

Taller 1: Extracción de Datos y Análisis de Sentimiento en E-commerce

Grupo 6

Carlos Mauricio Arteaga
Miguel Ángel Pablos
María Alejandra Olarte

El presente informe detalla el proceso de extracción de datos de la sección de celulares en la plataforma Mercado Libre, seguido de un análisis de los comentarios asociados con los productos recopilados y los principales hallazgos en el proceso.

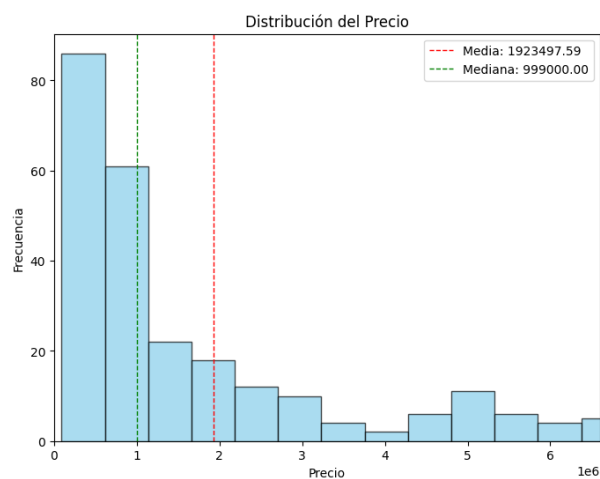
1. Explicación del proceso de extracción y herramientas utilizadas.

El script desarrollado para scrapear la plataforma de Mercado Libre consta de los siguientes pasos:

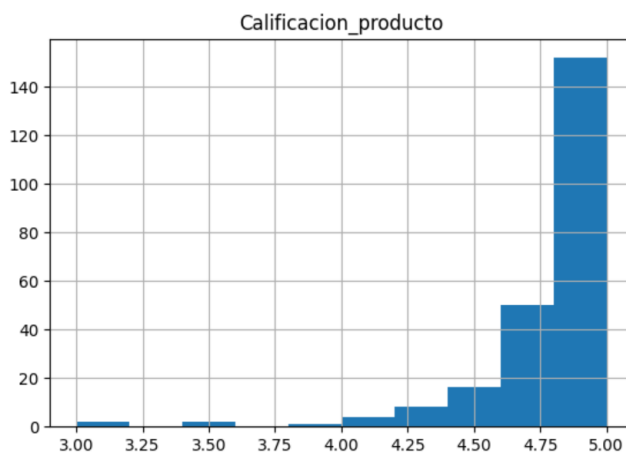
- Introducir una lista de URLs de donde se va a raspar la información.
- Inspeccionar manualmente el código HTML de la página para identificar las etiquetas, los elementos y las clases en las cuales queremos iterar para obtener la información.
- Generamos un código para iterar sobre el listado de productos para obtener inicialmente la marca, características, calificación del producto, precio y el enlace relacionado para cada producto.
- Generamos un código para iterar en las URLs de los productos en búsqueda de la información faltante, como el nombre del vendedor, la calificación del vendedor, la cantidad de comentarios y los comentarios.
- Con esta información obtenemos un DataFrame.
- Generamos una función de limpieza del DataFrame donde utilizamos métodos, expresiones regulares, eliminación de espacios, conversión de tipo, selección de variables entre otras para obtener un Dataframe limpio.
- Para el análisis de sentimientos, primero se tradujeron los comentarios al inglés y teniéndolos en este idioma se utilizó la librería TextBlob.
- Utilizamos las librerías requests para hacer peticiones a páginas web, BeautifulSoup para hacer web scraping, pandas para manejo de datos, URLError para manejo de errores de URL, re para manejo de expresiones regulares, word_tokenize para tokenizar palabras, matplotlib y seaborn para graficar, TextBlob para el análisis de sentimientos, nltk para el procesamiento de los comentarios con métodos como PorterStemmer, GoogleTranslator para traducir del español al inglés para realizar el análisis de sentimientos y subjetividad.

