INFORME FINAL SOBRE EL SCRAPEO DE DATOS BRUTOS OBTENIDOS EN MEDIOS INFORMATIVOS DIGITALES ESPAÑOLES EN X (TWITTER), FACEBOOK Y PORTALES WEB

Proyecto Hatemedia (PID2020-114584GB-I00), MCIN/ AEI /10.13039/501100011033

Proyecto ejecutado por:

Universidad Internacional de La Rioja

Responsables del Proyecto:

Elias Said-Hung (IP), Julio Montero-Díaz (IP)

Equipo investigador y colaboradores, a cargo de la validación de la información expuesta en este informe:

Almudena Ruiz, Xiomara Blanco, Daniel Pérez, Oscar De Gregorio, Juan José Cubillas Mercado.

Informe elaborado con la participación de la empresa colaboradora:

possible

Índice

1. PROCESO DE SCRAPEADO 2

1.1. Facebook 2

1.2. X (Twitter) 7

1.3. Web 8

2. PROBLEMAS QUE SE HAN PRESENTADO DURANTE EL PROCESO DE RECOPILACIÓN DE DATOS 10

2.1. Facebook 10

2.2. X (Twitter) 10

2.3. Web 11

3. ACCIONES PARA RECUPERAR LOS DATOS MISSING 12

3.1. Facebook 12

3.2. X (Twitter) 13

3.3. Web 14

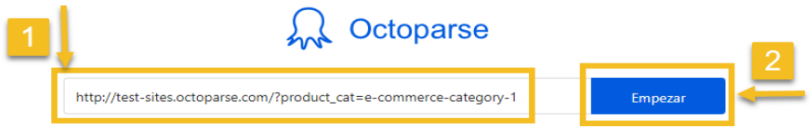
4. DEPURACIÓN Y RECUENTO DE DATOS 15

1. PROCESO DE SCRAPEADO

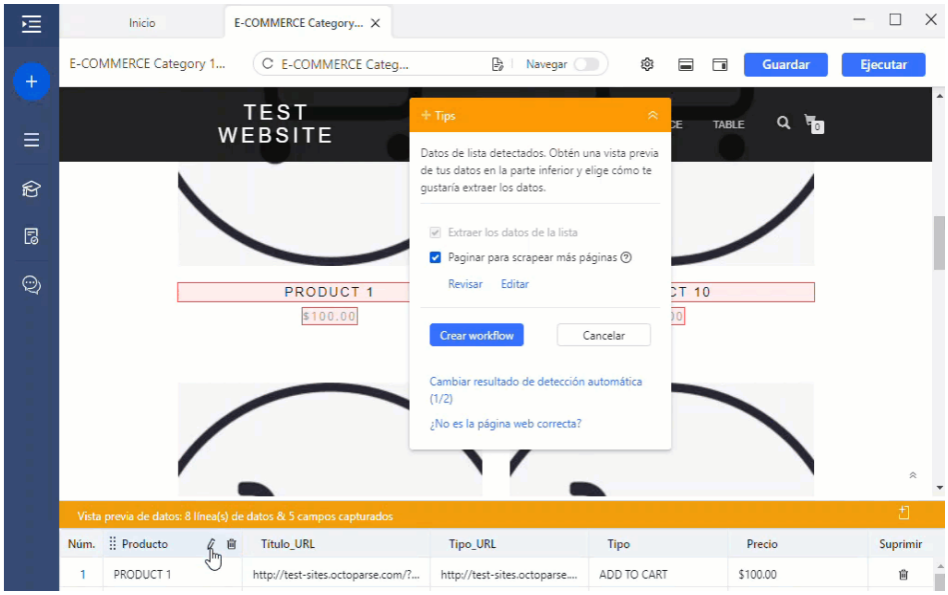
1.1. Facebook

La herramienta de scrapping utilizada para la plataforma Facebook fue “Octoparse”. Es un software de extracción de datos web visual que simula la operación humana para interactuar con páginas web. El flujo de trabajo funciona mediante la creación de tareas. A continuación, se muestra un ejemplo:

