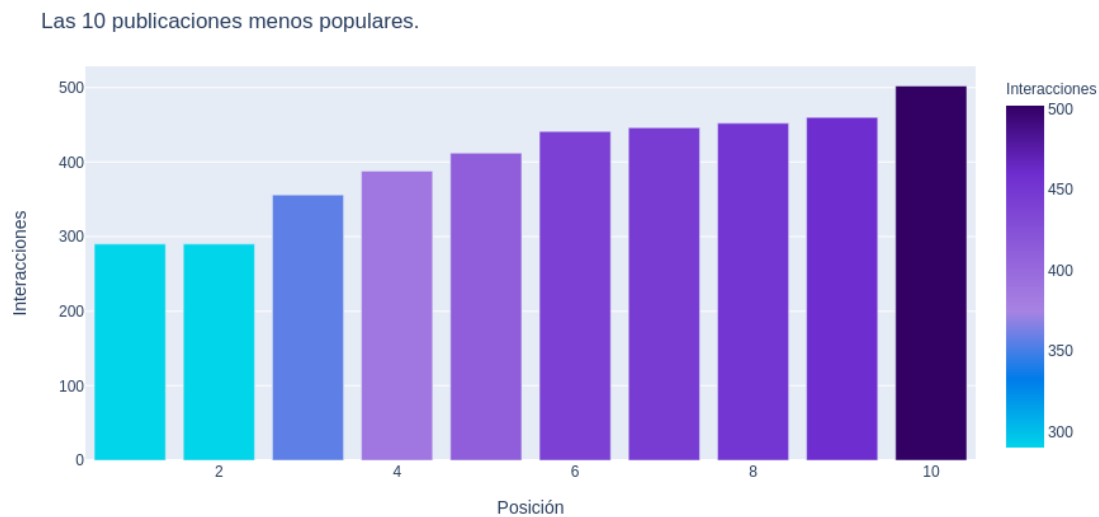


Ilustración 31: Estudio de las publicaciones peor valoradas de la cuenta de Audi en España durante cuatro semanas.

En las ilustraciones 32 y 33 se pueden observar las capturas de las dos publicaciones mencionadas anteriormente. Ambas son **vídeos de más de treinta segundos** de duración y han sido publicadas en noviembre y diciembre de 2020, respectivamente. La peor valorada situada a la izquierda hace referencia a la posibilidad de añadir o eliminar funcionalidades de un automóvil mediante un teléfono móvil. Mientras que la segunda publicación que menos interacciones ha recibido presenta uno de los modelos eléctricos de la marca. Visualizando el resto de publicaciones del *ranking* representado, podemos confirmar que la gran mayoría se corresponden con **vídeos de mayor duración**, además de otro tipo de publicaciones relativas a los **automóviles eléctricos** que está desarrollando la marca Audi. En la siguiente sección podremos apreciar la hostilidad que genera esta última temática en algunos de los seguidores de esta cuenta y cómo afecta al contenido que publica.

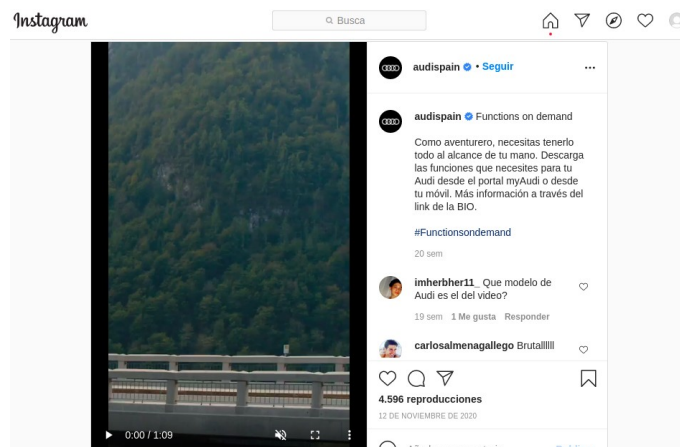


Ilustración 32: Captura de la publicación peor valorada de la cuenta de Audi en España.

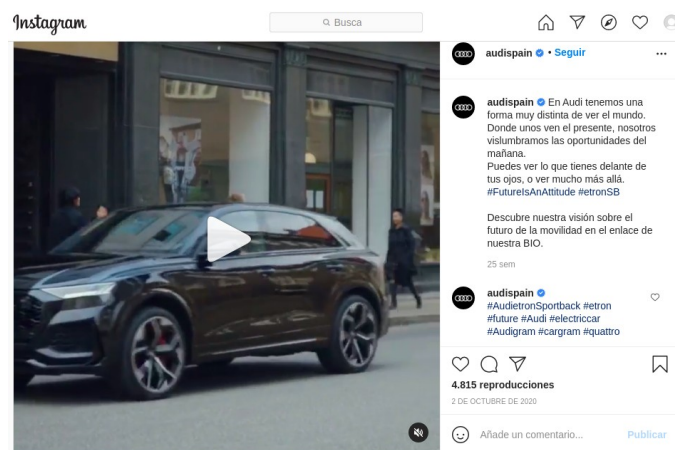


Ilustración 33: Captura de la segunda publicación peor valorada de la cuenta de Audi en España.

