Procesamiento de Reseñas de Trustpilot mediante Scraping Web

Daniel Sanchez Trujillo
Universidad Mayor de San Andrés
La Paz, Bolivia
Correo: daniel.sanchez@agetic.gob.bo

Ludwing Young
Universidad Mayor de San Andrés
La Paz, Bolivia
Correo: ludwing.young@gmail.com

Sergio Agreda
Universidad Mayor de San Andrés
La Paz, Bolivia
Correo: sergioagreda21@outlook.com

***Resumen*—Este informe presenta el diseño e implementación de un sistema automatizado para la extracción, limpieza y consolidación de reseñas obtenidas desde la plataforma Trustpilot. El proceso desarrollado integra técnicas de *web scraping*, normalización y validación de datos, así como la generación de un repositorio estructurado en formato JSON que facilita su análisis posterior. Adicionalmente, se elaboró un dashboard interactivo mediante Streamlit, que permite explorar tendencias de calificaciones, analizar el sentimiento asociado a los textos y visualizar patrones clave a través de indicadores y nubes de palabras. La solución resultante constituye una herramienta integral para el estudio comparativo de empresas y categorías, habilitando un análisis claro y accesible de la percepción del usuario.**

***Index Terms*—Scraping, Trustpilot, Análisis de Sentimiento, Reseñas, Limpieza de Datos, Streamlit.**

I. Introducción

Trustpilot es una plataforma en línea que permite a los consumidores compartir opiniones y experiencias sobre diversas empresas. En el presente proyecto se desarrolló un sistema automatizado para extraer reseñas de empresas pertenecientes a las siguientes categorías:

- Bancos
- Seguros de Viaje
- Concesionario de autos
- Tiendas de ropa

Se seleccionaron las 10 empresas de cada categoría con mayor cantidad de reseñas para obtener una muestra representativa y focalizada. Dentro de estas categorías, se eligió realizar un análisis más detallado sobre el sector *Tiendas de ropa*, específicamente sobre las empresas **TWOTHIRDS** y **ZARA**. Esta decisión se tomó con el propósito de mostrar de manera ilustrativa cómo el dashboard permite explorar diferencias en percepción, sentimiento, patrones de reseñas y comportamiento temporal entre dos marcas del mismo rubro, pero con perfiles de reputación claramente distintos. Las visualizaciones obtenidas sirven como ejemplo del potencial de la herramienta desarrollada para efectuar análisis comparativos entre empresas o sectores específicos.

El proceso se compone de varias etapas fundamentales:

- **Extracción de Datos:** La etapa de extracción consistió en recopilar información directamente desde la plataforma objetivo mediante técnicas de raspado web automatizado. Este proceso permitió obtener datos crudos de manera sistemática y reproducible, capturando elementos clave como títulos, contenidos, fechas y calificaciones de las reseñas. La automatización garantizó una recolección eficiente y homogénea, reduciendo errores manuales y asegurando la disponibilidad de datos actualizados para su posterior procesamiento.
- **Limpieza y Validación:** El proceso incluyó una etapa de limpieza y validación de los datos extraídos, garantizando que la información recolectada desde el sitio web fuese consistente, utilizable y estuviera libre de errores comunes generados durante el scraping. Para ello se normalizaron cadenas de texto, se eliminaron caracteres inválidos, se homogenizó el formato de fechas y se reconstruyeron aquellos campos incompletos. Asimismo, se aplicaron validaciones básicas para asegurar que cada reseña contara con los elementos mínimos necesarios, como título, contenido, fecha y calificación. Esta fase permitió transformar datos crudos en un conjunto estructurado adecuado para análisis posteriores.
- **Generación de Salidas:** Una vez depurados, los datos fueron organizados y exportados en estructuras listas para su consumo por otras herramientas del proyecto. Se generaron archivos en formato JSON limpio y, cuando fue necesario, estructuras adicionales que preservan la jerarquía empresa $\rightarrow$ categoría $\rightarrow$ reseñas. Esta salida estandarizada permite integrar los datos con el pipeline de análisis, visualización y carga a la base de datos, asegurando compatibilidad con módulos posteriores del sistema y facilitando su reutilización en otros procesos analíticos o dashboards.