2. Resultados del análisis de sentimiento.

En el momento en que se realizó el ejercicio¹, se recopiló información de 258 dispositivos móviles publicados, destacando como la marca más vendida el Samsung Galaxy (18%), seguido por el Xiaomi Redmi (10%) y el iPhone (8,5%). El precio promedio de estos celulares fue de \$1.923.000. Cabe resaltar que el 50% de los dispositivos tiene un precio inferior a \$1.000.000.



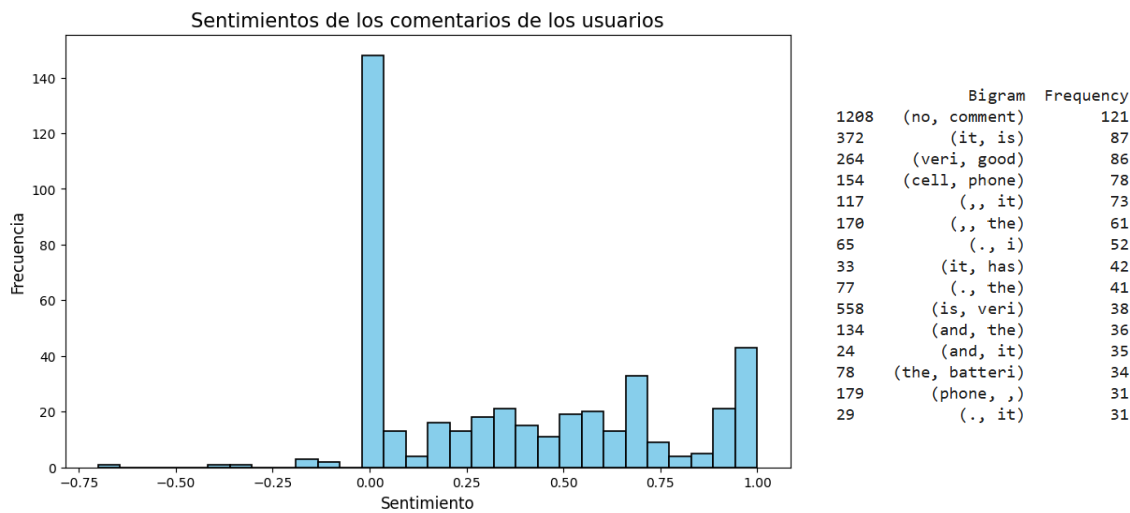
Por su parte, el promedio de la calificación de los productos fue 4,7 sobre 5, donde la mayoría de los productos tienen una calificación superior a 4,5 como se muestra en el siguiente histograma.



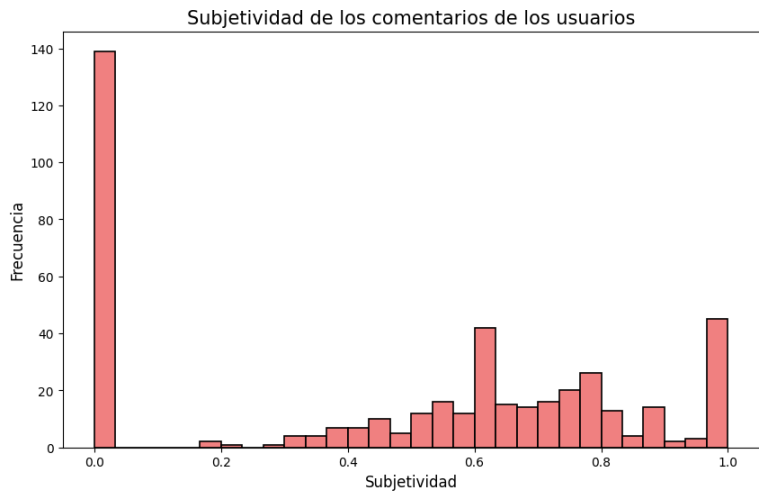
¹ Como se menciona en otro apartado del informe, la web de Mercado Libre es dinámica y van cambiando los productos que se muestran, por lo tanto esto corresponde a la consulta realizada el 21 de noviembre 6pm.

Teniendo en cuenta principalmente estas características, se realizó un análisis de sentimientos a los tres primeros comentarios que aparecen en la página web de Mercado Libre para cada producto, logrando analizar 532 comentarios. El indicador promedio de sentimiento obtenido fue de 0.36, lo que sugiere una tendencia ligeramente positiva en los comentarios, aunque no marcadamente optimista. En general, el tono refleja una postura moderadamente favorable.