6.5. Análisis de sentimientos basados en texto

En este quinto análisis se pretende realizar un estudio acerca de la **intencionalidad** que se esconde tras los títulos asociados a las publicaciones, como tras las opiniones que redactan los usuarios de la red social. El primer caso se puede visualizar en la ilustración 34, en el que he utilizado el conjunto de datos completo que cuenta con una duración de **cuatro semanas**. Tal y como podemos apreciar en la siguiente captura, más del **80% de los títulos analizados han sido clasificados como positivos**. Este resultado era de esperar puesto que cuando cualquier marca desea publicitar sus productos, por norma general suele utilizar términos y estructuras gramaticales positivas con el fin de evocar buenas sensaciones en los usuarios que visualizan su contenido.

Análisis de sentimientos de los títulos de las publicaciones.

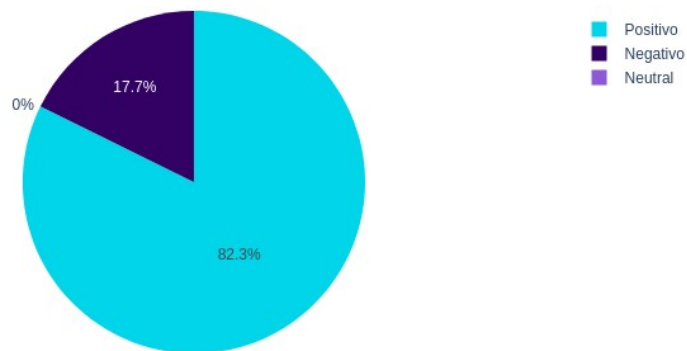


Ilustración 34: Análisis de sentimientos de los títulos de las publicaciones de Audi en España durante cuatros semanas.

A continuación procedo a aplicar dos análisis de sentimientos adicionales, aunque en este caso sobre **dos conjuntos de comentarios diferentes** con una misma duración de **siete días**, considerando una media de ciento cuarenta comentarios por publicación y cien publicaciones por día, es decir, cada población dispone de media de unos **98.000 comentarios**. En la ilustración 35 el rango temporal se encuentra comprendido entre el 1 y el 7 de **noviembre** de 2020, mientras que en la ilustración 36 se aprecia el segundo análisis realizado el período se encuentra entre el 1 y el 7 de **diciembre** de 2020. El objetivo de escoger estos dos intervalos de fechas consiste en descubrir si existe alguna diferencia sustancial en los sentimientos que proyectan los usuarios de Instagram durante un período normal y durante la campaña de Navidad. Tal y como se puede apreciar, en el primer análisis de noviembre se ha clasificado un **porcentaje más alto de comentarios positivos** que durante el estudio realizado en diciembre. Acompañando a este decremento de textos positivos, se encuentra el **aumento de comentarios clasificados como negativos** durante el mes de diciembre. Este hecho, por lo general, contrasta con el ambiente que se suele respirar en fechas cercanas a Navidad, donde la sociedad tiende a dejarse llevar por las emociones y la sensación global se corresponde más con la felicidad y el agradecimiento.

Análisis de sentimientos de los comentarios de las publicaciones.

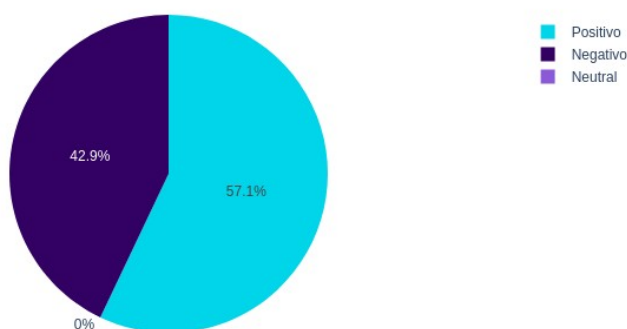


Ilustración 36: Análisis de sentimientos de los comentarios de las publicaciones de Audi en España durante una semana del mes de diciembre.

Por ello, investigando sobre el contenido de los comentarios clasificados como negativos he descubierto dos aspectos interesantes que parecen ser la razón de este aumento de la negatividad durante la campaña de Navidad. En primer lugar he encontrado una gran cantidad de comentarios, como los que a continuación se pueden visualizar en la tabla 5, relativos al **descontento de los seguidores** de la marca por el lanzamiento de diversos **automóviles eléctricos**, en especial el que correrá en el próximo Dakar de 2022.

Comentario	Sentimiento	Grado de confianza
<i>Eléctrico=caca la vaca</i>	Negativo	0,76
<i>Donde haya un buen cambio de marchas con su palanca, que se quiten botoncitos de micromachine..</i>	Negativo	0,99
<i>Hasta que no me den la autonomía de un coche de combustión interna yo no me pasaré al mundo eléctrico con un coche que me ofrece 1256 km entrando en la reserva de autonomía a mí no me sale rentable... siempre he tenido Audi pero a día de hoy sus nuevas tecnologías no me ofrecen lo que yo pido</i>	Negativo	0,98
<i>@audispain me refiero a que no abandoneis la gasolina y gasoil, siendo referentes en WEC y DTM sería estúpido pasarse a lo eléctrico, la Formula E no tiene público</i>	Negativo	0,99

Tabla 6. Ejemplos de comentarios negativos recopilados en el mes de diciembre sobre la cuenta de Audi España en Instagram.