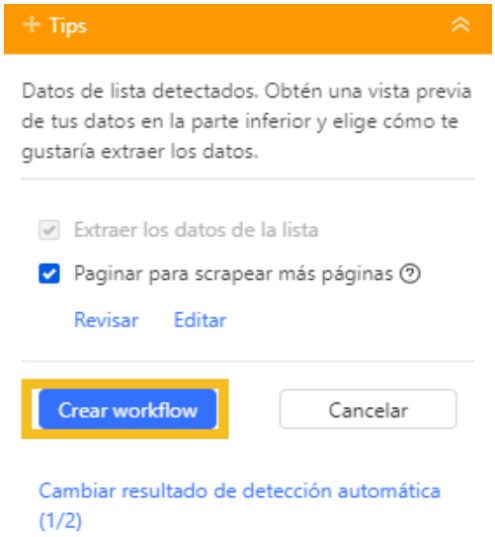
1. Se crea una tarea especificando la url en cuestión, para el caso de Facebook se accede a la página de cada periódico, por ejemplo: https://www.facebook.com/20minutos.es/?locale=es_ES



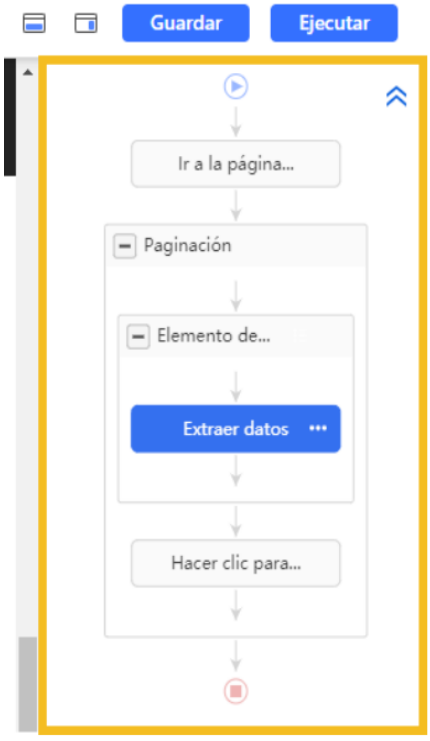
2. Octoparse cargará la información de la página automáticamente y detectará los campos automáticamente también. Hay que verificar dichos campos en la sección de vista previa y seleccionar los que se quieren guardar. Estos se resaltan en la página. También es posible cambiar el nombre de los campos seleccionados o eliminar aquellos que no sean necesarios.



3. Una vez configurados los campos, se crea un workflow (flujo de trabajo):



4. Octoparse generará un flujo de trabajo basado en los campos detectados y en la configuración que se haya aplicado.



5. Una vez creado, se puede editar para cambiar alguna configuración, o se puede ejecutar la tarea que empezará con el proceso de scrapeo. Cuando se inicie dicha tarea pedirá un inicio de sesión para el que se debe disponer de una o varias cuentas de facebook ya que es propenso a bloqueos.

Como se ha mencionado anteriormente dicha herramienta en su versión gratuita es muy propensa a bloqueos por lo que tras un tiempo de scrapeo con ella, se decidió desecharla.

Para evitar bloqueos y continuar con el proceso de scraping se decidió recurrir a "Facepager", una herramienta similar a Octoparse.

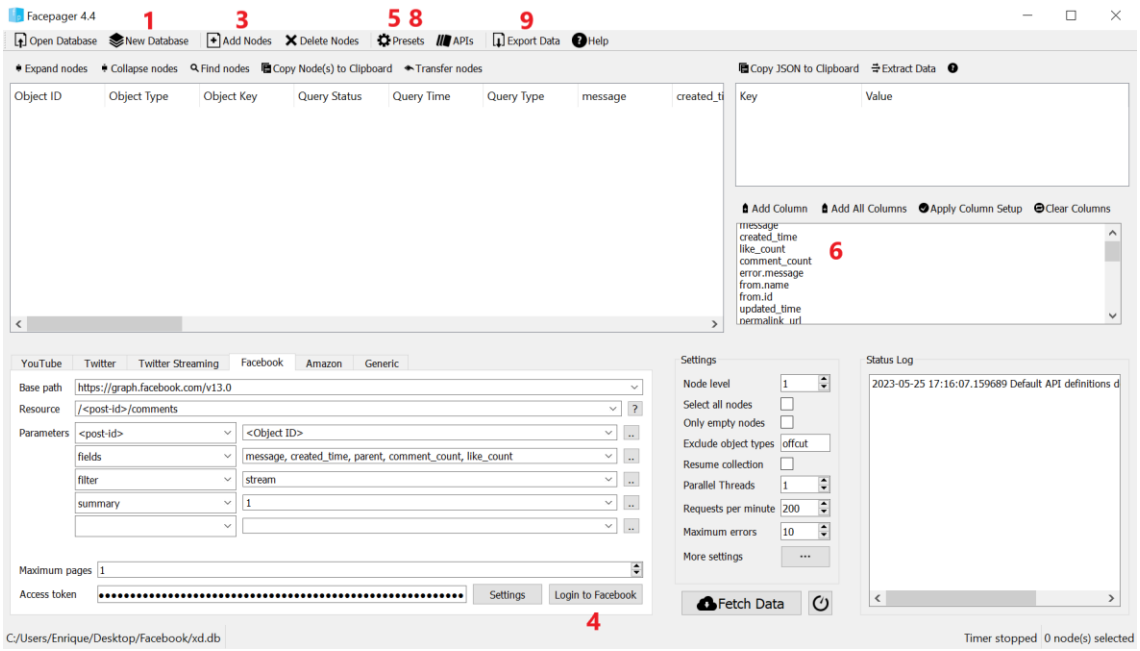
Con Facepager, los usuarios pueden definir parámetros de búsqueda y seleccionar tipos de datos que desean extraer, como publicaciones, comentarios, reacciones y detalles de perfil de los usuarios. La herramienta utiliza la API de Facebook para acceder a los datos públicos disponibles y permite filtrar los resultados por fecha, palabras clave y otros criterios.

