

# Imágenes desgarradas: el uso de *scrapers* en investigación social en Instagram sobre cáncer

## Scattered Images: scrapers and the representation of cancer on Instagram

### *Imagens rasgadas: o uso de scrapers na pesquisa social sobre câncer no Instagram*

**Miguel Varela-Rodríguez**, Universidad de Valladolid, Valladolid, España  
([miguel.varela@uva.es](mailto:miguel.varela@uva.es))

**Miguel Vicente-Mariño**, Universidad de Valladolid, Valladolid, España  
([miguel.vicente@uva.es](mailto:miguel.vicente@uva.es))

**RESUMEN** | El uso de redes de contenido visual como Instagram está bien documentado en la comunicación en salud, especialmente el análisis de contenido para estudiar las imágenes. Sin embargo, esta metodología supone un reto ante las crecientes dificultades en el acceso y un marco legal y de actuación muy limitados. Basado en los postulados de la sociología visual, este artículo explora una metodología para obtener datos de Instagram mediante el uso de *scrapers*, revisando las necesidades técnicas y las implicaciones éticas en el uso de este tipo de herramientas. Se analiza la distribución de imágenes acompañadas por la etiqueta #SacaPecho, creada por la Asociación Española Contra el Cáncer con ocasión del Día Internacional de la Lucha Contra el Cáncer (19 de octubre de 2020). El uso de *scrapers* permite obtener referencias de más de 7000 imágenes en poco tiempo. El trabajo permite entender las herramientas al alcance de la investigación social para acceder a datos relevantes en Instagram y propone un debate sobre las posibilidades éticas en este ámbito.

**PALABRAS CLAVE:** investigación en redes sociales; comunicación en salud; sociología visual; cáncer; métodos computacionales en investigación social.

#### FORMA DE CITAR

Varela-Rodríguez, M. & Vicente-Mariño, M. (2021). Imágenes desgarradas: el uso de *scrapers* en investigación social en Instagram sobre cáncer. *Cuadernos.info*, (49), 72-97.  
<https://doi.org/10.7764/cdi.49.27809>

---

**ABSTRACT** | *The use of visual social networks like Instagram in health communications is well-documented, especially when it comes to the analysis of specific contents to study images. However, this methodology poses a challenge due to the growing difficulties in access and a very limited legal and action framework. Based on the assumptions of visual sociology, this article explores a methodology to obtain data from Instagram using scrapers, offering a revision of its technical and ethical implications. Through this method, we analyse the distribution of Instagram posts carrying the hashtag #SacaPecho (a Spanish expression that refers to standing up for or showing support to a cause), created by the Spanish Society Against Cancer (Asociación Española Contra el Cáncer) on the occasion of the International Day Against Cancer (October 19, 2020). We obtain data from more than 7,000 images in a matter of minutes. Our work sheds light on the tools available for researchers to access data from Instagram and proposes a debate on the ethical possibilities to implement them.*

**KEYWORDS:** social media research; health communication; visual sociology; cancer; computational methods in social science.

---

**RESUMO** | A utilização de Instagram e de redes semelhantes de conteúdo visual na comunicação sobre saúde está bem documentada, especialmente na análise de conteúdo. Contudo, as crescentes limitações no acesso a esta informação representam um desafio. Baseado na sociologia visual, este artigo explora uma metodologia para obter dados da Instagram através da utilização de scrapers. Revemos as necessidades técnicas e as implicações éticas da utilização deste tipo de ferramenta. Através desta metodologia, analisamos a distribuição de imagens acompanhadas pela hashtag #SacaPecho (“mostre o seu apoio”) na Instagram. A campanha foi criada pela Associação Espanhola Contra o Cancro por ocasião do Dia Internacional contra o Câncer (19 de outubro de 2020). A utilização de scrapers permite a obtenção de referências de mais de 7.000 imagens num curto espaço de tempo. O trabalho permite uma compreensão dos mecanismos à disposição da investigação social para acessar a dados relevantes no Instagram e propõe um debate sobre as possibilidades éticas nesta área.

**PALAVRAS CHAVE:** investigação em redes sociais; comunicação em saúde; sociologia visual; câncer; métodos computacionais na investigação social.

## INTRODUCCIÓN

A pesar de los grandes avances en el estudio de las redes sociales en la última década, la obtención y el análisis automatizados de imágenes es, en muchos casos, un reto para la investigación social. Desde que Facebook y otros proveedores cerraran el acceso a sus plataformas tras el caso Cambridge Analytica en 2018 (Bruns, 2019; Rogers, 2018), se ha dificultado la búsqueda y el uso de imágenes de redes para la investigación, por la dificultad técnica y distintas consideraciones éticas. A menudo, los investigadores se ven envueltos en un proceso manual de identificación, selección y procesamiento, invirtiendo grandes cantidades de tiempo en recortar, copiar y pegar las imágenes en sus computadoras. Mientras unos buscan establecer relaciones de asociación con Facebook a través de concursos –una tarea costosa, como se discute más adelante– otros se ven en la tesitura de utilizar herramientas que aceleran su trabajo, pero que en muchos casos rompen los Términos de Uso de plataformas como Instagram. Algunos abogan por abandonar este tipo de trabajo hasta que el clima sea más favorable. Entretanto, las imágenes continúan siendo parte fundamental de los modos de expresión e intercambio en la sociedad contemporánea.

En este contexto, este artículo revisa el uso de una técnica sencilla para buscar y obtener imágenes en Instagram como alternativa viable tras el cierre de las API públicas en 2018 mediante dos herramientas de *scraping* (o rascado de datos) conocidas como Instaloader e Instalooter. Para ello, se vale de un caso práctico: la campaña de la Asociación Española Contra el Cáncer (AECC) en torno al Día Mundial de la Lucha Contra el Cáncer con el hashtag #SacaPecho. Este tipo de campañas genera grandes cantidades de contenidos visuales en línea. Se exploran los pasos específicos para diseñar una búsqueda de imágenes subidas a Instagram con una etiqueta (*hashtag*) en fechas determinadas, se observan las consideraciones éticas y se plantean alternativas.

Los objetivos son demostrar las oportunidades al alcance de investigadores sociales para acceder a imágenes de redes y estimular un debate sobre las dificultades (técnicas y éticas) asociadas al uso de este tipo de herramientas y el rol de las empresas proveedoras de plataformas como Instagram.

## MARCO TEÓRICO: MODOS DE VER EN EL MUNDO DIGITAL

### La imagen como fuente de información sociológica

El estudio de las imágenes y de las ciencias sociales tienen una relación larga y compleja, facilitada en gran manera por la antropología a principios del siglo XX. La democratización en el acceso a las cámaras fotográficas hacia fines de la década de 1960, el abaratamiento de la impresión en color y el rol fundamental de los medios

visuales en la guerra de Vietnam (Berger, 2016; Sontag, 2017) incidieron en el encuentro entre el estudio de la imagen y la sociología. Esto abrió un espacio para algunos sociólogos que “rechazaban los paradigmas dominantes en la investigación y la teoría [sociales]” (Harper, 1988, p. 59).

Poco a poco, lo visual se convertiría en un modo de contar historias, pero también en una fuente de información clave para (re)construirlas. Ambas son funciones básicas que explora la sociología visual desde la Sociedad Internacional para la Sociología Visual (International Visual Sociology Association) en 1981, y que permean gran parte del trabajo con imágenes en las ciencias sociales (Harper, 1996; Pauwels, 2015), especialmente desde el análisis computacional.

### **La imagen en redes y en Instagram, más allá del análisis de contenido**

Con la llegada de la fotografía democratizada, esta se convirtió en una actividad común a millones de personas y familias, hasta el punto de convertirnos en “yonquis a las imágenes” (Sontag, 2017, p. 33). La imagen se transformó en un bien producido y consumido en masa. Los primeros sociólogos visuales son conscientes de esta situación y la aprovechan como una oportunidad. Trabajan con una multitud de fotografías encontradas, archivos familiares y otros bancos de imágenes para revisar los significados y las dinámicas contenidas en la imagen<sup>1</sup>. La importancia de la imagen en redes sociales va más allá de lo representado e incluye también la representación (Pauwels, 2015). Mientras que el análisis de lo representado se refiere a revisar lo que muestran las imágenes, el estudio de la representación tiene que ver con las decisiones a la hora de producir las imágenes y sus modos de circulación.

Rose (2019) se refiere a esta diferenciación como los cuatro ámbitos de estudio de la imagen. En su enfoque, la imagen puede estudiarse desde sus contenidos, su modo de producción, su modo de circulación o su modo de visualización. Más allá de buscar qué se representa en la imagen, Rose se preocupa por quién la ha producido, cómo, por qué, cómo la comparte y en qué circunstancias se acercan los consumidores a esa imagen.