Finalmente, la segunda explicación que he encontrado para el aumento de los comentarios negativos durante el análisis realizado en diciembre, es que sendos analizadores no son capaces de identificar la **ironía de ciertos emoticonos**, como el fuego. Este comportamiento lo he podido observar personalmente, como usuaria de Instagram, en la propia plataforma puesto que en algunos casos la compañía **oculta ciertos comentarios por ser considerarlos dañinos o spam**. Este hecho indica que aún queda bastante trabajo por realizar para que un algoritmo sea capaz de identificar el contexto en el que se están utilizando tanto los términos como los emoticonos. Algunos de los ejemplos representativos de esta situación, en la que la intención del texto es positiva aunque se ha clasificado como negativo, se pueden observar en la ilustración 37.

Comentario	Sentimiento	Grado de confianza
Queremos el rs3 ya porfavorr 🙄🙄🙄	Negativo	0,53
👊🙄🙄	Negativo	0,76
Brutal	Negativo	0,99
🔥🔥	Negativo	0,94
🔥🔥🔥🙄	Negativo	0,76

Ilustración 37: Comentarios irónicos encontrados en las publicaciones de la cuenta de Audi en España.

6.6. Análisis de patrones de conducta

Finalmente, a partir de los análisis de sentimientos realizados en la sección anterior, a continuación se efectúa un último estudio para obtener los **patrones de comportamiento que muestran los seguidores** de la cuenta de Audi España en Instagram. El principal objetivo que se persigue con este análisis consiste en traducir los resultados obtenidos de los analizadores de sentimientos, a la conducta que demuestra la comunidad de seguidores de una cuenta en particular, durante un determinado período de tiempo. De este modo, se puede confeccionar una especie de **termómetro social** con el que medir el nivel de agrado o de descontento de los usuarios que consumen sus publicaciones.

En el caso de la marca de automóviles, podemos observar en la ilustración 38 que durante el mes de **noviembre** disponían de un **mayor número de seguidores que aportaban comentarios positivos**. Este hecho acompaña a los resultados proporcionados durante el mismo período en la sección anterior, en la que se habían contabilizado más de un 60% de opiniones positivas. Por lo tanto, en este rango temporal se puede afirmar que el contenido publicado por la cuenta de Audi España en Instagram tenía un impacto predominantemente positivo en su comunidad de seguidores.

Análisis de los patrones de comportamiento de los miembros de la comunidad.

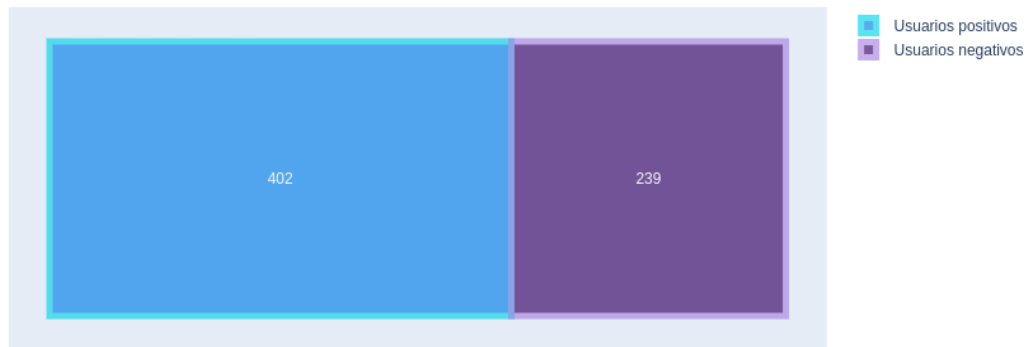


Ilustración 38: Análisis de los patrones de comportamiento de la cuenta de Audi en España durante noviembre.

Por el contrario, durante el mes de **diciembre** el número de comentarios clasificados como negativos aumentó considerablemente, como se ha estudiado en la sección anterior, lo que se refleja en un **aumento de la negatividad** entre los seguidores que publicaron sus opiniones en el contenido subido por la compañía. Esta conclusión se puede apreciar en la ilustración 39, en la que se visualiza cómo el número de usuarios con opiniones mayoritariamente positivas ha disminuido en pos de los usuarios que realizan un mayor número de críticas. Gracias a este tipo de análisis, la compañía podría invertir, de manera eficiente, los recursos necesarios para realizar los estudios que determinen los motivos del desequilibrio asociado al termómetro social generado por la herramienta.

Análisis de los patrones de comportamiento de los miembros de la comunidad.

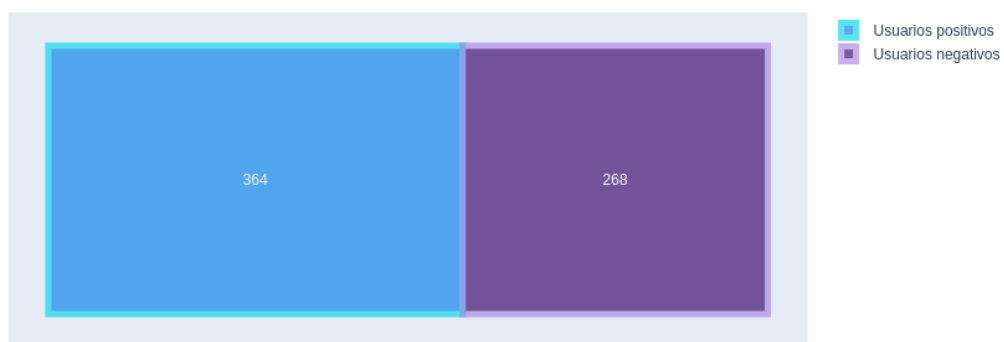


Ilustración 39: Análisis de los patrones de comportamiento de la cuenta de Audi en España durante diciembre.

