UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DISEÑO DE UNA HERRAMIENTA PARA LA CARACTERIZACIÓN FINANCIERA DE USUARIOS DE REDES SOCIALES

ALBERTO SÁNCHEZ LÓPEZ

2017

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Título: Diseño de una herramienta para la caracterización financiera de usuarios de redes sociales

Autor: D. Alberto Sánchez López

Tutor: D. Juan Carlos Yelmo García

Departamento: DIT

MIEMBROS DEL TRIBUNAL

Presidente: D. Juan Antonio de la Puente Alfaro

Vocal: D. Juan Carlos Yelmo García

Secretario: D. Miguel Ángel de Miguel Cabello

Suplente: D. José María del Álamo Ramiro

Los miembros del tribunal arriba nombrados acuerdan otorgar la **calificación** de: ……

Madrid, a de de 2017

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DISEÑO DE UNA HERRAMIENTA PARA LA CARACTERIZACIÓN FINANCIERA DE USUARIOS DE REDES SOCIALES

ALBERTO SÁNCHEZ LÓPEZ

2017

RESUMEN

Los datos están presentes cada vez más en nuestras vidas, y sin embargo apenas nos damos cuenta de sus aplicaciones. Ya existen ejemplos exitosos de monetización, compañías puramente digitales que generan valor a partir de los datos, siendo estos el centro de su actividad. La oferta de herramientas a terceros para la explotación de datos constituye una de las principales propuestas de valor.

La evolución que está sufriendo nuestra cultura y sociedad en la última década ha cambiado la forma de comunicarse e informarse abriendo una nueva era en la que las redes sociales constituyen una de las principales fuentes de información. En estos medios destaca la existencia de determinados personajes o entidades, conocidos como *influencers*, capaces de influir con trascendencia en la forma de actuar y pensar de los usuarios que se encuentran en su red.

Las nuevas tecnologías están cambiando la propuesta de valor de los productos y servicios financieros existentes. Esta tendencia denominada *Fintech*, describe cómo los negocios de siempre asimilan las ideas innovadoras mientras las *startups* irrumpen en el sector. El marketing de este movimiento es un proceso estratégico, en el que el lanzamiento de nuevos productos a un mercado inestable y competitivo, es primordial para alcanzar el éxito.

El objetivo de este proyecto es lograr una ventaja competitiva para el marketing de aplicaciones *Fintech* exprimiendo las oportunidades del marco contextual descrito mediante la elaboración de un caso real de estudio. Para ello, se desarrolla una herramienta con capacidad para extraer datos de *Twitter*, procesar la información obtenida e inferir datos estadísticos almacenándolos en forma de *dataset* y, analizar las colecciones de datos siguiendo una metodología basada en experimentos.

El resultado que proporciona el caso de estudio, es un listado que clasifica a los usuarios más influyentes en *Twitter,* relacionados con la temática, en función de una serie de métricas basadas en la confianza, influencia y relevancia.

PALABRAS CLAVE

Ciencia de los datos, Fintech, Influencer, Marketing, R, REST API, Twitter

SUMMARY

Data is increasingly present in our lives, however, we are hardly aware of the uses they have. There are already successful examples of monetization, purely digital companies that generate value from the data, being the center of their activity. Offering tools to third parties for the exploitation of data constitutes one of the main value propositions.

The progress of our culture and society in the last decade has changed the way of communicating and informing, starting a new era in which social networks are one of the main sources of information. In these media, it stands out the existence of certain characters or entities, known like influencers, with the ability to influence with transcendence in the way of acting and thinking of users inside the network.

New technologies are changing the value proposition of existing financial products and services. This trend, called Fintech, describes how usual business assimilate innovative ideas whereas startups break into the industry. The marketing of this movement is a strategic process, in which the launch of new products to an unstable and competitive market, is essential to achieve success.

The aim of this project is to achieve a competitive advantage for Fintech's marketing applications by taking advantage of the contextual framework opportunities described by working out a real study case. For this purpose, a tool is developed with the capacity to extract data from Twitter, process the information obtained and infer statistical data by storing them in dataset form, and analyze the data collections following a methodology based on experiments.

The result that provides the study case, is a list that classifies the most influential Twitter users, related to the topic, based on metrics in terms of confidence, influence and relevance.

KEYWORDS

Data science, Fintech, Influencer, Marketing, R, REST API, Twitter

ÍNDICE DEL CONTENIDO

1. GLOSARIO Y FIGURAS 1

2. INTRODUCCIÓN 3

2.1. Motivación 3

2.2. Objetivos 3

3. ESTADO DEL ARTE 4

3.1. Contexto 4

3.1.1. Medios sociales 4

3.1.2. Influencers y líderes de opinión 5

3.1.3. Fintech 6

3.1.4. Marketing y Branding 7

3.2. Ciencia de los datos 8

3.3. Análisis de datos 9

3.3.1. Definición 9

3.3.2. Investigación cuantitativa y cualitativa 9

3.3.3. Procedimiento de análisis cuantitativo 10

3.3.4. Estadística descriptiva e inferencial 11

3.4. Herramientas para el análisis de datos 12

3.4.1. R y RStudio 12

3.4.2. Twitter API 13

3.4.3. AI Applied API 17

4. CASO DE ESTUDIO: "IDENTIFICACIÓN DE INFLUENCERS PARA EL MARKETING DE FINTECH" 19

4.1. Introducción al caso 19

4.1.1. Motivación e interés 19

4.1.2. Descripción del caso 19

4.1.3. Caracterización de influencers Fintech 19

4.1.4. Metodología y estructura 20

4.2. Obtención de datos 21

4.2.1. Información relevante 21

4.2.2. Estrategia de recopilación 22

4.2.3. Procesado y limpieza 24

4.3. Modelo analítico y puntuación 25

4.3.1. Factores, subfactores y métricas 25

4.3.2. Sistema de puntuación 29
4.4. Experimentos 31
4.4.1. Experimento 1: Evaluación general de usuarios 31
4.4.2. Experimento 2: Evaluación de datos estadísticos inferidos de sus tweets 33
4.4.3. Experimento 3: Evaluación de datos estadísticos inferidos de sus seguidores y demografía 35
4.5. Ejecución y desarrollo con la herramienta 36
4.5.1. Estructura y preparación 36
4.6. Análisis e interpretación de resultados 36
4.6.1. Fuente de obtención de usuarios 36
4.6.2. Puntuaciones de los experimentos 38
4.6.3. Observación de candidatos 40
5. CONCLUSIONES Y LÍNEAS FUTURAS 51
5.1. Conclusiones 51
5.2. Líneas futuras 52
6. BIBLIOGRAFÍA 53
7. ANEXOS 55
7.1. Manual de ejecución y obtención de datos 55

1. GLOSARIO Y FIGURAS

3rd party apps	Aplicaciones de terceros desarrolladas para funcionar en cualquier SO
API	(*Application Programming Interface*) Interfaz de programación de aplicaciones
CRAN	(*Comprehensive R Archive Network*) Red archivos de documentación
Dataset	Colección de datos normalmente tabulada
Endpoint	Punto de entrada a un servicio, un proceso o un destino de cola en una arquitectura orientada a servicios
FAV/Favorite	Marcar un tweet como especial
Fintech	Contracción de las palabras inglesas *finance* y *technology*.
IDE	(Integrated Development Environment) Entorno de desarrollo integrado
Influencer	Líder de opinión con capacidad de influencia.
JSON	(*JavaScript Object Notation*) Formato de texto ligero para el intercambio de datos
Query	Consulta realizada contra una base de datos
REST	(Representational State Transfer) Transferencia de estado representacional
RT/Retweet	Compartir tweet de otra persona
Timeline	Forma de mostrar una lista de eventos de forma cronológica
Token	Clave que contiene las credenciales de seguridad del login

Tabla 1. Áreas que engloba Fintech....6
Tabla 2. Diferencias investigación cuantitativa y cualitativa. 9
Tabla 3. Formas de acceso a los token Twitter....13
Tabla 4. Límites endpoint según el tipo de autorización....14
Tabla 5. Endpoints relevantes REST API Twitter....16
Tabla 6. Endpoints relevantes de la REST API de AI Applied 18
Tabla 7. Información relevante extraída de los endpoints. 22
Tabla 8. Listado de palabras clave 23
Tabla 9. Factor, subfactor, métrica de confianza 26
Tabla 10. Factor, subfactor, métrica de capacidad de influencia....27
Tabla 11. Factor, subfactor y métrica de relevancia para la campaña....29
Tabla 12. Desarrollo del experimento 1 32
Tabla 13. Lista de palabras clave localización geográfica....32
Tabla 14. Lista de palabras clave relación campaña....33
Tabla 15. Desarrollo del experimento 2 34
Tabla 16. Lista de palabras clave relación campaña....34
Tabla 17. Lista de palabras clave early adopter....34
Tabla 18. Lista de palabras clave finanzas 34
Tabla 19. Desarrollo del experimento 3 35
Tabla 20. Lista de palabras clave relación red followers 36
Tabla 21. Coeficientes de importancia por experimento y factor....40

Tabla 22. Síntesis información @realmadrid....42
Tabla 23. Síntesis información @ShawnMendes....43
Tabla 24. Síntesis información @Pontifex_es....43
Tabla 25. Síntesis información @Rubiu5....45
Tabla 26. Síntesis información @rgarciadelacruz....47
Tabla 27. Síntesis información @martin_scto....48
Tabla 28. Síntesis información @ComasRodriguez....48

Ilustración 1. Uso y actividad de Internet y Redes Sociales en España....4
Ilustración 2. Tipo de actividad del sector Fintech en España....6
Ilustración 3. Actividad financiera mundial Fintech (2010 - 2015)....7
Ilustración 4. Diagrama de Venn de la ciencia de datos por Drew Conway....8
Ilustración 5. Proceso de análisis de datos....10
Ilustración 6. Diagrama del proceso de obtención de datos....22
Ilustración 7. Endpoints empleados en la obtención de datos por etapas....23
Ilustración 8. Gráfico en estrella puntuación factores....30
Ilustración 9. Desarrollo de experimentos....31
Ilustración 10. Representación del nº de usuarios en función de la fuente de obtención "stored_users"....37
Ilustración 11. Representación del nº de usuarios en función de la fuente de obtención "statistic_users"....37
Ilustración 12. Representación del nº de usuarios en función de la fuente de obtención "followers_users"....38
Ilustración 13. Nº de usuarios por rango de puntuaciones E1....38
Ilustración 14. Nº de usuarios por rango de puntuaciones E2....39
Ilustración 15. Nº de usuarios por rango de puntuaciones E3....39
Ilustración 16. Nº de usuarios por rango de puntuaciones E1, E2, E3....40
Ilustración 17. Clasificación de candidatos según el factor confianza....41
Ilustración 18. Gráfico en estrella de candidatos según el factor confianza....41
Ilustración 19. Clasificación de candidatos según el factor capacidad de influencia....44
Ilustración 20. Gráfico en estrella de candidatos según el factor capacidad de influencia....45
Ilustración 21. Clasificación de candidatos según el factor relevancia para la campaña....46
Ilustración 22. Gráfico en estrella de candidatos según el factor relevancia para la campaña....46
Ilustración 23. Clasificación de candidatos total....49
Ilustración 24. Gráfico en estrella de candidatos total....50

Ecuación 1. Media....11
Ecuación 2. Moda....11
Ecuación 3. Mediana....11
Ecuación 4. Varianza....12
Ecuación 5. Desviación típica....12
Ecuación 6. Cálculo de la puntuación de cada factor....29
Ecuación 7. Cálculo de la puntuación de cada subfactor....29
Ecuación 8. Cálculo de la puntuación de cada métrica....30
Ecuación 9. Puntuación de métricas especial ajustada al caso de estudio....30

2. INTRODUCCIÓN

2.1. MOTIVACIÓN

Los usuarios de redes sociales y otros servicios en la red proporcionan de forma intencionada o inadvertida gran cantidad de información de carácter personal que puede utilizarse de manera individual para determinar hábitos sociales o de consumo, datos de localización y contacto, perfil de solvencia financiera, etc. Los datos también pueden agregarse para determinar correlaciones que permiten deducir tendencias de negocio, impacto social de programas de televisión, difusión de epidemias, etc.

Asimismo, existen personajes o entidades en los medios sociales, que son considerados *influencers*. Estos usuarios tienen la capacidad de influir en la forma de actuar o pensar de un determinado sector de la población.

Las tecnologías aplicadas al sector financiero (*Fintech*) han experimentado un crecimiento excepcional en los últimos años y a día de hoy es uno de los principales intereses de las empresas del sector financiero. El marketing de esta tendencia es un asunto del que se puede obtener beneficio ya que existe una clara necesidad de penetración en el mercado.

2.2. OBJETIVOS

- Estudio general de técnicas y metodologías de análisis de datos y contenidos para su compresión y uso. Se elaborará un marco conceptual general y una breve taxonomía y caracterización de cada una de las técnicas y de las principales metodologías.
- Estudio de técnicas de análisis estadístico de datos: recolección, análisis e inferencia.
- Aplicación de técnicas de representación y visualización de datos.
- Aprendizaje del lenguaje R y entorno RStudio para la manipulación de datos y aplicación al análisis estadístico.
- Aplicación del marco conceptual y el entorno de herramientas para el análisis de datos de carácter de interés financiero en redes sociales. En particular, se trabajará con *datasets* de mensajes, usuarios y puntuaciones de la red social *Twitter* para estudiar la idoneidad de usuarios desde la perspectiva de su interés para proveedores de servicios F*intech* enfocado a campañas de marketing y selección de usuarios prescriptores para nuevos servicios y aplicaciones.
- Redacción de una memoria que cumpla con la normativa de la ETSIT-UPM e incluya: Glosario, Introducción, Estado del arte, Caso de estudio, Conclusiones, Líneas de estudio futuras, Bibliografía y Anexos.

3. ESTADO DEL ARTE

3.1. CONTEXTO

3.1.1. MEDIOS SOCIALES

Los **medios sociales** *(social media en inglés)* son «un grupo de aplicaciones basadas en Internet que se desarrollan sobre los fundamentos ideológicos y tecnológicos de la Web 2.0, y que permiten la creación y el intercambio de contenidos generados por el usuario» [1]. De manera más sencilla, se pueden definir como «plataformas online orientadas a la interacción humana síncrona y asíncrona con un alcance global y local» [2].

Internet ha permitido el avance de estas plataformas que soportan interacciones entre usuarios de cualquier tipo. Actualmente, encontramos redes sociales para compartir videos, imágenes, mensajes, estados de ánimo, proyectos, entradas de blog, información profesional, relaciones personales, etc.

Las redes sociales constituyen hoy por hoy una fuente inagotable de intercambio de información, intereses y hasta crecimiento empresarial, ya que muchas de ellas se han convertido en un vehículo de posicionamiento de mercado. [3]

En España [4], la **digitalización** llega con 46,09 millones de población, 35,71 millones de usuarios activos en Internet y *22 millones (48% de la población) de usuarios activos en medios sociales*. El tiempo medio de uso de Internet es de 3h y 47min, dedicando *1h 36min a las redes sociales*.

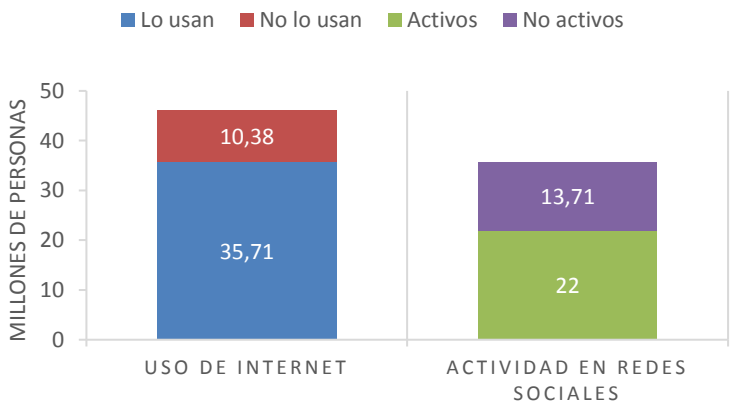


Ilustración 1. Uso y actividad de Internet y Redes Sociales en España

Entre las redes sociales más populares (obviando los sistemas de Mensajería/Chat) se encuentran Facebook, Twitter, Google+, Instagram, LinkedIn y Pinterest, ordenadas según la relación usuarios/actividad.

Estas complejas estructuras pueden ser representadas y estudiadas por medio de programas que acceden a sus datos mediante las *API*. Se capturan, representan y analizan los datos para

obtener cualquier tipo de información que pueda resultar relevante, haciendo visible estadísticas y comportamientos de la sociedad.

3.1.1.1. Twitter

Twitter es una red social global gratuita, online y sin un modelo de negocio claro que combina elementos de blog, mensajes de texto, y contenidos multimedia. Fue creada en 2006 y desde entonces ha llegado a obtener 313 millones de usuarios activos [5]. Tal ha sido su éxito en España que la Real Academia Española (*RAE*) ha incorporado las siguientes palabras al diccionario "*tuit*", "*tuitero*" y "*tuitear*".

Los usuarios escriben mensajes de texto o *tweets,* de 140 caracteres a cualquiera que haya elegido recibir mensajes de ese usuario. Cada *tweet* puede contener texto simple, *#hashtags*, @usuarios, $símbolos, enlaces, imágenes, vídeos, localizaciones, etc., lo que permite compartir contenido con diferentes objetivos. Estos propósitos pueden ser desde interactuaciones, noticias, servicios y actividad hasta funciones de marketing y publicidad.

Esta red social es el punto de unión de otros medios, permitiendo compartir contenidos de terceros. Convirtiéndose así, en el factor determinante que genera interés para el caso de estudio, puesto que se pueden conseguir conjuntos de datos extensos y diversidad para el análisis.

3.1.2. *INFLUENCERS* Y LÍDERES DE OPINIÓN

Un ***influencer*** es una persona que, en un grupo determinado, ejerce una mayor influencia por su estatus de experto y fuente fiable.

Un **líder de opinión** viene dado por su conocimiento sobre la materia y su notoriedad. Sin embargo, no tiene necesariamente la capacidad de influir en las opiniones de los demás.

