Challenges and Solutions in Using Web Scraping and AI for Automated Sentiment Analysis

Resumen– El análisis de sentimientos basado en inteligencia artificial (IA) y *web scraping* es una herramienta clave para extraer y evaluar opiniones en plataformas digitales. Sin embargo, su implementación enfrenta desafíos en la recolección, procesamiento y análisis de datos. Restricciones como bloqueos de *bots*, CAPTCHAs y cambios estructurales en los sitios web dificultan la extracción eficiente de información. Además, los datos suelen ser no estructurados o ruidosos, lo que complica su procesamiento.

La precisión de los modelos de IA también representa un reto, ya que pueden verse afectados por sesgos, ambigüedad lingüística, sarcasmo y variaciones idiomáticas. Para mitigar estos problemas, se emplean APIs oficiales, *proxies*, rotación de IPs y técnicas avanzadas de preprocesamiento. Asimismo, la implementación de aprendizaje profundo y el uso de conjuntos de datos más representativos han optimizado la detección de emociones. Este artículo examina estos desafíos y estrategias para mejorar la precisión del análisis automatizado de sentimientos.

Palabras Clave. Análisis de sentimientos, aprendizaje automático, inteligencia artificial, *Web scraping*

Abstract– Sentiment analysis based on artificial intelligence (AI) and web scraping is a key tool for extracting and evaluating opinions on digital platforms. However, its implementation faces challenges in data collection, processing, and analysis. Restrictions such as bot blocks, CAPTCHAs, and structural changes to websites make it difficult to efficiently extract information. In addition, data is often unstructured or noisy, which complicates its processing. The accuracy of AI models is also a challenge, as they can be affected by bias, linguistic ambiguity, sarcasm, and language variations. To mitigate these problems, official APIs, proxies, IP rotation, and advanced preprocessing techniques are employed. Likewise, the implementation of deep learning and the use of more representative datasets have optimized the detection of emotions. This article examines these challenges and strategies for improving the accuracy of automated sentiment analysis

Keywords. Artificial intelligence, machine learning, Sentiment analysis, *Web scraping*

I. INTRODUCCIÓN

En la era digital, las opiniones se expresan en redes sociales, foros y plataformas comerciales. El uso de *web scraping* y análisis de sentimientos basado en inteligencia artificial (IA) permite extraer información clave para el

marketing, la política y la atención al cliente. No obstante, su implementación enfrenta desafíos técnicos, éticos y legales que requieren estrategias especializadas para garantizar su eficacia y cumplimiento normativo.

El *web scraping* permite la recolección masiva de datos, pero enfrenta restricciones como bloqueos de bots y cambios estructurales en los sitios web. Asimismo, el análisis de sentimientos con IA presenta desafíos en la limpieza y preprocesamiento de datos, la interpretación contextual, el sarcasmo y la precisión de los modelos de clasificación.

A su vez, el *web scraping* se ha consolidado como una técnica clave en los sistemas de recuperación de información, dado que facilita una obtención eficaz de datos de la amplia extensión de la web. Entender los métodos, los obstáculos, técnicos, legales y las implicaciones éticas del *web scraping* resulta fundamental para desarrollar sistemas de recuperación de información responsables y eficientes que proporcionen a los usuarios datos valiosos y pertinentes [1] [2] [3].

El análisis de sentimientos es una herramienta clave en la era digital para comprender emociones y opiniones en línea. Se aplica en marketing, para evaluar la percepción de marca; en redes sociales, para monitorear tendencias; y en estudios de opinión, proporcionando información valiosa que facilita la toma de decisiones en diversas industrias.

En este sentido se presenta este artículo resultado parcial de una investigación que se centra en la problemática de combinar eficazmente *web scraping* e IA para el análisis automatizado de sentimientos. El objetivo principal es identificar los principales obstáculos que surgen al integrar estas dos disciplinas, desde las limitaciones técnicas del *scraping* hasta los desafíos del procesamiento del lenguaje natural. A partir de esta identificación, se busca proponer soluciones prácticas y eficientes que permitan superar dichas barreras y mejorar la precisión y confiabilidad del análisis de sentimientos automatizado.