Capítulo 7

Conclusiones y trabajo futuro

En este último capítulo se recopilan las conclusiones que se han podido extraer durante la realización de este proyecto y la experimentación con la herramienta desarrollada. La primera que me gustaría destacar es la **dificultad de encontrar una fuente de información** medianamente funcional y capaz de proveer un volumen de datos suficiente. Desde hace algunos años, se está incrementando el valor asociado a la información tanto generada como proporcionada por los usuarios en diferentes plataformas sociales. Gracias a esta nueva entidad, se ha generado un **nuevo mercado** en el que la competencia es feroz, y por tanto las empresas se sienten obligadas a implementar y mejorar las medidas de seguridad oportunas para proteger su valiosa información. Es por ello por lo que la mayoría de los propietarios de las redes sociales han **transformado, e incluso eliminado, sus API públicas** con el fin de dotarlas de más robustez y de reducir la cantidad de datos a los que se puede acceder de forma gratuita.

La segunda conclusión a subrayar consiste en el aumento de la influencia comercial que aparece en la mayoría de las redes sociales, aunque especialmente en Instagram. Tal es así, que en una de sus últimas actualizaciones de 2021 se ha incluido una **nueva pestaña que recopila algunas tiendas online**, las cuales se encuentran accesibles directamente para los usuarios de esta plataforma. De este modo, cada vez más las empresas son conscientes de la visibilidad que aportan las comunidades virtuales y por tanto, invierten una mayor cantidad de recursos humanos, monetarios y temporales en mejorar su apariencia y aumentar su presencia con el objetivo de alcanzar un público más amplio e internacional.

Por todos estos motivos, herramientas como las que se plantean en este proyecto son cada vez **más demandadas** por todo tipo de entidades con el objetivo de conocer cuáles son los tipos de contenidos que mejor funcionan, qué estrategias están siguiendo el resto de organizaciones de la competencia y si están consiguiendo el éxito esperado, cómo reaccionan las comunidades de usuarios frente a nuevos productos antes de lanzarlos al mercado, lo que puede ayudar a evaluar los riesgos que conlleva y a estimar un margen de beneficio aproximado, e incluso detectar con qué usuarios sería productivo realizar una colaboración estudiando su capacidad de difusión, su presencia y relevancia en una o varias redes sociales, lo que comúnmente conocemos como **influencers**.

Como **ideas futuras** para continuar desarrollando y mejorando este proyecto, se plantean las siguientes mejoras y nuevas funcionalidades.

- Integración de un **mayor número de fuentes de información** con el objetivo de realizar tanto análisis de forma independiente como en conjunto.
- Diseño, entrenamiento y validación de **modelos inteligentes** para el análisis de sentimientos basado en texto con el objetivo de mejorar la identificación de los términos y emoticonos utilizados de forma irónica.
- Integración de **nuevos análisis** capaces de estudiar las propiedades de las comunidades de seguidos y seguidores estructurándolos en **grafos**, con el objetivo de identificar el grado de centralidad de los nodos, la cohesión de las comunidades, entre otros.

- Implementación de una nueva funcionalidad capaz de **generar informes** con las estadísticas generadas y la información recopilada de la cuenta analizada para su descarga o envío a través de correo electrónico.

Bibliografía

1. Universidad de Bhubaneswar (India), Techniques and Business Analytics , Brojo Kishore Mishra, Deepannita Hazra, Kahkashan Tarannum, Manas Kumar, Business Intelligence using Data Mining, 2016, https://www.researchgate.net/publication/315918436_Business_Intelligence_using_Data_Mining_techniques_and_Business_Analytics
2. Techopedia. Definition – What does Social Network mean?. 2017. <https://www.techopedia.com/definition/4838/social-network>
3. Computer Hope. Social network. 2020. <https://www.computerhope.com/jargon/s/socinetw.htm>
4. Digit. Söuvik Das. The origin and history of social media. 2016. <https://www.digit.in/features/internet/the-origin-and-history-of-social-media-31655.html>
5. Christopher McFadden, A Chronological History of Social Media, 2020, <https://interestingengineering.com/a-chronological-history-of-social-media>
6. 1stWebDesigner. Editorial Team. The history of Social Networking: How it all began!. 2016. <https://1stwebdesigner.com/history-of-social-networking/>
7. Jasmin Leete. What is Usenet? Complete Guide to Usenet & How to Use It in 2020. 2020. <https://www.vpnmentor.com/blog/what-is-usenet-complete-guide/>
8. Centro Virtual de Convenciones de Salud. María Teresa Abreu, Ester Regalado Miranda, Elsa Regalado Miranda. Internet Relay Chat (IRC). <http://www.cencomed.sld.cu/node/38>
9. Ciarán Mc Mahon. Why do we 'like' social media?. 2015. https://thepsychologist.s3.eu-west-2.amazonaws.com/articles/pdfs/0915mcma.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIA3JJOMCSR35UA6UU%2F20200827%2Feu-west-2%2Fs3%2Faws4_request&X-Amz-Date=20200827T182907Z&X-Amz-SignedHeaders=Host&X-Amz-Expires=10&X-Amz-Signature=8f7d11b8ecedf77cefb9c93c52393e75777e200f1a6db6108317fff6e51c8ea52
10. Petter Bae Brandtzaeg, Jan Heim. Why people use social networking sites. 2009. https://www.researchgate.net/publication/221095501_Why_People_Use_Social_Networking_Sites
11. Jimmie Manning. Definition and Classes of Social Media. 2014. https://www.researchgate.net/publication/290514612_Definition_and_Classes_of_Social_Media
12. Biteable. The 7 different types of social media. <https://biteable.com/blog/the-7-different-types-of-social-media/>
13. Universidad de Victoria, Jan Kietzmann. Universidad de Simon Fraser, Kristopher Hermkens e Ian Paul McCarthy. Universidad de Manitoba, Bruno Silvestre. Social Media? Get serious! Understanding the Functional building Blocks of Social Media. 2011. https://www.researchgate.net/publication/227413605_Social_Media_Get_Serious_Understanding_the_Functional_Building_Blocks_of_Social_Media
14. Universidad de Oslo. Petter Bae Brandtzaeg, Jan Heim. A typology of social networking sites users. 2011. https://www.researchgate.net/publication/220131874_A_typology_of_social_networking_sites_users