Estos personajes deben ser [6]:

- ***Carismático***: en el sentido de ser atractivos y creíbles para el sector al que se dirigen
- Poseer ***conocimientos específicos y superiores al promedio***: la credibilidad está respaldada por la habilidad que este tiene para entender el tema de interés.
- ***Identificarse con el público*****:** deben tener un vínculo poderoso y suelen ser elegidos entre los miembros de cada grupo.
- ***Reconocimiento*****:** la imagen de un líder debe ser fácilmente reconocible por su grupo.

De este modo, estos personajes son los receptores de los medios de comunicación. Analizan y procesan la información que reciben para transmitirla a su público de influencia.

Este procedimiento que se desarrolla en dos etapas (medios de comunicación → líderes de opinión/*influencers* → público) es conocido como la "***Teoría de dos pasos***" [7].

3.1.3. *FINTECH*

El origen de ***Fintech*** es el acrónimo de las palabras *Financial Technology* Se define como la industria económica formada por compañías que usan la tecnología para crear sistemas financieros más eficientes [8]. Por extensión, se puede usar para nombrar a las empresas especializadas en las nuevas tecnologías que quieren hacerse con una parte del mercado dominada por las grandes compañías.

Las iniciativas que propone, están enfocadas a las necesidades del consumidor. Definiendo de este modo, aquellos nuevos servicios financieros basados en la innovación tecnológica.

A nivel global, destacan las siguientes áreas:

Banca móvil	Big data y modelos predictivos	Compliance
Crowdfunding	Crowdlending	Criptomonedas y monedas alternativas
Forex (mercado de divisas)	Gestión automatizada de procesos y digitalización	Gestión del riesgo
Pagos y transferencias	Préstamos P2P	Insurtech (Seguros)
Seguridad y privacidad	Servicios de asesoramientos financiero	Trading

Tabla 1. Áreas que engloba Fintech

Fintech surgió con fuerza en España recientemente, «el 74% de las *startups* se han creado entre 2011 y 2014» [9]. Destacando que, el 50% de las empresas tienen un enfoque directo al cliente (***B2C**, Business to Consumer*), mientras que el 37% se dedica a los negocios entre empresas (***B2B**, Business to Business*), con el 13% restante incluyendo servicios para ambos. En el siguiente gráfico se presenta el **mapa del sector** en España [10]:

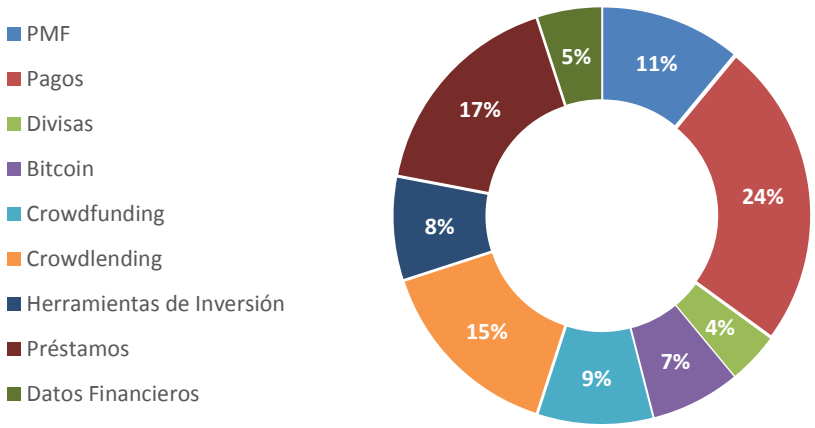


Ilustración 2. Tipo de actividad del sector Fintech en España

Según un estudio publicado por *Accenture* [11], la **financiación** de las empresas *Fintech* sigue creciendo a nivel internacional, alcanzando una cifra de 5.300 millones de dólares en el primer trimestre de 2016, un 67% superior respecto al mismo periodo el año anterior.

Descartando de este modo un posible "pinchazo" de la considerada burbuja y continuando con su tendencia al alza a nivel mundial:

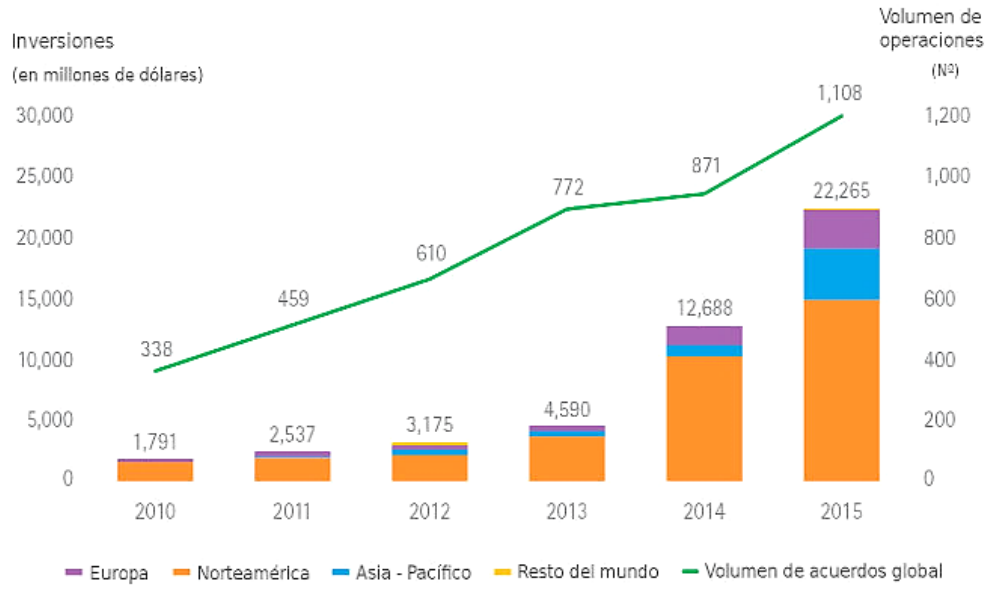


Ilustración 3. Actividad financiera mundial Fintech (2010 - 2015)

Para aprovechar este fenómeno en auge, las empresas necesitarán apoyarse en buenas campañas de marketing, siendo éstas, unas de las principales motivaciones que impulsan a realizar este caso de estudio.

3.1.4. MARKETING Y BRANDING

Por **marketing** entendemos el conjunto de técnicas y estudios que tratan de mejorar la comercialización de un producto o servicio, mientras que, el **branding** es la técnica que emplean las agencias para tratar de resaltar el poder de una marca, es decir, los valores de la empresa que les diferencian con respecto de sus competidores intentando causar el mayor impacto en el mercado. [12]

Las empresas, ya sean *startups* o compañías consolidadas, requieren de campañas de marketing y branding tanto para el lanzamiento de nuevos productos y servicios, como para la reputación de la marca. De este modo las motivaciones principales serán la de captar la atención del mayor público *target*, competir contra aplicaciones similares y fidelizar sus usuarios o clientes.

3.1.4.1. Marketing de influencers

El **marketing de *influencers*** [13] ha llamado la atención de los profesionales por su comprobada efectividad y resultados. Esta estrategia, consiste en la ***identificación*** de líderes de opinión con capacidad de influencia relevantes para una marca y servicio, para su posterior ***contacto*** y ***acuerdo de objetivos***. La identificación no es una tarea sencilla, ya que no es necesariamente una persona famosa o un periodista, si no que puede ser un personaje activo en redes sociales con gran visibilidad y efecto *viral*. El fin de esta técnica es la de "usar" a estos

personajes para *obtener una actitud favorable* en un producto o servicio y *mejorar la imagen de la marca*.

Un estudio realizado por *Augure* [14], avala los resultados de este tipo de campañas. Afirma, que prácticamente la mitad de las empresas encuestadas, destinan entre un 1% y un 20% de su presupuesto en este tipo de publicidad. Además, el 69% de los profesionales encuestados determinaron que esta estrategia es efectiva y que se obtienen resultados favorables.

3.2. CIENCIA DE LOS DATOS

La **ciencia de datos** o *data science* es un campo interdisciplinario que involucra los procesos y sistemas para extraer conocimiento o un mejor entendimiento de grandes volúmenes de datos en sus diferentes formas (estructurados o no estructurados) y formatos [15].

El científico *Drew Conway* planteó una representación gráfica de las áreas que abarca utilizando un diagrama de Venn [16]:

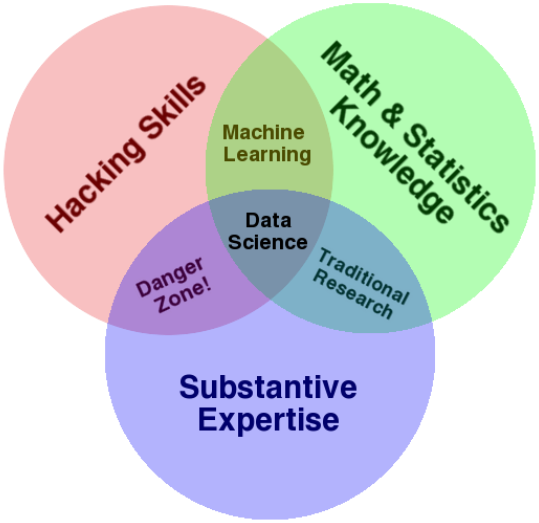


Ilustración 4. Diagrama de Venn de la ciencia de datos por Drew Conway

La ciencia de datos engloba 3 áreas esenciales que corresponden con:

- **Habilidades informáticas** (*Hacking Skills*): aptitudes para extraer, ordenar, analizar y manipular los datos de las fuentes de información utilizando diferentes lenguajes de programación.
- **Conocimientos estadísticos y matemáticos** (*Math and Statistics Knowledge*): competencia en la interpretación y en el procesado de los datos obtenidos con las herramientas oportunas.
- **Experiencia en el entorno** (Substantive Expertise): conocimiento y destreza del entorno que motiva a plantear nuevos escenarios e hipótesis.

Para poder hablar de ciencia de datos, tendrán que incluirse estas tres áreas sin excepción, en caso contrario estaríamos hablando de otras tendencias:

- *Machine Learning*: no se tiene conocimiento del entorno de trabajo, en el que el resultado probablemente no tenga una finalidad concreta.
- Investigación tradicional (*Traditional Research*): carece de habilidades informáticas, que no permiten el manejo de información y procesado rápido.
- Zona comprometida (*Danger zone!*): no se posee conocimientos matemáticos y estadísticos, pudiendo procesar los datos de forma incorrecta.

3.3. ANÁLISIS DE DATOS

3.3.1. DEFINICIÓN

El **análisis de datos** es el proceso de investigación, limpieza, transformación y modelado de datos que tiene por objetivo el hallazgo de información valiosa, la resolución de problemas y la toma de decisiones [17].

3.3.2. INVESTIGACIÓN CUANTITATIVA Y CUALITATIVA

La **investigación cuantitativa** tiene por objetivo estudiar las propiedades y fenómenos cuantificables y sus relaciones. Emplea modelos matemáticos, teorías e hipótesis que buscan responder a preguntas tales como cuál, dónde y cuándo.

La **investigación cualitativa** se emplea para recolectar datos no cuantificables, con el propósito de explorar y describir la realidad. Busca explicar las razones de los diferentes aspectos respondiendo a preguntas como el por qué y el cómo. Se basa en la toma de pequeñas muestras como la observación de grupos de población reducidos. [18]

A continuación, se detallan las principales diferencias de estos tipos de investigación:

Investigación cuantitativa	Investigación cualitativa
Basada en la inducción probabilística del positivismo lógico	Centrada en la fenomenología y comprensión
Medición penetrante y controlada	Observación naturista sin control
Objetiva	Subjetiva
Inferencias más allá de los datos	Inferencias de sus datos
Confirmatoria, inferencial, deductiva	Exploratoria, inductiva y descriptiva
Orientada al resultado	Orientada al proceso
Datos "sólidos y repetibles"	Datos "ricos y profundos"
Generalizable	No generalizable
Particularista	Holista
Realidad estática	Realidad dinámica

Tabla 2. Diferencias investigación cuantitativa y cualitativa.

3.3.3. PROCEDIMIENTO DE ANÁLISIS CUANTITATIVO

El proceso consta de una serie de etapas que se resumen a continuación [19]:

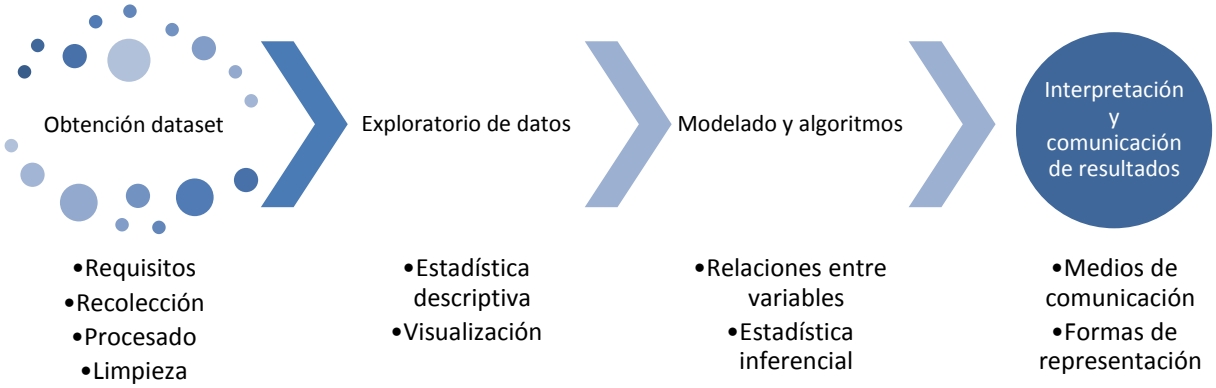


Ilustración 5. Proceso de análisis de datos

3.3.3.1. Requisitos, recolección, procesado y limpieza

La primera fase es la búsqueda de fuentes de información de las que se obtendrán los datos cumpliendo con una serie de **requisitos**, unidad experimental o segmento relacionado con el caso de estudio. Una vez localizadas, se procederá a la extracción y **recolección** de datos. Estos datos serán **procesados** de forma ordenada en tablas y se iniciará una fase de **limpieza** en la que se previenen y corrigen errores, incluyendo, duplicados, caracteres no legibles, etiquetas no relevantes, segmentación de columnas, ortografía, etc.

3.3.3.2. Exploratorio de datos

Una vez finalizadas las etapas anteriores, comienza una fase de análisis con el objeto de entender los mensajes contenidos en los datos. Se apoya en la **estadística descriptiva** y en la **visualización** para comprender los datos.

3.3.3.3. Modelado y algoritmos

Se aplican fórmulas y algoritmos sobre los datos para encontrar **relaciones entre variables** (correlación o causalidad). Se basa en técnicas de **estadística inferencial** para el rastreo de relaciones.

3.3.3.4. Interpretación y comunicación de resultados

Una vez realizado el análisis pertinente, se interpretan los resultados obtenidos en las fases anteriores y se estudian las diversas formas de presentarlas a los usuarios.

3.3.4. ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

La **estadística** se ocupa de los métodos científicos para recolectar, organizar, resumir, presentar y analizar datos, así como sacar conclusiones válidas y tomar decisiones con base a este análisis. Existen dos tipos de estadística [20]:

- La **estadística descriptiva o deductiva** trata el recuento, ordenación y clasificación de los datos obtenidos de las observaciones. Incluyendo la construcción de tablas, gráficos y cálculo de parámetros.
- La **estadística inferencial o inductiva** utiliza los resultados de la estadística descriptiva y se apoya en el cálculo de probabilidades para la obtención de conclusiones sobre una población a partir de resultados obtenidos de una muestra.

3.3.4.1. Medidas de tendencia central

Este tipo de medidas permiten identificar y ubicar el punto (valor) alrededor del cual tienden a reunirse los datos [21].

- **Media**: es una medida de tendencia central calculada a partir de la suma aritmética de un conjunto de valores.

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \quad \left| \begin{array}{l} \bar{x} = \text{media} \\ \text{n} = \textit{número de valores} \\ x_i = \textit{valores} \end{array} \right.$$

Ecuación 1. Media

- **Moda**: es el valor que se repite con mayor frecuencia en una distribución de datos.

$$\text{M} = L_i + \left(\frac{D_1}{D_1 + D_2}\right) A_i$$

$L_i = inferior\ de\ la\ clase\ modal$
$D_1 = delta\ \ de\ frec.abs.modal\ y\ frec.abs.premodal$
$D_2 = delta\ de\ frec.abs.modal\ y\ fre.abs.postmodal$
$A_i = amplitud\ del\ intervalo\ modal$

Ecuación 2. Moda

- **Mediana**: representa el valor de la variable de posición central en un conjunto de datos ordenados.

$$\text{M} = x_{\frac{(n+1)}{2}} \quad \left| \begin{array}{l} \text{n} = \textit{número de valores impar} \\ \text{M} = \textit{mediana} \\ x = \textit{valor} \end{array} \right. \qquad \text{M} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \quad \left| \begin{array}{l} \text{n} = \textit{número de valores par} \\ \text{M} = \textit{mediana} \\ x = \textit{valor} \end{array} \right.$$

Ecuación 3. Mediana

3.3.4.1. Medidas de dispersión

Este tipo de medidas permiten reconocer cuanto se dispersan los datos alrededor del punto central, es decir, indican en qué medida se desvían las observaciones alrededor de su promedio aritmético [22].

- **Varianza**: identifica la diferencia promedio que hay entre cada uno de los valores respecto a su punto central (media).

$$\sigma^2 = \frac{1}{N}\left(\sum X_i - \bar{\mu}\right)^2 \quad \left| \begin{array}{l} \sigma^2 = varianza \\ \mathrm{N} = número\ de\ observaciones \\ X_i = valores \\ \bar{\mu} = media\ poblacional \end{array} \right.$$

Ecuación 4. Varianza

- **Desviación típica**: permite determinar el promedio aritmético de fluctuación de los datos respecto a su punto central o media.

$$\sigma = \sqrt{\sigma^2} \quad \left| \begin{array}{l} \sigma^2 = varianza \\ \sigma = desviación\ típica \end{array} \right.$$

Ecuación 5. Desviación típica

3.4. HERRAMIENTAS PARA EL ANÁLISIS DE DATOS

3.4.1. R Y RSTUDIO

3.4.1.1. R (lenguaje de programación)

R [23] es un entorno y lenguaje de programación enfocado al **análisis estadístico.** Es una implementación de software libre del lenguaje S con soporte de alcance estático. Se trata de uno de los lenguajes más empleados en la investigación por la comunidad estadística, siendo popular en otros campos tales como la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras.