El artículo analiza los desafíos en la integración de *web scraping* e IA para el análisis automatizado de sentimientos y propone soluciones para mejorar su eficiencia y fiabilidad. Además, explora estrategias innovadoras para optimizar la extracción, procesamiento y análisis de datos textuales en entornos digitales en constante evolución.

II. REVISIÓN DE LA LITERATURA

La información disponible en la web se presenta en datos estructurados, no estructurados y semiestructurados, cada uno con características distintas. Los datos estructurados siguen un formato definido, como bases de datos y hojas de cálculo, facilitando su análisis automatizado. Los datos no estructurados, como publicaciones en redes sociales, videos y

correos electrónicos, carecen de una organización predefinida, lo que dificulta su procesamiento. Por otro lado, los datos semiestructurados, como archivos JSON o XML, contienen etiquetas o metadatos que aportan cierta organización sin ser completamente rígidos los cuales requieren de procesar este tipo de datos [4] [5].

El auge de redes sociales como Facebook, Twitter y YouTube ha transformado la vida humana, permitiendo a los usuarios compartir videos, difundir mensajes y expresar opiniones, lo que fomenta una comunidad global más conectada e interconectada [6].

En consonancia con esto, Internet se erige como el principal canal de comunicación a nivel mundial, facilitando la transmisión de datos entre millones de usuarios y entidades diversas [7]. La vasta cantidad de información que almacena la web, en formatos diversos como páginas web, correos electrónicos y redes sociales, la convierte en la base de datos más amplia y variada jamás creada

Además, La *World Wide Web* es un espacio de información global con millones de datos a los que se puede acceder a través de Internet, donde la extracción información de la web que se denomina extracción de datos web o raspado de datos web [5].

En este sentido el *web scraping* es esencial para la recopilación masiva de datos, el monitoreo de competidores, el seguimiento de tendencias y el análisis de sentimientos, aunque es crucial respetar los términos de servicio y la privacidad del usuario [8]. En el ámbito de la fijación de precios en línea, el análisis del traspaso del tipo de cambio ofrece una comprensión profunda de cómo los cambios en los costos marginales se reflejan en los precios al consumidor, aspecto clave para las estrategias de fijación de precios empresariales [9].

A su vez, la combinación de técnicas de *web scraping* con el procesamiento del lenguaje natural (NLP) ha surgido como un enfoque poderoso para desbloquear información más profunda a partir de datos textuales no estructurados [10]. Las técnicas de PNL se utilizan habitualmente para obtener información y descubrir información relevante que no se haya explotado previamente. La extracción de información, la recuperación o el descubrimiento de conocimientos son algunos de los campos dentro del PNL que intentan transformar datos no estructurados en conocimiento procesable [11].

Por otro lado, *Beautiful Soup* [12], *Selenium* [13] y *SpaCy* [14] son herramientas fundamentales en la extracción y procesamiento de datos web y texto. *Beautiful Soup* facilita el *web scraping* al analizar y extraer información de documentos HTML y XML de manera eficiente. *Selenium* permite la automatización de navegación web, interactuando con páginas dinámicas mediante control de eventos como clics y desplazamientos. Por otro lado, *SpaCy* es una biblioteca avanzada de procesamiento de lenguaje natural (NLP) que optimiza tareas como tokenización, análisis sintáctico y reconocimiento de entidades, mejorando el análisis de sentimientos y la comprensión de texto [15].