15. Marysol Villeda. Randy McCamey. Universidad de Calabar, Eyo Emmanuel Essien. Universidad Federal de Wukari, Christian Amadi. Use of Social Networking Sites for Recruiting and Selecting in the Hiring Process. 2019. https://www.researchgate.net/publication/331022707_Use_of_Social_Networking_Sites_for_Recruiting_and_Selecting_in_the_Hiring_Process
16. Elizabeth C. Alexander. Deanna R. D. Mader. Fred H. Mader. Using Social Media During the Hiring Process: A Comparison Between Recruiters and Job Seekers. https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1203&context=ama_proceedings
17. Valentinas Navickas. Adriana Grenčíková. Jana Španková. The Use of Social Media Job Search. 2019. https://www.researchgate.net/publication/330876734_The_Use_of_Social_Media_Job_Search
18. Universidad de Sunderland, Vipin Nadda. Sumesh Dadwal. A. Firdous. Social Media Marketing. 2015. https://www.researchgate.net/publication/297056488_Social_media_marketing
19. Rubathee Nadaraja. Universidad de Ciencia y Tecnología de Malasia, Rashad Yazdanifard. Social Media Marketing SOCIAL MEDIA MARKETING: ADVANTAGES AND DISADVANTAGES. 2013. https://www.researchgate.net/publication/256296291_Social_Media_Marketing_SOCIAL_MEDIA_MARKETING_ADVANTAGES_AND_DISADVANTAGES
20. Rohit Bansal. Rana Zehra Masood. Varsha Dadhich. Social Media Marketing – A Tool of Innovative Marketing. 2014. https://www.researchgate.net/publication/318225418_Social_Media_Marketing-A_Tool_of_Innovative_Marketing
21. Warren Jolly. The 6 Most Effective Types of Social Media Advertising in 2020. <https://www.bigcommerce.com/blog/social-media-advertising/#1-facebook-advertising>
22. Adobe Spark. Los 7 principales sitios de redes sociales que debes tener en cuenta en 2020. <https://spark.adobe.com/es-ES/make/learn/top-social-media-sites/>
23. Medium. Yasu, Top 10 Social Media APIs: Twitter, Facebook, Instagram, and many more. 2019. <https://medium.com/rakuten-rapidapi/top-10-social-media-apis-twitter-facebook-instagram-and-many-more-5c13262c61fe>
24. Medium. Perry Eising. What exactly is an API?. 2017. <https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f>
25. Red Hat. ¿Qué son las API y para qué sirven?. <https://www.redhat.com/es/topics/api/what-are-application-programming-interfaces>
26. Facebook for Developers. API Graph de Instagram. <https://developers.facebook.com/docs/instagram-api/>
27. Facebook for Developers. Instagram Basic Display API. <https://developers.facebook.com/docs/instagram-basic-display-api>
28. GitHub. Repositorio público de la API Instaloader. <https://github.com/instaloader/instaloader>
29. GitHub. Repositorio público de la API LevPasha Instagram. <https://github.com/LevPasha/Instagram-API-python>
30. IGI Global. What is Intelligent Data Analysis. <https://www.igi-global.com/dictionary/intelligent-data-analysis/15023>
31. Universidad de Konstanz, Michael R Berthold. Universidad de Arizona, Paul R. Cohen. Intelligent Data Analysis : Reasoning About Data. Xiaohui Liu. https://www.researchgate.net/publication/220605045_Intelligent_Data_Analysis_Reasoning_About_Data
32. Keith D. Foote. A Brief History of Analytics. 2018. <https://www.dataversity.net/brief-history-analytics/#>

