Análisis de Sentimientos

Autor:
Christian Servando Garza González

Programa:
Procesamiento y Clasificación de Datos

Fecha:
23 de enero de 2025

Índice general

Capítulo 1

Introducción

El análisis de sentimientos es una técnica que permite interpretar y clasificar automáticamente la actitud de los usuarios frente a determinado producto o tema, basándose en el lenguaje natural utilizado en sus reseñas. De esta manera, se pueden obtener indicadores valiosos sobre satisfacción, críticas recurrentes y aspectos que los jugadores consideran relevantes.

En este reporte, se presentará el proceso de análisis de sentimientos aplicado a un conjunto de reseñas de videojuegos obtenido de un archivo denominado `Video_Games.jsonl.gz`. Haremos uso de librerías de Python, como `pandas`, `nltk`, y el léxico AFINN para cuantificar la polaridad del texto (sentimientos positivos o negativos) de las opiniones. Posteriormente, se relacionarán dichos valores de sentimiento con la calificación (*rating*) proporcionada en cada reseña, con el fin de investigar la existencia de patrones o correlaciones significativas.

Los objetivos principales de este estudio son:

- Determinar la distribución de los sentimientos en las reseñas de un videojuego específico (filtrando por `parent_asin`).
- Explorar la correlación entre la calificación (*rating*) y la valoración léxica de cada reseña.
- Identificar posibles discrepancias entre el *rating* otorgado y el sentimiento expresado en el texto.
- Visualizar y comparar las palabras más frecuentes en reseñas con sentimientos muy positivos y muy negativos, así como en calificaciones altas y bajas.

Capítulo 2

Descripción de los Datos

El conjunto de datos utilizado proviene de un archivo llamado `Video_Games.jsonl.gz`, el cual contiene reseñas de videojuegos en formato JSON (*JavaScript Object Notation*). A continuación, se describen las principales características de dichos datos:

2.1. Volumen y Estructura Interna

Cada registro (*row*) dentro del archivo incluye diversos campos, algunos de los cuales pueden variar entre reseñas debido a diferencias en las fuentes de donde provienen. Sin embargo, se identificaron las siguientes columnas más relevantes:

- `parent_asin`: Identificador principal del producto (juego).
- `title`: Título del videojuego o del artículo reseñado.
- `rating`: Calificación numérica que otorga el usuario al videojuego (del 1 al 5).
- `text`: Comentario o reseña escrita por el usuario.
- `images`: Lista de URLs o metadatos relacionados con imágenes asociadas a la reseña.

Adicionalmente, el conjunto de datos puede contener otros campos (como `reviewerID`, `helpful` o fechas de la reseña). Sin embargo, para el objetivo

de este análisis se han priorizado las columnas mencionadas, especialmente aquellas relacionadas con la calificación numérica (`rating`) y el contenido textual (`text`), indispensable para el proceso de *sentiment analysis*.

2.2. Filtros Aplicados

Dado que el estudio se centra en un título específico (por ejemplo, *Elden Ring*) o en un subconjunto de videojuegos de interés, se aplican diversos filtros para refinar el conjunto de datos:

- Se seleccionan solo aquellos registros que contengan en el `title` el nombre del videojuego de interés, sin distinguir mayúsculas o minúsculas (`case-insensitive`).
- De igual forma, se pueden filtrar registros por `parent_asin` específico, de modo que el análisis no se mezcle con reseñas de otros productos.

Tras la aplicación de estos criterios, el número de reseñas se reduce a un subconjunto más gestionable y coherente con los objetivos de este trabajo, evitando así la contaminación de resultados con datos de videojuegos diferentes.

Capítulo 3

Metodología

En esta sección se describe el procedimiento seguido para llevar a cabo el análisis de sentimientos sobre las reseñas de videojuegos. El proceso se ha dividido en varias etapas, cada una de las cuales se expone a continuación:

3.1. Limpieza de Datos

Las reseñas en texto crudo (`text`) a menudo contienen *ruido* que dificulta su análisis. Por ello, se aplicaron varios pasos de limpieza:

1. **Eliminación de etiquetas HTML y caracteres especiales:** Se sustituyeron secuencias como `<br />` y otros marcadores por espacios en blanco.

2. **Remoción de caracteres no alfanuméricos:** Se filtraron signos de puntuación y símbolos que no aportaban significado al sentimiento.

3. **Estandarización de mayúsculas y minúsculas:** Todo el texto se convirtió a minúsculas para un tratamiento más homogéneo.

3.2. Tokenización y Remoción de Stopwords

Una vez limpio el texto, se procedió a la *tokenización*, es decir, la división de cada reseña en palabras individuales (*tokens*). Posteriormente, se eliminaron las denominadas *stopwords* o palabras muy comunes (por ejemplo, *the, is, a, an*), que en la práctica no ofrecen valor semántico para el análisis de

la polaridad. Para esta labor se emplearon las funciones y diccionarios de *stopwords* disponibles en la biblioteca `nltk` de *Python*.