En redes sociales, el volumen de imágenes disponibles resulta abrumador, con más de 95 millones subidas cada día a Instagram (Abutaleb, 2016); así, el estudio de los modos de producción y circulación cobra mayor relevancia y complejidad. Hand (2016) asegura que “[p]odríamos decir que el número de fotografías ‘encontradas’ implica que debemos abandonar cualquier idea de analizarlas

---

1. Ver, como ejemplo en redes sociales, Manovich (2016), González-Requena (1988, 2013), con sus ejemplos en cine, o Abril-Curto (2013) sobre comunicación política.

individualmente” (p. 216). Autores como Manovich (2018) y los seguidores del análisis cultural se adaptan a esta situación con soltura. Además del contenido y la forma de la imagen, les interesan sus medios de producción y, especialmente, sus medios de circulación. En su propuesta por una analítica cultural, entienden que el volumen de imágenes es tal que una categorización individual pierde sentido y limita la capacidad de análisis del investigador, que vive “en la sombra de un mundo de redes sociales, recomendaciones, apps e interfaces que utilizan la analítica de medios” (Manovich, 2018, p. 2).

Este modo de acercarse a las redes, favoreciendo un análisis de los modos de producción y circulación, sin intervenir sobre los textos individuales, ha tenido especial calado en las redes con interfaces de programación abiertas, como Twitter, pero su uso con Instagram es limitado. En parte, como explica Rose, porque las imágenes permiten identificar a los individuos representados, lo que plantea problemas éticos (Rose, 2019), y en parte, sin embargo, por cuestiones económicas.

En Twitter o YouTube, los estudios han sido particularmente fructíferos cuando se ha tratado de analizar volumen y contenido en la comunicación política (Kligler-Vilenchik et al., 2020), en los estudios de género (Giraldo-Luque et al., 2018; Regueira, Ferreiro, & Vila, 2020) –aplicando técnicas de *web scrapping* al estudio de la representación de la mujer en YouTube o el *screen scrapping* para analizar más de 100 mil tuits en el movimiento #NiUnamenos–, así como durante la pandemia de la COVID 19 (Babvey et al., 2020; Cinelli et al., 2020; Orduña-Malea et al., 2020). En campos como el marketing, el análisis computacional de redes es una práctica consolidada (Harrigan et al., 2020; Purba et al., 2020; Tafesse & Wood, 2020).

Estos métodos cuentan con un alto nivel de desarrollo y con opciones de código abierto y gratuitas (Arcila-Calderón et al. 2019), y reducen la carga de trabajo y el tiempo necesarios para obtener y estudiar datos de redes sociales.

En Instagram, sin embargo, el trabajo no ha seguido esta evolución, sino un enfoque marcadamente cualitativo. Por ejemplo, en el análisis de *selfies* extraídas a mano (Abidin, 2016; Couture Bue, 2020), o a través de la entonces disponible API de Instagram (Jang et al., 2015). En 2016, Munk y sus colegas afirmaban que “la investigación social en Instagram está en su infancia” (2016, p. 2), y animaban al uso de Instagram más allá de la imagen específica y hacia el análisis de redes sociales. Si bien existen trabajos que desarrollan métodos para obtener información automáticamente de Instagram, son previas al cierre de las API de las redes sociales en 2018 (Highfield & Leaver, 2014) o no exploran las implicaciones éticas (Dewi et al., 2019). Tampoco responden a la emergencia de los algoritmos en las redes sociales como un momento definitorio de nuestra era (Schie, Westra, & Schäfer, 2017).

A pesar de estas dinámicas, la investigación social sobre imágenes en redes sociales continúa limitada por dinámicas comerciales y debates éticos, por el bloqueo en el acceso, y por la indefinición e incertidumbre sobre lo que es y no es aceptable al estudiarlas.

### **El scraping como método de obtención de imágenes y el debate ético que suscita**

Una de las técnicas para obtener imágenes de fuentes en línea surgida del advenimiento de las redes sociales y la explosión productiva en imágenes desde principios de los 2000 es el *scraping* –o rascado, en una traducción tosca al castellano–. El término hace referencia al modo de operar de estas herramientas: basándose en una serie de parámetros, arrastran o rascan todo el contenido de una página web que concuerda con dichos parámetros y los pone al alcance del usuario.

Los métodos del *scraping* los describen en detalle Marres y Weltevrede (2013), Hirschev (2014) y en el contexto de YouTube, Regueira y sus colegas (2020). Las herramientas de *scraping* pueden acceder a una determinada página web (como por ejemplo la página de Instagram dedicada a la etiqueta #cancer) y descargar todo su contenido, incluyendo las imágenes, el texto que las acompañan y sus metadatos, mediante una conexión HTML. Este proceso se conoce como *web scraping*, donde la herramienta conecta directamente con la base de datos del proveedor (Instagram, en nuestro estudio). Un método alternativo es el llamado *screen scraping*: en lugar de conectar con la base de datos, la herramienta captura todo el contenido cargado en la pantalla del usuario. Para navegadores como Google Chrome, la propia tienda de extensiones de Google incluye herramientas para hacer este tipo de extracción.

En general, el *scraping* es una técnica con una reputación cuestionable, a menudo vinculada con prácticas ilegales (Marres & Weltevrede, 2013). No en vano es una técnica prohibida en algunas plataformas, incluida Instagram según sus términos de uso (Instagram, 2018), y que se ha visto severamente limitada tras el caso Cambridge Analytica en 2018. Hay varias razones para ello, adscritas fundamentalmente a dos ámbitos de riesgo: para el usuario y para el proveedor de la plataforma.

Lo primero tiene que ver con cuestiones clave en los estudios sociales con imágenes, que ya reconocía Becker (1974) en su propuesta para una sociología visual: la privacidad, el consentimiento y la inclusión. Antropólogos, etnógrafos y sociólogos no son extraños a estos debates. Muchos trabajos exploran las implicaciones éticas del uso de imágenes en la investigación social (Markham & Buchanan, 2012, 2017; Prosser et al., 2008; Rose, 2019). Varios exploran este tema en relación con Instagram y las redes sociales específicamente (Henricks, 2017; Marres & Weltevrede, 2013; Walsh & Baker, 2017; White & Boatwright, 2020).

Además de consideraciones éticas, Prosser y sus colegas (2008) invitan a los investigadores a revisar el marco legal de su investigación. Para Instagram, incluye revisar los Términos de uso y cerciorarse de que el uso de estas técnicas está permitido (no lo está). Mientras, Schie y sus colegas (2017) plantean el proceso de investigación en tres fases, en las que llaman a “un momento de razonamiento ético” (p. 196): diseño de la investigación, en la que se atacan las cuestiones relacionadas con los posibles sujetos (de las imágenes, en nuestro caso), exploración segura de los datos, con una documentación clara de los pasos seguidos y almacenar los datos en un entorno protegido, y el proceso de obtención final.

Es aquí donde entra el segundo ámbito de riesgo: la propiedad de las imágenes, la intervención de las entidades comerciales como Facebook, Google y otras empresas de gran calado, y la protección de su modelo de negocio. El debate en torno al rol de estas empresas a la hora de delimitar qué es y qué no es investigable es agitado. Esta problemática está recogida en los marcos éticos de Townsend y Wallace (2016) y de Prosser y sus colegas (2008) que arrancan su modelo con la inclusión de consideraciones legales.

Más allá del argumento ético, el entorno legal en la obtención de imágenes de redes sociales complica el trabajo de los investigadores, especialmente desde que proveedores como Facebook cerrasen sus API tras el caso Cambridge Analytica. De hecho, Hirschey detalla la variedad de opciones a disposición de grandes empresas, tales como “sobrecargar los recursos legales de los *scrapers* [entendidos como individuos u organizaciones que utilizan la técnica de *scraping*] y rápidamente dejarlos en la bancarrota” (2014, p. 925). Se convierte, pues, en una batalla legal y empresarial que poco tiene que ver con la ética de investigación, tal y como denuncian Rogers (2018) o Bruns (2019), y en la que los investigadores deben bailar con la lógica económica.

### **El cáncer y la educación en salud como ejemplo de la importancia de la imagen en línea**

La imagen ha jugado un rol especialmente importante en el ámbito de la comunicación y la educación en salud. Más específicamente, la comunicación en torno al cáncer. Ya en 1978, Susan Sontag reflexionaba sobre el impacto que tenían las imágenes sobre las pacientes con cáncer de mama (Sontag, 1978). Igualmente, Ehrenreich (2001) hace una crítica feroz al culto al rosa con el que se encontró tras su diagnóstico de cáncer de mama. La imagen del cáncer se ha estudiado en revistas especializadas y en medios de comunicación (Grant & Hundley, 2008), en campañas de concienciación (Cartwright, 1998) y, más recientemente, en redes sociales como Instagram (Basch & MacLean, 2019). Predominan los trabajos dedicados al análisis del discurso (Döbrössy et al., 2020; Gibson et al., 2016; Sutton et al., 2018; Zhang et al., 2019), a las dinámicas de viralización (Noar et al., 2018; Wang et al., 2019), o a la respuesta fisiológica ante las imágenes (Chou et al., 2020).