La inteligencia artificial ha crecido exponencialmente, con aplicaciones en aprendizaje automático, procesamiento del lenguaje natural y visión por computadora. *Frameworks* como *TensorFlow*, *PyTorch* y *Keras* permiten desarrollar modelos de IA para clasificación, regresión y reconocimiento de imágenes y texto. Su integración con *web scraping* y análisis de datos no solo facilita la extracción y procesamiento de información, sino que también permite predecir tendencias, identificar patrones y generar conocimiento valioso [16]. Estas técnicas optimizan la recolección y análisis de datos estructurados en HTML, mejorando la eficiencia en la extracción inteligente de información [17].

III. DESAFÍOS EN EL USO DE WEB SCRAPING PARA EL ANÁLISIS DE SENTIMIENTOS

Las restricciones de acceso en el uso de *web scraping* se presentarán en tablas, abordando temas como bloqueos de bots, cambios en la estructura de los sitios web, calidad inconsistente de los datos y variabilidad lingüística, factores que dificultan la extracción y el análisis preciso de sentimientos de manera automatizada.

La tabla I están con los desafíos con el acceso y restricciones en *web scraping*:

TABLA I
ACCESOS Y RESTRICCIONES

Desafío	Descripción
robots.txt y políticas de <i>scraping</i> :	Algunos sitios prohíben el <i>scraping</i> (robots.txt).
Políticas de uso restringidas	Términos de servicio pueden prohibir el <i>scraping</i> .
Bloqueo de <u>bots</u> y CAPTCHAs:	Los sitios usan mecanismos para detectar y bloquear
Cambios en la estructura de los sitios web	La volatilidad de las páginas web y la necesidad de adaptabilidad en los <i>scrapers</i> .

En la tabla II, se presentan desafíos del procesamiento de texto.

TABLA II
PROCESAMIENTO DE TEXTO

Desafío	Descripción
Ruido en los datos	Los datos obtenidos pueden contener información irrelevante como anuncios, etiquetas HTML o metadatos.
Normalización de texto	Diferencias en mayúsculas, acentos y variaciones lingüísticas afectan la interpretación del sentimiento.

Eliminación de <i>stopwords</i>	Palabras sin carga semántica pueden generar ruido en el análisis.
Tokenización inconsistente	Fragmentar correctamente palabras y frases es clave para evitar errores en la interpretación.
Manejo de emojis y símbolos	Los emojis pueden cambiar el significado de un mensaje.
Ambigüedad en expresiones	Algunas palabras pueden tener múltiples significados en el contexto.
Lenguaje informal y abreviaciones	Expresiones informales, errores ortográficos o abreviaciones dificultan la interpretación automática.
Idioma y dialectos	Variaciones lingüísticas afectan el análisis al modelo no estar entrenado en distintos idiomas o jergas.

La tabla III expone los desafíos con los modelos de Análisis de Sentimientos.

TABLA III
MODELOS

Desafío	Descripción
Elección del modelo adecuado	Modelos basados en reglas, ML o DL tienen diferentes fortalezas y limitaciones.
Falta de datos etiquetados	Se necesitan grandes volúmenes de datos etiquetados para entrenar modelos precisos.
Desbalance de clases	En <i>datasets</i> con más opiniones neutrales, los modelos pueden sesgarse hacia esa categoría.
Interpretabilidad del modelo	Modelos de <i>Deep Learning</i> pueden ofrecer predicciones precisas pero difíciles de explicar.
<i>Overfitting</i> en el entrenamiento	Modelos sobre ajustados a datos específicos pueden fallar en datos nuevos.

En la tabla IV visualiza los desafíos de Precisión y contexto.

TABLA IV
PRECISIÓN Y CONTEXTO

Desafío	Descripción
Detección de sarcasmo e ironía	La IA tiene dificultades para interpretar el sarcasmo sin información contextual.

Desafío	Descripción
Análisis de contexto	Palabras pueden cambiar de significado según el contexto
Opiniones mixtas en un solo texto	Un texto puede contener opiniones positivas y negativas al mismo tiempo.
Polaridad extrema vs. neutralidad	Detectar matices intermedios en opiniones es un reto.
Influencia del dominio temático	Un modelo entrenado en un dominio específico puede fallar
Sesgo en los datos	Datos recolectados pueden reflejar sesgos y afectar la imparcialidad del análisis.
Evolución del lenguaje	La forma en que las personas expresan opiniones cambia con el tiempo, afectando la precisión.