33. Smithsonian Magazine. Joseph Stromberg. Herman Hollerith's Tabulating Machine. 2011. <https://www.smithsonianmag.com/smithsonian-institution/herman-holleriths-tabulating-machine-2504989/>
34. GutCheck. A History of Data Collection, Storage, and Analysis. 2018. <https://www.gutcheckit.com/blog/a-history-of-data/>
35. Masashi Miyazaki. A Brief History of Data Analysis. 2020. <https://www.flydata.com/blog/a-brief-history-of-data-analysis/>
36. Universidad Técnica Federico Santa María, Carlos Valle Vidal. "Análisis Inteligente de Datos: Introducción". 2009. <https://www.inf.utfsm.cl/~cvalle/INF-390/Introduccion.pdf>
37. Jim Frost. The Importance of Statistics. <https://statisticsbyjim.com/basics/importance-statistics/>
38. Laura Mora. Qué es Big Data: fases y elementos. "Análisis Inteligente de Datos: Introducción". 2016. <https://www.ve.com/es/blog/que-es-big-data-fases-elementos>
39. Redacción PowerData. Procesos de datos: sus fases y la generación de valor. 2016. <https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/procesos-de-datos-sus-fases-y-la-generacion-de-valor>
40. Kit Smit. The Best Free and Paid Social Media Analytics Tools. 2019. <https://www.brandwatch.com/blog/social-media-analytics-tools/>
41. Facebook for Business. Learn More About the People that Matter to Your Business with Facebook Audience Insights. 2020. <https://www.facebook.com/business/news/audience-insights>
42. Iris Vermeren. The Best Twitter Analytics Tools. 1. Twitter Analytics. 2019. <https://www.brandwatch.com/blog/twitter-analytics-tools/>
43. Kit Smith. The Best Instagram Analytics Tools. 1. Instagram Insights. 2020. <https://www.brandwatch.com/blog/instagram-analytics-tools/>
44. William Comcowich. Disadvantages of Native Social Media Analytics. 2018. https://glean.info/disadvantages-of-native-social-media-analytics/?doing_wp_cron=1600016102.9604120254516601562500
45. Brent Barnhart. 10 of the best social media analytics tools for marketers. <https://sproutsocial.com/insights/social-media-analytics-tools/>
46. Kit Smith. The Best Free and Paid Social Media Analytics Tools. 2019. <https://www.brandwatch.com/blog/social-media-analytics-tools/>
47. Iason Demiros. The limitations of social media analytics. 2016. <https://blog.qualia.ai/the-limitations-of-social-media-analytics-d3f4ee6521db>
48. KPI Partners New Team, Traditional vs. Agile Software Development Methodologies, 2018. <https://www.kpipartners.com/blog/traditional-vs-agile-software-development-methodologies>
49. ProofHub, Sandeep Kashyap, Traditional vs Agile Project Management Method: Which One is Right for Your Project?. <https://www.proofhub.com/articles/traditional-vs-agile-project-management>
50. Universidad Católica Argentina. Maida, Esteban Gabriel, Pacienza, Julián. Metodologías de desarrollo software. <https://repositorio.uca.edu.ar/bitstream/123456789/522/1/metodologias-desarrollo-software.pdf>
51. FactorialBlog, Marina Camacho, El coste de un trabajador para la empresa [+fórmula], 2020 <https://factorialhr.es/blog/coste-empresa-trabajador/>
52. Gobierno de España, Seguridad Social, Información General, <http://www.seg-social.es/wps/portal/wss/internet/Trabajadores/CotizacionRecaudacionTrabajadores/10721/10957/583>
53. Gobierno de España, Fondo de Garantía Salarial O.A., <https://www.mites.gob.es/fogasa/faqs.html>
54. Gobierno de España, Formación Profesional para el empleo. Aspectos generales, https://www.mites.gob.es/es/Guia/texto/guia_4/contenidos/guia_4_10_1.htm

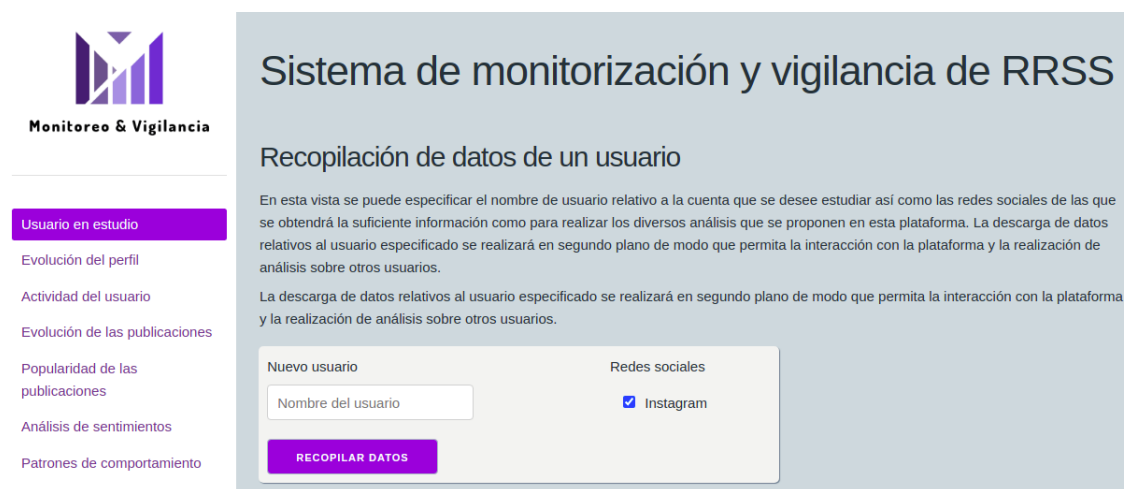
55. Microsoft, Serie de máquinas virtuales, <https://azure.microsoft.com/es-es/pricing/details/virtual-machines/series/>
56. Charles Leifer, librería *huey* de Python, <https://huey.readthedocs.io/en/latest/>
57. SuHun Han, librería *googletrans* de Python, <https://pypi.org/project/googletrans/>
58. LuShan, librería *google-trans-new* de Python, <https://pypi.org/project/google-trans-new/>
59. CJ Hutto, librería *vader-sentiment* de Python, <https://pypi.org/project/vader-sentiment/>
60. Pahul Preet Singh Kohli, Sentiment Analysis - Methods and Pre-Trained Models Review, 2020, <https://pahulpreet86.github.io/sentiment-analysis-methods-and-pre-trained-models-review/>
61. MonkeyLearn, Sentiment Analysis: A Definitive Guide, 2020, <https://monkeylearn.com/sentiment-analysis/>
62. sloria, librería *TextBlob* de Python, <https://pypi.org/project/textblob/>
63. Documentación sobre la librería *TextBlob*, Advanced Usage: Overriding Models and the Blobber Class, https://textblob.readthedocs.io/en/dev/advanced_usage.html
64. Parthvi Shah, Sentiment Analysis using TextBlob, 2020, <https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524>
65. Alan Akbik, librería *flair* de Python, <https://pypi.org/project/flair/>
66. Imad Dabbura, Character-Level Language Model, 2018, <https://towardsdatascience.com/character-level-language-model-1439f5dd87fe>
67. Chris Parmer, librería *dash* de Python, <https://pypi.org/project/dash/>
68. Pallets, Flask, web development, one drop at a time, 2010, <https://flask.palletsprojects.com/en/2.0.x/>
69. Plotly, Plotly Python Open Source Graphing Library, <https://plotly.com/python/>
70. Facebook, React, Una biblioteca de JavaScript para construir interfaces de usuario, <https://es.reactjs.org/>
71. Krasimir Hristozov, 2019, MySQL vs PostgreSQL -- Choose the Right Database for Your Project, <https://developer.okta.com/blog/2019/07/19/mysql-vs-postgres>