Facepager realiza el siguiente proceso para el scraping de datos:

1. Crear una base de datos: Antes de comenzar, es necesario crear una base de datos en Facepager. Esto permitirá luego organizar y almacenar los datos extraídos de Facebook.
2. Obtener el Facebook ID de la página: Para iniciar el proceso de scraping, se debe obtener el Facebook ID de la página objetivo. Se puede lograr accediendo a un enlace específico y proporcionando la URL de la página. Este paso es fundamental para identificar la página de la cual se desean extraer datos.
3. Añadir un nuevo nodo: Una vez obtenido el Facebook ID, se debe agregar un nuevo nodo en Facepager. Se logra al ingresar el Facebook ID obtenido anteriormente. Este nodo representará la página de la cual se extraerán los datos.
4. Iniciar sesión en Facebook: Para acceder a los datos de Facebook a través de Facepager, es necesario iniciar sesión en una cuenta de Facebook. Esto se puede hacer haciendo clic en el botón correspondiente dentro de la herramienta. Es importante contar con una cuenta de Facebook

válida para llevar a cabo el proceso de scraping.

5. Scraping de posts: Después de completar los pasos anteriores, se puede comenzar el proceso de scraping. Para extraer los posts de la página, se debe ir al botón de "presets" en la parte superior de Facepager y seleccionar la opción "Facebook Get posts". Esto configurará automáticamente los parámetros necesarios para obtener los posts.
6. Configuración de campos adicionales: Para obtener datos específicos de cada post, es necesario agregar campos adicionales al parámetro de "fields". Algunos campos comunes incluyen: mensaje (message), autor (from), fecha de creación (created_time), fecha de actualización (updated_time), URL del post (permalink_url) y cantidad de veces compartido (shares). Si se desea extraer posts dentro de un rango de fechas específico, también se pueden añadir los parámetros "since" y "until".
7. Obtener los datos: Una vez configurados los parámetros según los requisitos, se debe seleccionar el ID del objeto y hacer clic en "Fetch Data" para obtener los datos de los posts.
8. Scraping de comentarios: Después de extraer los datos de los posts, se puede proceder a obtener los comentarios asociados a cada uno. Para ello, se debe seleccionar nuevamente el botón de "presets" y aplicar el preset correspondiente para obtener los comentarios. Luego, seleccionar todos los comentarios y hacer clic en "Fetch Data" nuevamente para extraerlos.
9. Exportar los datos: Una vez finalizado el proceso de scraping y con los datos deseados, se pueden exportar en formato CSV. Es importante utilizar la coma (",") como delimitador al exportarlos para asegurar una estructura adecuada del archivo.



Pero con Facepager se han presentado problemas similares a los de Octoparse. Por ello, se ha optado finalmente por una API de pago que proporciona la empresa Axesso.

Con esta API de Axesso hemos habilitado una solución temporal que nos ha permitido realizar la extracción de noticias y comentarios de diversos periódicos. Para gestionar este proceso al igual que X (Twitter) , se hará uso de Airflow para automatizar las tareas de extracción.