En la tabla V presenta algunas soluciones y mejoras prácticas ante las restricciones de acceso en el uso de *web scraping*, abordando estrategias como el uso de *proxies*, APIs oficiales, detección de cambios estructurales, limpieza de datos y modelos adaptativos, las cuales optimizan la extracción y mejoran la precisión del análisis automatizado de sentimiento

TABLA V
SOLUCIONES Y MEJORES PRACTICAS

Solución	Descripción
Estrategias avanzadas de <i>web scraping</i>	Uso de proxies y rotación de IPs: Para evitar bloqueos. Scraping con Selenium o Puppeteer: Para sitios con JavaScript dinámico. APIs oficiales: Priorizar el uso de APIs para un acceso más estable y legal.
Técnicas de preprocesamiento avanzadas	Uso de librerías como <i>BeautifulSoup</i> , <i>Scrapy</i> y diccionarios especializados. Detección de sarcasmo e ironía: Implementación de algoritmos específicos
Modelos de IA más robustos	Uso de Transformers (BERT, Roberta, etc.): Modelos de lenguaje contextualizados. El entrenamiento con conjuntos de datos diversos y balanceados. para reducir sesgos.
Evaluación y métricas	Para evaluar el rendimiento de los modelos. Como para asegurar la generalización de los resultados

III. METODOLOGÍA

Para el desarrollo del proyecto de investigación “*Explorando la Fusión: Integración de Web scraping e*

Inteligencia Artificial para Recopilación y Análisis de Datos” origen de esta publicación se trabajan las siguientes fases:

A. Identificación de fuentes de datos

El primer paso fue identificar los sitios web relevantes de los que se extraerán los datos. Estos sitios se seleccionaron en función de su relevancia para el tema de investigación.

B. Diseño e implementación de técnicas de Web scraping

Se trabajó en el diseño e implementación de técnicas de *Web scraping* para extraer datos de los sitios web identificados. Este proceso implica la creación de *scripts* de extracción de datos que puedan navegar por estos sitios web y recopilar la información necesaria.

C. Preprocesamiento de datos

Una vez recopilados los datos, se llevó a cabo un proceso de preprocesamiento para limpiar y estructurar los datos. Esto puede implicar la eliminación de datos irrelevantes o duplicados, la corrección de errores y la transformación de los datos en un formato adecuado para el análisis.

D. Diseño e implementación de técnicas de Inteligencia Artificial

Se estudian diferentes modelos e implementan técnicas de Inteligencia Artificial para analizar los datos recopilados. Esto implica el uso de algoritmos de aprendizaje automático para identificar patrones y tendencias en los datos.

E. Interpretación y presentación de los resultados

Finalmente, los resultados del análisis se interpretan y presentarán de una manera que sea comprensible y útil para los interesados.

Cada uno de estos pasos se lleva a cabo de manera sistemática y rigurosa, asegurando que los resultados de la investigación sean válidos y confiables. Así mismo, se tomarán medidas para garantizar que todas las actividades de investigación cumplan con las normas éticas y legales pertinentes. Actualmente todavía se viene trabajando en las dos últimas fases del desarrollo del proyecto.

El estudio se desarrolla en fases interrelacionadas, empleando *web scraping* e inteligencia artificial para el análisis automatizado de sentimientos. Se identifican fuentes relevantes como sitios web, foros y redes sociales, extrayendo datos mediante bibliotecas como *BeautifulSoup* y *Scrapy*, garantizando eficiencia y cumplimiento de las políticas de privacidad.