Al analizar el histograma, se observa una concentración de valores en torno a cero, lo que se explica, en parte, porque 121 productos no tenían comentarios. En estos casos, el texto analizado fue "no comments", como se evidencia también en la frecuencia de los bigramas.



Por su parte, el indicador de subjetividad en promedio fue 0,47, lo que muestra una tendencia subjetiva, aunque no totalmente inclinada a opiniones o percepciones. Esto comentarios contienen especialmente opiniones y percepciones, pero también podrían incluir algunos elementos objetivos. El histograma muestra una concentración en cero, lo que puede estar explicado, en parte, por el número de comentarios con “no comments”.



En línea con los indicadores anteriormente descritos, se crean bigramas que se obtienen de los comentarios en español, luego de eliminar las “stopwords” y realizar el proceso de lematización. Se encuentra que los bigramas más frecuentes son combinaciones de palabras que contienen un calificativo positivo, lo que está en línea con el indicador de sentimiento.

	Bigram	Frequency
8	(excelent, producto)	19
190	(buen, producto)	18
32	(batería, dura)	17
89	(gama, media)	11
626	(buen, celular)	9
243	(buen, equipo)	9
233	(funciona, bien)	8
177	(calidad, precio)	8
176	(relación, calidad)	7
364	(excelent, celular)	7
366	(buena, cámara)	7
276	(mejor, celular)	6
892	(excelent, teléfono)	6
33	(dura, día)	6
73	(duración, batería)	6

Analizando los resultados obtenidos para el bigrama con las palabras en español encontramos varios datos relevantes relacionados con el comportamiento de compra y las características importantes en la decisión de compra como buen producto, duración de batería, calidad precio, buena cámara; estos resultados se relacionan con el resultado de la distribución de sentimientos.

3. Observaciones relevantes sobre la calidad de los datos y posibles sesgos.

La página de MercadoLibre tiene una capa dinámica que dificulta la extracción de los datos, hay un componente aleatorio agregado al HTML que ocasiona que cada vez que se ejecute el script los datos obtenidos varíen, esto puede ser por algún tipo de filtrado de la página con cambios como venta de celulares, comentarios recientes o actualizaciones del producto. También podría ser una medida de MercadoLibre para que “scrapear” la página no sea tan fácil.

La fuente de datos obtenida tuvo que pasar por varios procesos de limpieza para convertirlos en datos estructurados para su correcto análisis.

Dentro de los posibles sesgos más relevantes que encontramos en la extracción de la información está el hecho de que MercadoLibre solo muestra una selección de comentarios destacados y para ir más profundo en los comentarios se requiere de una interacción de despliegue en el navegador que no logramos simular con los conocimientos adquiridos hasta el momento. No obstante, para fines académicos los tres comentarios por producto nos permiten hacer el ejercicio de análisis de sentimiento solicitado.

★★★★★

Duración de la batería

★★★★★

Durabilidad


★★★★★

Opiniones destacadas


104 comentarios

★★★★★

21 may. 2024




Le compré el celular a mi esposa, hizo el cambio de iphone a samsung y la verdad la diferencia es notable. Ya no tiene que estar cargando el iphone cada rato. La cámara trasera es excelente y la cámara frontal es buena. El rendimiento que ofrece es notable al usar varias apps, también resiste salpicaduras lo cual es bueno para cuando llueve. En resumen, es mejor que las porcelanas del iphone.

Es útil  47

★★★★★


11 abr. 2024

Muy bueno producto.


Es útil  11

★★★★★


09 jul. 2024



Un celular muy lindo, de gran rendimiento y por supuesto con una buen relación calidad precio.


Es útil  7

Mostrar todas las opiniones



Compra en tiendas fuera de Mercado Libre

MI MEDELLINSHOP




Celular Realme 9i Dual Sim 128 Gb - 4...