II. Especificaciones del equipo

El desarrollo y las pruebas del cuaderno de raspado se realizaron en un equipo con las siguientes características:

Cuadro I
Características del entorno de ejecución

Parámetro	Especificación
Sistema Operativo	Windows 11, 64 bits
Arquitectura	x64
Procesador	12th Gen Intel Core i7-12650H
Memoria RAM	16 GB

A continuación se detallan las librerías utilizadas para el desarrollo del sistema de análisis de reseñas, incluyendo herramientas para visualización, procesamiento de datos, análisis de sentimiento y automatización. Las versiones corresponden al entorno de ejecución utilizado durante el desarrollo del proyecto.

Cuadro II
LIBRERÍAS UTILIZADAS EN EL PROYECTO Y SUS VERSIONES

Librería	Versión
streamlit	1.51.0
pandas	2.3.3
plotly	6.4.0
nltk	3.9.2
wordcloud	1.9.4
numpy	2.2.6
transformers	4.56.2
selenium	4.38.0
webdriver_manager	4.0.2
pytz	2025.2
sqlalchemy	2.0.44
json	2.0.9
re	2.2.1

Estas especificaciones aseguran un entorno estable para la ejecución de Selenium y Jupyter Notebook, así como suficiente memoria para manejar la creación y manipulación de *DataFrames* de tamaño moderado.

III. FUNDAMENTOS TEÓRICOS Y TECNOLÓGICOS

El enfoque metodológico de este proyecto, que utiliza el Procesamiento de Lenguaje Natural (NLP) para analizar reseñas de texto, se alinea con una robusta línea de investigación académica. La literatura reciente valida que el análisis de datos no estructurados (como las reseñas) es crucial para entender la satisfacción del cliente.

Por ejemplo, Le et al. [1] desarrollaron un marco predictivo para el comercio electrónico que, de manera similar a este proyecto, utiliza modelos de NLP (específicamente BERT y Bi-GRU) para realizar análisis de sentimiento a nivel de aspecto en las reseñas [1]. Su trabajo demuestra cómo la integración de los sentimientos extraídos del texto con metadatos (como el precio) permite predecir con alta precisión la satisfacción general [1]. Del mismo modo, Sun et al. [2] utilizaron deep learning para analizar cómo las expresiones emocionales en las reseñas de Yelp se correlacionan con la satisfacción, validando que el tono emocional es un predictor clave [2].

La elección de modelos basados en Transformers está respaldada. Un estudio comparativo de Aftan & Shah [3] encontró que un modelo Transformer (AraBERT) superó significativamente a las arquitecturas tradicionales de CNN y RNN en la clasificación de la satisfacción del cliente a partir de tuits en árabe [3]. Esto subraya la capacidad de los modelos Transformer para capturar matices del lenguaje y jerga que otros modelos no perciben [3].

Si bien este proyecto se centra en el análisis textual, la literatura del comercio electrónico también destaca la importancia de los datos estructurados. Estudios como los de Zaghloul et al. [4] y Wong & Marikannan [5] identificaron que factores logísticos, como el tiempo de entrega y la precisión del pedido, son impulsores fundamentales de la satisfacción [4], [5]. Las reseñas analizadas en este informe (especialmente las de ZARA) probablemente capturan las consecuencias sentimentales de estos factores operativos.

Finalmente, el objetivo de este proyecto de convertir texto no estructurado en métricas visuales (nubes de palabras, puntuaciones de sentimiento) es un precursor de enfoques de vanguardia. Teichert & Shah [6] proponen el uso de Modelos de Lenguaje de Gran Escala (LLMs) para transformar reseñas en çonstructos"teóricos estructurados, como la 'calidad percibida' y el 'valor percibido' [6]. Su hallazgo de que la 'calidad del servicio' es el impulsor dominante en plataformas de entrega [6] resuena con los hallazgos de este proyecto, donde las quejas de ZARA se centran en procesos (devoluciones, servicio al cliente) y los elogios a TWOTHIRDS se centran en el producto (calidad).