3.4.1.2. RStudio (entorno)

RStudio [24] es un entorno de desarrollo integrado (***IDE***) para R. Incluye una consola, editor de sintaxis para el apoyo de ejecución del código y herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

3.4.1.3. Bibliotecas

R incluye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo o traficación. Para este trabajo serán relevantes las librerías (documentadas en **CRAN**): “httr”, “jsonlite”, “fmsb”, “ggplot2”, “moodest” y “xlsx”.

3.4.2. TWITTER API

3.4.2.1. Portal para desarrolladores

La red social *Twitter* ofrece una **plataforma para desarrolladores** [25] en la que se encuentra una gran variedad de documentación, herramientas y *API* a disposición de cualquier usuario.

Destacan los módulos *Fabric* (herramientas de desarrollo para aplicaciones móvil), *Twitter for Websites* (widgets, botones y herramientas de scripting), *Cards* (tarjetas para mostrar contenido adicional en un tweet), *OAuth* (tipo de conexión de usuarios para enviar peticiones seguras), *REST API* (provee acceso para leer y escribir información), *Streaming API* (permite el envío continuo de respuestas de la *REST API* sobre una conexión HTTP larga y abierta), *Ads API* (proporciona a los socios una manera de integrar la gestión de la publicidad de *Twitter* en su producto), *MoPub* (servidor de intercambio de publicidad móvil).

Se tratará únicamente la documentación de los módulos ***OAuth*** y ***REST API***, ya que son las de interés para nuestro proyecto.

3.4.2.2. Portal de aplicaciones

El **portal de aplicaciones** [26] gestiona las aplicaciones asociadas a una cuenta de usuario. En él se puede crear una aplicación, usando las credenciales de la cuenta de *Twitter*, y obtener acceso a las claves de autenticación OAuth *Consumer Key*/*Consumer Secret.* Éstas se usan para realizar la conexión de la aplicación a la *API*. Además, se encuentran los *token* de acceso, *Access Token*/*Access Token Secret*, que permiten establecer conexión del usuario, para obtener funcionalidades de usuario extra. Adicionalmente, existen otros métodos de obtener acceso a los *token* de usuario que pueden ser de gran utilidad en función de la aplicación a desarrollar:

Inicio de sesión con Twitter - mediante un formulario user/password
3-legged OAuth - igual que el anterior, pero se pide acceso por cada petición
PIN-based OAuth - mediante un código PIN
xAuth - a partir de user/password sin uso de acceso web
OAuth Echo - útil cuando existe un tercero involucrado

Tabla 3. Formas de acceso a los token Twitter

3.4.2.3. Seguridad y Acceso

Twitter utiliza ***OAuth*** [27] para proporcionar acceso a su *API* enviando peticiones seguras y autorizadas. Con *OAuth* los usuarios no comparten contraseñas con ***3rd party applications***.

Existen dos modelos de autenticación en la versión actual (*API v1.1's Authentication Model*):

- **Application-User Authentication**: autenticación tanto de la aplicación como del usuario mediante un *token* de acceso, garantizando todos los métodos en relación con el usuario autorizado. El número de peticiones está limitado *por usuario.*

- **Application-Only Authentication**: la aplicación crea las peticiones sin un contexto de usuario. No soporta algunos métodos, en concreto, *POST*, conexiones a *Streaming endpoints*, *Geo endpoints*, acceso a Mensajes directos y credenciales de la cuenta. El número de peticiones está limitado por aplicación.

3.4.2.4. Limitaciones

La *API* de *Twitter* está limitada por el número de peticiones que se pueden enviar a los *endpoints*. Existen dos tipos de restricciones según el método de autenticación empleado, por usuario o por aplicación. Éstas son independientes y la cantidad de accesos es diferente. Se gestiona en intervalos de tiempo de 15 minutos, es decir, una vez consumidos los accesos al *endpoint*, no se podrán realizar más peticiones hasta que transcurran los 15 minutos.

La tabla de **límites** [28] de los *endpoints* necesarios para este proyecto se representan a continuación:

Method/Endpoint	Family	Request per 15 min UserAuth	Request per 15 min AppAuth
GET search/tweets	search	180	450
GET users/search	users	900	-
GET users/lookup	users	900	300
GET users/show	users	900	900
GET statuses/user_timeline	users	900	1500
GET friends/list	friends	15	15
GET followers/list	followers	15	15

Tabla 4. Límites endpoint según el tipo de autorización

Como se puede observar en la tabla, las autenticaciones *application-user* y *application-only* tienen limitaciones diferentes según el *endpoint*. Es recomendable usar la más apropiada para cada tipo de petición o incluso una combinación de ambas.

El *endpoint GET application/rate_limit_status* devuelve información acerca de dichos límites, en concreto, las peticiones disponibles y restantes de cada familia de recursos.

3.4.2.5. Acceso y obtención de datos en R

El paquete ***httr*** [29] de R permite establecer conexiones y obtener datos de una *API REST*, simplificando el proceso de autenticación y el de petición de datos a un *endpoint*.

La librería está organizada en torno a 5 métodos HTTP: *GET, PATCH, POST, HEAD, PUT y DELETE*. Cada petición devuelve una respuesta objeto con información relevante en formato *html, xml, json, png o jpeg*. Soporta autenticación *OAuth 1.0* (LinkedIn, Twitter and Vimeo) y OAuth 2.0 (Facebook, Github, Google).

Para el **acceso a la *API* de *Twitter***, se utilizan las funciones que ofrece el paquete *httr* de *OAuth 1.0*, las cuales se describen a continuación:

```
#Crea una aplicación OAuth.
#appname: "Twitter"; key: consumer_key; secret: consumer_secret
oauth_app(appname, key, secret = NULL)
```

```
#Firma una petición OAuth
#app: oauth_app; token: access_token, secret: access_token_secret
sign_oauth1.0(app, token = NULL, token_secret = NULL, ...)
```

La firma para la autenticación *application-user* se obtiene rellenando todos los parámetros indicados en las funciones anteriores, mientras que para la autenticación *application-only* tan sólo usa el *consumer_key* y el *consumer_secret*, ya que los *tokens* son propios de la autenticación de usuario.

Una vez obtenida la firma, se accede a los contenidos de la *API* usando las funciones *GET* y *POST* que proporciona la librería y soporta *Twitter*.

```
#Obtener contenido dada una url endpoint
#url: "https://api.twitter.com/1.1/"; config: sig (sign_oauth)
GET(url = NULL, config = list(), ..., handle = NULL)
```

Esta función devuelve información en un archivo **JSON** que se debe parsear a un objeto en R con las funciones que proporciona la librería ***jsonlite*** [30].

3.4.2.6. Endpoints (REST API)

Twitter ofrece una ***REST API* documentada** [31] que proporciona acceso a los datos, tanto de lectura como de escritura. Los *endpoints* son los enlaces a los que se pueden mandar peticiones para que realicen alguna operación obteniendo una respuesta. Esta *API* acepta peticiones HTTP de tipo *GET* y *POST*, cuyos parámetros se encuentran en la misma *url* y las respuestas serán devueltas en formato JSON. Existe una lista de 93 *endpoints* documentados, de los cuales, analizamos los que son de relevancia para este proyecto.

GET search/tweets	Devuelve una lista de tweets que guarden relación con la *query* introducida como parámetro.
	q (required): acepta una query de hasta 500 caracteres. ***lang (optional):*** restringe resultados al lenguaje indicado. ***result_type (optional)***: especifica el tipo de tweet pudiendo ser: *mixed, recent o popular*. ***count (optional)***: número de resultados, hasta 100 resultados. ***until (optional)***: devuelve resultados anteriores a una fecha YYYY-MM-DD, hasta 7 días de antigüedad. ***since_id (optional)***: devuelve resultados con un ID superior al indicado, es decir, más recientes. ***max_id (optional)***: devuelve resultados con un id inferior al indicado, es decir, más antiguos.

GET users/search	Proporciona una búsqueda relevante de cuentas públicas de Twitter, que guarden relación (id, nombre, descripción…) con una *query* introducida como parámetro.
	q (required): acepta una query de hasta 500 caracteres. ***count (optional)***: número de resultados, hasta 20 resultados.
GET users/lookup	Devuelve información acerca de los usuarios buscados a partir de su *user_id* o su *screen_name.* Hasta 100 resultados por búsqueda.
GET users/show	Devuelve información acerca de un usuario buscado a partir de su user_id o su screen_name.
GET statuses/user_timeline	Devuelve los 200 *tweets* más recientes del usuario indicado a partir de su *user_id* o su *screen_name.*
	screen_name (optional): usuario a buscar. **since_id (optional):** devuelve resultados con un ID superior al indicado, es decir, más recientes. ***max_id (optional)***: devuelve resultados con un id inferior al indicado, es decir, más antiguos. **count (optional):** número de resultados, hasta 200 resultados. **trim_user (optional):** marcar false para que no devuelva información del usuario **exclude_replies (optional):** marcar true, para prevenir duplicados. **contributor_details (optional):** marcar false para que no devuelva información de los contribuidores **include_rts (optional):** marcar false, para que no incluya los RT de otros usuarios.
GET friends/list **GET followers/list**	Devuelve una lista de hasta 200 personas seguidas o que siguen a un usuario. (*Friends/followers*)
	screen_name (required): usuario a buscar. **cursor (semi):** -1 primera página, orden en páginas de los seguidos. **count (optional):** número de resultados, hasta 200 resultados. **skip_status (optional):** marcar 1, para que no incluya statuses. **inlcude_user_entities (optional):** marcar false, para no incluir entities.

Tabla 5. Endpoints relevantes REST API Twitter

A modo de ejemplo, escribimos un ejemplo de cómo sería el *endpoint* y el resultado obtenido:

```
#Devuelve una lista de 5 tweets que contenga la palabra “twyp” obteniendo los resultados más recientes y en español.

https://api.twitter.com/1.1/users/search.json?q=twyp&lang=es&result_type=recent&count=5
```

Para obtener listas de resultados mayores a los que están limitadas cada petición a un *endpoint,* se debe hacer uso de los parámetros *max_id* y *since_id*, que permiten indicar a partir de que id y hasta que id se devuelven los resultados. De la misma forma se recurre al parámetro *cursor* para obtener listados de resultados de páginas no repetidas.

3.4.3. AI APPLIED API

Esta *API* permite **analizar textos** en línea para obtener probabilidades de resultados que no se pueden obtener de forma directa de las redes sociales, tales como palabras clave, género, edad, sentimientos, etc.

3.4.3.1. Acceso y límites

Las peticiones se realizan a los *endpoints* a través de peticiones a la siguiente *url* en la que se indica el nombre del endpoint deseado y un *request* en formato *JSON* con los parámetros a introducir, obteniendo una respuesta con los resultados.

```
http://api.ai-applied.nl/api/name_endpoint/request={...}
```

Dentro del *request* hay que introducir el *api_key* de acceso, con la cual se pueden hacer 5000 peticiones de forma gratuita.

3.4.3.2. Endpoints (REST API)

AI Applied ofrece una ***REST API* documentada** con los *endpoints* (enlaces a los que se pueden mandar peticiones para que realicen alguna operación obteniendo una respuesta). Esta *API* acepta peticiones HTTP de tipo *GET* y *POST*, cuyos parámetros se encuentran en la misma *url* y las respuestas serán devueltas en formato *JSON*. Existe una lista de 7 *endpoints* documentados, de los cuales, analizamos los que son de relevancia para este proyecto.

- **Demographics API** [32]

La demografía se puede entender como estadística cuantificable dado un segmento. Con el uso de esta *API*, se puede estimar la **edad** y el **género** de un *usuario* y un *texto*.

Dependiendo del lenguaje y del contexto, la *API* estima el género como “male”, “female” o “unknown” con una precisión del 65%, y la edad como “12-20 years”, “21-30 years”, “31-40 years”, “41-55 years”, “56-65 years” con un 58% de precisión.

- **Sentiment Analysis API** [33]

Con el análisis de sentimientos que ofrece, se puede detectar la **actitud, opinión y sentimiento** de un texto escrito.

Dependiendo del lenguaje y del contexto, estima el sentimiento como “positive”, “negative”, “neutral” o “unknown” con una precisión del 80%

A continuación, describimos los *endpoints* con sus correspondientes parámetros de entrada:

GET demographics_api	Devuelve el género del usuario introducido y la edad, junto con el factor de confianza de ambos.
	api_key (required): token de acceso ***text (required):*** texto a analizar ***user (required)***: nombre de usuario ***language_iso (required)***: idioma (“spa”) ***id (required)***: id para múltiples peticiones
GET demographics_api	Devuelve la positividad o negatividad del mensaje junto con su factor de confianza del texto introducido
	api_key (required): token de acceso ***text (required):*** texto a analizar ***classifier (required)***: “default” o “subjective” ***language_iso (required)***: idioma (“spa”) ***id (required)***: id para múltiples peticiones

Tabla 6. Endpoints relevantes de la REST API de AI Applied

4. CASO DE ESTUDIO: "IDENTIFICACIÓN DE *INFLUENCERS* PARA EL MARKETING DE *FINTECH*"

4.1. INTRODUCCIÓN AL CASO

4.1.1. MOTIVACIÓN E INTERÉS

En una empresa el "darse a conocer" y el "prestigio" pueden ser tan importantes como la calidad de sus productos y servicios, por tanto, dependen de las técnicas de **marketing** para penetrar y dominar el mercado.

Entre las redes sociales existe un mundo de los denominados ***influencers***, que está dando mucho de qué hablar en estos aspectos, y desde hace tiempo, se explotan en los negocios para obtener mejores resultados.

La tendencia del ***Fintech*** ha experimentado un crecimiento excepcional estos últimos años y a día de hoy es uno de los principales intereses de las empresas del sector financiero.

4.1.2. DESCRIPCIÓN DEL CASO

El caso de estudio que se propone es el de la "**Identificación de *influencers* para el marketing de *Fintech***".

La propuesta consiste en el rastreo de personajes influyentes adecuados para la promoción y el testeo de aplicaciones relacionadas con *Fintech.* Estos individuos se encontrarán en las redes sociales, en concreto, la investigación se centra en *Twitter*.

Una vez identificados, se podrán clasificar según su relevancia para la campaña y escoger los sujetos afines para dar a conocer las aplicaciones, formular críticas constructivas, realizar valoraciones positivas y promocionar la marca.

4.1.3. CARACTERIZACIÓN DE *INFLUENCERS* FINTECH

En este apartado se caracterizan los individuos que son motivo de búsqueda analizando sus principales atributos, cualidades y enfoques.

Los **atributos** que personifican a un *influencer* en el área *Fintech,* serán aquellos propios de un líder de opinión, con intereses, redes e influencias en los sectores de las finanzas, economía, las nuevas tecnologías, el mundo digital, etc.

Las **cualidades básicas** que determinarán la naturaleza de un *influencer* y serán motivo de análisis son los siguientes:

- Confianza
- Credibilidad
- Reconocimiento
- Reputación
- Capacidad de influencia
- Calidad de la red
- Carisma

El **enfoque que acondiciona** estos personajes al sector *Fintech*:

- Relevancia
- Adecuación
- Conocimientos del tema

4.1.4. METODOLOGÍA Y ESTRUCTURA

4.1.4.1. Primeros pasos

Los primeros pasos antes de comenzar la búsqueda y el análisis serán: fijar un claro objetivo y establecer las herramientas en función de las necesidades.

El **objetivo** en este caso es la búsqueda de *influencers* adecuados para realizar campañas de marketing de aplicaciones relacionadas con el sector financiero que son tendencia en la actualidad (*Fintech*).

Dado que se tratará la búsqueda de individuos que cumplan una serie de requisitos, como son la confianza y la capacidad de influencia dentro de las redes sociales, nos centramos en *Twitter* como red social para el análisis. Se ha elegido por ser el punto de unión de muchas otras redes sociales, por su cuota de usuarios activos y por la propia funcionalidad y finalidad de la misma.

Se ha empleado la **herramienta** *RStudio* para la extracción, análisis y representación de datos de la *API REST* que proporciona *Twitter y otras aplicaciones externas* para la elaboración del estudio.

4.1.4.2. Obtención de datos

Tras realizar una observación de la diferente información que se requiere y puede obtener, junto con las formas de consecución, se tratará de **recopilar e inferir toda la información posible** que resulte de interés para el estudio en forma de *datasets*. Conseguida la información básica, se realizarán cálculos estadísticos con la ayuda de las funciones que proporciona *R,* aplicaciones externas y las *API*, obteniendo nuevos datos relevantes para el análisis.

4.1.4.3. Modelo analítico y sistema de puntuación

En el **modelo analítico** quedan definidos los factores, subfactores y métricas asociados a una interpretación de carácter argumentativa. Algunos de estos subfactores no podrán ser

analizados por falta de recursos y quedarán planteados. Cada uno de estos elementos se valorará de forma apropiada mediante el **sistema de puntuación** elaborado y explicado en ese punto.

4.1.4.4. Experimentación

Se ejecutarán tres **experimentos,** siguiendo el modelo analítico planteado, distribuidos en función de los tipos de datos a analizar. En el primer experimento se tratarán los datos relacionados con la información que se obtiene directamente de los usuarios. En el segundo, tras filtrar entre los mejores del primer experimento, se estudiarán los datos estadísticos relacionados con la publicación de *tweets* de cada candidato. Y en el tercero, se estudiarán los datos vinculados a los seguidores de cada usuario.

4.1.4.5. Ejecución y desarrollo con la herramienta

En este apartado se explica cómo se ha ejecutado y desarrollado el estudio en forma de manual o documentación paso a paso. Asimismo, se indican los datos que se han ido obteniendo y donde se almacenan en cada paso mediante la ejecución de las funciones.