Apéndice 1

Manual de usuario

En este apéndice se pretende mostrar las diferentes opciones que se encuentran disponibles en la herramienta desarrollada. Al acceder a la página principal, podemos apreciar la existencia de un **menú lateral** en el que aparecen todas las secciones disponibles. En cada una de ellas, se explica detalladamente la funcionalidad que representa y las instrucciones que se deben aplicar para llevarla a cabo.

A continuación en la ilustración 40 se puede visualizar la primera sección que se corresponde con la **recopilación de datos** sobre un usuario en particular, el cual se puede especificar en la caja de información cuyo título es *Nuevo usuario*. Asimismo, también cabe la posibilidad de seleccionar las fuentes de información que se desean utilizar para la descarga de información, de entre las que se encuentran disponibles actualmente.



The screenshot displays a web application interface. On the left is a purple sidebar with the logo 'Monitoreo & Vigilancia' and a list of menu items: 'Usuario en estudio' (highlighted in purple), 'Evolución del perfil', 'Actividad del usuario', 'Evolución de las publicaciones', 'Popularidad de las publicaciones', 'Análisis de sentimientos', and 'Patrones de comportamiento'. The main content area has a light blue header with the title 'Sistema de monitorización y vigilancia de RRSS'. Below this is a section titled 'Recopilación de datos de un usuario'. It contains two paragraphs of text explaining the data collection process. At the bottom is a form with two columns: 'Nuevo usuario' with a text input field labeled 'Nombre del usuario', and 'Redes sociales' with a checked checkbox for 'Instagram'. A purple button labeled 'RECOPILAR DATOS' is positioned below the form fields.

Ilustración 40: Sección de recopilación de datos sobre un usuario en particular y un conjunto de fuentes de información.

En la ilustración 41 se muestran los **filtros** que se deben completar para realizar un análisis sobre la evolución de un perfil en concreto. En primer lugar se especifica el **usuario** que se desea estudiar para que, a continuación, se seleccione la **red social** de la que se desea obtener su información. Posteriormente, se seleccionan las **fechas** de inicio y de fin de entre las disponibles, en función de la información disponible procedente del usuario y la fuente escogida. Finalmente, se pulsa sobre el botón *Analizar* para que se lleve a cabo el análisis sobre la evolución del perfil del usuario seleccionado a partir de la información recuperada en base a la configuración establecida. Una vez ha finalizado, se mostrarán los resultados de manera gráfica tras el apartado de los filtros, tal y como se puede observar en la ilustración 42. Todo lo explicado para esta pantalla se puede aplicar a los análisis de la actividad del usuario, la evolución del interés de las publicaciones y los patrones de comportamiento.

Usuario en estudio

Evolución del perfil

Actividad del usuario

Evolución de las publicaciones

Popularidad de las publicaciones

Análisis de sentimientos

Patrones de comportamiento

Sistema de monitorización y vigilancia de RRSS

Análisis de la evolución del perfil

El principal objetivo de este análisis reside en estudiar si existe una relación directa entre el número de seguidores, seguidos y contenido publicado. De este modo podrá conocer si a mayor número de publicaciones, existe un mayor número de usuarios que se suscriben a la cuenta para continuar recibiendo el nuevo contenido que se genere. Asimismo tendrá la oportunidad de observar si existe un equilibrio entre el número de seguidores y el número de cuentas que sigue con el objetivo de determinar si el uso de la cuenta es exclusivo para dar a conocer su contenido, o si por el contrario, también la utiliza para conectar con otros miembros de la red social.