A continuación, se emplean algoritmos de procesamiento de lenguaje natural (PLN) basados en IA, como Transformers y modelos preentrenados (BERT, GPT), para analizar el contenido extraído. Estos modelos se ajustan y validan mediante técnicas de aprendizaje supervisado para clasificar el sentimiento expresado en los textos (positivo, negativo, neutral). Además, se aplica un enfoque de análisis semántico para mejorar la precisión del modelo, incorporando correcciones ortográficas y semánticas.

El estudio también explora los principales desafíos asociados con la calidad de los datos obtenidos mediante *web scraping*, como la ambigüedad lingüística, la variabilidad de los lenguajes y los errores en la recolección de datos. Finalmente, se proponen soluciones basadas en el uso combinado de IA, validación manual y herramientas de corrección automatizada.

IV. RESULTADOS

A continuación, se presenta los hallazgos parciales hasta el momento, mostrando tanto avances como dificultades en el proceso. La implementación de *web scraping* y técnicas de inteligencia artificial (IA) ha permitido obtener una base de datos sustancial proveniente de diversas fuentes en línea, incluyendo redes sociales y foros especializados. A través del proceso de *scraping*, se ha logrado recolectar un volumen significativo de texto, aunque se han encontrado dificultades relacionadas con la calidad y consistencia de los datos. Algunas páginas web presentan estructuras irregulares que dificultan la extracción eficiente, lo que requiere ajustes continuos en el código para optimizar la recolección.

La propuesta busca optimizar el análisis de sentimientos para medir la satisfacción del cliente en tecnología y comercio electrónico, mejorando la experiencia del usuario y reduciendo tiempos de respuesta. Para ello, se desarrollará un modelo de clasificación eficiente, priorización de comentarios, análisis de texto y evaluación del modelo para garantizar su efectividad.

Además, las correcciones semánticas y ortográficas han mejorado significativamente la calidad del análisis, reduciendo el número de errores relacionados con la interpretación de palabras mal escritas o empleadas. A pesar de estos avances, se sigue trabajando en optimizar los modelos para mejorar su precisión y adaptabilidad a textos más complejos.

La propuesta se centra en optimizar el análisis de sentimientos para medir la satisfacción de clientes en productos tecnológicos y comercio electrónico, con el objetivo de mejorar la experiencia del usuario y reducir los tiempos de respuesta. Para ello, se establecieron objetivos generales y específicos. Entre los generales, destacan la creación de un modelo de clasificación eficiente y la priorización de comentarios. Los objetivos específicos incluyen la implementación de técnicas de análisis de texto y la evaluación del modelo propuesto, garantizando su efectividad en la detección y clasificación de opiniones.

Para llevar a cabo la investigación, se emplearon diversos métodos, incluyendo la vectorización de textos utilizando TF-IDF y la implementación de modelos de clasificación como *Naive Bayes* Multinomial, Árbol de Decisión, Máquina de Soporte Vectorial (SVM), *Gaussian Naive Bayes*, *Random Forest* y *LinearSVC*. Se realiza una validación cruzada para evaluar el rendimiento de los modelos, utilizando métricas como precisión, *recall* y *f1-score*. Por ahora, los resultados obtenidos muestran que el modelo de Máquina de Soporte

Vectorial logró el mejor rendimiento, con una precisión de 0.95 y un $f1$ -score de 0.95 , evidenciando su capacidad para manejar la clasificación de los comentarios de manera efectiva.

VI. CONCLUSIONES

El uso de *web scraping* e inteligencia artificial (IA) para el análisis automatizado de sentimientos representa una herramienta poderosa para la extracción y evaluación de opiniones en entornos digitales. Sin embargo, su implementación enfrenta múltiples desafíos que van desde restricciones en el acceso a los datos hasta problemas de interpretación semántica y precisión en los modelos de análisis.

Las restricciones impuestas por robots.txt, bloqueos de bots y CAPTCHAs requieren estrategias avanzadas, como el uso de proxies, rotación de agentes de usuario y enfoques basados en APIs oficiales. Asimismo, la variabilidad en la estructura de los sitios web demanda técnicas robustas de adaptación y monitoreo continuo.