4.1.4.6. Análisis e interpretación de resultados

Se realiza un análisis completo del estudio tratando de aportar veracidad y validez mediante la interpretación de resultados relacionados con la fuente de obtención de usuarios, las puntuaciones globales y la observación de candidatos.

4.2. OBTENCIÓN DE DATOS

4.2.1. INFORMACIÓN RELEVANTE

La *API* de *Twitter* nos permite obtener información de múltiples *endpoints*. Como se ha visto en el apartado de la *REST API*, nos centramos sobre todo en aquellos relacionados con la obtención de información de tweets y de usuarios.

GET search/tweets	Ofrece información acerca de los tweets.
	`Id, Fecha de creación, @Usuario dueño RT, Número de RT, Número de Favoritos, Texto, Lenguaje, Hashtags, Nº de hashtags, Nombres de usuario de menciones, Nº de menciones, Símbolos, URLs, Media, Dispositivo, Tipo de tweet, Nombre de usuario, @Usuario`
GET users/lookup **GET users/search**	Ofrece información acerca de los usuarios.
	`Id, Fecha de creación, Nombre de usuario, @Usuario, Número de seguidores, Número de seguidos, Número de listas, Localización, Descripción, Número de favoritos, Cuenta verificada, Número de tweets, Lenguaje`

GET statuses/user_timeline	Ofrece información acerca de los tweets más recientes del usuario.
	`Id, Fecha de creación, Número de RT, Número de FAV's, Número de Hashtags, Número de menciones, Texto, Source`
GET followers/list	Ofrece información acerca de los usuarios que siguen a otro usuario.
	`Id, Nombre de usuario, @Usuario, Descripción, Número de tweets, Cuenta verificada, Número de seguidores, Número de seguidos, Lenguaje`

Tabla 7. Información relevante extraída de los endpoints.

A partir de todos estos datos, con funciones y *API* externas se pueden **inferir otros datos** como la edad, el género, positividad, estadísticos, frecuencias, etc. que serán de gran utilidad para nuestro estudio.

4.2.2. ESTRATEGIA DE RECOPILACIÓN

La **estrategia** seguida para la recopilación de datos sigue el siguiente modelo:

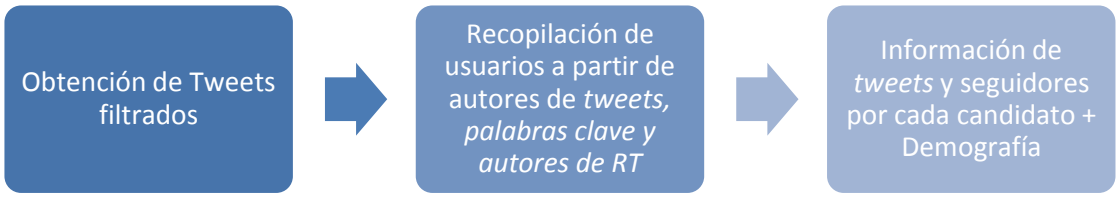


Ilustración 6. Diagrama del proceso de obtención de datos

En **primer lugar,** obtendremos un listado de ***tweets* filtrados** en función de nuestras necesidades usando el *endpoint* **GET search/tweets**. Dado que el estudio se basa en la búsqueda de *influencers* para el marketing de servicios en España, filtraremos los tweets por **idioma** ES y mediante una lista de **palabras clave** que guarde relación con la temática de nuestro estudio. Las palabras clave se han determinado y clasificado por tendencias, aplicaciones actuales que podrían ser candidatos de la campaña de marketing, bancos principales y otros términos relacionados con la búsqueda. Pese a que el idioma de búsqueda es español muchos términos de la lista están en inglés porque no significa que no puedan aparecer palabras en otro idioma.

Tendencias	Aplicaciones	Bancos	Otros términos
fintech	fintonic	santander	finances / finanzas
earlyadopter	twyp	bbva	payment / pago
startup	bizum	caixa	mobile / movil

digital	imaginbank	bankia	bank / banca
app	mooverang	ing	technology / tecnologia
bitcoin	wallo		cash / dinero
business			wallet / cartera

Tabla 8. Listado de palabras clave

En segundo lugar, una vez almacenados los *tweets* filtrados en un *dataset* (con la información que nos otorga el *endpoint* correspondiente), iremos almacenando paralelamente en otro *dataset* todos los usuarios con su respectiva información de interés como se indica en la **segunda etapa** del diagrama. Utilizando el *endpoint* **GET users/lookup** se obtienen los **usuarios que han publicado los *tweets* en su *timeline* y los autores originales de los *retweets*** (en el caso de que el tweet publicado sea RT de otro usuario)**.** Con el *endpoint* **GET users/search** se obtienen usuarios a partir de la lista de **palabras clave** en función de la similitud que guarde con su descripción y nombre. Hasta este momento sólo habremos obtenido datos que proporciona la *API* de forma directa.

Por último, procesaremos los **datos inferidos acerca de los tweets de cada usuario** tales como la positividad de los mensajes, estadísticos, frecuencia de palabras clave y fuentes con ayuda de otros *endpoints* y *API* externas. A continuación, obtendremos **información estadística acerca de sus seguidores** de la misma forma, junto con la **demografía**. Esta última etapa se explica con mayor detalle en el punto de "Procesado y limpieza".

La ***API* sólo permite** 450, 900, 900 peticiones de *tweets*, *usuarios por id* (usado dos veces en cada iteración) y *usuarios por palabras clave* respectivamente cada 15 minutos. Al obtener los datos desde R observamos que no limita ningún factor en la **primera y segunda etapa** ya que tarda más de 15 minutos en consumir todas las peticiones.

En la siguiente ilustración se presentan los diferentes *endpoints* de cada etapa con su finalidad y limitación:

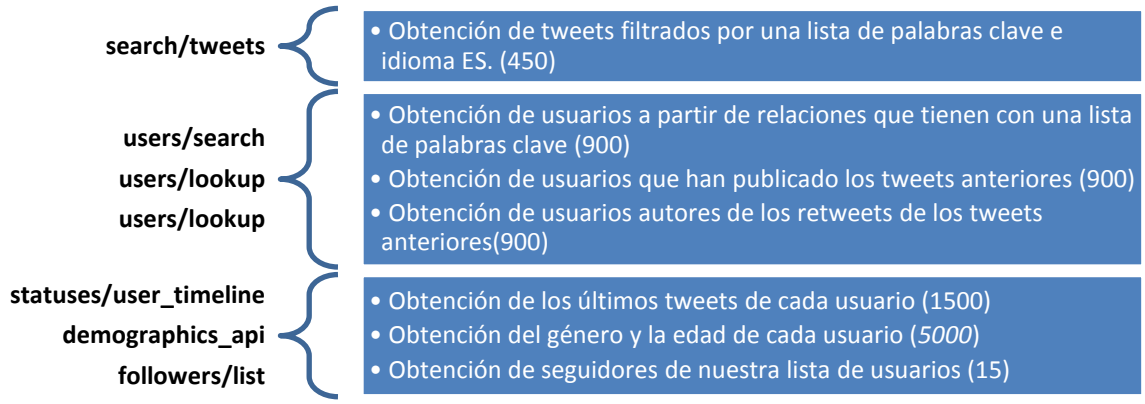


Ilustración 7. Endpoints empleados en la obtención de datos por etapas

4.2.3. PROCESADO Y LIMPIEZA

La información que nos proporciona la *API* no regresa todos los resultados de la forma esperada y por tanto habrá que tratar los textos para almacenar la información deseada. Además, se obtendrán una serie de *tweets* y usuarios duplicados que también habrá que optimizar. Para ello se han creado unas funciones que simplifican las operaciones de limpieza, eliminación de duplicados, *tweets* que provienen de *retweets*.

Los datos que no se obtienen directamente de la *API*, se obtendrán de la siguiente forma:

4.2.3.1. Estadísticos

Por cada usuario desde el *endpoint* **GET statuses/user_timeline** obtenemos una lista de los últimos tweets y con ayuda de las funciones que ofrece R realizaremos los cálculos almacenándolos en un *dataset* nuevo de usuarios ya filtrado. Este *endpoint* está limitado a 1500 peticiones por usuario cada 15 min, pero tampoco es un factor limitante, salvo el tiempo de obtención, porque R requiere más tiempo para la obtención de los datos correspondientes a esta **tercera etapa**.

- A partir del número de RT's, FAV's, Hashtags y Menciones de cada tweet de los últimos disponibles de cada usuario, obtendremos los ***estadísticos*: media, moda, mediana, cuantiles, varianza y desviación típica**.
- Observando la fuente/dispositivo de cada *tweet* por usuario, calculamos la **moda** del dispositivo más frecuente.
- Analizando los textos de cada uno de sus tweets y **comparándolos** con nuestra lista de palabras clave determinaremos la media de palabras clave de interés por tweet.

Emplearemos el *endpoint* **GET followers/list** del cual obtenemos una lista con información de los seguidores más recientes de cada usuario. Dado que está **limitado** a 15 peticiones cada 15 minutos, no se pueden obtener estos datos de una gran cantidad de usuarios.

- A partir del número de *tweets*, *followers*, *friends* de cada seguidor de los últimos disponibles, obtendremos los ***estadísticos*: media, moda, mediana, cuantiles, varianza y desviación típica**.
- Observando el lenguaje de cada *seguidor*, calculamos la **moda** del idioma más frecuente entre sus seguidores.

4.2.3.2. API externa y aplicaciones

- El **género y la edad** se puede obtener a partir del nombre de usuario y a la descripción de la cuenta con la *API AI Applied*. En concreto, introduciendo estos datos en el *endpoint* ***Demographics API,*** devuelve una lista con el género y la edad junto con su factor de confianza. Las llamadas a esta *API* sí serán un **factor limitante** ya que tan sólo se pueden realizar 5000 peticiones que no se restauran con el tiempo.
- Esta misma *API* permite el análisis de textos obteniendo de cada uno el **sentimiento** positivo, negativo, neutro o inidentificado.

4.3. MODELO ANALÍTICO Y PUNTUACIÓN

4.3.1. FACTORES, SUBFACTORES Y MÉTRICAS

Se propone dividir el estudio por **factores**, **subfactores** y **métricas**. De este modo se engloban una serie de características a factores determinados para obtener una visión global de los resultados.

A cada factor le corresponden varios subfactores que lo definen, mientras que cada subfactor puede implicar a varios factores. Las métricas asociadas a cada subfactor indican la forma de obtención de valor, que, a su vez, una misma métrica puede implicar a varios subfactores.

Se han determinado tres factores: **confianza**, **capacidad de influencia** y **relevancia para la campaña**.

Algunos de los subfactores no se podrán obtener por falta de información, o por no poder realizar una estimación razonable. Estos se indicarán por el interés y el valor que aportarían, pese a estar descartados para el caso de estudio, pudiendo ser añadidos en un futuro.

4.3.1.1. Confianza

La **confianza** se define como la seguridad para emprender una acción difícil o comprometida. La seguridad sobre una persona se adquiere a partir del conocimiento de la misma y de sus actos. En esta situación no se suele dar el caso de conocimiento directo y por tanto, centraremos la confianza sobre el reconocimiento y credibilidad de los demás. En consecuencia, los subfactores deberán aportar un valor que incremente la confianza hacia esa persona:

FACTOR: **CONFIANZA**		
SUBFACTOR	**MÉTRICA**	**INTERPRETACIÓN**
Identidad verificada	API Twitter | Metadato	La red social *Twitter* verifica que es quien dice ser realizando las comprobaciones necesarias
Seguido	API Twitter | Número de followers	El reconocimiento se puede inferir en relación al número de sus seguidores, basando la confianza en la de los demás
Usuario activo	API Twitter | Número de seguidos, listados, *tweets*, favoritos	Descartamos que un usuario sea un "Bot" cuando es activo en estos aspectos
Datos personales (Número de teléfono)	*API Twitter | Descripción cuenta personal Twitter*	*El hecho de proporcionar un teléfono móvil para verificar la cuenta ya proporciona una cierta seguridad*
Identidad personal y profesional	*API Twitter | Descripción cuenta personal Twitter*	*Al proporcionar cualquier tipo de información personal como edad, empleo, estudios genera una cierta convicción*
Calidad Tweets	API Twitter | Número medio de RT's y FAV's	La forma de valorar la calidad de los tweets será a partir de la media del número de RT's

		y FAV's de estos. Los usuarios que al hacen RT y FAV depositan su confianza en ellos
Interactuación	API Twitter | Número medio de menciones	Las personas que interactúan con otras generan más seguridad
Participación en campañas previas	*API Twitter | Análisis de Tweets*	*La participación en otro tipo de campañas previas nos garantiza en cierto modo que es un buen candidato*
Reputación	*API Externa | AI-Applied API, análisis de sentimientos de tweets en los que se nombra al usuario*	*El sentimiento de los tweets en los que se nombra al usuario en cuestión, indica el aprecio de los demás sobre este individuo*
Veracidad de seguidores	API Twitter | Análisis de la actividad de los seguidores	Al comprobar la veracidad de los seguidores, se puede concluir que son reales y no son falsos usuarios comprados o "Bots". A partir de la actividad se puede comprobar su veracidad
Género	API Externa | AI-Applied API, análisis de nombres de usuario y texto para la estimación del género	Si la *API* empleada consigue obtener el género del usuario con un factor de confianza suficientemente alto se puede inferir que la cuenta es personal
Recursividad de confianza	*API Twitter | Análisis recursivo de la red*	*Repetición de algunos de los subfactores para los usuarios de su red*

Tabla 9. Factor, subfactor, métrica de confianza

Los subfactores que no se han podido alcanzar han sido los *datos personales (número de teléfono), la identidad personal y profesional, la participación en campañas previas, la reputación y la recursividad.* La *API* no proporciona información de los números de teléfono ni de la identidad profesional y la única forma de obtenerlo sería en el caso de que el propio usuario lo muestre en su descripción. Por otra parte, no se ha conseguido obtener un listado con los tweets en los que se nombre al usuario a analizar por la limitación que tiene la *API*. Tampoco se ha encontrado la forma de analizar si un usuario ha participado en campañas previas ya que no es realizable programáticamente hablando. El análisis recursivo de la red tampoco se ha podido llevar a cabo debido a las limitaciones que presenta la *API*.

4.3.1.2. Capacidad de influencia

El factor **capacidad de influencia** se define como la aptitud de una persona o cosa para determinar o alterar la forma de pensar o de actuar de alguien. Basaremos la influencia de los candidatos en función de los datos que nos proporciona la red social.

FACTOR: **CAPACIDAD DE INFLUENCIA**		
SUBFACTOR	**MÉTRICA**	**INTERPRETACIÓN**
Alcance	API Twitter | Número de seguidores (*followers*)	El número de seguidores es sinónimo al número de lectores

Actividad en la red	API Twitter | Número de seguidos, listados, tweets, favoritos	Un usuario activo en la red será más influyente que otro que apenas participa
Alcance recursivo	API Twitter | Número medio de RT's	Cuando otros usuarios comparten (RT) el *tweet* de un usuario pasa a una nueva red de lectores. Por ello, se calcula el número medio de RT's que obtiene cada usuario
Alcance cualitativo	API Twitter | Número medio de FAV's	El hecho de que un tweet obtenga FAV's implica un interés adicional que se traduce normalmente en calidad
Intención de ser encontrado	API Twitter | Media de hashtags	Un usuario hace uso de hashtags porque quiere ser encontrado a partir de esas palabras. Luego, el uso de hashtags indica cierta preferencia por ser encontrado
Ruido	*API Twitter | Número de respuestas a los tweets y comentarios que genera*	*En el momento en el que un tweet comienza a ser comentado por otros usuarios genera expectación y adquiere valor adicional*
Sentimiento de las interacciones	*API Twitter | Análisis de sentimiento de las interacciones y menciones con otros usuarios.*	*Si las interacciones y las menciones tienen una actitud positiva implican un sentimiento de acuerdo entre estos y por tanto serán más influyentes*
Actividad de los seguidores	API Twitter | Análisis de la actividad de los seguidores	Cuando los seguidores son activos en la red, son más atractivos que otros a los que les siguen cuentas sin ningún tipo de actividad
Alcance de los seguidores	API Twitter | Número de *followers* de los seguidores	El número de seguidores vuelve a ser sinónimo al número de lectores en el caso de que los primeros compartan los *tweets*
Recursividad de influencia	*API Twitter | Análisis recursivo de la red*	*Repetición de algunos de los subfactores para los usuarios de su red*

Tabla 10. Factor, subfactor, métrica de capacidad de influencia

Los subfactores que no se han podido alcanzar han sido el de *ruido, sentimiento de las interacciones y recursividad.* La *API* dificulta la opción de analizar los tweets respuesta y además el modelo de análisis llevado para el estudio junto a sus límites no facilita la tarea, pese a ello sería de gran valor obtener resultados acerca de los subfactores de ruido y sentimiento de las interacciones. Las herramientas de análisis de sentimientos no son del todo fiables ya que sólo se basan en si las palabras tienen una connotación positiva o negativa olvidando la intencionalidad de la oración completa. El análisis recursivo de la red tampoco se ha podido llevar a cabo debido a las limitaciones que presenta la *API*.

4.3.1.3. Relevancia para la campaña

La relevancia se define como la importancia o significación que destaca de algo. El factor **relevancia para la campaña** es por tanto la importancia que una característica supone para el cometido.