Mediante estos trabajos se percibe la importancia creciente de las redes sociales como ámbito de consumo de información relacionada con el cáncer, hasta el punto de haberse convertido en el segundo medio de información para pacientes (Blanch-Hartigan & Viswanath, 2015). Pardo (2019) destaca el rol que tienen las redes en la construcción de una imagen más humana y menos dominada por la medicina y la despersonalización, y reflexiona sobre los ámbitos de producción y circulación que ya destacaba Rose.

Sobre el cáncer versan importantes trabajos de fotografía y vídeo documentales, como los de Jo Spence (1986), Angel Merendino<sup>2</sup> o Jeremy Nicholl<sup>3</sup>.

De nuevo, trabajos documentales y analíticos comparten un enfoque: el del análisis de la imagen individual. A través del análisis de discurso y de contenido, desarrollan las historias del cáncer en un esfuerzo casi etnográfico, añadiendo valor narrativo. Sin embargo, en exploraciones preliminares realizadas por los autores del presente artículo en 2019, se han detectado más de 40 millones de imágenes con palabras clave asociadas al cáncer solo en Instagram. Tal volumen de imágenes relacionadas con el cáncer puede abrumar cualquier intento de este tipo de análisis en redes sociales.

## **METODOLOGÍA**

Para el propósito de este artículo, se exploran las imágenes acompañadas por la etiqueta #SacaPecho en Instagram durante el 19 de octubre de 2020, Día internacional de la lucha contra el cáncer de mama. Dicha etiqueta fue lanzada por la Asociación Española Contra el Cáncer (Asociación Española Contra el Cáncer, 2020), una de las organizaciones sin fines de lucro más grandes del país y con un gran calado en la imagen social y financiamiento de la investigación contra esta enfermedad. #SacaPecho, parte de su campaña para concienciar sobre el cáncer de mama ese año, logró miles de impresiones en las primeras horas. Utilizaremos la técnica de *scraping* para analizar las dinámicas de publicación en torno a esta etiqueta y su fecha clave, detallando cada paso seguido. Esta utiliza el lenguaje de programación Python, para obtener los datos.

### **Herramientas utilizadas e instalación**

*Python (Miniconda) y Pip*

Python es un lenguaje de programación con el que muchos investigadores sociales están familiarizados como alternativa a R para el análisis estadístico. Se

---

2. <https://www.mywifesfightwithbreastcancer.com/>

3. <http://mybeautifulcancer.com/>



puede descargar como paquete completo utilizando Miniconda, un instalador sencillo que incluye la última versión disponible, además de otras herramientas útiles. Aunque existen otros modos de instalar Python, Miniconda es, posiblemente, el más intuitivo y sencillo. Está disponible, sin costo, en su página web<sup>4</sup>. En un PC, Miniconda aparecerá instalado como Anaconda Prompt (miniconda 3)<sup>5</sup>.

Pip es un instalador de paquetes para Python –una herramienta utilizada para añadir componentes que utilizan este lenguaje y que añaden funcionalidades como la producción de gráficos a partir de los datos obtenidos o la definición de parámetros de búsqueda–. Pip permite descargar códigos desarrollados por terceros e instalarlos con una sencilla línea de texto a través del Anaconda Prompt (Miniconda 3) instalado previamente. Está integrado en Miniconda y lo utilizaremos para instalar nuestras herramientas de *scraping*.

Para instalar ambas herramientas, basta con descargar el paquete de instalación de la página web y seguir los pasos indicados.

### *Jupyter Notebook*

El entorno de programación Python se puede implementar desde el símbolo del sistema o desde un procesador de texto dedicado, una opción más sencilla y conveniente. En este caso, utilizaremos Jupyter Notebook, una herramienta desarrollada específicamente para facilitar el trabajo con Python. Se trata de un editor de texto de uso intuitivo. Para instalarlo, basta con:

1. Abrir Anaconda Prompt (Miniconda 3), instalado previamente. Abrirá una pantalla negra en la que podremos introducir texto.
2. Introducir este código:

```
conda install -c conda-forge jupyterlab
```

Con esto, Jupyter Notebook quedará instalado y listo para utilizar. Más adelante se detalla cómo arrancarlo.

### *Instaloder e Instalooter*

Instaloder e Instalooter son herramientas de *scraping* que conectan directamente con la base de datos de Instagram, sin necesidad de utilizar una interfaz oficial. A pesar de romper los Términos de uso de Instagram, ambos están, a fecha de redacción de este artículo, disponibles a través de Pip, y se actualizan con regularidad. Los pasos para su instalación son sencillos:

---

4. <https://docs.conda.io/en/latest/miniconda.html>

5. A fecha de redacción, con la versión Python 3.9.

1. Abrir Anaconda Prompt (Miniconda 3).
2. Introducir los siguientes códigos:

```
pip install instaloader6
```

```
pip install instalooter
```

Con esto, quedan instaladas las herramientas de scraping. Hay más información disponible en las páginas de las herramientas:

<https://instaloader.github.io/installation.html>

<https://instalooter.readthedocs.io/en/stable/install.html>

### **Paso 1. Construcción de una carpeta para trabajar**

Una vez instaladas las herramientas de trabajo, el primer paso consiste, sencillamente, en crear una carpeta en el directorio deseado. En ella se guardarán todos los datos obtenidos. En nuestro ejemplo:

```
C:/Users/Test/Imágenes/SacaPecho
```

### **Paso 2. Diseño de la búsqueda**

Utilizando el proceso de diseño de la investigación detallado por van Schie y sus colegas (2017), proponemos diseñar una búsqueda de etiquetas empleando las herramientas de *scraping* instaladas en tres pasos:

- 2.1. Definir la etiqueta a estudiar: nos interesan las imágenes acompañadas de #SacaPecho por el reconocimiento que tiene la Asociación Española Contra el Cáncer y el alto volumen de imágenes relacionadas con el cáncer de mama observado en fases exploratorias.
- 2.2. Definir las fechas de búsqueda: el 19 de octubre de 2020, Día internacional de la lucha contra el cáncer de mama, clave en la campaña #SacaPecho.
- 2.3. Obtener una idea de cuántas publicaciones llevan dicha etiqueta. Acceder a la página de Explore para #SacaPecho seleccionado para entender el volumen general de imágenes: <https://www.instagram.com/explore/tags/sacapecho/>. 12.792 posts al 30 de octubre de 2020.

---

**6.** Instaloader a menudo resulta en errores si se utiliza en un navegador desde el que hemos accedido recientemente a Instagram. La mayoría de errores se resuelven actualizando Instaloader a su última versión (`pip install --upgrade`), actualizando Python (`conda update python`), evitando las descargas masivas y, simplemente, esperando unas horas en caso de bloqueo.

Hasta aquí llega la información que proporciona Instagram en su motor de búsqueda. Para cualquier paso adicional necesitaremos de una herramienta de *scraping* como Instaloader o Instalooter.

### Paso 3. Lanzar Jupyter Notebook en la carpeta deseada

Con los parámetros de búsqueda definidos, en este tercer paso podremos obtener una lista con los metadatos de todas las publicaciones (o *posts*) hechas entre dos fechas determinadas y que lleven una etiqueta específica. Para ello, utilizaremos el editor de texto Jupyter Lab. Lanzamos Jupyter-Lab a través de Anaconda Prompt:

```
[Abrir Anaconda Prompt (Miniconda)]
```

```
cd c:/Users/Test/Images/Instaloader
```

```
jupyter-lab
```

Al introducir ambos comandos, se abrirá el navegador de Internet por defecto con el editor Jupyter Lab. Una vez dentro, seleccionamos *Notebook > Python* y estamos listos para programar la búsqueda.

### Paso 4. Descarga de metadatos

Para obtener los metadatos de todas las imágenes con la etiqueta #SacaPecho subidas el 19 de octubre de 2020, se desarrolló un código utilizando ambas herramientas de *scraping*. El código completo está disponible bajo solicitud a los autores, y está compuesto de dos partes: una primera parte realiza la búsqueda de metadatos y los registra. La segunda parte descarga una muestra de imágenes. Así, el código permite definir parámetros como la etiqueta de búsqueda, el rango de fechas y un archivo en el que guardar los datos obtenidos. Asimismo, se han incorporado ya todos los paquetes de Python necesarios. Es importante señalar que el código desarrollado solamente descarga datos de perfiles públicos<sup>7</sup>.