Estos estudios proporcionan una base sólida para el presente proyecto, validando la metodología de usar NLP y modelos de Transformers para analizar reseñas de plataformas como Trustpilot y extraer información accionable sobre la percepción y satisfacción del cliente.

El desarrollo del proyecto se sustentó en un conjunto de bibliotecas y herramientas clave dentro del ecosistema de análisis de datos, automatización web y procesamiento del lenguaje natural. Cada componente cumplió una función específica dentro del flujo general, que abarca la extracción, limpieza, transformación y análisis de reseñas provenientes de la plataforma Trustpilot.

Para la construcción del dashboard interactivo se empleó Streamlit [12], que permitió diseñar una interfaz flexible y dinámica para la visualización de métricas, nubes de palabras y tendencias. La manipulación y estructuración del dataset se realizó mediante Pandas [13], herramienta fundamental para cargar, combinar, transformar y filtrar la información procesada. Las visualizaciones interactivas se generaron con Plotly [14], permitiendo representar patrones temporales y distribuciones de calificaciones de forma intuitiva.

El procesamiento del lenguaje natural se apoyó en NLTK [15], utilizado principalmente para gestionar listas de stopwords y normalizar texto, y en WordCloud [16] para generar las representaciones visuales de términos más frecuentes. Para el análisis de sentimiento se integraron modelos preentrenados mediante la biblioteca Transformers de HuggingFace [18], posibilitando evaluar automáticamente la polaridad textual de las reseñas.

La etapa de extracción de datos utilizó Selenium [19] en conjunto con WebDriver Manager [20] para la automatización del navegador durante el scraping. Asimismo, la librería Pytz [21] garantizó la correcta estandarización de fechas y zonas horarias dentro del conjunto de datos. Como apoyo adicional, herramientas estándar del lenguaje Python como `json` [22] y `re` [23] fueron utilizadas para la gestión estructural de los datos y la validación de patrones textuales.

Finalmente, NumPy [17] proporcionó capacidades numéricas esenciales durante el preprocesamiento de datos y la elabo-

ración de indicadores empleados en el análisis descriptivo. En conjunto, este ecosistema tecnológico permitió implementar un flujo robusto, reproducible y escalable para el análisis integral de reseñas en línea.

IV. Metodología y contenido del cuaderno

El cuaderno `scrapp.ipynb` se organiza en las siguientes secciones:

IV-A. Estructura de Directorios del Datalake

Antes de iniciar el proceso de consolidación de datos, se definió una estructura que permite gestionar los datos de manera ordenada conforme avanzan en el flujo de procesamiento, desde su extracción inicial hasta su uso final en análisis o visualizaciones.

La creación de estas carpetas se realizó de forma automatizada mediante un script en Python, el cual verifica la existencia de los directorios y los genera en caso de no encontrarse. Las carpetas definidas fueron las siguientes:

- **1. LANDING ZONE**: Contiene los datos crudos tal como fueron extraídos del proceso de raspado web. En esta zona se almacenan los archivos CSV y JSON iniciales correspondientes a cada categoría, sin modificaciones ni limpieza previa.
- **2. REFINED ZONE**: Alberga los datos procesados y estructurados. Aquí se guardan los archivos resultantes de las etapas de limpieza, validación y reorganización, incluyendo las versiones fusionadas de los datasets CSV y JSON.
- **3. CONSUMPTION ZONE**: Aquí se encuentra también el *dashboard* interactivo desarrollado para la visualización de reseñas, métricas y análisis de sentimiento.

Esta organización en capas facilita la trazabilidad de los datos y mantiene un flujo claro desde la ingestión hasta su consumo, permitiendo así un manejo más controlado y reproducible del conjunto de información.

IV-B. Generación de Dataset

Con el objetivo de unificar la información obtenida desde distintas categorías, se implementó un proceso de consolidación tanto para los datos almacenados en formato CSV como para aquellos disponibles en formato JSON.

En el caso de los archivos CSV, se cargaron los datasets generados para cada una de las categorías (Bancos, Seguros de viaje, etc.), se incorporó una columna denominada `categoria` con el valor correspondiente y se reordenaron las columnas para que dicha variable apareciera al inicio. Posteriormente, ambos conjuntos de datos fueron concatenados en un único archivo denominado `dataset_reviews.csv`.