1.2. X (Twitter)

El proceso de scrapeo de X (Twitter) se basa en el uso de la API de X (Twitter) y la plataforma Airflow. El uso de la API y de Airflow permite que se puedan extraer datos de X (Twitter) de una manera automatizada. Cada tarea de scrapeo se configura con parámetros específicos, como fechas de inicio y finalización, y el medio de comunicación del cual se desea obtener los tweets.

Airflow, por su parte, es una plataforma de flujo de trabajo que permite programar y coordinar tareas en un entorno distribuido. En el código, se define un DAG (Direct Acyclic Graph) que representa el flujo de trabajo. El DAG para el scrapeo de X (Twitter) que se ha desarrollado incluye varias tareas que cumplen diferentes funciones en el proceso de scrapeo de datos de X (Twitter) .

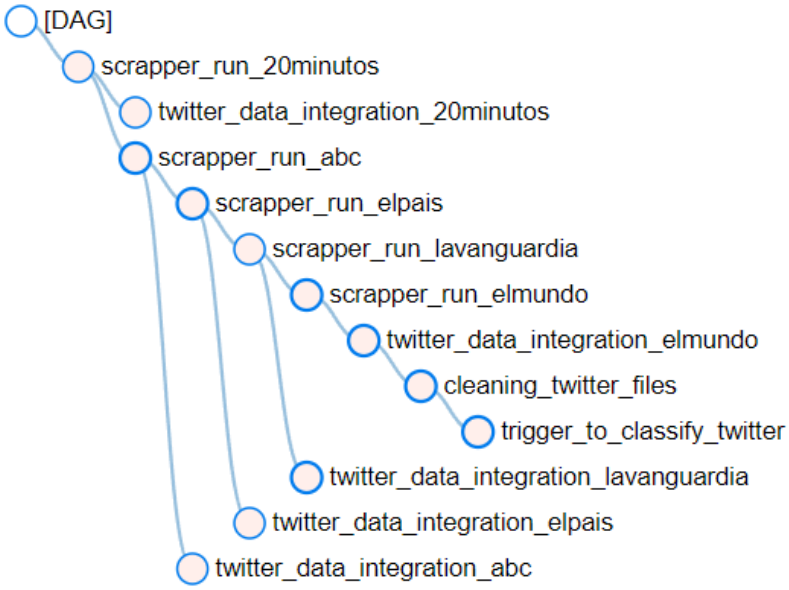
En primer lugar, se definen tareas de scrapeo que utilizan la API de X (Twitter) para buscar y obtener tweets y replies de diferentes medios de comunicación. Esta tarea se encarga de recopilar toda la información del día antes de su ejecución.

En segundo lugar, se incluyen tareas de integración a la base de datos de Django que se encargan de tomar los datos de X (Twitter) y realizar cualquier transformación necesaria antes de almacenarlos en la base de datos de Django.

Además, se incluyen tareas específicas para limpiar los archivos temporales generados durante el proceso de scrapeo. Estás tareas garantizan que los archivos temporales que se han utilizado se eliminen una vez utilizados para evitar la comunicación de archivos no deseados.

En general, el DAG proporciona una estructura organizada que permite ejecutar las tareas de scrapeo, integración y limpieza de datos de X (Twitter); esto asegura que el proceso de scrapeo se realice de manera eficiente y que los datos obtenidos se almacenen en la base de datos de manera automatizada.

A continuación se muestra de manera gráfica todas las tareas que contienen el DAG que se ha desarrollado para scrapear datos de X (Twitter) :