La primera celda del código ejecuta la búsqueda y el registro de metadatos, según los parámetros introducidos por el usuario. Resulta tentador extraer la mayor cantidad de datos posible, pero puede resultar en el bloqueo de la aplicación. Así, es recomendable hacer una primera búsqueda limitada a metadatos básicos como la ID de la imagen (su número identificador) y la fecha de subida. El usuario puede editar los parámetros de forma sencilla cambiando las siguientes líneas de código:

---

7. La descarga de perfiles privados es posible si el usuario accede a su propia cuenta de Instagram a través de Python, algo que no se explora en este artículo.

```
#Definimos el hashtag (desired_tag)
```

```
desired_tag="sacapecto [u otra etiqueta]"
```

```
#Definimos el tamaño de muestra (sample_size). Se utilizará para descargar una muestra de imágenes más adelante. En ningún caso deben descargarse más de 100 en una prueba, para evitar sobrecargar el sistema o ser bloqueado por Instagram.
```

```
sample_size=50[u otro número]
```

```
#Definimos las fechas (DESDE y HASTA).
```

```
DESDE = datetime(2020, 10, 19) # Más lejano a hoy, inclusivo
```

```
HASTA = datetime(2020, 10, 20) # Más cercano a hoy, no inclusivo
```

```
#Definimos la carpeta de destino.
```

```
destination="[DIRECTORIO COMPLETO DE LA CARPETA DE TRABAJO]"
```

```
#Definimos el archivo donde se guardarán los resultados con la extensión .txt. Importante no olvidar cambiarlo si se va a lanzar otro trabajo, para evitar sobrescribir los resultados.
```

```
file_to_record="[NOMBRE DESEADO PARA EL ARCHIVO].txt"
```

```
También es posible limitar la búsqueda a videos, historias, imágenes o los llamados reels (vídeos de formato corto). Asimismo, es posible modificar los metadatos que deseamos descargar:
```

```
else: #los datos a solicitar se indican aquí debajo
```

```
print(post.shortcode,",",post.date,",",post.profile,",",post.likes)8
```

## Paso 5. Descarga de imágenes

Una vez obtenidos los metadatos, podemos descargar una muestra de imágenes para comprobar que se están recogiendo los resultados deseados. Este quinto paso está contenido en la segunda celda de código que acompaña

---

<sup>8</sup>. Estos parámetros están recogidos en la documentación de *Instaloader*, pero los más comunes para una búsqueda sencilla serán `post.shortcode` (un número identificador de la imagen) y `post.date` (fecha de subida). Cualquier otro parámetro ralentizará la búsqueda y deberá ser editado para evitar exponer detalles de las cuentas estudiadas.

al artículo, y utiliza la herramienta Instaloooter, que permite fijar un número máximo de descargas entre dos fechas. Sugerimos un máximo de 50, para evitar sobrecargar el sistema.

Desde Jupyter, podemos lanzar ambos códigos en una sola sesión y empezar a obtener resultados rápidamente. El código completo se encuentra documentado y contiene parámetros sencillos que cambiar para modificar la búsqueda.

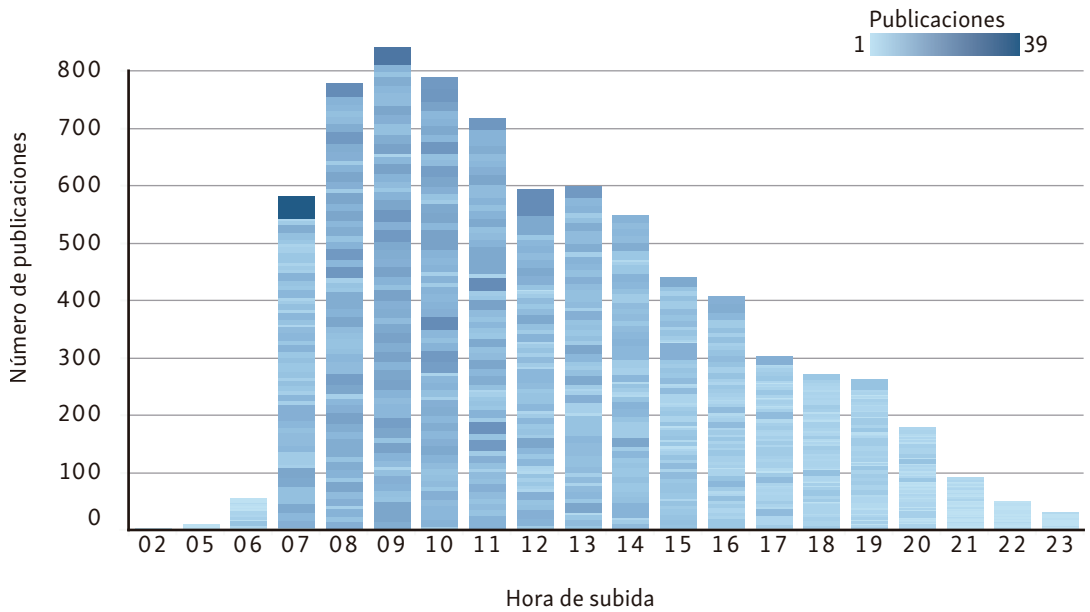
## RESULTADOS

La implementación del código desarrollado ha permitido obtener los datos de 7556 imágenes subidas a Instagram el 19 de octubre de 2020 con la etiqueta #SacaPecho. Una primera búsqueda, limitada a la ID de la imagen y la fecha y hora de subida, tardó 12 minutos en producir un archivo de texto con la información deseada. Una búsqueda más extensa, incluyendo el número de Me gusta y el perfil que la publicó, tardó más de cinco horas. Asumimos que esta discrepancia se debe a que Instaloader espacia las solicitudes de este tipo de datos para evitar que Instagram bloquee su acceso.

De la representación visual de los datos obtenidos podemos extraer la distribución horaria de las publicaciones.

El gráfico 1 muestra el número de publicaciones en franjas de una hora, detallando el minuto de subida. Las publicaciones con la etiqueta #SacaPecho se mantienen a lo largo del día, con su mayor pico entre las 09:00 y las 10:00. Se observa una mayoría de publicaciones a las horas en punto (07:00, 08:00, 09:00, etc.), con el mayor número de publicaciones simultáneas a las 07:00, lo que sugiere que un volumen importante pueden ser publicaciones programadas para subirse a horas específicas, ya sea por organizaciones o particulares. No en vano se observan aquí publicaciones de los distintos perfiles (nacional y provinciales) de la propia Asociación Española Contra el Cáncer, posiblemente programados como parte de la campaña.

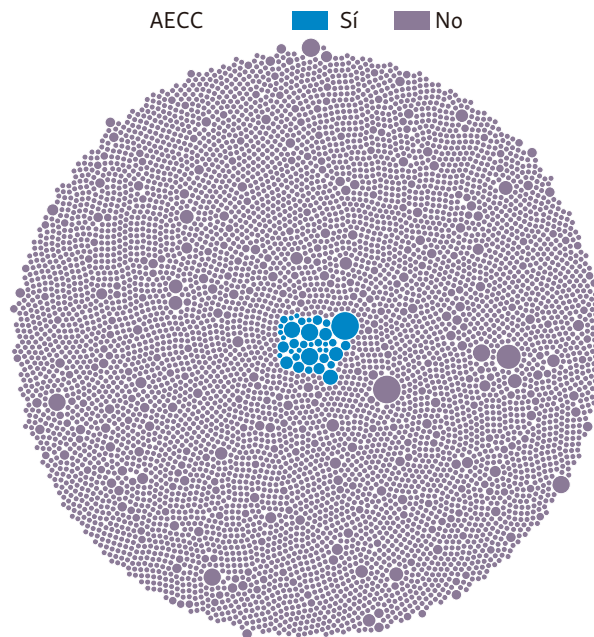
Utilizando los metadatos adicionales, se obtuvo el nombre de la cuenta que subió cada imagen, con un total de 6840 cuentas diferentes. Resulta relevante observar la distribución de publicaciones por cuenta (figura 1), contrastando el volumen de publicaciones de cuentas asociadas a la AECC (que contienen AECC en su nombre) y por cuentas externas. Esto sugiere una gran capacidad de extensión de la campaña.



Las subdivisiones indican la distribución por minuto de subida. Un color más intenso y un tamaño mayor indica un mayor número de imágenes subidas en el minuto indicado, comenzando por el minuto 00 arriba y terminando por el minuto 59 junto al eje X.