Para los archivos JSON, se procedió de manera análoga: se cargaron las estructuras que contenían la información de las empresas y sus reseñas, se añadió la clave `categoria` a cada objeto y, mediante una estructura ordenada, se reorganizaron las claves para garantizar que la categoría se ubicara en la primera posición. Finalmente, las listas fueron combinadas y almacenadas en un único archivo consolidado, `dataset_reviews.json`.

Este proceso de integración permite reunir información proveniente de distintas fuentes en un repositorio unificado, garantizando consistencia estructural y facilitando su posterior análisis y explotación dentro del pipeline de procesamiento de datos.

IV-C. Limpieza, Validación y Almacenamiento de Datos

En esta etapa se desarrolla un proceso completo para transformar los datos de reseñas obtenidos mediante raspado web en un conjunto estructurado, limpio y listo para su uso en etapas posteriores del proyecto. El enfoque seguido combina ajustes de formato, normalización de texto, corrección de fechas y generación de un archivo unificado en formato JSON dentro de la zona refinada del datalake.

1. Configuración inicial

En primer lugar, se definen las rutas de los archivos de entrada y salida dentro de la estructura del datalake y se establece la zona horaria de Bolivia (*America/La_Paz*). Esto asegura que todo el procesamiento posterior se realice de manera consistente respecto al contexto temporal local.

2. Normalización de valores numéricos

A continuación, se estandarizan los campos numéricos clave, en particular la puntuación asociada a cada empresa. Para ello se eliminan textos adicionales, se corrige el formato decimal y se verifica que los valores resultantes sean válidos. De este modo, las puntuaciones quedan listas para ser interpretadas y comparadas en análisis posteriores.

3. Ajuste y estandarización de fechas

El cuaderno también se encarga de transformar las fechas originales de las reseñas a un formato estándar y de ajustarlas a la zona horaria de Bolivia. A partir de esta conversión se generan dos campos diferenciados: la fecha local y la hora local. Esto facilita la lectura e interpretación de la información temporal, así como su uso en estudios de comportamiento a lo largo del tiempo.

4. Limpieza del contenido textual

Los textos correspondientes a títulos y contenidos de las reseñas se someten a un proceso de limpieza que incluye la conversión a minúsculas, la eliminación de acentos y la depuración de caracteres especiales o símbolos que no aportan información relevante. Con ello se obtiene un corpus de texto homogéneo, más adecuado para tareas de análisis posteriores, como exploración de palabras frecuentes o análisis de sentimiento.

5. Estructuración por empresa y reseña

Sobre cada empresa y sus reseñas se aplican reglas de validación que garantizan la presencia y coherencia de los campos más importantes. Se revisan y estandarizan atributos como la categoría, el nombre, la ubicación y la página web de la empresa, y en el caso de las reseñas se comprueba que existan identificadores, texto y calificaciones numéricas válidas. Asimismo, se eliminan campos redundantes o innecesarios para simplificar la estructura final.

6. Generación del archivo depurado

Una vez completados los pasos de limpieza y validación, el conjunto de datos resultante se guarda en un único archivo JSON dentro de la zona `2_REFINED_ZONE`. Este archivo consolida la información de empresas y reseñas en un formato uniforme y coherente, sirviendo como fuente principal para las etapas posteriores de análisis y para la construcción del dashboard ubicado en la `3_CONSUMPTION_ZONE`.

En conjunto, este flujo convierte los datos crudos obtenidos del raspado web en un recurso confiable y estructurado, listo para su explotación dentro del ecosistema del proyecto.

V. Ejecución y actualización del proceso

Actualmente, la actualización del repositorio de datos y la visualización de resultados se realiza de manera controlada por el usuario, siguiendo una secuencia de pasos sencilla pero reproducible. En primer lugar, cuando se requiere incorporar nuevas reseñas al sistema, se ejecuta el cuaderno `scrapp.ipynb`. Este cuaderno se encarga de realizar el raspado de las categorías definidas, aplicar las etapas de limpieza y validación descritas previamente y generar como resultado el archivo depurado `dataset_reviews_limpio.json` dentro de la carpeta `datalake/2_REFINED_ZONE`.