Para ello es necesario que indique la fecha de inicio y de fin entre las que obtener la información necesaria para llevar a cabo este análisis sobre el usuario especificado dentro de una red social concreta.

Fecha de inicio	Fecha final	Usuario a estudiar	Red social
<input type="text" value="27-10-2020"/>	<input type="text" value="27-10-2020"/>	<input type="text" value="Audi Spain"/>	<input type="text" value="Instagram"/>
ANALIZAR			

Ilustración 41: Filtros a completar para realizar un análisis sobre la evolución del perfil de un usuario.

Usuario en estudio

Evolución del perfil

Actividad del usuario

Evolución de las publicaciones

Popularidad de las publicaciones

Análisis de sentimientos

Patrones de comportamiento

Sistema de monitorización y vigilancia de RRSS

Análisis de la evolución del perfil

El principal objetivo de este análisis reside en estudiar si existe una relación directa entre el número de seguidores, seguidos y contenido publicado. De este modo podrá conocer si a mayor número de publicaciones, existe un mayor número de usuarios que se suscriben a la cuenta para continuar recibiendo el nuevo contenido que se genere. Asimismo tendrá la oportunidad de observar si existe un equilibrio entre el número de seguidores y el número de cuentas que sigue con el objetivo de determinar si el uso de la cuenta es exclusivo para dar a conocer su contenido, o si por el contrario, también la utiliza para conectar con otros miembros de la red social.

Para ello es necesario que indique la fecha de inicio y de fin entre las que obtener la información necesaria para llevar a cabo este análisis sobre el usuario especificado dentro de una red social concreta.

Fecha de inicio	Fecha final	Usuario a estudiar	Red social
<input type="text" value="27-10-2020"/>	<input type="text" value="14-12-2020"/>	<input type="text" value="Audi Spain"/>	<input type="text" value="Instagram"/>
ANALIZAR			

Evolución del número de seguidores, seguidos y publicaciones.



Ilustración 42: Resultados gráficos del análisis sobre la evolución del perfil de un usuario.

Finalmente, a continuación se muestran dos capturas de las diferentes configuraciones que podemos encontrar para los análisis de popularidad de las publicaciones y los análisis de sentimientos, respectivamente. En el primero de ellos, como se observa en la ilustración 43, existe un parámetro más relativo al **orden ascendente o descendente** de las publicaciones con el objetivo de mostrar el *ranking* de las mejores o las peores, respectivamente. Mientras que en la ilustración 44, el parámetro adicional que se puede configurar permite elegir si el análisis de sentimientos se realizará sobre un conjunto de comentarios o de títulos de publicaciones.

Usuario en estudio

Evolución del perfil

Actividad del usuario

Evolución de las publicaciones

Popularidad de las publicaciones

Análisis de sentimientos

Patrones de comportamiento

Sistema de monitorización y vigilancia de RRSS

Análisis de la popularidad de las publicaciones

Este estudio realiza un análisis acerca de las interacciones que han recibido las publicaciones de un usuario concreto durante un período de tiempo con el objetivo de presentar las características de las diez publicaciones más o menos populares. Dependiendo del medio social escogido, las métricas serán diferentes puesto que cada red social dispone de un conjunto de interacciones comunes pero también contiene algunas que solo aparecen de forma particular.

Para ello es necesario que indique la fecha de inicio y de fin entre las que obtener la información necesaria para llevar a cabo este análisis sobre el usuario especificado dentro de una red social concreta. Asimismo, también podrá seleccionar si desea visualizar las diez publicaciones más o menos populares.

Fecha de inicio	Fecha final	Usuario a estudiar	Red social
<input type="text" value="27-10-2020"/>	<input type="text" value="14-12-2020"/>	<input type="text" value="Audi Spain"/>	<input type="text" value="Instagram"/>
Ranking de publicaciones			
<input type="text" value="Más populares"/>			
<input type="button" value="ANALIZAR"/>			

Ilustración 43: Filtros a completar para realizar un análisis sobre la popularidad de las publicaciones de un usuario.

Usuario en estudio

Evolución del perfil

Actividad del usuario

Evolución de las publicaciones

Popularidad de las publicaciones

Análisis de sentimientos

Patrones de comportamiento

Sistema de monitorización y vigilancia de RRSS

Análisis de la popularidad de las publicaciones

Este estudio realiza un análisis acerca de las interacciones que han recibido las publicaciones de un usuario concreto durante un período de tiempo con el objetivo de presentar las características de las diez publicaciones más o menos populares. Dependiendo del medio social escogido, las métricas serán diferentes puesto que cada red social dispone de un conjunto de interacciones comunes pero también contiene algunas que solo aparecen de forma particular.

Para ello es necesario que indique la fecha de inicio y de fin entre las que obtener la información necesaria para llevar a cabo este análisis sobre el usuario especificado dentro de una red social concreta. Asimismo, también podrá seleccionar si desea visualizar las diez publicaciones más o menos populares.

Fecha de inicio	Fecha final	Usuario a estudiar	Red social
<input type="text" value="27-10-2020"/>	<input type="text" value="14-12-2020"/>	<input type="text" value="Audi Spain"/>	<input type="text" value="Instagram"/>
Ranking de publicaciones			
<input type="text" value="Más populares"/>			
<input type="button" value="ANALIZAR"/>			

Ilustración 44: Filtros a completar para realizar un análisis de sentimientos sobre los títulos o los comentarios de las publicaciones de un usuario.