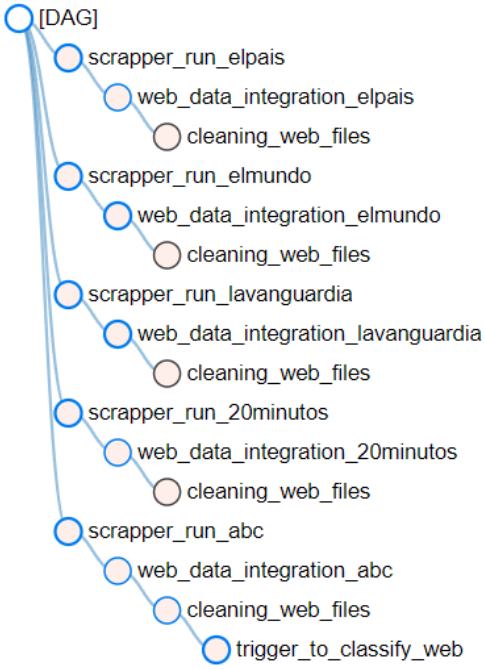


1.3. Web

El proceso de scrapeo web se basa en el uso de las bibliotecas requests, lxml y random_user_agent en Python. Estas bibliotecas permiten realizar solicitudes HTTP, obtener el contenido HTML de las páginas web, analizarlo y extraer los datos necesarios. Además, el random_user_agent ayuda a generar agentes de usuario aleatorios para evitar la detección de bots.

En este caso, se utiliza la plataforma de Airflow, al igual que el scrapeo para los datos de X (Twitter) . Se configura un flujo de trabajo (DAG) en Airflow, con prácticamente las mismas tareas que las que contiene el DAG para el scrapeo de X (Twitter): se trata tareas específicas para cada medio de comunicación a scrapear, tareas de integración de los datos en la base de datos y tareas de eliminación de archivos temporales.

Como en el caso anterior, se muestra de manera gráfica todas las tareas que contienen el DAG:



Como se puede observar en este gráfico, en el scraping web es posible realizar el proceso de scrapeo de manera paralela, a diferencia de X (Twitter), donde la API limita la extracción de datos a un solo medio de comunicación a la vez.

Esto significa que en el proceso de scraping web, se puede realizar la extracción de datos de múltiples medios de comunicación al mismo tiempo, de forma simultánea y en paralelo. Esto permite obtener información de varios medios de manera eficiente y acelerar el proceso de scraping en comparación con las limitaciones de la API de X (Twitter) .

2. PROBLEMAS QUE SE HAN PRESENTADO DURANTE EL PROCESO DE RECOPILACIÓN DE DATOS

2.1. Facebook

Para la extracción de datos de *Facebook* se encontraron tres problemas con las herramientas empleadas.

- Violación de términos de servicio: al realizar la extracción de datos en *Facebook* uno de los problemas más comunes es la eliminación de cuentas por la violación de los términos de servicio de la plataforma, ya que se prohíbe la extracción de datos personales y contenido protegido por derechos de autor de manera no autorizada, aún siendo esta pública.
- Detección de actividades anómalas: *Facebook* utiliza algoritmos muy avanzados para detectar actividades en su plataforma, incluyendo la extracción de datos de manera automática. Esto concluye en el bloqueo temporal o permanente de la cuenta empleada, algo que ha sucedido con bastante frecuencia.
- Modificación constante de la estructura de la página: esto resulta en la obsolescencia de las configuraciones de extracción de datos, dejando estas de funcionar.

2.2. X (Twitter)

Desde febrero del año 2023, el uso de la API de *X (Twitter)* (ahora llamada *X*) fue restringido. Se dejó de ofrecer una versión gratuita y ofertaron varios planes de pago que varían entre los 100$ mensuales (API Basic) hasta los 42.000$ mensuales (API Enterprise), teniendo, en función del pago, una serie de privilegios de uso u otros.

El método que se empleaba para la extracción de datos para el proyecto de Hatemedia utilizaba una API que dependía de un modelo llamado *Academics*, cuyo uso se restringió de manera gradual. Esto produjo que los datos extraídos fuesen nulos o insuficientes durante el período de eliminación de dichos privilegios hasta que se retomó la extracción con métodos de pago.

Estos fallos no se detectaron inicialmente porque la herramienta de automatización estaba programada para continuar la ejecución ante cualquier error y se pudiera generar un archivo *json* con los datos. No estaba previsto un error de autorización de la API, solamente contemplados los errores 429 por *request limit*. Esto provocó que aunque la ejecución se marcaba como exitosa a pesar de no serlo, se generaban documentos, pero estaban todos vacíos.

2.3. Web

El principal problema detectado a la hora de extraer datos de las páginas web de los periódicos tuvo lugar en el periódico 20 Minutos. La página web de este medio dispone de algún tipo de *firewall* que impide que, tras varias extracciones de datos, se pueda acceder al servicio desde un mismo punto, todo esto para evitar ataques de denegación de servicios o DDOS. Es por ello que obtener datos de este periódico es extremadamente difícil.