\$ 899.900

3 cuotas de \$ 299.967 con 0% interés

Envío gratis

MI MEDELLINSHOP




Apple iPhone 12 Pro Max (128 Gb) - Azul

\$ 3.999.900

3 cuotas de \$ 899.967 con 0% interés

Envío gratis

MI MEDELLINSHOP




Apple iPhone XR 64 Gb

\$ 1.299.000

3 cuotas de \$ 389.700 con 0% interés

Envío gratis

MI MEDELLINSHOP



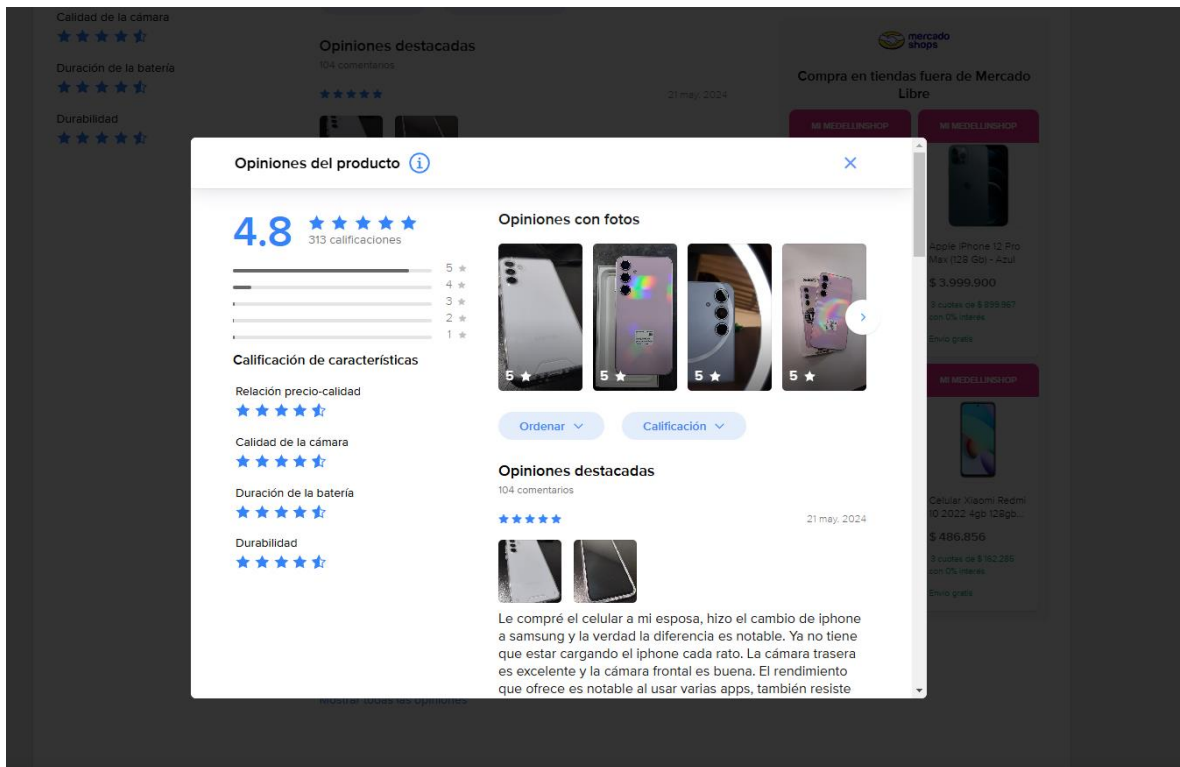
Celular Xiaomi Redmi 10 2022 4gb 128gb...