Una vez que los datos han sido actualizados y almacenados en la zona refinada, el usuario puede consultar los resultados a través del *dashboard* desarrollado en Streamlit, ubicado en la `3_CONSUMPTION_ZONE`. Para ello, basta con ejecutar el siguiente comando desde la línea de comandos, situándose en la raíz del proyecto:

```
streamlit run datalake/3_CONSUMPTION_ZONE/app.py
```

Este comando inicia la aplicación web de Streamlit, que carga el dataset limpio y permite explorar las reseñas, visualizar métricas descriptivas y analizar el sentimiento asociado a las opiniones de los usuarios. De este modo, la combinación entre la ejecución del cuaderno y el lanzamiento del *dashboard* proporciona un flujo claro y reutilizable para la actualización y consulta periódica de la información.

VI. Dashboard y análisis de las empresas

El tablero incorpora varios elementos que trabajan de forma integrada para ofrecer una visión completa del comportamiento de las reseñas. En primer lugar, se dispone de un conjunto de filtros que permiten seleccionar categoría, empresa, año y mes, adaptando todos los gráficos y métricas al contexto elegido. Esta interacción flexible habilita análisis comparativos y un seguimiento más preciso a lo largo del tiempo.

En la parte superior se presentan indicadores clave que resumen el estado general del conjunto de datos filtrado, como el número de empresas, el total de reseñas y la calificación promedio. Estos valores brindan una referencia rápida sobre la magnitud y calidad de la información disponible.

El dashboard también incorpora una nube de palabras que destaca los términos más frecuentes en los comentarios, permitiendo identificar rápidamente los temas más mencionados por los usuarios. Complementariamente, se incluye un indicador visual de sentimiento que resume la percepción global mediante un dial que clasifica el sentimiento promedio como negativo, neutro o positivo.

Además, se ofrece una tabla interactiva que muestra el detalle de cada reseña junto con su clasificación de sentimiento, facilitando la revisión puntual de los comentarios. A esto se suma una visualización de la distribución de calificaciones, que permite analizar la proporción de reseñas según las puntuaciones otorgadas por los usuarios.

Finalmente, el dashboard presenta gráficos temporales que muestran la evolución del número de reseñas por día y la distribución diaria de las calificaciones en formato de barras apiladas. Estas vistas permiten detectar patrones, identificar fechas con mayor actividad y evaluar cómo varía la percepción del usuario a lo largo del tiempo.

En conjunto, estas funcionalidades convierten al dashboard en una herramienta completa para el análisis de opiniones, brindando al usuario una plataforma clara, flexible y orientada al descubrimiento de información relevante.

VI-A. Análisis de la empresa TWOTHIRDS

Para profundizar en el comportamiento de una empresa particular dentro del conjunto de datos, se realizó un análisis focalizado sobre TWOTHIRDS, una marca perteneciente a la categoría *Tienda de ropa*. Esta revisión permite identificar patrones de satisfacción, temas recurrentes en las reseñas y tendencias en la distribución de calificaciones durante el periodo estudiado.

VI-A1. Dinámicas generales de opinión: En el conjunto filtrado correspondiente al año 2025, TWOTHIRDS presenta un volumen moderado de reseñas, pero con una clara inclinación hacia valoraciones positivas. El indicador de sentimiento, basado en modelos de análisis semántico, muestra un promedio superior a 80/100, lo que sugiere una percepción altamente favorable por parte de los clientes.

VI-A2. Principales temas mencionados: La Figura 1 muestra la nube de palabras generada a partir de los títulos y contenidos de las reseñas. Se observa una fuerte presencia de términos asociados a calidad del producto, experiencia de compra y aspectos positivos del servicio. Palabras como *calidad*, *perfecto*, *pedido* y *bien* reflejan que la percepción del usuario se centra principalmente en atributos favorables del producto.

Figura 1. Nube de palabras correspondiente a las reseñas de TWOTHIRDS.