Otro problema que sucede de manera bastante habitual con los cinco periódicos elegidos es que la página realice cambios y modificaciones en su estructura HTML, comprometiendo el funcionamiento de la herramienta empleada para extraer los datos de la web. Esto afecta tanto a la recuperación de datos, al modificar la hemeroteca de los periódicos, como a la obtención de contenido regular.

3. ACCIONES PARA RECUPERAR LOS DATOS MISSING

3.1. Facebook

Para extraer datos de Facebook se empezó utilizando la herramienta *Octoparse*, una herramienta de extracción de datos web que permite a los usuarios recopilar información de sitios web de manera automatizada para convertirla en datos estructurados. Su uso fue temporalmente limitado por incumplir esta herramienta las políticas de *Facebook*.

En su lugar, se buscaron alternativas como *Facepager*, una herramienta de código abierto utilizada para la extracción de datos de redes sociales, incluyendo *Facebook*. *Facepager* permitía recopilar información de páginas públicas de Facebook de manera automatizada y convertirla en datos estructurados que podían ser exportados, pero, a lo largo del tiempo, fue imposible extraer datos sin utilizar la API pública de *Facebook*.

Para evitar problemas de bloqueo de cuentas se empleó su API de *Graph*. Se trata de una API gratuita extremadamente limitada para el uso requerido, puesto que no se debe utilizar para estos fines. Aún así, fue solicitada en el pasado y su acceso fue denegado.

En vista a estas limitaciones, se decide tomar medidas temporales y utilizar la API proporcionada por Axesso y el servicio proporcionado por exportcomments para poder continuar con la extracción de datos. Sin embargo, se encontraron dificultades significativas en cuanto a la consistencia de los resultados obtenidos mediante estas plataformas.

Por una parte, después de probar el tier de pago de la API de Axesso, se tomó la decisión de no continuar con esta opción debido a la falta de coherencia en los datos proporcionados al llevar a cabo la extracción. De manera similar, al probar la herramienta exportcomments, se constató que no era posible automatizar

el proceso de extracción, y además, los resultados obtenidos tampoco cumplían los criterios de consistencia necesarios.

Estas limitaciones hicieron que se tomase la decisión de no continuar con el scrapeo de Facebook ya que, actualmente, no existe una herramienta que arroje resultados consistentes, confiables y de calidad con los que se pueda asegurar la eficacia en el proceso.

3.2. X (Twitter)

Se procedió a la contratación de la API en *tier Basic*. Esto permitía mantener actualizada la información de manera regular, pero no obtener información más allá de 10 días. Ante un error desconocido, se tuvo acceso a la *tier Pro* de la API de X (Twitter), que ya permitió poder acceder a contenido de meses anteriores de los que no se disponía información y recuperarla.

Se comenzó por la realización de una prueba con el código ya empleado, se ajustaron los parámetros de extracción de datos para obtener la mayor cantidad posible de información, que a la vez fuese equitativa entre todos los meses faltantes. Tras unos ligeros cambios se empleó el código ya utilizado para obtener información de manera regular y poder buscar en aquellas fechas de las que había interés (periodo de enero 2021 a julio 2022). Se crearon tres aplicaciones para extraer la mayor cantidad de datos. Desde el día 10 de agosto, estuvieron funcionando las tres aplicaciones las 24 horas del día los 7 días de la semana. Así se obtuvo una muestra de cada periódico para cada mes faltante.

Por las limitaciones temporales, solamente se pudo tener la herramienta en funcionamiento durante 14 días y resultó insuficiente para conseguir toda la información necesaria. Se obtuvo solo una muestra escasa para el mes de mayo del año 2023 (en El País y La Vanguardia). Fue imposible obtener más información durante los meses de junio y julio de ese mismo año, porque hubo que detener manualmente la herramienta a las 23:59 del día 23 de agosto de 2023.

3.3. Web

Se han hecho modificaciones para hacer al extractor de datos adaptable a la evolución de la estructura de la página web, pero este problema no solventa el ocasionado por el periódico 20 Minutos por lo que se siguen investigando posibles formas de actualizar el scraper para adaptarlo a la nueva estructura de la página.

4. DEPURACIÓN Y RECUENTO DE DATOS

Se explica ahora la metodología utilizada para realizar el depurado y posterior recuento de los datos recabados por los procedimientos señalados en puntos anteriores. El proceso de depurado y recuento mantuvo el siguiente protocolo:

1. Las carpetas con los datos *scrappeados* contienen varios ficheros JSON y CSV separados por meses en distintas carpetas del tipo 01_ENERO, 02_FEBRERO, 03_MARZO, etc. El primer paso consiste en juntar todos estos JSON y escribirlos a archivos CSV (diferenciando entre comentarios y publicaciones). Posteriormente se unen a los demás ficheros CSV que también contiene la carpeta.