**Gráfico 1. Hora y minuto de subida de publicaciones a Instagram con la etiqueta #SacaPecho el 19 de octubre de 2020**

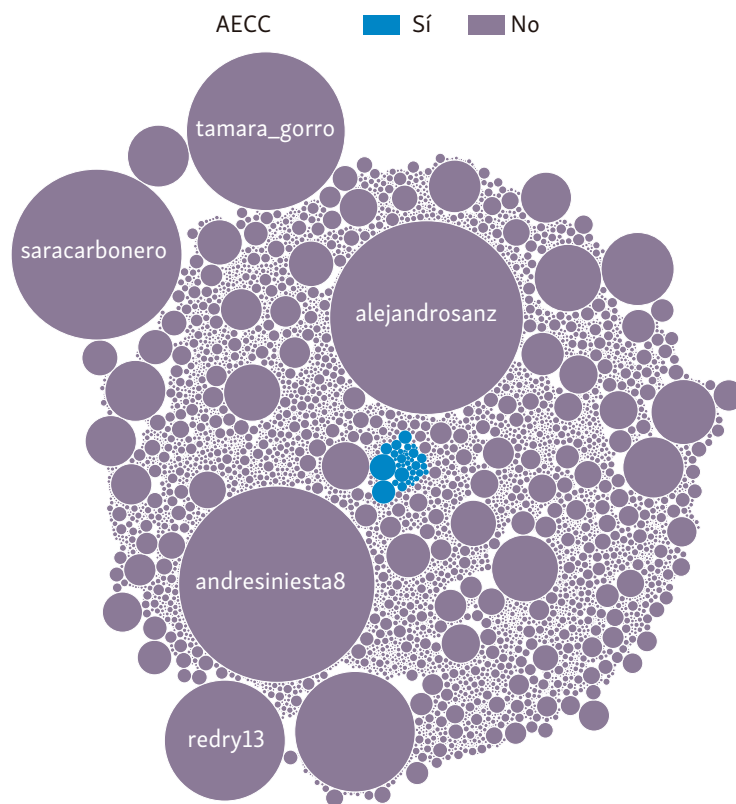
*Fuente: Elaboración propia con base en los datos obtenidos de Instagram.*



Se muestran agrupados en un color distinto los perfiles que contienen la palabra AECC en su nombre. Una burbuja de mayor tamaño indica un mayor número de publicaciones.

**Figura 1. Distribución por número de publicaciones en Instagram el 19 de octubre de 2020 con la etiqueta #SacaPecho, agrupados por perfil**

*Fuente: Elaboración propia.*



AECC se refiere a perfiles que contienen esas siglas en su nombre.

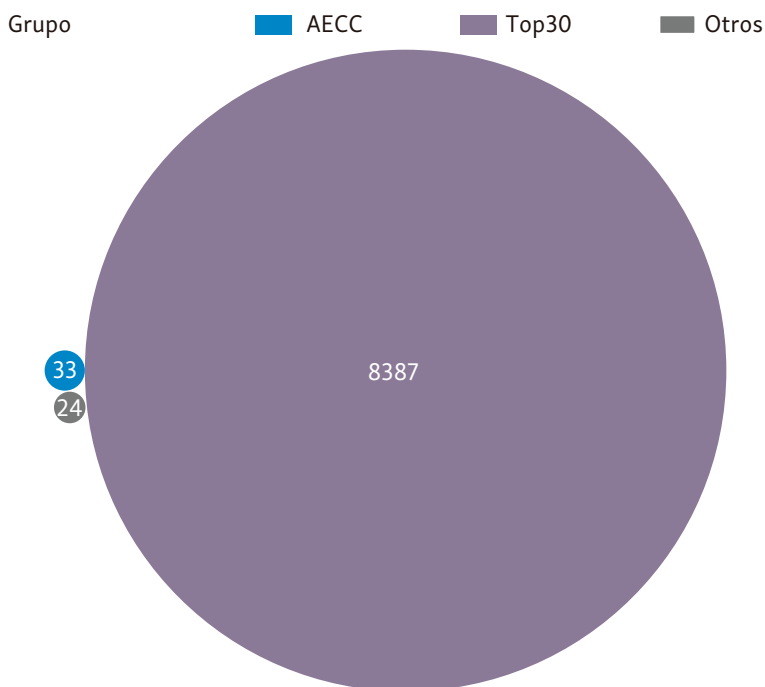
**Figura 2. Número de me gusta recibidos por las publicaciones en Instagram con el hashtag #SacaPecho hechas el 19 de octubre de 2020, agrupados por perfil**

*Fuente: Elaboración propia.*

También es destacable el número de Me gusta que acumulan las imágenes dependiendo del momento de publicación y, especialmente, según la cuenta origen de la publicación (figura 2).

En este sentido, es innegable la presencia de lo que podría clasificarse como *superinfluencers*: cuentas que alcanzan un nivel de visibilidad y Me gusta extremadamente alto, especialmente comparado con los que obtienen las publicaciones de usuarios medio y de la propia AECC. Con estos datos, se obtuvo un listado del top 30 de cuentas según la suma de Me gusta y se comparó con los otros dos grupos (figura 3).

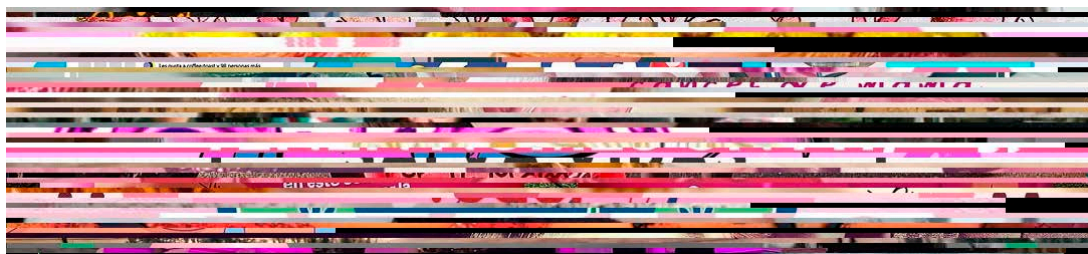
Salvo algunas excepciones institucionales como equipos de fútbol, la mayoría de estos *superinfluencers* son cuentas de personas famosas o influyentes: cantantes, escritores, deportistas, o actores. Algunos de ellos han sufrido algún tipo de cáncer y representan esa personalización de la imagen de la enfermedad explorada anteriormente. En conjunto, multiplican el alcance de las cuentas asociadas con la AECC por un factor de 75.



El top 30 se refiere a los 30 perfiles que más Me gusta recibieron, mientras que AECC se refiere a perfiles que contienen esas siglas en su nombre.

**Figura 3. Mediana del número de me gusta recibidos en las publicaciones con la etiqueta #SacaPecho hechas el 19 de octubre de 2020, agrupados en tres perfiles**

*Fuente: Elaboración propia.*



**Figura 4. Composición de 50 imágenes de muestra publicadas en Instagram con el hashtag #SacaPecho el 19 de octubre de 2020**

*Fuente: Elaboración propia.*

Utilizando la segunda parte del código desarrollado, se obtuvo una muestra de publicaciones que permite dibujar una imagen general de su contenido. Para ilustrar este paso, se descargó una muestra de 50 imágenes de la lista original, seleccionadas aleatoriamente. Todas ellas incluyen el rosa como color dominante, construyendo una distribución del color acorde con las impresiones de Ehrenreich (2001) ya señaladas. La figura 4 muestra la distribución de colores, de manera que se puedan apreciar las características generales pero que no sea posible identificar las imágenes individuales.



Un análisis de contenido revela que las publicaciones de muestra priorizan los contenidos tradicionalmente relacionados con el cáncer de mama en la cultura popular, destacando las imágenes en situaciones cotidianas: lazos, pañuelos y mascarillas de color rosa; diagramas sobre las posibles manifestaciones del cáncer al tacto en una autoexploración (para detectar posibles señales de alerta); la promoción del deporte y la actividad física en general, o mujeres en pose de lucha (con el brazo flexionado en señal de fuerza) son algunos de los elementos más repetidos. Cabe mencionar que el análisis de contenido del material fotográfico puede ser objeto de aplicación de otros métodos computacionales, capaces de procesar la información de esos archivos de modo automático para profundizar en el sentido colectivo del flujo masivo de circulación de mensajes visuales.

### **DISCUSIÓN: LAS ALTERNATIVAS AL SCRAPING**

Más allá de las limitaciones técnicas, el uso de estas herramientas plantea dificultades éticas significativas. Existen alternativas al *scraping*, sin duda. La primera es, como la describe Bruns (2019), retirarse. Aceptar que el acceso a imágenes encontradas en Instagram es hostil y difícil, y dedicarse a redes con acceso más sencillo (Flickr, Twitter, o Reddit). Esto negaría la evidencia de que Instagram es, hoy en día, la mayor red de imágenes y con una gran influencia en dinámicas sociales, de marketing o, en nuestro caso, en la comunicación y la educación en salud.

La segunda es el trabajo manual. Muchos de los trabajos en investigación social centrados en lo representado siguen este método. Realizar búsquedas manuales y extraer capturas de pantalla es un mecanismo utilizado por Abidi (2016), Cinelli et al. (2020) o Couture Bue (2020). Entraña un problema: entre las múltiples variables de clasificación de Instagram (como la ubicación, los hashtags, el nombre de usuario o la fecha y la hora), solamente los hashtags pueden explorarse con cierta facilidad. Desde la página de búsqueda Explore de Instagram, los investigadores pueden navegar por las distintas etiquetas asociadas a las imágenes. Por ejemplo, buscando la etiqueta *#cancer* pueden verse todas las imágenes que llevan dicha etiqueta (en la fecha de redacción de este artículo, más de 17 millones). Esto, además, presenta dos limitaciones: la primera es que Instagram prioriza el contenido destacado, es decir el contenido que más atención ha recibido en la red social en los últimos días, basándose en una serie de algoritmos de difícil interpretación. La segunda es que las imágenes se muestran en una página infinita, sin categorización alguna y sin acompañamiento de metadatos, además de en miniatura, por lo que el investigador tendrá que ir imagen por imagen haciendo recortes. En este caso, la recogida de datos en volumen puede ayudar a eliminar el sesgo de los algoritmos de Instagram, pero nos devuelve a un debate técnico-ético.

