UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Desarrollo y Gestión de Software

Extracción de Conocimiento en Bases de Datos

II.3. Reporte de solución de caso de estudio de técnicas de limpieza de datos

IDGS91N

PRESENTA:

T.S.U. Hugo Uriel Chaparro Estrada

DOCENTE:

Enrique Mascote

Chihuahua, Chih., 12 sept 2025

1. Introducción

La limpieza de datos es una etapa crítica en cualquier proyecto de análisis de datos, ya que garantiza la calidad, integridad y coherencia de la información que se utilizará para la toma de decisiones. En la actualidad, el volumen de datos disponibles ha aumentado exponencialmente gracias a las redes sociales, los dispositivos móviles, sensores inteligentes y plataformas digitales. Sin embargo, no todos estos datos están preparados para un análisis inmediato. Muchos contienen errores, inconsistencias, valores ausentes o formatos inadecuados, lo que puede afectar gravemente los resultados del análisis y conducir a interpretaciones erróneas.

Este reporte tiene como objetivo presentar un caso de estudio detallado sobre el proceso de limpieza de datos recolectados desde plataformas sociales, específicamente opiniones de usuarios sobre un servicio de entrega a domicilio. A través del análisis de la procedencia de los datos, la clasificación de sus tipos y fuentes, y la aplicación de técnicas de limpieza, se pretende demostrar la importancia de este proceso como paso previo esencial al análisis de datos.

2. Procedencia de los datos

El conjunto de datos utilizado en este caso de estudio fue recolectado a partir de diversas plataformas sociales y web, incluyendo Twitter, Facebook y Google Maps. La recolección se llevó a cabo durante un periodo de quince días y consistió en la recopilación automatizada de opiniones de usuarios sobre la empresa "FastBox Delivery", una compañía ficticia que ofrece servicios de entrega de paquetes a nivel nacional.

Estos datos provienen de distintas fuentes clasificadas según su origen:

- **Datos generados por humanos**: Se refiere a información que ha sido producida voluntariamente por usuarios, como comentarios, publicaciones, valoraciones y reseñas. Este tipo de datos representa directamente las percepciones, experiencias y sentimientos de las personas hacia el servicio recibido. Por ejemplo, frases como “Muy rápido y seguro” o “Mi pedido nunca llegó” fueron extraídas como parte del dataset.

- **Datos web y de redes sociales**: Incluyen todos los textos, imágenes, puntuaciones y metadatos extraídos desde sitios web públicos o APIs. Estos datos son generados en entornos digitales y están sujetos a variaciones constantes debido a la naturaleza dinámica de las plataformas.

- **Datos máquina a máquina (limitados)**: Aunque no constituyen la mayoría del dataset, algunos metadatos como la hora exacta de la publicación o el dispositivo usado se generan automáticamente por los sistemas informáticos y complementan la información principal.

Esta combinación de datos humanos y digitales permite realizar un análisis completo del comportamiento del usuario, pero también representa un reto, ya que los datos generados en redes sociales tienden a ser informales, no estructurados y con alta presencia de ruido o irrelevancia.

La procedencia de los datos también implica ciertas responsabilidades éticas, especialmente cuando se trabaja con información que puede ser sensible o identificable. Por lo tanto, en este estudio se aplicaron técnicas de anonimización de nombres de usuario y se excluyeron comentarios con datos personales explícitos.

3. Tipos y fuentes de datos

El conjunto de datos analizado presenta una estructura heterogénea, lo que significa que contiene diferentes tipos de variables y orígenes. A continuación, se detallan sus clasificaciones según dos criterios fundamentales: **por tipo de dato** y **por fuente de recolección**.

3.1 Clasificación por tipo de dato

- **Datos cualitativos no estructurados**: Constituyen la mayoría del conjunto y se refieren a los comentarios escritos por los usuarios. Estos textos son altamente subjetivos y requieren técnicas de procesamiento del lenguaje natural para su análisis. Frases como “Estoy decepcionado con la atención al cliente” o “Excelente trato del repartidor” son ejemplos de este tipo de dato.

- **Datos cuantitativos discretos**: Incluyen elementos como el número de “me gusta”, compartidos o retweets, así como las calificaciones en estrellas. Son datos numéricos, contables, que permiten análisis estadísticos directos.

- **Datos nominales**: Variables categóricas sin orden específico, como el género del usuario (si se conoce), la ciudad desde donde comenta o el tipo de dispositivo (Android, iOS, PC).