FACTOR: **RELEVANCIA PARA LA CAMPAÑA**		
SUBFACTOR	**MÉTRICA**	**INTERPRETACIÓN**
Localización geográfica	API Twitter | Metadato	La localización geográfica suele ser una necesidad para la campaña
Idioma	API Twitter | Metadato	El idioma del *influencer* suele ser una necesidad para la campaña
Relación campaña	API Twitter | Palabras clave en la descripción	En la descripción del usuario se suelen indicar sus predilecciones e información personal, por ello, si coincide con la temática, adquiere un valor positivo
Interés general en la temática de la campaña	API Twitter | Análisis de tweets, palabras clave	Un *influencer* que tenga interés por la campaña tendrá una valoración positiva. Se obtiene a partir de un análisis de palabras clave
Interés por nuevas tecnologías (early adopter)	API Twitter | Análisis de tweets, palabras clave	Un *influencer* que tenga interés por las nuevas tecnologías tendrá una valoración positiva. Se obtiene a partir de un análisis de palabras clave
Interés por las finanzas (finances)	API Twitter | Análisis de tweets, palabras clave	Un *influencer* que tenga interés por las finanzas tendrá una valoración positiva. Se obtiene a partir de un análisis de palabras clave
Dispositivos usados	API Twitter | Metadato	Según la necesidad de la campaña se puede valorar el uso de un dispositivo u otro. Discriminando dispositivos u herramientas que sirvan para analíticas de *Twitter*
Relación de la red (friends)	*API Twitter | Lista de usuarios a los que sigue, cuentas relacionadas con fianzas u otros influencers*	*En el caso de que un usuario siga cuentas relacionadas con la temática de la campaña adquiriría un valor positivo*
Relación de la red (followers)	API Twitter | Relación de sus *followers* con los temas indicados	En el caso de que a un usuario le siga gente influyente relacionada con la temática de la campaña adquiriría una valoración positiva.
Idioma de su red	API Twitter | Moda del idioma de sus seguidores (*followers*)	Será importante para la campaña que los usuarios objetivos hablen el idioma indicado.
Edad	API Externa | AI-Applied API	Según la necesidad de la campaña se puede valorar un rango de edades más que otro

Género	API Externa | AI-Applied API	Según la necesidad de la campaña se puede valorar un género u otro. Discriminando también género desconocido por ser una cuenta no personal o corporativa
Recursividad de relevancia	*API Twitter | Análisis recursivo de la red*	*Repetición de algunos de los subfactores para los usuarios de su red*

Tabla 11. Factor, subfactor y métrica de relevancia para la campaña

Los subfactores que no se han podido alcanzar han sido los de *relación de la red (friends) y recursividad*. El análisis de los seguidos de cada usuario no aporta demasiado valor con la estimación empleada y tan sólo sería útil para este subfactor. El análisis recursivo de la red tampoco se ha podido llevar a cabo debido a las limitaciones que presenta la *API*.

4.3.2. SISTEMA DE PUNTUACIÓN

Se ha optado por la elaboración de un sistema de puntuación con el objetivo de evaluar las habilidades tomadas como factores de los *influencers*. Los resultados obtenidos se almacenarán en una tabla con una fila por cada usuario para su posterior representación.

4.3.2.1. Cálculo de puntuación

La **puntuación de cada uno de los factores** vinculados a un usuario irá de 1 a 10, asimismo, cada subfactor obtendrá otra valoración de 1 a 10 que se multiplicará por un ***coeficiente de interés*** (de 0 a 1). La suma de todos los coeficientes de interés de cada factor se ajustará a 1, por lo que, la suma de los subfactores multiplicados por el coeficiente será como máximo 10.

$$f_i = \sum_{n=1} (s_n * coef_n) \quad \left| \begin{array}{l} 0 \leq f, s_n \leq 10 \\ 0 \leq coef_n \leq 1 \\ \\ f_i = factor \\ i = \text{número de factor} \\ \text{n} = número\ de\ subfactor \\ s_n = puntuación\ subfactor \\ coef_n = coeficiente\ subfactor \end{array} \right.$$

Ecuación 6. Cálculo de la puntuación de cada factor

La **evaluación de los subfactores** se determina realizando la suma aritmética de las distintas métricas involucradas.

$$s_n = \frac{1}{j}\left(\sum_{z=1} m_z\right) \quad \left| \begin{array}{l} j = \text{número de métricas del subfactor} \\ z = \text{número de métrica} \\ \text{n} = número\ de\ subfactor \\ s_n = puntuación\ subfactor \\ m_j = puntuación\ métrica \end{array} \right.$$

Ecuación 7. Cálculo de la puntuación de cada subfactor

El cálculo de **puntuación de las métricas** será relativo. La puntuación máxima la marcará el usuario que obtenga el valor óptimo al que se le atribuye un 10, y de forma proporcional, los demás usuarios obtendrán las puntuaciones en esa métrica siguiendo la siguiente fórmula:

$$m_z = \frac{v}{v_{zopt}} * 10$$

$z =$ número de métrica
$m_z = puntuación\ métrica$
$v_{zopt} = valor\ óptimo$
$v = valor$

Ecuación 8. Cálculo de la puntuación de cada métrica

Puntuación de métricas especial ajustada al estudio. La puntuación máxima la marcará también el usuario que obtenga el valor óptimo, sin embargo, los usuarios que estén por encima de la media de los **XXX**Y óptimos valores se les otorgará una puntuación de 5 puntos, mientras que el resto de la puntuación será proporcional al que marca el máximo en el caso de estar en la franja superior o al que marca la media en el caso de estar en la franja inferior. XXXY son las cifras del número total de los candidatos y para el cálculo de la media utilizaremos todas menos la última. Con esto conseguimos equilibrar las puntaciones en el caso de que existan puntuaciones muy altas para pocos usuarios.

$m_z = 5 + \frac{v}{v_{zopt}} * 5$	$m_z = \frac{v}{mean_{opt}} * 5$
$Si\ v > mean$	$Si\ v \leq mean$
$z =$ número de métrica $m_z = puntuación\ métrica$ $v_{zopt} = valor\ óptimo$ $v = valor$	$z =$ número de métrica $m_z = puntuación\ métrica$ $mean_{opt} = media$ $v = valor$

Ecuación 9. Puntuación de métricas especial ajustada al caso de estudio

4.3.2.2. Representación: Gráfico en estrella.

La representación final de resultados por cada usuario será en forma de **gráfico en estrella**. Se ha optado por este modelo ya que es muy representativa para evaluar resultados a simple vista, pudiendo elegir entre un resultado u otro en función de cada factor.

A continuación, se muestra un ejemplo de cómo será la representación de estos factores por cada usuario:

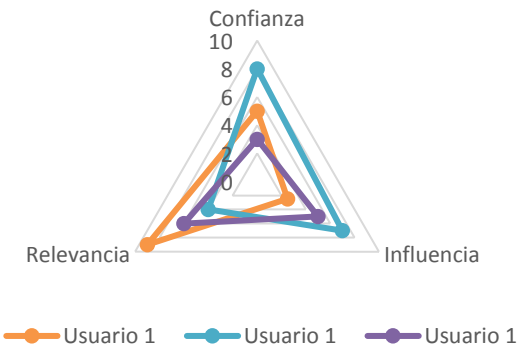


Ilustración 8. Gráfico en estrella puntuación factores

4.4. EXPERIMENTOS

Los experimentos se han **dividido en función de las fases** que se han experimentado para la obtención de datos. Mientras avanzamos en la obtención de datos específicos, como pueden ser la recolección de tweets o de seguidores de cada usuario, los límites son mayores y por ello habrá que filtrar en cada fase los mejores candidatos hasta reducir los resultados a una lista razonable que satisfaga los límites.

De este modo, el desarrollo de experimentos se planteará en el siguiente orden:

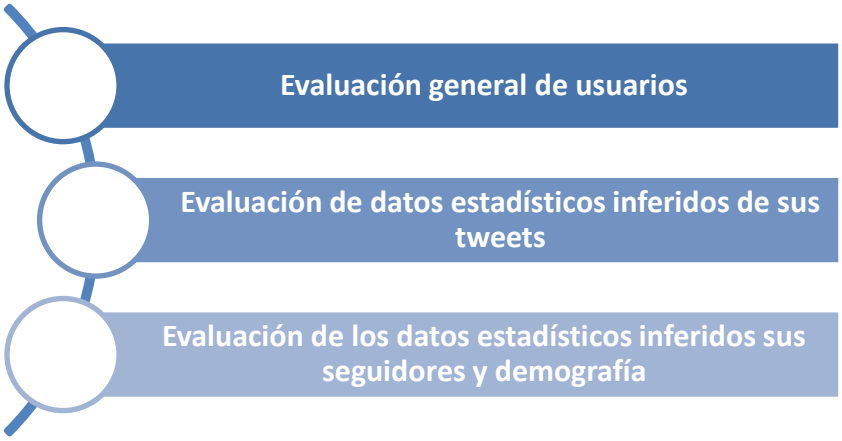


Ilustración 9. Desarrollo de experimentos

4.4.1. EXPERIMENTO 1: EVALUACIÓN GENERAL DE USUARIOS

4.4.1.1. Descripción

En este experimento se puntúan los diferentes datos que nos proporciona la *API* de *Twitter* de los usuarios, obtenidos de las 3 formas vistas anteriormente (a partir del listado de tweets, obteniendo los usuarios que han publicado los *tweets* en su *timeline* y los autores originales de los *RT's,* y por búsqueda de palabras clave en usuarios). Estos datos se obtienen de forma directa sin necesidad de recurrir a estadísticos u otros *endpoints.*

Una vez finalizado el experimento se determinarán los usuarios más prometedores que serán los sujetos de análisis en el próximo experimento, filtrando de esta forma el número de candidatos por las limitaciones que presentan las siguientes formas de obtención de datos.

4.4.1.2. Factores, subfactores, métricas y coeficientes

A continuación, se analizan los diferentes factores y subfactores que hacen referencia a cada usuario. Cada subfactor tiene asignado un **coeficiente de interés** que está parametrizado y se podrá modificar a posteriori.

Con los datos obtenidos en este caso podemos inferir que un usuario aporta **confianza** cuando es conocido por un público general, presenta una cierta actividad en la red y si puede demostrar que es quien dice ser. La **capacidad de influencia** viene determinada por la cantidad de personas que leen sus opiniones junto con la actividad que tiene en la red. Y es **relevante para la campaña** si se encuentra en la localización geográfica indicada (palabras clave), habla el idioma objetivo ("es") y si la descripción que aporta está relacionada (palabras clave) con los términos de la campaña.

En la siguiente tabla se detalla en profundidad cuales han sido las métricas para calcular los subfactores con sus correspondientes coeficientes:

Factores	Subfactores	Métricas	Coef.
Confianza	Seguido	Número de seguidores (*followers*)	0,5
	Usuario activo	Número de seguidos (*friends*) Número de listados Número de tweets favoritos Número de tweets	0,2
	Identidad verificada	Usuario verificado	0,3
Capacidad de influencia	Alcance	Número de seguidores (*followers*)	0,7
	Actividad de la red	Número de seguidos (*friends*) Número de listados Número de tweets favoritos Número de tweets	0,3
Relevancia para la campaña	Localización geográfica	Comparación de palabras clave de localización geográfica	0,4
	Idioma	Idioma del usuario ("es")	0,3
	Relación campaña	Relación de la campaña con la descripción mediante comparación de palabras clave	0,3

Tabla 12. Desarrollo del experimento 1

4.4.1.1. Listados de palabras clave

Localización geográfica					
spain	alicante	guadalajara	almeria	albacete	granada
españa	castellon	toledo	valladolid	murcia	cataluña
madrid	lugo	avila	huesca	andalucia	castilla
barcelona	pontevedra	burgos	teruel	salamanca	leon
valencia	orense	palencia	zaragoza	jaen	mancha
vasco	pamplona	segovia	oviedo	cordoba	rioja
galicia	vitoria	zamora	mallorca	sevilla	aragon
asturias	bilbao	gerona	tenerife	huelva	extremadura
cantabria	ceuta	lerida	santander	cadiz	baleares
navarra	melilla	tarragona	cuenca	malaga	canarias

Tabla 13. Lista de palabras clave localización geográfica

Relación campaña					
fintech	imaginbank	mobile	dinero	business	banca
earlyadopter	mooverang	bank	cartera	fintonic	tecnologia
startup	wallo	technology	twyp	bankia	bitcoin
digital	santander	cash	bizum	ing	caixa
app	bbva	wallet	finances	payment	finanzas

Tabla 14. Lista de palabras clave relación campaña

4.4.2. EXPERIMENTO 2: EVALUACIÓN DE DATOS ESTADÍSTICOS INFERIDOS DE SUS TWEETS

4.4.2.1. Descripción

En este experimento se puntúan los diferentes datos que se pueden inferir de forma estadística de los tweets que ha publicado cada usuario a partir del listado filtrado en el experimento anterior.

Una vez finalizado el experimento se determinarán de nuevo los usuarios más prometedores que serán los sujetos de análisis para el próximo experimento. Habrá que filtrar de forma exhaustiva, ya que el modelo de obtención de datos para el siguiente ensayo es muy limitado.

4.4.2.2. Factores, subfactores, métricas y coeficientes

Análogamente, se estudian los factores y subfactores relacionados con los tweets de cada candidato.

Con los datos obtenidos en este caso podemos deducir que un usuario aporta **confianza** cuando interactúa con otros usuarios y la gente deposita su confianza compartiendo y guardando sus *tweets*. La **capacidad de influencia** viene determinada por el alcance recursivo y cualitativo que llegan a alcanzar sus comentarios en la red junto con su intención de llamar la atención en determinados temas mediante el uso de *hashtags*. Y es **relevante para la campaña** si muestra interés por la temática de la campaña en general, las nuevas tecnologías y las finanzas. Además, puede ser relevante en función del dispositivo habitual de uso.

En la siguiente tabla se detalla en profundidad cuales han sido las métricas para calcular los subfactores con sus correspondientes coeficientes:

Factores	Subfactores	Métricas	Coef.
Confianza	Calidad Tweets	Número medio de *RT* por tweet Número medio de *FAV's* por tweet	0,7
	Interactuación	Número medio de menciones por tweet	0,3
Capacidad de influencia	Alcance recursivo	Número medio de *RT's* por tweet	0,5

	Alcance cualitativo	Número medio de *FAV's* por tweet	0,3
	Intención de ser encontrado	Media de hashtags	0,2
Relevancia para la campaña	Interés general en la temática de la campaña	Media de palabras clave por análisis de tweets	0,3
	Interés por nuevas tecnologías (early adopter)	Media de palabras clave por análisis de tweets	0,3
	Interés por las finanzas (finances)	Media de palabras clave por análisis de tweets	0,3
	Dispositivos usados	Valoración en función del dispositivo usado	0,1

Tabla 15. Desarrollo del experimento 2

4.4.2.1. Listados de palabras clave

Relación campaña					
fintech	imaginbank	mobile	dinero	business	banca
earlyadopter	mooverang	bank	cartera	fintonic	tecnologia
startup	wallo	technology	twyp	bankia	bitcoin
digital	santander	cash	bizum	ing	caixa
app	bbva	wallet	finances	payment	finanzas

Tabla 16. Lista de palabras clave relación campaña

Early adopter					
early	windows	learning	prueba	automatico	salida
adopter	os	big	aprender	auto	innovar
alpha	app	data	datos	ciencia	probar
beta	web	scientist	aplicacion	innovate	tech
android	iot	science	cientifico	new	internet
ios	machine	test	aprendizaje	nuevo	thing

Tabla 17. Lista de palabras clave early adopter

Finanzas					
finance	banco	accion	caida	transfer	apostar
bank	credit	current	subida	virtual	open
fintech	moneda	asset	fall	online	abierto
cash	coin	passive	drop	finanza	inflacion
bitcoin	invest	activo	rise	conta	cae
bolsa	stock	corriente	interes	account	sube
invertir	money	pasivo	cajero	direct	empresa
inversion	amortiza	acreedor	transferencia	bet	cooperativa

Tabla 18. Lista de palabras clave finanzas

4.4.3. EXPERIMENTO 3: EVALUACIÓN DE DATOS ESTADÍSTICOS INFERIDOS DE SUS SEGUIDORES Y DEMOGRAFÍA

4.4.3.1. Descripción

En este experimento se puntuarán los diferentes datos estadísticos que obtenemos acerca de los seguidores de cada candidato junto a sus datos demográficos. Sacaremos conclusiones del tipo de si sus seguidores son usuarios reales, el alcance que tendría en el caso de que estos compartieran los *tweets* y la relación con la temática.

4.4.3.2. Factores, subfactores, métricas y coeficientes

Análogamente a los dos experimentos anteriores, se analizan los factores y subfactores relacionados con el estudio de los seguidores y la demografía de los candidatos.

Con los datos obtenidos en este caso podemos inferir que un usuario aporta **confianza** cuando se demuestra la veracidad de los seguidores y si la cuenta del usuario es personal. La **capacidad de influencia** se determina mediante la actividad y el alcance de sus seguidores en el caso de compartir sus *tweets*. Y es **relevante para la campaña** si sus seguidores hablan el idioma, la cuenta es personal y sus descripciones guardan relación con los términos de la campaña. Para ciertas campañas será importante que los influencers sean de un determinado género o sus edades estén comprendidas en un rango, sin embargo, en este, no es relevante y les otorgaremos la misma puntuación a todos.

En la siguiente tabla se detalla en profundidad cuales han sido las métricas para calcular los subfactores con sus correspondientes coeficientes:

Factores	Subfactores	Métricas	Coef.
Confianza	Veracidad de seguidores	Número medio de seguidores (*followers*) Número medio de seguidos (*friends*) Número medio de tweets	0,7
	Género	Verificación si es cuenta personal mediante el género	0,3
Capacidad de influencia	Actividad de los seguidores	Número medio de seguidos (*friends*) Número medio de tweets	0,3
	Alcance de los seguidores	Número medio de seguidores (*followers*)	0,7
Relevancia para la campaña	Relación de la red (followers)	Media de palabras clave por análisis de descripción de seguidores	0,3
	Idioma de su red	Moda del idioma de sus seguidores (*followers*)	0,3
	Edad	Puntuación en función de la edad	0,1
	Género	Puntuación en función del género y verificación si es cuenta personal	0,3

Tabla 19. Desarrollo del experimento 3

4.4.3.1. Listados de palabras clave

Relación red followers									
invertir	inversion	bank	bbva	current	drop	account	movil	iot	datos
bitcoin	earlyadopter	banca	banco	asset	machine	direct	early	rise	aplicacion
amortiza	aprendizaje	test	credit	passive	interes	bet	adopter	fall	cientifico
caixa	mooverang	pago	wallo	money	caida	online	windows	cae	moneda
digital	santander	bolsa	app	accion	learning	conta	empresa	web	aprender
fintech	imaginebank	bankia	cartera	stock	acreedor	virtual	inflacion	ios	science
fintonic	technology	cash	ing	activo	cajero	apostar	alpha	big	payment
startup	cooperativa	mobile	finance	subida	business	finanza	sube	os	prueba

Tabla 20. Lista de palabras clave relación red followers

4.5. EJECUCIÓN Y DESARROLLO CON LA HERRAMIENTA

4.5.1. ESTRUCTURA Y PREPARACIÓN

El estudio se desarrolla siguiendo el modelo de obtención de datos junto a los experimentos asociados. Este proceso está dividido en **10 fases de ejecución de la herramienta implementada** mediante la introducción de una serie de parámetros, los cuales son necesarios para realizar cada uno de los experimentos.