Como respuesta, algunos investigadores, como Giraldo-Luque y sus colegas (2018) abogan por utilizar herramientas de *screen scraping*, tales como las extensiones para navegadores como Google Chrome. Con esto se perdería cualquier información adicional sobre la imagen, y su legalidad sigue sin estar clara.

Una tercera alternativa es el trabajo a través de la API oficial de Instagram. Las API (*Application Programming Interface* o Aplicación de Interfaz de Programación) son aplicaciones de programación provistas por las plataformas, interfaces oficiales mediante las cuales los investigadores pueden solicitar formalmente a la plataforma obtener los datos necesarios. Highfield y Leaver (2014) proveen una metodología detallada para este proceso, aunque redactada en 2014, antes de que Instagram cerrara su API. Desde entonces, el proceso para obtener permiso para utilizar la API de Instagram es confuso. Rogers (2018) y Bruns (2019) señalan retrasos constantes en el procesamiento de las solicitudes, cambios en los formatos de las API, y concursos públicos para acceder a ellas atados a fechas difíciles de alcanzar. De hecho, Rogers (2018) señala que los requisitos para trabajar con API oficiales como la de Facebook son poco razonables, con un alto coste de transacción, y apunta a un retorno al *web scraping*.

No obstante, el trabajo de Zarei y sus colegas (2020) es un ejemplo prometedor de las posibilidades de establecer una asociación con Instagram para obtener datos a través de su API oficial. Desarrollan una base de (meta)datos obtenidos en Instagram y relacionados con la COVID19. Aunque no detallan cómo lo lograron, es razonable asumir que obtuvieron permiso explícito de Facebook (como propietaria de Instagram) para implementar su búsqueda y su herramienta de *crawling* (un método similar al *scraping*) gracias a la relevancia social de su tema de investigación y el empuje por facilitar la investigación sobre la pandemia. En principio, podría abrir la puerta a otros posibles trabajos.

La cuarta opción es esperar que el uso de estas técnicas pase inadvertido. Bruns (2019) y Rogers (2018) debaten esta posibilidad, no sin subrayar las dificultades que supone. La hemos visto utilizada en Himawan y sus colegas (2020), Dewi y colegas (2019), Nobles y colegas (2020), o Harrigan y colegas (2020). Sin embargo, como el propio Bruns indica, cuantos más investigadores utilicen estas herramientas, mayor será su visibilidad y, por consiguiente, más alta la probabilidad de que Facebook las detecte y bloquee.

Esto nos lleva a lo que Hirschev llama una aceptación pragmática del *scraping*. Considera que los proveedores deben reconocer el beneficio de permitirlo desde un punto de vista económico, y solo tomar acción legal si “(1) el *scraper* representa una amenaza para el negocio del proveedor o (2) el proveedor tiene un caso lo suficientemente fuerte para ganar un posible litigio” (2014, p. 901).

Hirschey hacía estas recomendaciones en 2014, consciente de que el entorno legal no era enteramente favorable a los proveedores, pues no existía una regulación específica para la práctica del *scraping*. Naturalmente, el autor no hace referencia al uso de datos para avanzar en la investigación y el desarrollo social, pues su argumento es puramente económico. Eso nos deja a nosotros, los investigadores sociales, en la tesitura de vernos, aún más, sujetos a la lógica económica en la investigación si queremos colaborar con proveedores como Facebook.

### **Limitaciones y conclusiones del estudio**

El *scraping* es un método con ventajas y desventajas y, en general, un ámbito difícil para la investigación social. De forma muy sencilla, con varias líneas de código y utilizando herramientas de código abierto y completamente gratuitas, hemos podido obtener información clave para entender la distribución del cáncer en redes sociales. Si bien es cierto que la investigación en Instagram no está aún al nivel de trabajos en redes como Twitter o Reddit (Lama et al., 2019; Record et al., 2018; Schrading et al., 2015) –fundamentalmente, debido a la opacidad corporativa de Facebook–, queda patente que existen herramientas a disposición de los investigadores para continuar aprendiendo sobre esta red.

Desde un punto de vista técnico y ético, Instaloader e Instalooter son herramientas con limitaciones severas. Aunque facilita la obtención de datos, el *scraping* como tal no da una respuesta fácil al complejo proceso de determinar cuáles son las etiquetas que debemos explorar ni a cómo maximizar la representatividad de los datos. Esto se resuelve con un proceso de diseño como el que describen Schie y sus colegas (2017), con lo que la técnica es aquí un elemento de apoyo más que de respuesta.

Asimismo, el diseño de estas herramientas dificulta las búsquedas complejas, donde podamos cruzar variables como distintos hashtags o distintos rangos de fecha. Tampoco es posible hacer una búsqueda por ubicación, pues Instagram no provee esos datos. Es un método dado a la frustración, ya que los bloqueos por parte de Instagram son comunes, como lo es también la ralentización intencionada de los resultados.

Por último, aunque están bien documentadas, las herramientas requieren cierto nivel de conocimiento de programación en el lenguaje Python, y están expuestas a un gran riesgo de desaparecer, a la vista de los cambios y las limitaciones impuestas por Facebook en los últimos años. Herramientas similares, como Instagram PHP Scraper, Django-Instagram o Instagram-Scraper enfrentan retos semejantes. Opciones comerciales, como Meta Eyes o Vurku no han sido estudiadas en un ámbito académico.

El empleo de estos recursos permite la escalabilidad del análisis, ya que el volumen de información susceptible de ser procesado alcanza dimensiones amplias. Sin embargo, el proceso de *scraping* debe buscar conexiones con otros métodos computacionales, que profundicen en el análisis automatizado de las imágenes obtenidas. La producción y circulación masiva de imágenes en el ecosistema digital provoca que los equipos de investigación deban capacitarse en la combinación de diversos métodos computacionales que permitan no solamente recopilar y almacenar, en este caso, los archivos fotográficos, sino que avancen en la codificación, el análisis y el procesamiento estadísticos de los aspectos más objetivables de la comunicación visual. El diálogo entre las posibilidades que ofrece la ciencia de la computación y el largo recorrido completado por la sociología visual o los estudios en comunicación visual durante las últimas décadas se presenta como una vía fructífera para la investigación durante la próxima década.

Nuestros resultados parecen indicar que el uso de *scrapers* puede ayudar a los investigadores sociales a mantener el pulso de las redes, acelerando su trabajo. La variedad de usuarios que ha utilizado la etiqueta #SacaPecho y, especialmente, el carácter individual de los *superinfluencers*, sugiere que la dinámica de personalización y apropiación del discurso de la enfermedad en redes sociales –en oposición, tal y como señala Pardo, a un discurso exclusivamente médico-profesional– está muy presente en Instagram. Esto no hace sino subrayar la utilidad de esta red para entender el discurso social.

Obtener este tipo de datos a una velocidad mucho mayor de la que podríamos alcanzar manualmente permite hacer una clasificación del lugar de circulación y del lugar de la audiencia (Rose, 2019) de las imágenes relacionadas con el cáncer. En el caso de #SacaPecho, la información gana en relevancia al tratarse de una campaña claramente delimitada, con una etiqueta específica y una fecha de búsqueda. No obstante, también es posible obtener resultados similares para etiquetas que abarcan más, como #CáncerDeMama, con rangos temporales más amplios. A partir de ahí, podríamos hacer una distribución por tipos de cáncer, visualizar los resultados y buscar correlaciones entre su visibilidad, sus características visuales y sus datos epidemiológicos.

Las avenidas que se abren para la investigación empírica son tan abundantes como atractivas en el análisis de la imagen en redes sociales digitales, pero los dilemas éticos y las dificultades técnicas coexistirán en tanto que la dimensión económica de las plataformas que las albergan no encuentre una relación mutuamente beneficiosa con la investigación académica y su naturaleza necesariamente crítica.