El preprocesamiento del texto es esencial en el análisis de sentimientos para eliminar ruido, normalizar el lenguaje y manejar expresiones ambiguas. Además, los modelos de IA deben seleccionarse y entrenarse con datos representativos para reducir sesgos y mejorar la precisión, incluso en casos de sarcasmo o lenguaje coloquial.

En conclusión, la integración efectiva de *web scraping* e IA requiere una combinación de enfoques técnicos y éticos para garantizar la fiabilidad y validez de los resultados en un ecosistema digital en constante cambio.

AGRADECIMIENTO/RECONOCIMIENTO

Los autores agradecen al Instituto Técnico Central de la Escuela Tecnológica (ETITC) el apoyo otorgado para la realización de este trabajo en el marco del proyecto de investigación en curso.

Finalmente, nos gustaría agradecer a los profesores del programa de Doctorado en Ingeniería de Sistemas e Informática de la Universidad de Zaragoza (España) por la orientación y formación recibidas durante nuestra investigación. Su apoyo ha sido fundamental para mejorar sustancialmente nuestra redacción en publicaciones científicas.

REFERENCIAS

- [1] A. Dreyer y J. Stockton, «data scraping”: A primer for counseling clients,» *New York Law Journal*, vol. 7, pp. 1-3, 2013.
- [2] A. Dreyer y J. Stockton, «Internet “data scraping”: A primer for counseling clients,» *New York Law Journal*, vol. 7, pp. 1-3, 2013.
- [3] J. Snell y N. Menaldo, «Web scraping in an era of big data 2.0,» *Bloomberg Law News*, 2016.
- [4] V. Krotov y L. Johnson, «Big web data: Challenges related to data, technology, legality, and ethics,» *Business Horizons*, vol. 66, n° 4, pp. 481-491, 2023.
- [5] A. J. Nagengast, D. Bursian y J. O. Menz, «Dynamic pricing and exchange rate pass-through: Evidence from transaction-level data,» *European Economic Review*, p. 103662, 2021.
- [6] Maranda et al., «State behavioral health agency website references to evidence-based program registers,» *Evaluation and Program Planning*, p. 101906, 2021.
- [7] Kshitija et al., «A Survey on Web Content Mining and Extraction of Structured and Semistructured Data,» *2008 First International Conference on Emerging Trends in Engineering and Technology*, pp. 543-546, 2008.
- [8] Parvez et al., «Analysis Of Different Web Data Extraction Techniques,» *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp. 1-7, 20218.
- [9] V. Krotov y L. Johnson, «Big web data: Challenges related to data, technology, legality, and ethics,» *Business Horizons*, vol. 66, n° 4, pp. 481-491, 2023.
- [10] R. Mijangos-Espinosa, H. Estrada-Esquivel y Y. H. Pérez, «Uso de técnicas de Web Scraping para obtención automática de bases de datos en la Web,» *Res. Comput. Sci*, pp. 143-157, 2022.
- [11] Pichiyani et al., «Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis,» *Procedia Computer Science*, vol. 230, pp. 193-202, 2023.
- [12] «Beautiful Soup Documentation,» Beautiful Soup 4.12.0, 2024. [En línea]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. [Último acceso: 30 1 2025].
- [13] Selenium, «Selenium automates browsers. That's it!,» 2025. [En línea]. Available: <https://www.selenium.dev/>. [Último acceso: 30 1 2025].
- [14] spaCy, «Industrial-Strength Natural Language Processing,» 2025. [En línea]. Available: <https://spacy.io/>. [Último acceso: 30 1 2025].
- [15] A. Duygu, *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*, Packt Publishing, 2021.
- [16] Chandradeep et al., «Web Scraping Techniques and Its Applications,» *A Review. 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pp. 1-8, 2023.
- [17] M. Ranjani, «An intelligent web spider for online e-commerce data extraction,» *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 332-339, 2018.