La herramienta se prepara para la ejecución cargando todos los *scripts* ".R" disponibles que conforman el programa en el entorno global. Será necesario instalar los paquetes de las librerías utilizadas mediante el comando:

```
install.packages("httr","jsonlite","modeest","fmsb","ggplot2","rJava","xlsx")
```

También será necesario introducir los *tokens* de acceso a la *API* de *Twitter* en el archivo "parameters.R". Una vez instaladas las librerías e introducidos los *tokens* comenzamos el estudio mediante la ejecución del programa. Toda la documentación necesaria para la ejecución se encuentra en el manual del **anexo**.

4.6. ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

4.6.1. FUENTE DE OBTENCIÓN DE USUARIOS

Los *datasets*: "**stored_users"** (información general), "**statistic_users"** (información del análisis de *tweets*) y "**followers_users"** (información del análisis de seguidores y demografía), contienen datos de **177.467**, **4.993**, **615 usuarios** respectivamente. Cada uno abarca la

información del anterior más la nueva añadida, con los usuarios filtrados en función de las puntuaciones obtenidas en los experimentos. Por tanto, el *dataset* "followers_users" posee información completa de 615 usuarios.

Como se ha visto con anterioridad, existen tres formas de obtener usuarios: por búsqueda directa mediante palabras clave que caractericen el nombre o la descripción del usuario (***Palabras clave***); o por búsqueda de usuarios en el listado de *tweets*: usuario que publica el *tweet* en su *timeline (**Publicación de tweets**) y* autor original del *RT* (***Autores de RT's***).

A continuación, se muestran datos estadísticos acerca de la fuente de obtención de usuarios de cada *dataset*:

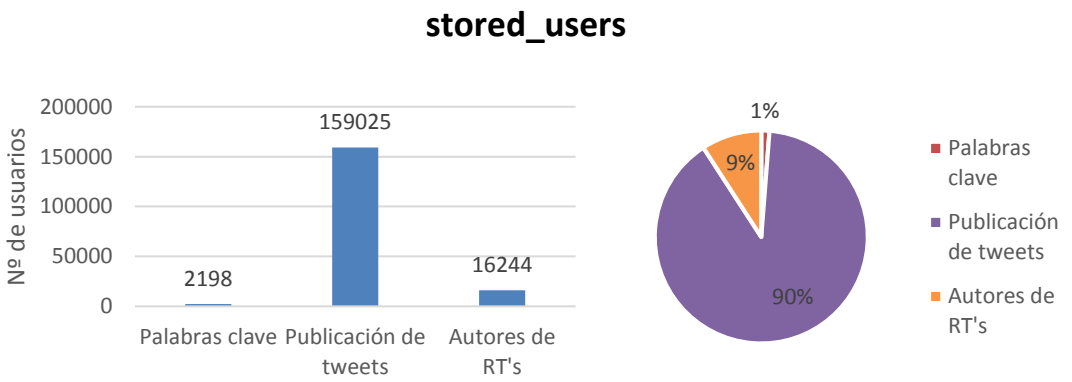


Ilustración 10. Representación del nº de usuarios en función de la fuente de obtención "stored_users"

De los 177.467 usuarios, el 99% se han obtenido a partir del listado de *tweets* ya sean usuarios que publican los tweets o usuarios autores de los RT's, mientras que tan solo el 1% se han obtenido a partir de la busqueda de palabras clave en usuarios.

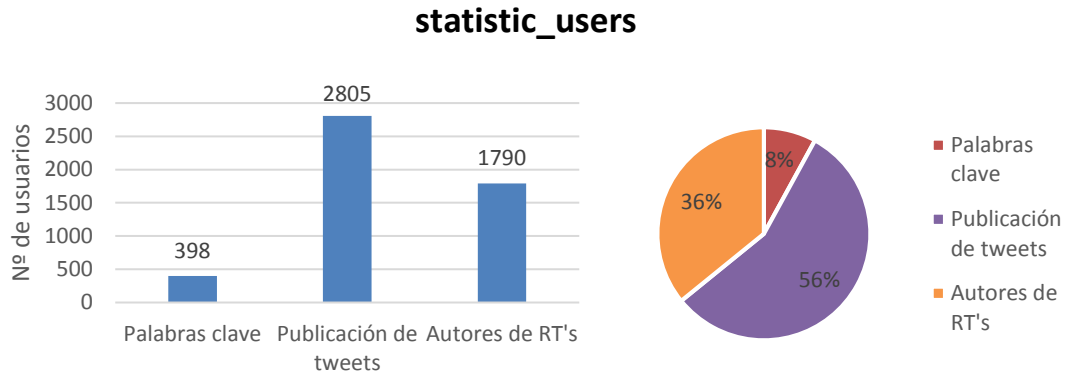


Ilustración 11. Representación del nº de usuarios en función de la fuente de obtención "statistic_users"

Tras la ejecución del primer experimento y análisis de los 4.993 usuarios correspondientes a este *dataset*, la relación de la fuente de obtención de usuarios varía notablemente. Esto se debe a que los usuarios obtenidos por palabras clave poseen mayor relevancia para la campaña y los usuarios autores de RT's poseen mayor capacidad de influencia estadísticamente hablando.

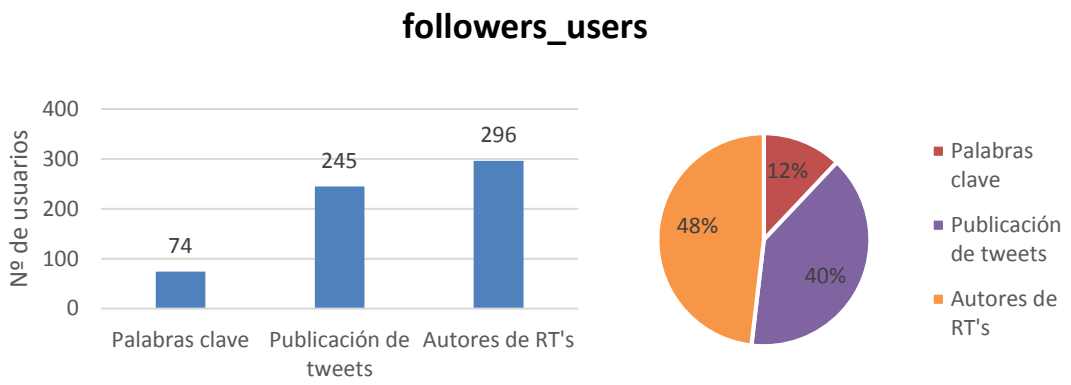


Ilustración 12. Representación del nº de usuarios en función de la fuente de obtención "followers_users"

Por último, tras la ejecución del segundo experimento y análisis de los 615 usuarios correspondientes a este *dataset*, la relación de la fuente de obtención de usuarios vuelve a variar en favor de los usuarios obtenidos por palabras clave y autores de RT's por el mismo motivo.

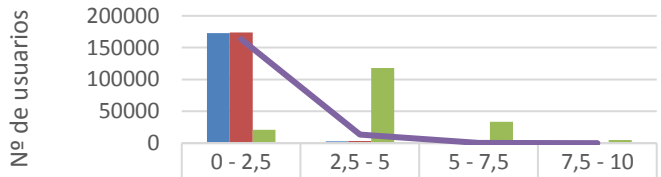
De este modo, se confirma que las tres formas de obtención de usuarios son perfectamente válidas para este estudio y que unas poseen mayor relevancia que otras.

4.6.2. PUNTUACIONES DE LOS EXPERIMENTOS

Por cada uno de los *datasets* vistos en el apartado anterior (**stored_users**, **statistic_users** y **followers_users**) se realiza un experimento que sirve de filtro para las siguientes fases de obtención de datos. Los experimentos, tal y como se han visto, son los encargados de aplicar las métricas agrupadas en subfactores y factores para la evaluación de los usuarios.

En este apartado comprobaremos la validez de las puntuaciones otorgadas en cada experimento. Para ello, se representa gráficamente el número de usuarios que se encuentran en cada uno de los cuatro rangos de puntuación. Estas puntuaciones figuran para cada uno de los factores y para la suma total de estos.

Con los resultados del **primer experimento,** vinculado a la información general de los usuarios, obtenemos la siguiente figura:

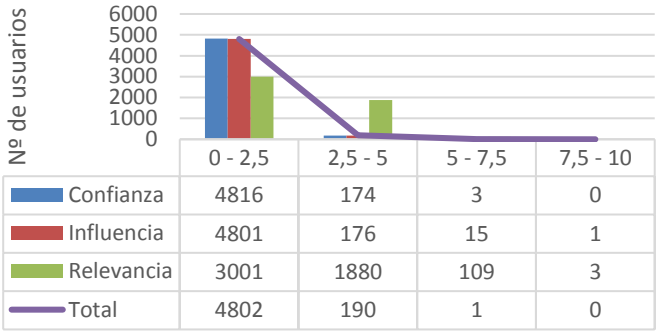


	0 - 2,5	2,5 - 5	5 - 7,5	7,5 - 10
Confianza	172830	3000	1634	3
Influencia	173747	3631	88	1
Relevancia	21103	118130	33435	4799
Total	163306	13668	492	1

Ilustración 13. Nº de usuarios por rango de puntuaciones E1

De los 177.467 usuarios totales, la mayoría de los usuarios obtienen puntuaciones superiores a 2,5 y 5 en cuanto a la relevancia para la campaña, puesto que, la obtención de estos usuarios se realizó en función de un listado de palabras clave que guardan relación con la campaña y filtrando por idioma "es" de *tweets*. Mientras que, el 2,6% y el 2% de los usuarios supera la puntuación de 2,5 en relación a la confianza y capacidad de influencia.

Del **segundo experimento,** relacionado con el análisis estadístico de los *tweets* de cada candidato, obtenemos el siguiente gráfico:

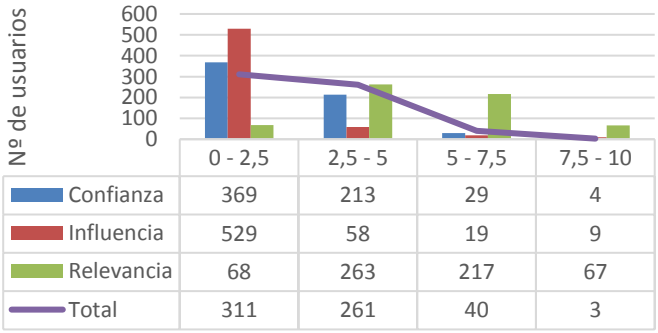


	0 - 2,5	2,5 - 5	5 - 7,5	7,5 - 10
Confianza	4816	174	3	0
Influencia	4801	176	15	1
Relevancia	3001	1880	109	3
Total	4802	190	1	0

Ilustración 14. Nº de usuarios por rango de puntuaciones E2

En este caso, se analizan los 4993 mejores candidatos del primero experimento. El efecto que se produce en el anterior experimento se mantiene en el análisis de los tweets de cada uno de los candidatos con un 40% de usuarios superando una puntuación de 2,5 en relevancia para la campaña.

Con el **tercer experimento**, que analiza las estadísticas de los seguidores de cada candidato, obtenemos las siguientes estadísticas:



	0 - 2,5	2,5 - 5	5 - 7,5	7,5 - 10
Confianza	369	213	29	4
Influencia	529	58	19	9
Relevancia	68	263	217	67
Total	311	261	40	3

Ilustración 15. Nº de usuarios por rango de puntuaciones E3

Para los 615 usuarios analizados, se vuelve a repetir el mismo efecto que en los anteriores experimentos con respecto a la relevancia de la campaña. Sin embargo, la relación de confianza e influencia en comparación con los anteriores experimentos se ve incrementada. Esto se debe al objetivo, análisis de seguidores y demografía, que desvía en cierto modo el estudio de los anteriores experimentos.

Finalmente, uniendo los **resultados de los tres experimentos** se propone un coeficiente de interés a cada factor:

	Experimento 1	Experimento 2	Experimento 3
Confianza	0,4	0,4	0,2
Influencia	0,45	0,45	0,1
Relevancia	0,4	0,4	0,2

Tabla 21. Coeficientes de importancia por experimento y factor

Se ha otorgado mayor importancia al primer y segundo experimento por estar relacionadas con información personal y el análisis de los propios tweets del usuario. Por otro lado, el tercer experimento depende de los últimos seguidores de cada candidato, de los cuales no se puede inferir demasiada información con seguridad, y el peso de este experimento se lo lleva la identificación de personas físicas mediante el género.

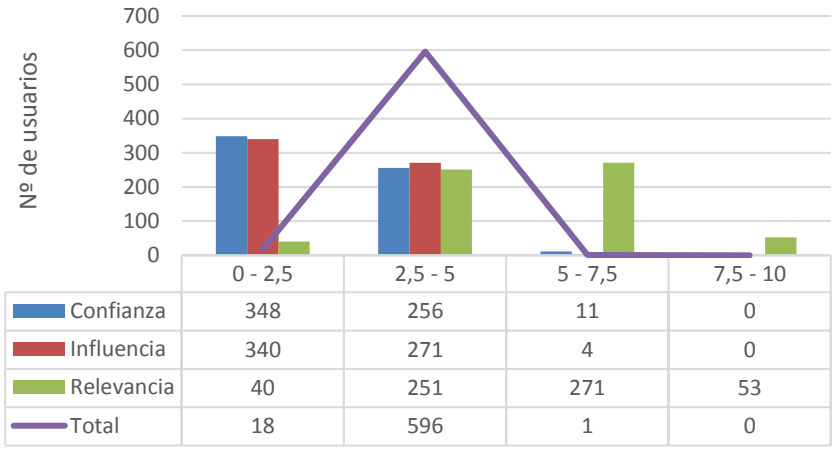


Ilustración 16. Nº de usuarios por rango de puntuaciones E1, E2, E3

De los 615 candidatos finales, podríamos afirmar que cualquiera de ellos podría cumplir las expectativas de *influencer* a pesar de los bajos valores. Las puntuaciones obtenidas dependen de demasiados factores y es prácticamente imposible encontrar un candidato que cumpla todos con solvencia. Por tanto, cada candidato tendrá sus puntos fuertes y habrá que valorar si se busca equilibrio entre los tres o se da más importancia a algún factor.

4.6.3. OBSERVACIÓN DE CANDIDATOS

Llegados a este punto, **se analizan los resultados finales que genera la herramienta en forma de candidatos y se comprueba su validez**. Esta observación se realiza clasificando los resultados en función de los puntos fuertes de cada usuario para cada factor y el equilibrio de los tres.

Por clasificación se diferencian 20 candidatos que han obtenido la mayor puntuación en un factor. A posteriori, se investigan las cuentas de *Twitter* de los más destacados para la verificación de resultados.

4.6.3.1. Clasificación por confianza

Se muestran a continuación las cuentas de los candidatos que han obtenido mayor puntuación en el factor **confianza.** Este factor nos indica, a partir de la confianza depositada en sus seguidores y en *Twitter*, la seguridad para emprender un contrato o negocio con esa persona.

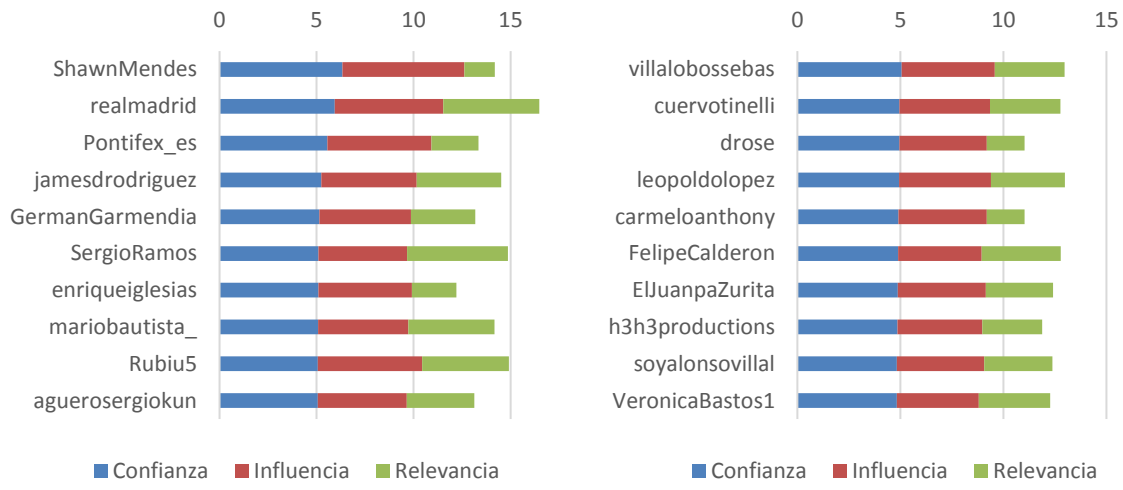


Ilustración 17. Clasificación de candidatos según el factor confianza

Todos los usuarios son **personas o entidades famosas a nivel nacional y global,** entre ellos, aparecen asociaciones, políticos, periodistas, *youtubers*, artistas y deportistas de élite. Sin embargo, no todos serán óptimos para la campaña, pese a cumplir con el factor de confianza, algunos podrían no alcanzar las expectativas de capacidad de influencia y relevancia para la campaña. Por ello, completamos el análisis introduciéndonos directamente en sus perfiles de *Twitter* y evaluando los resultados.