## REFERENCIAS

- Abidin, C. (2016). “Aren’t These Just Young, Rich Women Doing Vain Things Online?”: Influencer Selfies as Subversive Frivolity. *Social Media + Society*, 2(2), 2056305116641342. <https://doi.org/10.1177/2056305116641342>
- Abril Curto, G. (2013). *Cultura visual, de la semiótica a la política* (Visual culture, from semiotics to politics). Plaza y Valdés.
- Abutaleb, Y. (2016, June 21). Instagram’s user base grows to more than 500 million. *Reuters*. Retrieved from <https://www.reuters.com/article/us-facebook-instagram-users-idUSKCNOZ71LN>
- Arcila-Calderón, C., Ortega-Mohedano, F., Álvarez, M., & Vicente-Mariño, M. (2019). Análisis distribuido y supervisado de sentimientos en Twitter: Integrando aprendizaje automático y analítica en tiempo real para retos de dimensión big data en investigación de comunicación y audiencias (Distributed Supervised Sentiment Analysis of Tweets: Integrating Machine Learning and Streaming Analytics for Big Data Challenges in Communication and Audience Research). *Empiria. Revista de metodología de ciencias sociales*, (42), 113-136. <https://doi.org/10.5944/empiria.42.2019.23254>
- Asociación Española Contra el Cáncer (AECC). (2020, October 9). La AECC anima a “sacar pecho” contra el cáncer de mama. *AECC Noticias*. Retrieved from <https://www.aecc.es/es/actualidad/noticias/aecc-anima-sacar-pecho-contr-a-cancer-mama>
- Babvey, P., Capela, F., Cappa, C., Lipizzi, C., Petrowski, N., & Ramirez-Marquez, J. (2020). Using social media data for assessing children’s exposure to violence during the COVID-19 pandemic. *Child Abuse & Neglect*, 104747. <https://doi.org/10.1016/j.chiabu.2020.104747>
- Basch, C. H. & MacLean, S. A. (2019). Breast Cancer on Instagram: A Descriptive Study. *International Journal of Preventive Medicine*, 10(1), 166. [https://doi.org/10.4103/ijpvm.IJPVM\\_36\\_19](https://doi.org/10.4103/ijpvm.IJPVM_36_19)
- Becker, H. S. (1974). Photography and Sociology. *Studies in the Anthropology of Visual Communication*, 1(1), 3-26. Retrieved from <https://repository.upenn.edu/cgi/viewcontent.cgi?article=1003&context=svc>
- Berger, J. (2016). *Modos de ver: Basado en la serie de la televisión BBC en colaboración con John Berger* (Ways of Seeing: Based on the BBC Television Series). Barcelona, Spain: Gustavo Gili.
- Blanch-Hartigan, D. & Viswanath, K. (2015). Socioeconomic and sociodemographic predictors of cancer-related information sources used by cancer survivors. *Journal of Health Communication*, 20(2), 204-210. <https://doi.org/10.1080/10810730.2014.921742>
- Bruns, A. (2019). After the ‘APICalypse’: Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544-1566. <https://doi.org/10.1080/1369118X.2019.1637447>
- Cartwright, L. (1998). Community and the Public Body in Breast Cancer Media Activism. *Cultural Studies*, 12(2), 117-138. <https://doi.org/10.1080/095023898335500>
- Chou, W.-Y. S., Trivedi, N., Peterson, E., Gaysynsky, A., Krakow, M., & Vraga, E. (2020). How do social media users process cancer prevention messages on Facebook? An eye-tracking study. *Patient Education and Counseling*, 103(6), 1161-1167. <https://doi.org/10.1016/j.pec.2020.01.013>
- Cinelli, M., Quattrocioni, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 Social Media Infodemic. *Sci Rep*, (10), 16598. <https://doi.org/10.1038/s41598-020-73510-5>

- Couture Bue, A. C. (2020). The looking glass selfie: Instagram use frequency predicts visual attention to high-anxiety body regions in young women. *Computers in Human Behavior*, 108, 106329. <https://doi.org/10.1016/j.chb.2020.106329>
- Dewi, L.C., Meiliana, & Chandra, A. (2019). Social Media Web Scraping using Social Media Developers API and Regex. *Procedia Computer Science*, 157, 444-449. <https://doi.org/10.1016/j.procs.2019.08.237>
- Döbrössy, B., Girasek, E., Susánszky, A., Koncz, Z., Gyórfy, Z., & Bognár, V. K. (2020). «Clicks, likes, shares and comments» a systematic review of breast cancer screening discourse in social media. *PLOS ONE*, 15(4), e0231422. <https://doi.org/10.1371/journal.pone.0231422>
- Ehrenreich, B. (2001, November). Welcome to Cancerland: A Mammogram Leads to a Cult of Pink Kitsch. *Harper's Magazine*. Retrieved from <https://harpers.org/archive/2001/11/welcome-to-cancerland/>
- Gibson, F., Hibbins, S., Grew, T., Morgan, S., Pearce, S., Stark, D., & Fern, L. A. (2016). How young people describe the impact of living with and beyond a cancer diagnosis: Feasibility of using social media as a research method. *Psycho-Oncology*, 25(11), 1317-1323. <https://doi.org/10.1002/pon.4061>
- Giraldo-Luque, S., Fernández-García, N., & Pérez-Arce, J. C. (2018). La centralidad temática de la movilización #NiUnaMenos en Twitter (Thematic centrality on Twitter: The case of # NiUnaMenos mobilisation). *El profesional de la información*, 27(1), 96-105. Retrieved from <http://profesionaldelainformacion.com/contenidos/2018/ene/09.pdf>
- González Requena, J. (1988). *El discurso televisivo: Espectáculo de la posmodernidad* (The television discourse: Spectacle of postmodernity). Madrid, Spain: Cátedra.
- González Requena, J. (2013). El ser de las imágenes: De la teoría al análisis de imagen (The Being of Images: From Theory to Image Analysis). [www.gonzalezrequena.com](http://gonzalezrequena.com). Retrieved from <http://gonzalezrequena.com/textos-en-linea-0-2/libros-en-linea/el-ser-de-las-imagenes/>
- Grant, J. A. & Hundley, H. (2008). Fighting the Battle or Running the Race? *Visual Communication Quarterly*, 15(3), 180-195. <https://doi.org/10.1080/15551390802235578>
- Hand, M. (2016). Visuality in Social Media: Researching Images, Circulations and Practices. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE Handbook of Social Media Research Methods* (pp. 215-231). SAGE Publications Ltd.
- Harper, D. (1988). Visual Sociology: Expanding Sociological Vision. *The American Sociologist*, (19), 54-70. <https://doi.org/10.1007/BF02692374>
- Harper, D. (1996). Seeing Sociology. *The American Sociologist*, 27(3), 69-78. Retrieved from <https://www.jstor.org/stable/27698785>
- Harrigan, P., Daly, T. M., Coussement, K., Lee, J. A., Soutar, G. N., & Evers, U. (2020). Identifying influencers on social media. *International Journal of Information Management*, (56), 102246. <https://doi.org/10.1016/j.ijinfomgt.2020.102246>
- Henricks, S. C. (2017). Social Media, Publicly Available Information, and the Intelligence Community. *American Intelligence Journal*, 34(1), 21-31. Retrieved from <https://www.jstor.org/stable/26497113>

- Highfield, T. & Leaver, T. (2014). A methodology for mapping Instagram hashtags. *First Monday*, 20(1). <https://doi.org/10.5210/fm.v20i1.5563>
- Himawan, A., Priadana, A., & Murdiyanto, A. (2020). Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application. *IJID (International Journal on Informatics for Development)*, 9(2), 59-65. <https://doi.org/10.14421/ijid.2020.09201>
- Hirschev, J. (2014). Symbiotic Relationships: Pragmatic Acceptance of Data Scraping. *SSRN Electronic Journal*, (29). <https://doi.org/10.2139/ssrn.2419167>
- Instagram. (2018). *Condiciones de uso* (Terms of use). Retrieved from <https://help.instagram.com/581066165581870?ref=dp>
- Jang, J. Y., Han, K., Shih, P. C., & Lee, D. (2015). Generation Like: Comparative Characteristics in Instagram. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4039-4042. <https://doi.org/10.1145/2702123.2702555>
- Kligler-Vilenchik, N., Baden, C., & Yarchi, M. (2020). Interpretative polarization across platforms: How political disagreement develops overtime on Facebook, Twitter, and WhatsApp. *Social Media + Society*, 6(3), 2056305120944393. <https://doi.org/10.1177/2056305120944393>
- Lama, Y., Hu, D., Jamison, A., Quinn, S. C., & Broniatowski, D. A. (2019). Characterizing Trends in Human Papillomavirus Vaccine Discourse on Reddit (2007-2015): An Observational Study. *JMIR Public Health and Surveillance*, 5(1), e12480. <https://doi.org/10.2196/12480>
- Manovich, L. (2016). Instagram and Contemporary Image. *manovich*. Retrieved from [http://manovich.net/content/04-projects/151-instagram-and-contemporary-image/instagram\\_book\\_manovich\\_2017.pdf](http://manovich.net/content/04-projects/151-instagram-and-contemporary-image/instagram_book_manovich_2017.pdf)
- Manovich, L. (2018). *AI Aesthetics*. Moscow, Russia: Strelka Press.
- Markham, A. N. & Buchanan, E. (2012). Ethical Decision-Making and Internet Research: Recommendations from the AOIR Ethics Working Committee (Version 2.0). Association of Internet Researchers. Retrieved from <https://aoir.org/ethics/>
- Markham, A. N. & Buchanan, E. (2017). Research Ethics in Context: Decision-Making in Digital Research. In M. T. Schäfer & K. Es, van (Eds.), *The Datafied Society. Studying Culture through Data*. Amsterdam, The Netherlands: Amsterdam University Press.
- Marres, N. & Weltevrede, E. (2013). SCRAPING THE SOCIAL?: Issues in live social research. *Journal of Cultural Economy*, 6(3), 313-335. <https://doi.org/10.1080/17530350.2013.772070>
- Munk, A. K., Abildgaard, M. S., Birkbak, A., & Petersen, M. K. (2016). (Re-)Appropriating Instagram for Social Research: Three Methods for Studying Obesogenic Environments. *Proceedings of the 7th 2016 International Conference on Social Media & Society*, 1-10. <https://doi.org/10.1145/2930971.2930991>
- Noar, S. M., Leas, E., Althouse, B. M., Dredze, M., Kelley, D., & Ayers, J. W. (2018). Can a selfie promote public engagement with skin cancer? *Preventive Medicine*, (111), 280-283. <https://doi.org/10.1016/j.ypmed.2017.10.038>
- Nobles, A. L., Leas, E. C., Noar, S., Dredze, M., Latkin, C. A., Strathdee, S. A., & Ayers, J. W. (2020). Automated image analysis of instagram posts: Implications for risk perception and communication in public health using a case study of #HIV. *PLOS ONE*, 15(5), e0231155. <https://doi.org/10.1371/journal.pone.0231155>