\$ 486.856

3 cuotas de \$ 162.285 con 0% interés

Envío gratis

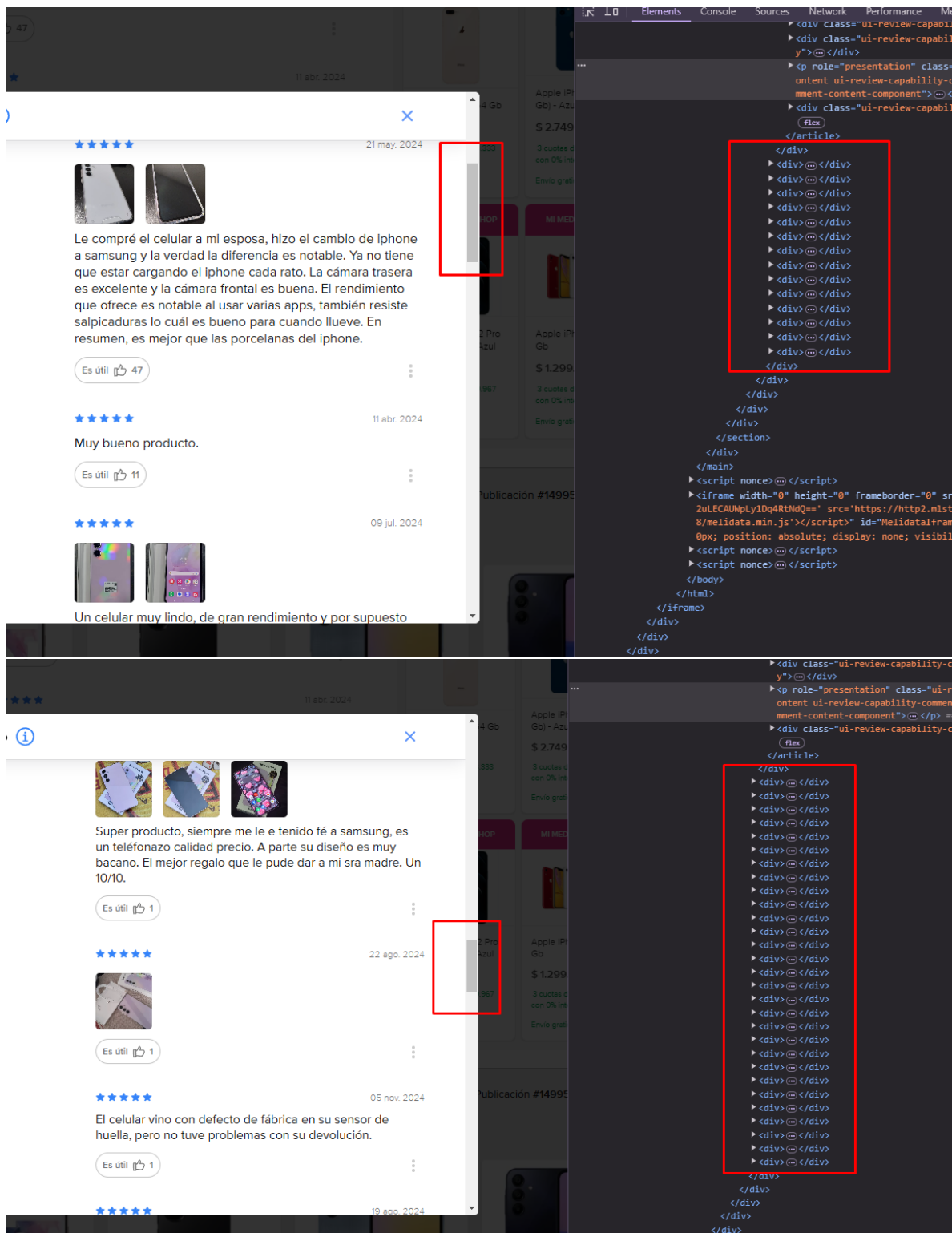
La base de datos de comentarios obtenidos podría tener un sesgo en el sentido que está compuesta por comentarios destacados, y estos normalmente tienden a ser positivos y derivan de una buena experiencia de compra. En parte este sesgo, podría explicar por qué el índice de sentimientos es positivo.



Como se observa en la anterior imagen el módulo interactivo de los comentarios está en una capa extra y con nuestro script no logramos acceder a esa capa más profunda que requiere de la interacción de la barra de desplazamiento.

En la siguiente imagen se muestra como la interacción con la barra de desplazamiento fue el limitante para obtener un mayor volumen de comentarios.

Intentamos realizar la extracción con la librería “selenium” y trabajando una simulación con “Chromium” pero aun así no logramos obtener los datos completos que nos hubieran dado un panorama más amplio de las opiniones de los clientes.



Solucionar este problema de extracción de datos es una posible área de mejora para próximos ejercicios dentro de la plataforma de Mercado Libre.