Las puntuaciones que obtienen los candidatos más destacados por cada factor se reflejan en la siguiente figura:

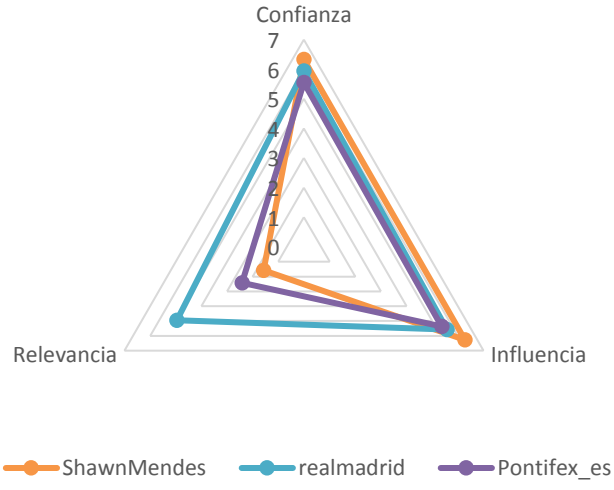


Ilustración 18. Gráfico en estrella de candidatos según el factor confianza

Se analizan las métricas de mayor relevancia a partir de los datos obtenidos con la herramienta para validar las puntuaciones. En la síntesis de resultados, las métricas asociadas a la **confianza** y **capacidad de influencia** (que coinciden por basar la confianza en los demás) son: *Seguidores, Nº de tweets, Media de RT's y Media de FAV's;* mientras que la **relevancia de la campaña** está vinculada a: *Descripción, Localización, Media de palabras clave, Dispositivo habitual, Idioma cuenta e Idioma seguidores*.

@realmadrid

Real Madrid C. F.		
Cuenta oficial del club deportivo de fútbol Real Madrid		
Localización	Madrid, Spain	
Seguidores	21.609.109	
Nº de tweets	54.659	
Media de RT's	825,61	
Media de FAV's	2238,49	
Media de palabras clave por *tweet*	General	0,04
	Nuevas tecnologías	0,56
	Finanzas	0,04
Dispositivo habitual	TweetDeck	
Idioma cuenta	es	
Idioma seguidores	es	
Género	-	

Tabla 22. Síntesis información @realmadrid

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, poseen valores muy altos en cuanto a nº de seguidores, media de RT's y media de FAV's.

Las métricas asociadas a la **relevancia para la campaña** cumplen los requisitos de localización e idioma.

Entrando en **valoración personal**, los equipos de fútbol son buenos medios para temas de marketing y publicidad, sin embargo, pueden llegar a ser muy costosos. Tal es la capacidad de influencia que la relación en temas de la campaña pierde relevancia mientras cumpla los requisitos de idioma y localización.

@ShawnMendes

Shawn Mendes		
Cuenta oficial del artista/cantante Shawn Mendes		
Localización	-	
Seguidores	8.419.134	
Nº de tweets	13.577	
Media de RT's	16.653,56	
Media de FAV's	47.817,4	
Media de palabras clave por *tweet*	General	0,31
	Nuevas tecnologías	0,19
	Finanzas	0,02
Dispositivo habitual	Twitter for iPhone	
Idioma cuenta	en	
Idioma seguidores	en	
Género	Masculino	

Tabla 23. Síntesis información @ShawnMendes

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, poseen valores muy altos en cuanto a nº de seguidores, media de RT's y media de FAV's.

Las métricas asociadas a la **relevancia para la campaña** no cumplen los requisitos de localización e idioma y la media de palabras clave por *tweet* es relativamente baja.

Entrando en **valoración personal**, este artista no cumple con las especificaciones requeridas en temas de relevancia para la campaña, en concreto por idioma y localización. Pese a ello, ha obtenido puntuaciones dominantes en los otros factores que demuestran su capacidad como *influencer*.

@Pontifex_es

Papa Francisco		
Cuenta oficial del personaje eclesiástico Papa Francisco		
Localización	Ciudad del Vaticano	
Seguidores	12.463.752	
Nº de tweets	1.043	
Media de RT's	9.270,5	
Media de FAV's	18.408,23	
Media de palabras clave por *tweet*	General	0,01
	Nuevas tecnologías	1,16
	Finanzas	0,03
Dispositivo habitual	TweetDeck	
Idioma cuenta	it	
Idioma seguidores	en	
Género	Masculino	

Tabla 24. Síntesis información @Pontifex_es

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, poseen valores muy altos en referencia al a nº de seguidores, media de RT's y media de FAV's.

Las métricas asociadas a la **relevancia para la campaña** no cumplen los requisitos de localización e idioma y la media de palabras clave por *tweet* es relativamente baja.

Entrando en **valoración personal**, este personaje eclesiástico es un usuario muy influyente y que inspira seguridad en sectores que no guardan ningún tipo de relación con las especificaciones requeridas como se demuestra en la puntuación obtenida en relevancia para la campaña.

4.6.3.2. Clasificación por capacidad de influencia

En este punto, se muestran los candidatos que han obtenido la mayor puntuación en el factor **capacidad de influencia.** Este factor nos indica, la aptitud de un *influencer* para determinar o alterar la forma de pensar o de actuar de un usuario.

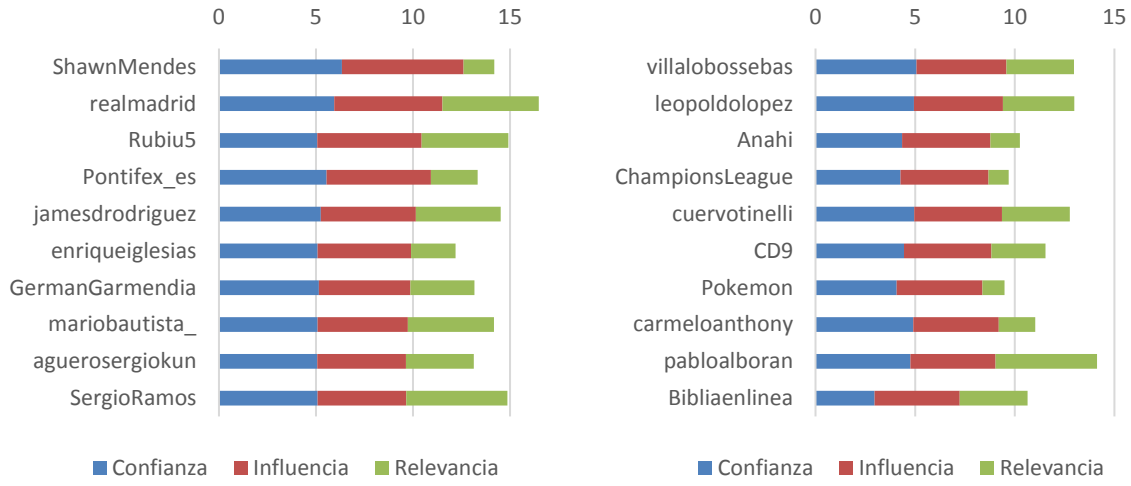


Ilustración 19. Clasificación de candidatos según el factor capacidad de influencia

La mayoría de usuarios son **personas o entidades que abarcan grandes sectores de influencia,** entre ellos, vuelven a aparecer asociaciones, políticos, periodistas, *youtubers*, artistas y deportistas de élite. A pesar de ello, no todos serán óptimos para la campaña si lo que se busca es abarcar sectores concretos. Por ello, completamos el análisis introduciendo los 3 individuos más destacados.

Las puntuaciones que obtienen los candidatos más destacados por cada factor se reflejan en la siguiente figura:

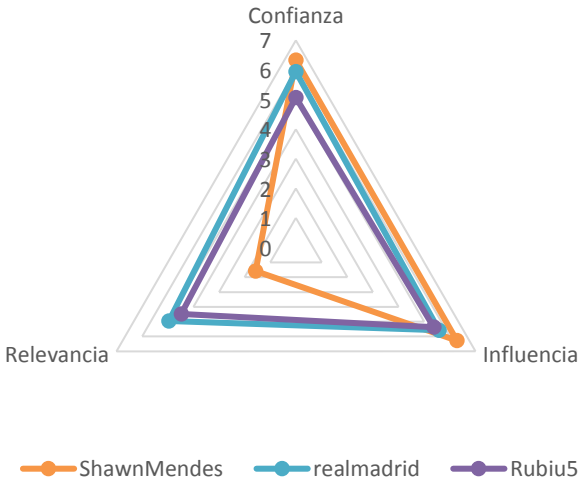


Ilustración 20. Gráfico en estrella de candidatos según el factor capacidad de influencia

Vuelven a aparecer el "Real Madrid C. F." y "Shawn Mendes" analizados en el apartado anterior, por lo que se estudia solamente el tercer individuo "elrubius".

@Rubiu5

elrubius		
Cuenta oficial del youtuber/gamer conocido como "elrubius"		
Localización	Madrid/Noruega	
Seguidores	8.086.072	
Nº de tweets	22.228	
Media de RT's	9816,85	
Media de FAV's	36568,56	
Media de palabras clave por *tweet*	General	0,04
	Nuevas tecnologías	0,61
	Finanzas	0,10
Dispositivo habitual	Twitter Web Client	
Idioma cuenta	es	
Idioma seguidores	es	
Género	-	

Tabla 25. Síntesis información @Rubiu5

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, poseen valores muy altos en cuanto a nº de seguidores, media de RT's y media de FAV's.

Las métricas asociadas a la **relevancia para la campaña** cumplen los requisitos de localización e idioma, además denota interés por las nuevas tecnologías.

Entrando en **valoración personal**, este *youtuber* que vive de sus seguidores, tiene un equilibrio entre confianza, influencia y relevancia. Aunque a no guarda relación con el área de las *Fintech*, está envuelto en las nuevas tecnologías y sus seguidores de perfil joven interesados en ese ámbito.

4.6.3.3. Clasificación por relevancia para la campaña

Este apartado presenta los candidatos que han obtenido mayor puntuación en el factor **relevancia para la campaña.** Nos indica la importancia de características que suponen para el cometido en concreto. Este estudio mostrará si realmente los subfactores planteados son correctos para encuadrar a los usuarios del sector *Fintech*, que denotan interés por las nuevas tecnologías y se encuentran en el mundo de las finanzas.

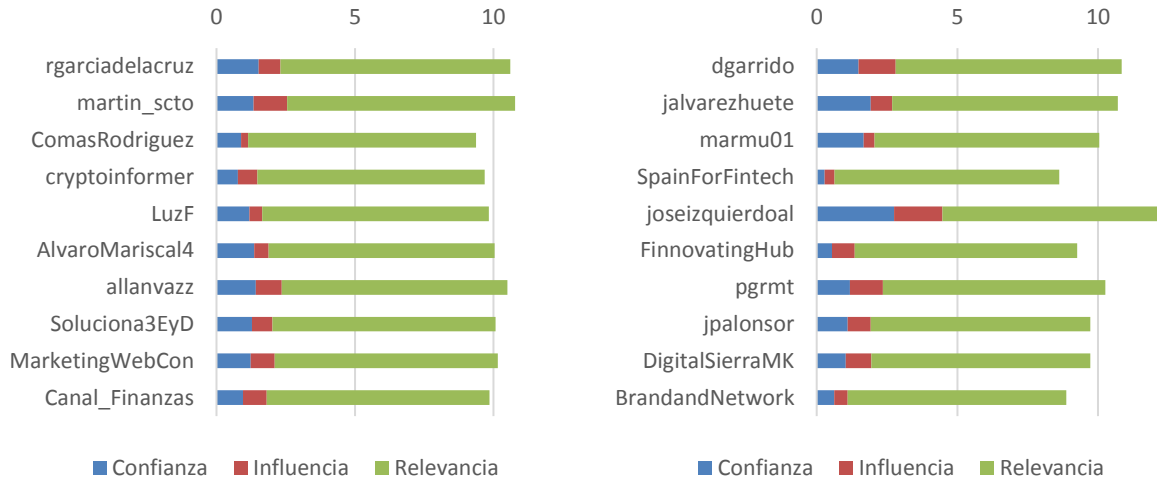


Ilustración 21. Clasificación de candidatos según el factor relevancia para la campaña

La mayoría de personas y entidades de la lista son **usuarios ligados a la temática,** entre ellos, aparecen asociaciones, empresarios, desarrolladores, emprendedores. A pesar de estar estrechamente relacionados con la campaña, no todos cumplirán los objetivos de confianza y capacidad de influencia por lo que completamos el análisis introduciendo los 3 individuos más destacados.

Las puntuaciones que obtienen los candidatos más destacados por cada factor se reflejan en la siguiente figura:

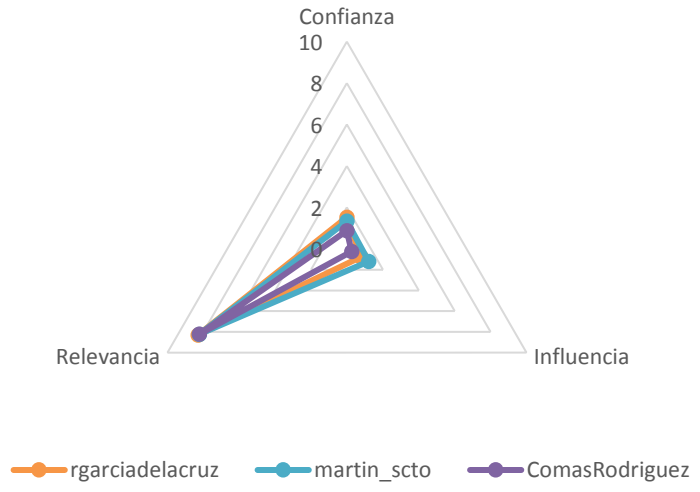


Ilustración 22. Gráfico en estrella de candidatos según el factor relevancia para la campaña

Seguidamente se comprueban los perfiles de estos usuarios:

@rgarciadelacruz

Rodrigo Garcia Cruz		
Cuenta de Rodrigo Garcia Cruz, CEO de la empresa *Finnovating*. Cofundador de *Fintech ventures* y vicepresidente de la asociación española de *Fintech* e *Insurtech*		
Localización	Spain	
Seguidores	2.624	
Nº de tweets	3.872	
Media de RT's	5,42	
Media de FAV's	5,65	
Media de palabras clave por *tweet*	General	1,29
	Nuevas tecnologías	1,38
	Finanzas	1,04
Dispositivo habitual	Twitter for Android	
Idioma cuenta	es	
Idioma seguidores	en	
Género	Masculino	

Tabla 26. Síntesis información @rgarciadelacruz

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, poseen valores bajos en cuanto a nº de seguidores, media de RT's y media de FAV's.

Las métricas asociadas a la **relevancia para la campaña** cumplen los requisitos de localización e idioma, y además denota un fuerte interés por la temática general, las nuevas tecnologías y las finanzas. Un inconveniente, sería el idioma mayoritario de sus seguidores.

Entrando en **valoración personal**, este individuo cumple a la perfección con los requisitos de la campaña, sin embargo, quizá no tenga la capacidad de influencia suficiente puesto que su red no es demasiado grande.

@martin_scto

Martin Schwarz		
Cuenta de Martin Schwarz, ganador del "Best CSS Award" por 5 veces y diseñador del año 2016		
Localización	Las Palmas, Islas Canarias	
Seguidores	1.729	
Nº de tweets	5.207	
Media de RT's	9,48	
Media de FAV's	14,95	
Media de palabras clave por *tweet*	General	1,90
	Nuevas tecnologías	2,33
	Finanzas	0,62
Dispositivo habitual	Twitter for iPhone	
Idioma cuenta	es	
Idioma seguidores	en	
Género	Masculino	

Tabla 27. Síntesis información @martin_scto

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, poseen valores bajos en cuanto a nº de seguidores, media de RT's y media de FAV's en comparación con los grandes *influencers*.

Las métricas asociadas a la **relevancia para la campaña** cumplen los requisitos de localización e idioma, y además denota un fuerte interés por la temática general, y las nuevas tecnologías. Un inconveniente, sería el idioma mayoritario de sus seguidores.

Entrando en **valoración personal**, este individuo cumple con los requisitos de la campaña, pero al igual que el anterior quizá no tenga la capacidad de influencia necesaria.

@ComasRodriguez

David Rodriguez		
Socio en *Arboribus*. Empresa definida como una nueva forma de entender la inversión y la financiación para empresas		
Localización	Barcelona	
Seguidores	967	
Nº de tweets	4.533	
Media de RT's	0,07	
Media de FAV's	0,14	
Media de palabras clave por *tweet*	General	0,79
	Nuevas tecnologías	0,66
	Finanzas	0,68
Dispositivo habitual	LinkedIn	
Idioma cuenta	es	
Idioma seguidores	es	
Género	Masculino	

Tabla 28. Síntesis información @ComasRodriguez

Los indicadores de **confianza y capacidad de influencia** a nivel de métricas, tienen valores muy bajos.

Las métricas asociadas a la **relevancia para la campaña** cumplen los requisitos de localización e idioma, y además denota interés por las finanzas.

Entrando en **valoración personal**, este individuo cumple con los requisitos de la campaña, pero al igual que los anteriores quizá no tenga la capacidad de influencia necesaria.

4.6.3.4. Clasificación en equilibrio

Tras la realización de las anteriores clasificaciones se ha hallado **un único listado con los nombres de usuario** como resultado final del proyecto. Para ello se ha calculado la media aritmética de los tres factores obteniendo y ordenando los resultados por la puntuación total.