- Orduña-Malea, E., Font-Julián, C. I., & Ontalba-Ruipérez, J.-A. (2020). Covid-19: Análisis métrico de vídeos y canales de comunicación en YouTube (Covid-19: metric analysis of videos and communication channels on YouTube). *Profesional de la Información*, 29(4), e290401. <https://doi.org/10.3145/epi.2020.jul.01>
- Pardo, R. (2019). Fotografía y enfermedad: Iconografías en transformación (Photography and disease: Iconographies in transformation). In M. Morcate & R. Pardo (Eds.), *La imagen desvelada: Prácticas fotográficas en la enfermedad, la muerte y el duelo* (The Unveiled Image: Photographic Practices in Illness, Death, and Grief) (pp. 19-60). Vitoria-Gasteiz, Spain: Sans Soleil Ediciones.
- Pauwels, L. (2015). *Reframing Visual Social Science: Towards a More Visual Sociology and Anthropology*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139017633>
- Prosser, J., Clark, A., & Wiles, R. (2008). *Visual Research Ethics at the Crossroads*. NCRM Working Paper. Manchester, Reino Unido: Realities, Morgan Centre. Retrieved from <http://eprints.ncrm.ac.uk/535/>
- Purba, K. R., Asirvatham, D., & Murugesan, R. K. (2020). Influence maximization diffusion models based on engagement and activeness on instagram. *Journal of King Saud University - Computer and Information Sciences*. <https://doi.org/10.1016/j.jksuci.2020.09.012>
- Record, R. A., Silberman, W. R., Santiago, J. E., & Ham, T. (2018). I Sought It, I Reddit: Examining Health Information Engagement Behaviors among Reddit Users. *Journal of Health Communication*, 23(5), 470-476. <https://doi.org/10.1080/10810730.2018.1465493>
- Regueira, U., Ferreiro, A. A., & Vila, S. D. (2020). La mujer en YouTube: Representación y participación a través de la técnica Web Scraping (Women on YouTube: Representation and participation through the Web Scraping technique). *Comunicar: Revista científica iberoamericana de comunicación y educación*, 63, 31-40. <https://doi.org/10.3916/C63-2020-03>
- Rogers, R. (2018). *Social Media Research After the Fake News Debacle* (Dataset). University of Salento. <https://doi.org/10.1285/I20356609V11I2P557>
- Rose, G. (2019). *Metodologías visuales: Una introducción a la investigación con materiales visuales* (Visual Methodologies. An Introduction to Researching with Visual Materials). Murcia, Spain: CENDEAC.
- Schie, van, G., Westra, I., & Schäfer, M. T. (2017). Get Your Hands Dirty: Emerging Data Practices as Challenge for Research Integrity. In M. T. Schäfer & K. Es, van (Eds.), *The Datafied Society. Studying Culture through Data*, 183-200. Amsterdam University Press. <https://doi.org/10.5117/9789462981362>
- Schrading, N., Ovesdotter Alm, C., Ptucha, R., & Homan, C. (2015). An Analysis of Domestic Abuse Discourse on Reddit. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2577-2583. <https://doi.org/10.18653/v1/D15-1309>
- Sontag, S. (1978). *Illness As Metaphor*. Farrar, Straus and Giroux.
- Sontag, S. (2017). *Sobre la fotografía* (On photography). Debolsillo.
- Spence, J. (1986). *Putting myself in the picture: A political, personal, and photographic autobiography*. Camden, United Kingdom: Camden Press.



- Sutton, J., Vos, S. C., Olson, M. K., Woods, C., Cohen, E., Gibson, C. B., Phillips, N. E., Studts, J. L., Eberth, J. M., & Butts, C. T. (2018). Lung Cancer Messages on Twitter: Content Analysis and Evaluation. *Journal of the American College of Radiology*, 15(1), 210-217. <https://doi.org/10.1016/j.jacr.2017.09.043>
- Tafesse, W. & Wood, B. P. (2020). Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy. *Journal of Retailing and Consumer Services*, (58), 102303. <https://doi.org/10.1016/j.jretconser.2020.102303>
- Townsend, L., & Wallace, C. (2016). *Social Media Research: A Guide to Ethics*. The University of Aberdeen. [https://www.gla.ac.uk/media/Media\\_487729\\_smxx.pdf](https://www.gla.ac.uk/media/Media_487729_smxx.pdf)
- Walsh, M. & Baker, S. A. (2017). The selfie and the transformation of the public-private distinction. *Information Communication and Society*, (20), 1185-1203. <https://doi.org/10.1080/1369118X.2016.1220969>
- Wang, X., Chen, L., Shi, J., & Peng, T.-Q. (2019). What makes cancer information viral on social media? *Computers in Human Behavior*, (93), 149-156. <https://doi.org/10.1016/j.chb.2018.12.024>
- White, C. L. & Boatwright, B. (2020). Social media ethics in the data economy: Issues of social responsibility for using Facebook for public relations. *Public Relations Review*, 46(5), 101980. <https://doi.org/10.1016/j.pubrev.2020.101980>
- Zarei, K., Farahbakhsh, R., Crespi, N., & Tyson, G. (2020). A First Instagram Dataset on COVID-19. *arXiv:2004.12226 [cs]*. Retrieved from <http://arxiv.org/abs/2004.12226>
- Zhang, J., Le, G., Larochelle, D., Pasick, R., Sawaya, G. F., Sarkar, U., & Centola, D. (2019). Facts or stories? How to use social media for cervical cancer prevention: A multi-method study of the effects of sender type and content type on increased message sharing. *Preventive Medicine*, 126, 105751. <https://doi.org/10.1016/j.ypmed.2019.105751>

### **FINANCIAMIENTO**

Este artículo ha sido realizado en el marco del proyecto de investigación “Verdad y ética en las redes sociales. Percepciones e influencias educativas en jóvenes usuarios de Twitter, Instagram y YouTube (Internética)”, financiado por la convocatoria de proyectos I+D+i del Ministerio de Ciencia, Innovación y Universidades de España de 2019 (PID2019-104689RB-100).

### **SOBRE LOS AUTORES**

**MIGUEL VARELA**, es sociólogo y Profesor Asociado en el Departamento de Sociología y Trabajo Social de la Universidad de Valladolid. Su trabajo de investigación se centra en el análisis del discurso visual, con especial atención a las redes sociales y a la tecnología digital aplicada al trabajo de paz y a la sociología de la salud. Con formación en fotografía documental. Con un MA en Fotoperiodismo y Fotografía Documental por la Universidad de las Artes de Londres y un MSc en Trabajo de Paz y Desarrollo por la Universidad Lineo, es miembro de la International Visual Sociology Association.

 <https://orcid.org/0000-0002-6443-9667>

**MIGUEL VICENTE-MARIÑO**, es Profesor Titular de la Universidad de Valladolid, donde dirige el Departamento de Sociología y Trabajo Social y coordina el Grupo de Investigación Reconocido en Ciencias Sociales Aplicadas. Doctor en Comunicación Audiovisual por la Universitat Autònoma de Barcelona, sus líneas de investigación se centran en el análisis de las audiencias y de los contenidos de los medios de comunicación masiva y de las redes sociales digitales, con especial atención a las innovaciones metodológicas y a las herramientas informáticas.

 <https://orcid.org/0000-0002-1467-3535>