- **Datos ordinales**: Presentes en escalas como la de estrellas (1 a 5), donde los valores tienen un orden inherente. También se puede inferir un orden en la clasificación del sentimiento de los comentarios (positivo, neutro, negativo).

- **Datos temporales**: Fechas y horas de publicación, fundamentales para observar patrones de comportamiento a lo largo del tiempo.

- **Datos geoespaciales**: Cuando se incluyen ubicaciones, permiten hacer análisis por región o zona.

3.2 Clasificación por fuente

- **Fuentes primarias**: Recolección directa desde las plataformas mediante APIs (por

ejemplo, la API de Twitter o Google Places) o técnicas de scraping. Son los datos más
cercanos a su origen y, por lo tanto, los más confiables en términos de fidelidad al usuario original.

- **Fuentes secundarias**: Cuando se utilizan bases de datos públicas ya curadas por otros investigadores o plataformas que comparten datasets para análisis de texto. Estos datos pueden estar preprocesados, pero requieren revisión para asegurarse de su calidad.

La mezcla de datos estructurados (fechas, estrellas, cantidad de likes) y no estructurados (texto libre) requiere un enfoque multidisciplinario, combinando estadística, minería de datos y procesamiento de lenguaje natural.

4. Técnicas de limpieza de datos

Una vez recolectados los datos, se identificaron diversos problemas que debían ser resueltos antes de proceder al análisis. A continuación, se describen las técnicas aplicadas y su fundamentación:

4.1 Manejo de valores nulos

Una de las primeras acciones fue identificar campos con valores faltantes. Por ejemplo, algunas publicaciones no incluían calificación en estrellas o ubicación. Para manejar estos casos se aplicaron:

- Eliminación de registros incompletos en variables clave.
- Imputación con la media o moda para variables numéricas no críticas. ●

Análisis del impacto de la eliminación en la representatividad del conjunto.

4.2 Detección de duplicados

Muchos usuarios publicaron el mismo comentario más de una vez o con ligeras variaciones.

Para identificar duplicados se usaron funciones de comparación exacta y técnicas de similitud de texto como la distancia de Levenshtein. Estos duplicados fueron eliminados para evitar distorsiones en la frecuencia de ciertas opiniones.

4.3 Corrección de errores de formato

Los comentarios extraídos presentaban caracteres mal codificados, palabras mal escritas y símbolos irrelevantes. Se aplicaron:

- Normalización de texto (uso de `unicodedata` en Python para estandarizar caracteres).
- Eliminación de símbolos, signos y URLs irrelevantes.
- Corrección ortográfica básica usando librerías como `textblob`.

4.4 Manejo de outliers

En el análisis exploratorio se detectaron usuarios con actividad anormal, como publicar más de 50 comentarios en una hora o calificaciones extremas sin justificación textual. Estos casos se clasificaron como valores atípicos y se evaluó su exclusión del conjunto.

4.5 Estandarización y transformación de variables

Para facilitar el análisis posterior:

- Las fechas se transformaron al formato ISO 8601.
- Las ciudades se unificaron bajo una misma nomenclatura.
- Las calificaciones fueron transformadas en categorías sentimentales (“positiva”, “neutral”, “negativa”) para su uso en modelos de análisis de sentimiento.

La limpieza de datos se realizó utilizando herramientas como **Pandas**, **NumPy**, **NLTK** y **Scikit-learn**, lo que permitió automatizar procesos y mantener consistencia.

5. Conclusiones

La limpieza de datos es una fase esencial dentro del ciclo de análisis de datos. Este caso de estudio demuestra que, a pesar de contar con grandes volúmenes de información disponibles en plataformas sociales, la calidad de los datos puede estar comprometida por errores humanos, automatismos, formatos inconsistentes y ausencia de validación.

A través de técnicas bien definidas, es posible transformar un conjunto de datos caótico en una fuente confiable de información, apta para análisis estadístico, minería de textos o entrenamiento de modelos de aprendizaje automático. Además, este proceso permite descubrir problemas sistemáticos en la generación de datos que pueden corregirse desde el origen.

El trabajo con datos provenientes de redes sociales representa un reto técnico y ético, pero también ofrece una valiosa oportunidad para obtener insights reales sobre la percepción del cliente y la reputación de marca en el entorno digital.

Referencias

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.

- Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

- Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.

- Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. Wiley-Interscience.