VI-A3. Distribución de calificaciones: La distribución de calificaciones (Figura 2) muestra que los puntajes más frecuentes corresponden a notas altas, destacando especialmente la calificación máxima. Esto confirma una tendencia consistente hacia la satisfacción del cliente. Solo una proporción mínima de reseñas refleja valoraciones neutras o ligeramente negativas.

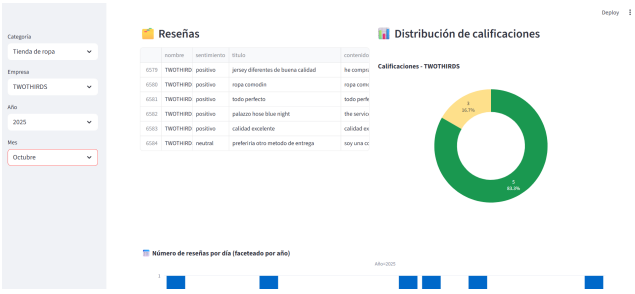

Figura 2. Distribución porcentual de calificaciones otorgadas a TWOTHIRDS.

VI-A4. Interpretación general: En conjunto, las métricas indican que TWOTHIRDS mantiene una reputación sólida basada en la calidad de sus productos y la experiencia positiva de compra. La predominancia de términos favorables en el análisis textual, sumada a un indicador de sentimiento muy alto, sugiere que la marca ha logrado establecer una relación de confianza y satisfacción sostenida con sus clientes. Este patrón se mantiene estable incluso al segmentar por meses o periodos específicos, lo que refuerza la consistencia del desempeño observado.

VI-B. Análisis de la empresa ZARA

Con el propósito de evaluar el comportamiento de otra empresa relevante dentro del conjunto de datos, se realizó un análisis focalizado sobre ZARA, también perteneciente a la categoría *Tienda de ropa*. Este examen permite identificar patrones de satisfacción, temas críticos en las reseñas y tendencias en la distribución de calificaciones registradas durante el periodo de estudio.

VI-B1. Dinámicas generales de opinión: En el conjunto filtrado para el año 2025, ZARA presenta un volumen considerable de reseñas, pero con una marcada inclinación hacia valoraciones negativas. El indicador de sentimiento ubica el promedio por debajo de 20/100, reflejando una percepción claramente desfavorable por parte de los clientes. Esta tendencia evidencia la existencia de experiencias recurrentemente problemáticas asociadas al servicio y a la gestión postventa.

VI-B2. Principales temas mencionados: La Figura 3 muestra la nube de palabras generada a partir de los títulos y contenidos de las reseñas. Se observa una fuerte presencia de términos vinculados a devoluciones, reembolsos, dificultades en la atención y problemas operativos. Palabras como *cliente*, *pedido*, *devolución* y *dinero* reflejan que las interacciones reportadas se centran mayoritariamente en experiencias negativas relacionadas con procesos de compra y postventa.

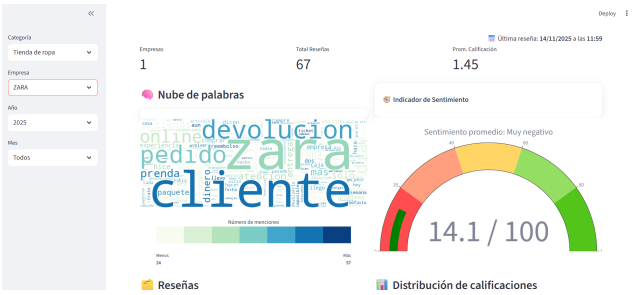

Figura 3. Nube de palabras correspondiente a las reseñas de ZARA.

VI-B3. Distribución de calificaciones: La distribución de calificaciones (Figura 4) revela que la gran mayoría de los puntajes otorgados corresponden a la calificación mínima. Las valoraciones positivas representan un porcentaje reducido y prácticamente marginal. Este comportamiento confirma que la experiencia del cliente con la marca presenta una tendencia predominantemente negativa en el periodo analizado.