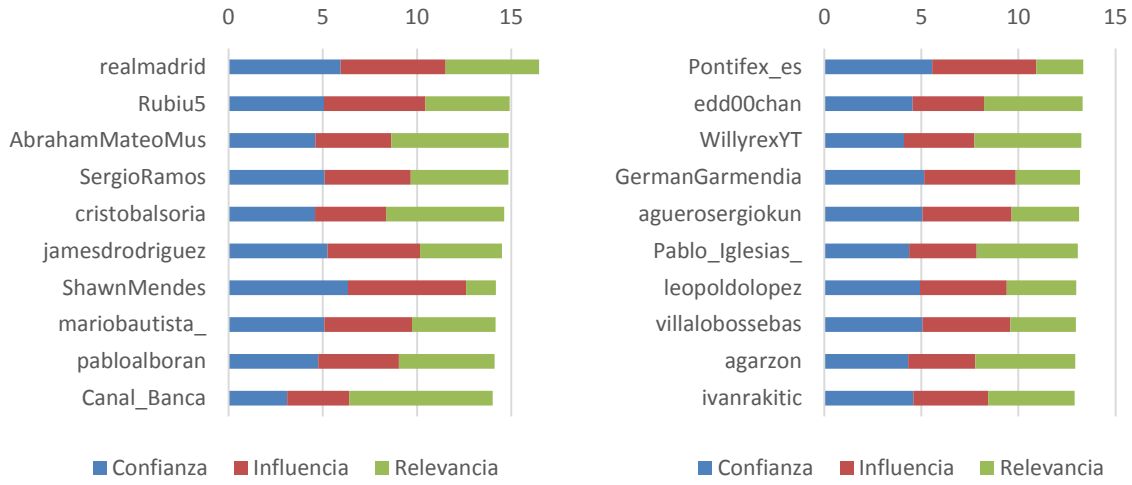


Ilustración 23. Clasificación de candidatos total

Las personas y entidades de la lista **poseen un equilibrio entre los tres factores.** Muchos de ellos se han analizado en los experimentos anteriores por lo que no se incide en las valoraciones individuales.

Las puntuaciones que obtienen los influencers por cada factor se reflejan en la siguiente figura:

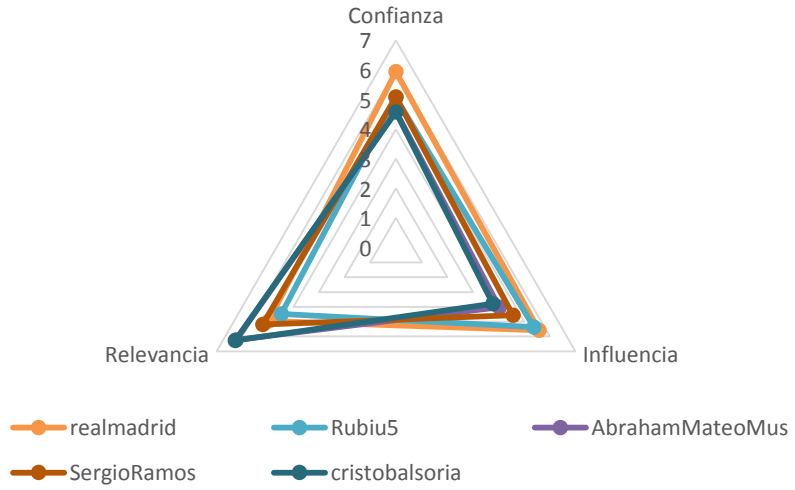


Ilustración 24. Gráfico en estrella de candidatos total

Como se aprecia en la figura, los usuarios alcanzan una puntuación alta y equilibrada en todos los factores. Aunque también aparecen algunos candidatos que han obtenido una puntuación extraordinaria en alguno de los factores sin flaquear en los demás.

5. CONCLUSIONES Y LÍNEAS FUTURAS

5.1. CONCLUSIONES

Todos los objetivos planteados en un primer momento se han cumplimentado satisfactoriamente. Asimismo, las nuevas propuestas e ideas que han surgido durante el desarrollo del trabajo, se han incorporado por el valor adicional que conllevaban.

En el **estado del arte** se ha elaborado un marco conceptual de las áreas que ponen en manifiesto el interés del caso de estudio. Seguidamente, se han planteado todos los conocimientos adquiridos que guardan relación con la ciencia y el análisis de los datos. Para ello, se han caracterizado cada una de las técnicas y metodologías que podrían suponer algún tipo de utilidad. Por último, se ha plasmado la descripción y aplicación de las herramientas empleadas en la obtención, manipulación y análisis de información.

Las técnicas y metodologías resultado del aprendizaje han sido puestas en práctica en el **caso de estudio.** En el mismo, se plantea un nuevo método de estudio, de invención propia, que comprende la obtención de datos, el modelo de análisis y la evaluación de resultados basada en la experimentación.

En consecuencia, el propio caso de estudio genera una serie de conclusiones que proponen autenticidad y validez al método creado:

Todas las formas de obtención de usuarios son perfectamente válidas, sin embargo, la de obtención de usuarios por autores de RT, devuelve resultados de mayor calidad en cuanto a confianza e influencia. Asimismo, todas las puntuaciones relativamente bajas son correctas ya que la multiplicidad de métricas que influyen en cada factor producen ese efecto.

Existen candidatos perfectamente relevantes que no forman parte de los datos obtenidos, esto se debe a periodos de inactividad que experimentan los usuarios en los días que se recolecta la información. Por tanto, sería conveniente ampliar o repetir la búsqueda en otras fechas para obtener resultados con mayor validez.

La elección final de los candidatos no tiene por qué basarse exclusivamente en la media de los factores, sino que se pueden filtrar los usuarios por los que resulten más relevantes.

Una vez tomada la decisión de cuáles son los candidatos óptimos para la campaña se entra en una fase de contacto y negociación. Es posible que no se llegue a un acuerdo y por tanto haya que descartarlo a posteriori.

La herramienta desarrollada y los *datasets* generados en el caso de estudio son accesibles a través del siguiente enlace:

https://github.com/sanxlop/tfg_etsit

5.2. LÍNEAS FUTURAS

Realización de un **estudio paralelo**, buscando diferentes objetivos en cuanto a la relevancia para la campaña, e ideando nuevos subfactores que impliquen nuevas métricas junto a una nueva parametrización.

Aplicación del modelo analítico a otras redes sociales y entornos, siguiendo la misma metodología, agregando las funciones y elementos de valor de las nuevas redes sociales y herramientas. Además, incluye un estudio sobre la relación de los candidatos más relevantes en la nueva red social en comparación con la de *Twitter*.

Se plantea una **aplicación web**, en el que cualquier usuario tiene acceso y puede plantear su temática de estudio para la obtención de usuarios influyentes. Para ello será necesario la implementación de una **interfaz gráfica de la aplicación** en HTML con R, parametrizando los valores adquiriendo mayor dinamismo y usabilidad, junto con un servidor de cómputo en el que se realicen las operaciones.

Valoración del impacto de los candidatos para la campaña junto con la forma de contacto y costes. En este caso se analizarán los diferentes *influencers* en función de la calidad/precio que supondría su contratación, llevando estas valoraciones a cabo a partir de las puntaciones obtenidas en los factores.

6. BIBLIOGRAFÍA

[1] A. K. y M. H. , Users of the world, unite! The challenges and opportunities of social media, Business Horizons, 2010.

[2] M. d. F. García, J. D. Alan y S.-C. S. , «Identifying the new Influencers in the Internet Era: Social Media and Social Network Analysis,» *Revista Española de Investigaciones Sociológicas,* nº 153, pp. 23-40, 2016.

[3] GorBrit, «Las Redes Sociales: Origen y evolución,» 2014. [En línea]. Available: https://gorbrit.wordpress.com/2014/06/24/las-redes-sociales-origen-y-evolucion/.

[4] We Are Social, «Slideshare,» Enero 2016. [En línea]. Available: http://www.slideshare.net/wearesocialsg/digital-in-2016/418.

[5] Twitter, «Q2 Letter to Shareholders,» 2016. [En línea]. Available: http://files.shareholder.com/downloads/AMDA-2F526X/2510368267x0x901385/664658CA-D1D8-4635-83F4-8C9D5A9A1F52.

[6] R. González, «Personal Influence: A 55 años de la irrupción de los líderes de opinión. Razón y Palabra,» *Razón y palabra,* 2011.

[7] P. Lazarsfeld y E. Katz, The personal influence, 1995.

[8] D. McAuley, «Wharton Fintech,» 2015. [En línea]. Available: https://medium.com/wharton-fintech/what-is-fintech-77d3d5a3e677.

[9] A. Rodríguez, «Hipertextual,» 27 Noviembre 2015. [En línea]. Available: https://hipertextual.com/2015/11/datos-fintech-espana.

[10] E. Arrieta, «Expansión,» 25 11 2015. [En línea]. Available: http://www.expansion.com/economia-digital/innovacion/2015/11/25/5654978222601de7068b4588.html.

[11] Accenture, «Fintech Evolving Landscape,» 2016. [En línea]. Available: http://www.fintechinnovationlablondon.co.uk/pdf/Fintech_Evolving_Landscape_2016.pdf.

[12] io Marketing, «El branding en las agencias de marketing,» [En línea]. Available: http://www.iomarketing.es/blog/branding-las-agencias-marketing/.

[13] C. Zuriguel, «Marketing de Influencers,» InboundCycle, 3 9 2014. [En línea]. Available: http://www.inboundcycle.com/blog-de-inbound-marketing/marketing-de-influencers-una-nueva-estrategia-cada-vez-m%C3%A1s-en-boga.

[14] Augure, «Launch Metrics,» 20 2 2014. [En línea]. Available: https://www.launchmetrics.com/es/recursos/blog/influencer-marketing-estatus-2014.

[15] D. A. Liu, «Research Methods | Data Science and Data Scientist,» IBM, 2015. [En línea]. Available: http://www.researchmethods.org/DataScienceDataScientists.pdf.

[16] D. Conway, «The Data Science Venn Diagram,» 2010. [En línea]. Available: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

[17] Wikipedia, «Data analysis,» [En línea]. Available: https://en.wikipedia.org/wiki/Data_analysis.

[18] «Investigacion cuantitativa vs cualitativa,» 2008. [En línea]. Available: http://es.slideshare.net/guest2bc00c/investigacion-cuantitativa-vs-cualitativa.

[19] «Data Analysis Wikipedia,» [En línea]. Available: https://en.wikipedia.org/wiki/Data_analysis.

[20] P. L. Eiroá, «Estadística descriptiva e inferencial,» 2013. [En línea]. Available: http://es.slideshare.net/marketing2009/estadstica-descriptiva-e-inferencial.

[21] SPSS, «Medidas de tendencia central,» [En línea]. Available: http://www.spssfree.com/curso-de-spss/analisis-descriptivo/media-mediana-moda-medidas-tendencia-central.html.

[22] SPSS, «Medidas de dispersión,» [En línea]. Available: http://www.spssfree.com/curso-de-spss/analisis-descriptivo/varianza-desviacion-medidas-de-dispersion.html.

[23] Wikipedia, «R (lenguaje de programación),» [En línea]. Available: https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n).

[24] Wikipedia, «RStudio,» [En línea]. Available: https://es.wikipedia.org/wiki/RStudio.

[25] Twitter, «Twitter Developers,» [En línea]. Available: https://dev.twitter.com.

[26] Twitter, «Twitter Apps,» [En línea]. Available: https://apps.twitter.com.

[27] Twitter, «OAuth,» [En línea]. Available: https://dev.twitter.com/oauth.

[28] Twitter, «Rate Limits: Chart,» [En línea]. Available: https://dev.twitter.com/rest/public/rate-limits.

[29] CRAN, «Package httr,» [En línea]. Available: https://cran.r-project.org/web/packages/httr/index.html.

[30] CRAN, «Package jsonlite,» [En línea]. Available: https://cran.r-project.org/web/packages/jsonlite/index.html.

[31] Twitter, «REST APIs,» [En línea]. Available: https://dev.twitter.com/rest/public.

[32] AI Applied, «Demographics API,» [En línea]. Available: http://ai-applied.nl/demographics-api/.

[33] AI Applied, «Sentiment Analysis API,» [En línea]. Available: http://ai-applied.nl/sentiment-analysis-api/.

7. ANEXOS

7.1. MANUAL DE EJECUCIÓN Y OBTENCIÓN DE DATOS

El programa se ejecuta desde el archivo **main.R** que contiene la función *main().* Se incluye la documentación de esta función para comprender su funcionamiento según se desarrolla el estudio.

```
## DESARROLLO AUTOMATICO
# @params: flag, first, repeatTimes, error, new, statUsers, followUsersSize,
firstAdd, lastAdd, auth, ntable
main <- function(flag, first = NA, repeatTimes = NA, error = NA, new = NA,
statUsers = NA, followUsersSize = NA, firstAdd = NA, lastAdd = NA, auth = NA,
ntable = NA)
```

7.1.1.1. Obtención y experimentación de tweets y usuarios

En la ejecución de la **primera fase** se obtiene una muestra en forma de *dataset* de 270.844 tweets, hallados a partir de un listado de palabras clave; y otra de 311.728 usuarios, conseguidos por búsqueda de palabras clave, usuarios que han publicado los *tweets* y autores originales de los *RT's*.

El *dataset* de tweets, **"stored_tweets",** incluye la siguiente información: *Id, Fecha de creación, @Usuario dueño RT, Número de RT, Número de Favoritos, Texto, Lenguaje, Hashtags, Nº de hashtags, Nombres de usuario de menciones, Nº de menciones, Símbolos, URLs, Media, Dispositivo, Tipo de tweet, Nombre de usuario y @Usuario*

Mientras que el *dataset* de usuarios, **"stored_users",** incluye: *Id, Fecha de creación, Nombre de usuario, @Usuario, Número de seguidores, Número de seguidos, Número de listas, Localización, Descripción, Número de favoritos, Cuenta verificada, Número de tweets, Lenguaje y la forma de obtención.*

```
Fase 1:
main(flag = 1, first = TRUE, repeatTimes = 100, error = FALSE, new = TRUE)

        flag = numeric #Nº de fase
        first = logical #Indica continuar busqueda sobre el MAX_ID
        repeatTimes = numeric #Nº de veces que se ejecuta la obtención de datos
        error = logical #Soluciona errores de la API
        new = logical #Indica crear nuevas tablas
```

En la **segunda fase** realizamos tareas de limpieza de duplicados quedando una muestra de **193.541 *tweets***, y un listado de **177.467 usuarios** a analizar**.** En este caso, los *tweets*

corresponden a las fechas del *2 al 5 de enero de 2017* y podrían añadirse nuevos resultados en caso de requerirlo.

A partir de la información adquirida de los usuarios, se ejecuta el **primer experimento** con la configuración descrita que evalúa los factores, subfactores y métricas relacionados. Los resultados de este experimento se almacenan en el *dataset* **"experiment1"**.

```
Fase 2:
main(flag = 2)

        flag = numeric #Nº de fase
```

7.1.1.2. Obtención y experimentación de datos estadísticos inferidos de sus tweets

Se ejecuta la **tercera fase** obteniendo un nuevo *dataset* **"statistic_users"** que incluye los datos de "stored_users" más las nuevas columnas vacías relacionadas con los datos estadísticos inferidos de los tweets de cada usuario. Las limitaciones de la *API* obligan a restringir esta parte del estudio por lo que tomaremos **4993 usuarios**, debido a que la obtención de esta información es costosa en cuanto a número de usuarios y tiempo.

```
Fase 3:
main(flag = 3, statUsers = 1000)

        flag = numeric #Nº de fase
        statUsers = numeric #Tamaño de la tabla
```

En la **cuarta fase** de ejecución, se obtienen y calculan los valores estadísticos tales como la *media, desviación típica, cuantiles, varianza y moda* de los *RT's, Fav's, hashtags, palabras clave y dispositivos habituales*, de los últimos 100 *tweets* de cada usuario y se almacenan en "statistitc_users".

```
Fase 4:
main(flag = 4, firstAdd = 1, lastAdd = 100)

        flag = numeric #Nº de fase
        firstAdd = numeric #Fila inicial
        lastAdd = numeric #Fila final
```

A partir de la información estadística inferida de los *tweets* de los usuarios, se ejecuta en la **quinta fase** el **segundo experimento** con la configuración descrita que evalúa los factores, subfactores y métricas relacionados. Los resultados de este experimento se almacenan en el *dataset* **"experiment2"**.

```
Fase 5:
main(flag = 5)

        flag = numeric #Nº de fase
```

7.1.1.3. Obtención y experimentación de datos estadísticos inferidos de sus seguidores y demografía

Se procede a ejecutar la **sexta fase** del proceso, que al igual que en la tercera fase se obtiene un nuevo *dataset* **"followers_users"** que incluye los datos de "stored_users" y "statistic_users" más las nuevas columnas vacías relacionadas con los datos estadísticos inferidos de los seguidores de cada usuario y la demografía. Las limitaciones de la *API* vuelven a restringir esta parte del estudio, por lo que tomaremos **615 usuarios**.

```
Fase 6:
main(flag = 6, followUsersSize = 1000)

        flag = numeric #Nº de fase
        followUsersSize = numeric #Tamaño de la tabla
```

En la **séptima fase** de ejecución, se obtienen y calculan los valores estadísticos tales como la *media, desviación típica, cuantiles, varianza y moda* de los *tweets, seguidores, seguidos, lenguaje y palabras clave,* de los últimos 200 seguidores de cada usuario junto con el género y edad. Estos datos se almacenan en "followers_users".

```
Fase 7:
main(flag = 7, firstAdd = 1, lastAdd = 15, auth = TRUE)

        flag = numeric #Nº de fase
        firstAdd = numeric #Fila inicial
        lastAdd = numeric #Fila final
        auth = logical #Indica el tipo de autorización
```

A partir de la información estadística inferida de los *seguidores* de los candidatos y la demografía, se ejecuta en la **octava fase** el **tercer experimento** con la configuración descrita que evalúa los factores, subfactores y métricas relacionados. Los resultados de este experimento se almacenan en el *dataset* **"experiment3"**.

```
Fase 8:
main(flag = 8)

        flag = numeric #Nº de fase
```

7.1.1.5. Obtención de puntuaciones finales

Una vez obtenidos los resultados de los tres experimentos, calculamos las puntaciones finales introduciendo el *coeficiente de importancia* de cada factor/experimento como parámetro, en la ejecución de la **novena fase**. El *dataset* con las puntaciones finales se almacenan en **"allExperiments"**.

```
Fase 9:
main(flag = 9)

        flag = numeric #Nº de fase
```

7.1.1.6. Exportación de datasets

La exportación de todos los *dataset* obtenidos se realiza en la **décima y última fase** de ejecución del programa. Para ello se introducen una serie de parámetros que indican el *dataset* a exportar y lo transforma a un archivo "xlsx" (formato Excel).

```
Fase 10:
main(flag = 10, ntable = 1)

        flag = numeric #Nº de fase
        ntable = numeric #Nº de tabla a exportar (1:8)
```