2. Con la totalidad de los datos en un mismo CSV, se procede a eliminar duplicados y filas con valores Nan.

3. Mediante un script se comprueba que las fechas de todos los mensajes de un CSV correspondan al mismo mes. Cuando la fecha de un mensaje no corresponda al mes del estudio, se identifican y seleccionan para trasladarlos posteriormente al fichero CSV adecuado (el que contiene los mensajes con la fecha adecuada).

4. Se eliminan duplicados de nuevo para asegurar que no existe información duplicada tras el traslado y organización de los mensajes en el CSV del paso anterior. Realizada esta segunda comprobación, quedan los datos sin duplicados y sin Nan dentro de la carpeta correspondiente a su mes.

5. Mediante un script se recorre el árbol de carpetas para leer los datos, eliminar emojis, caracteres no latinos y *stopwords*. El resultado se guarda sobreescribiendo el archivo que se ha limpiado.

Una vez realizada la depuración de los datos, se procede al recuento de comentarios y publicaciones distinguiendolos por mes.

Se obtuvieron así, depurados, el total de mensajes en bruto, el total de mensajes eliminados y el total de mensajes finales para los datos recabados del periodo enero 2021 a julio 2022.

2021				
Medio	**Soporte**	**Mensajes totales**	**Mensajes eliminados**	**Mensajes finales**
Vanguardia	Facebook	21,854	900	20954
	Twitter	155,237	11,963	143,274
	Web	1042866	608266	434600
El Pais	Facebook	25,876	1,240	24,636
	Twitter	645,638	147,311	498,327
	Web	1593222	901574	691648
El Mundo	Facebook	23,641	844	22797
	Twitter	995,233	117,489	877,744
	Web	3066496	1766737	1299759
ABC	Facebook	8,454	400	8,054
	Twitter	564,027	96,953	467,074
	Web	2488527	1231754	1256773
20 Minutos	Facebook	14,525	1,013	13,512
	Twitter	207,474	35,973	171501
	Web	1165200	759310	405890
		12,018,270	5,681,727	6,336,543

Número total de mensajes para cada medio y soporte - enero 2021 a diciembre 2021

2022				
Medio	**Soporte**	**Mensajes totales**	**Mensajes eliminados**	**Mensajes finales**
Vanguardia	Facebook	132,009	26,843	105,166
	Twitter	71,852	31,884	39,968
	Web	531904	301647	230257
El Pais	Facebook	185,424	31,929	153,495
	Twitter	187,524	66,650	120,874
	Web	1023349	478480	544869
El Mundo	Facebook	182,280	9,484	172796
	Twitter	146,718	75,192	71,526
	Web	1676977	776175	900802
ABC	Facebook	117,713	10,882	106,831
	Twitter	78,504	38,286	40,218
	Web	1327794	680930	646864
20 Minutos	Facebook	87,295	36,163	51,132
	Twitter	115,437	57,149	58,288
	Web	2167690	1546045	621645
		8,032,470	4,167,739	3,864,731

Número total de mensajes para cada medio y soporte - enero 2022 a julio 2022

Enero 2021 - Julio 2022				
Medio	Soporte	Mensajes totales	Mensajes eliminados	Mensajes finales
Vanguardia	Facebook	153,863	27,743	126,120
	Twitter	227,089	43,847	183,242
	Web	1,574,770	909,913	664857
El Pais	Facebook	211,300	33,169	178,131
	Twitter	833,162	213,961	619,201
	Web	2,616,571	1,380,054	1236517
El Mundo	Facebook	205,921	10,328	195593
	Twitter	1,141,951	192,681	949,270
	Web	4,743,473	2,542,912	2200561
ABC	Facebook	126,167	11,282	114,885
	Twitter	642,531	135,239	507,292
	Web	3,816,321	1,912,684	1903637
20 Minutos	Facebook	101,820	37,176	64,644
	Twitter	322,911	93,122	229,789
	Web	3,332,890	2,305,355	1027535
		20,050,740	9,849,466	10,201,274

Número total de mensajes para cada medio y soporte - enero 2021 a julio 2022