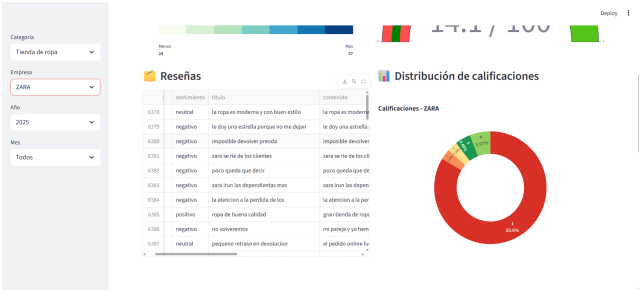

Figura 4. Distribución porcentual de calificaciones otorgadas a ZARA.

VI-B4. Interpretación general: En conjunto, las métricas muestran que ZARA enfrenta una percepción negativa sostenida por parte de sus clientes. La predominancia de términos asociados a problemas operativos y la baja puntuación promedio del indicador de sentimiento sugieren dificultades persistentes en procesos clave, como devoluciones, atención al cliente y manejo de pedidos. La consistencia de esta tendencia a lo largo del periodo evaluado indica que los usuarios experimentan insatisfacción recurrente, lo que afecta directamente la reputación de la marca en el entorno digital.

VII. Resultados y Conclusiones

El proceso desarrollado en este proyecto permitió construir un flujo completo de adquisición, depuración, estructuración y análisis de datos provenientes de reseñas de usuarios en la plataforma Trustpilot. A partir del cuaderno `scrapp.ipynb` y del dashboard implementado en `Streamlit`, se generó un sistema capaz de consolidar información de múltiples categorías, limpiarla de manera consistente y presentar indicadores clave que facilitan la interpretación de la experiencia del cliente.

VII-A. Resultados principales

En primera instancia, se estableció una estructura de almacenamiento basada en un *datalake*, organizada en tres zonas:

Landing, *Refined* y *Consumption*. Esta segmentación permitió mantener una separación clara entre datos sin procesar, datos limpios y datos listos para ser analizados en el dashboard. El cuaderno automatiza la creación de estas carpetas, así como la generación de los archivos finales en formato JSON estructurado.

Durante la etapa de procesamiento, se aplicaron diversas funciones de limpieza para estandarizar campos clave como la puntuación, la fecha y el texto de las reseñas. Se corrigieron formatos, se normalizó la zona horaria y se eliminaron caracteres especiales o ruidos en los campos textuales. Asimismo, se reestructuraron los objetos para uniformar la información a nivel de empresa y reseña, garantizando un dataset ordenado y preparado para su análisis posterior.

Posteriormente, se habilitó un análisis semántico mediante un modelo de sentimiento basado en *transformers*, lo que permitió asignar a cada reseña una etiqueta de sentimiento (positivo, neutral o negativo) y un puntaje numérico asociado. Esta información complementó el análisis cuantitativo tradicional basado en calificaciones, permitiendo evaluar tanto la puntuación explícita del usuario como su percepción textual.

Finalmente, se diseñó un dashboard interactivo que integra todos estos elementos. La herramienta permite filtrar por categoría, empresa, año y mes, y presenta indicadores clave como número de reseñas, promedio de calificaciones, nube de palabras, indicador de sentimiento, tabla de reseñas individuales, distribución de calificaciones y comportamiento diario de la actividad. Esto facilita una exploración visual e intuitiva del comportamiento de cada empresa, permitiendo comparar tendencias y detectar patrones críticos.

VII-B. Conclusiones

Los resultados obtenidos evidencian la eficacia del enfoque implementado para transformar datos crudos provenientes de la web en un sistema analítico completo y funcional. A lo largo del proceso se logró:

- Unificar y estructurar información heterogénea, obteniendo un dataset limpio, validado y consistente.
- Extraer patrones textuales relevantes mediante técnicas de análisis de sentimiento y análisis léxico.
- Identificar diferencias significativas entre empresas dentro de una misma categoría, como se observó en los casos analizados de TWOTHIRDS y ZARA.
- Desarrollar una herramienta flexible e interactiva que facilita la exploración de datos y permite ajustar el análisis a nuevas categorías, periodos o tipos de empresas.

Todo el código desarrollado en este proyecto se encuentra disponible en el repositorio de GitHub [7]. En conjunto, el proceso permitió consolidar un repositorio robusto de información y habilitar un entorno visual que favorece la toma de decisiones basada en evidencia. El diseño modular aplicado posibilita su futura ampliación mediante la integración de nuevos modelos de análisis, mayor automatización o incorporación de otras fuentes de datos.

REFERENCIAS

[1] Le et al., "Predictive model for customer satisfaction analytics in E-commerce sector using machine learning and deep learning." Disponible en: https://doi.org/10.1016/j.jjimei.2024.100295. [Accedido: 14-nov-2025].

[2] Sun et al., "Predicting and explaining customer satisfaction: A deep learning and sentiment analysis of emotional impacts." Disponible en: https://doi.org/10.1016/j.actpsy.2025.105597. [Accedido: 14-nov-2025].

[3] Aftan & Shah, "Using the AraBERT Model for Customer Satisfaction Classification of Telecom Sectors in Saudi Arabia." Disponible en: https://doi.org/10.3390/brainsci13010147. [Accedido: 14-nov-2025].

[4] Zaghloul et al., "Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches." Disponible en: https://doi.org/10.1016/j.jretconser.2024.103865. [Accedido: 14-nov-2025].

[5] Wong & Marikannan, "Optimising e-commerce customer satisfaction with machine learning." Disponible en: https://doi.org/10.1088/1742-6596/1712/1/012044. [Accedido: 14-nov-2025].

[6] Teichert & Shah, "From Reviews to Constructs: Using LLMs to Model Customer Satisfaction in Platform-Based Services." Disponible en: https://doi.org/10.1016/j.jretconser.2025.103182. [Accedido: 14-nov-2025].

[7] D. E. Sanchez, *Proyecto Modelo de Clasificación de Reseñas — Código Fuente*, GitHub, 2025. Disponible en: https://github.com/destandroid/proyecto_modelo_clasificacion_resenas/tree/main

[8] Trustpilot, "Consumer Reviews Platform." Disponible en: https://www.trustpilot.com. [Accedido: 11-nov-2025].

[9] SeleniumHQ, "Selenium WebDriver." Disponible en: https://www.selenium.dev. [Accedido: 11-nov-2025].

[10] Streamlit Inc., "Streamlit Documentation." Disponible en: https://docs.streamlit.io. [Accedido: 11-nov-2025].

[11] HuggingFace, "Transformers Library." Disponible en: https://huggingface.co. [Accedido: 11-nov-2025].

[12] Streamlit, "Streamlit Documentation." Disponible en: https://docs.streamlit.io. [Accedido: 11-nov-2025].

[13] Pandas Software, "Pandas Documentation." Disponible en: https://pandas.pydata.org. [Accedido: 11-nov-2025].

[14] Plotly Technologies Inc., "Plotly Python Graphing Library." Disponible en: https://plotly.com/python. [Accedido: 11-nov-2025].

[15] Natural Language Toolkit, "NLTK Documentation." Disponible en: https://www.nltk.org. [Accedido: 11-nov-2025].

[16] A. Mueller, "WordCloud for Python." Disponible en: https://github.com/amueller/word_cloud. [Accedido: 11-nov-2025].

[17] NumPy Developers, "NumPy Documentation." Disponible en: https://numpy.org. [Accedido: 11-nov-2025].

[18] HuggingFace, "Transformers Library." Disponible en: https://huggingface.co/docs/transformers. [Accedido: 11-nov-2025].

[19] SeleniumHQ, "Selenium WebDriver." Disponible en: https://www.selenium.dev. [Accedido: 11-nov-2025].

[20] Sergey Pirogov, "WebDriver Manager for Python." Disponible en: https://github.com/SergeyPirogov/webdriver_manager. [Accedido: 11-nov-2025].

[21] Pytz Contributors, "pytz: World Timezone Definitions." Disponible en: https://pypi.org/project/pytz. [Accedido: 11-nov-2025].

[22] Python Software Foundation, "json — JSON encoder and decoder." Disponible en: https://docs.python.org/3/library/json.html. [Accedido: 11-nov-2025].

[23] Python Software Foundation, "re — Regular expression operations." Disponible en: https://docs.python.org/3/library/re.html. [Accedido: 11-nov-2025].