3.3. Asignación de Sentimiento con Diccionario Léxico (AFINN)

Con las reseñas ya tokenizadas:

- Se utilizó el diccionario AFINN, que asocia a cada palabra del inglés un valor en una escala que va originalmente de -5 (muy negativo) hasta $+5$ (muy positivo).
- Para estandarizar el análisis, se transformó dicha escala a un rango de 1 a 5. La fórmula a utilizar proviene de la transformación lineal estándar para llevar un intervalo a otro. En el caso de AFINN, cada palabra tiene un puntaje que va desde -5 (muy negativo) hasta $+5$ (muy positivo); con ello, se tiene un intervalo de longitud 10. Si deseamos reescalar esos valores a un nuevo rango de 1 a 5 (longitud 4), aplicamos la siguiente fórmula:

$$\text{sentimiento\_transformado} \;=\; 1 \;+\; \frac{(\text{sentimiento} + 5)}{2{,}5},$$

  donde sentimiento es el valor de AFINN y sentimiento_transformado el nuevo valor normalizado.

Cada palabra de la reseña se mapeó a su valor léxico. Para cada reseña, se calculó posteriormente un promedio de estos valores, a fin de disponer de un indicador global del sentimiento.

3.4. Cálculo y Análisis de la Correlación

Una vez obtenidos el `rating` y el `sentimiento_transformado` para cada reseña, se calculó la correlación entre ambas variables. Este paso permitió cuantificar el grado en el que la calificación numérica del usuario se relaciona con el sentimiento expresado en su texto. Además, se generaron diferentes representaciones gráficas, como:

- Diagramas de dispersión (*scatter plots*) para observar visualmente posibles patrones o tendencias entre calificación y sentimiento.
- Histogramas para estudiar la distribución tanto del `rating` como de la métrica de sentimiento.
- Boxplots para identificar valores atípicos y observar la variabilidad dentro de cada valor de `rating`.

3.5. Identificación de Valores Atípicos

Durante el análisis, se prestó especial atención a las reseñas consideradas *inconsistentes*, es decir:

- Usuarios que otorgan calificaciones altas (por ejemplo, 4 o 5) pero cuyo texto arroja un sentimiento muy negativo (`sentimiento_transformado` bajo).
- Reseñas con calificación muy baja, pero alto puntaje de sentimiento.

Estos casos pueden obedecer a ironías, uso de lenguaje no convencional, o incluso errores de etiquetado en el texto o la calificación. Se extrajeron ejemplos para estudiarlos en más detalle, y se apuntaron como posibles limitaciones o áreas de mejora de la metodología.

3.6. Nubes de Palabras (*Word Clouds*)

Para complementar el análisis cuantitativo, se elaboraron nubes de palabras basadas en:

1. **Niveles de sentimiento:** Se crearon dos nubes de palabras, una para reseñas positivas y otra para reseñas negativas, empleando un umbral en la escala del `sentimiento_transformado`.
2. **Calificación numérica:** Se generaron nubes de palabras para calificaciones altas (por ejemplo, `rating` ≥ 4) y bajas (`rating` ≤ 2).

Estas nubes permitieron identificar términos frecuentes asociados a cada polaridad o calificación, brindando una perspectiva adicional sobre los aspectos más destacados en la opinión del usuario.

Capítulo 4

Resultados

4.1. Distribución de Calificaciones y Sentimientos

Se inició el estudio describiendo la distribución tanto de la variable `rating` (calificaciones) como de la métrica `sentimiento_transformado`. A continuación, se señalan algunos aspectos relevantes:

- **Calificaciones (`rating`)**: La mayoría de las reseñas se concentraron en valores altos (4 y 5), sugiriendo que el videojuego analizado cuenta en general con una percepción positiva por parte de los usuarios. No obstante, también se observaron un número reducido de reseñas con puntuaciones bajas (1 y 2), reflejando un cierto grado de insatisfacción.
- **Sentimiento Transformado**: Al llevar los valores AFINN de $[-5, 5]$ al rango $[1, 5]$, se obtuvo una distribución con una ligera tendencia hacia valores intermedios y altos. Esto coincide en buena medida con las calificaciones altas, pero hubo casos donde las palabras utilizadas en la reseña denotaban un sentimiento medio o bajo pese a la calificación numérica positiva.

En la Figura 4.1 se muestra un histograma representativo de la variable `sentimiento_transformado`, donde se aprecia su rango de variación y los intervalos en que se concentra la mayor densidad de reseñas. De igual forma, en la Figura 4.2 puede observarse la distribución de las calificaciones otorgadas por los usuarios.

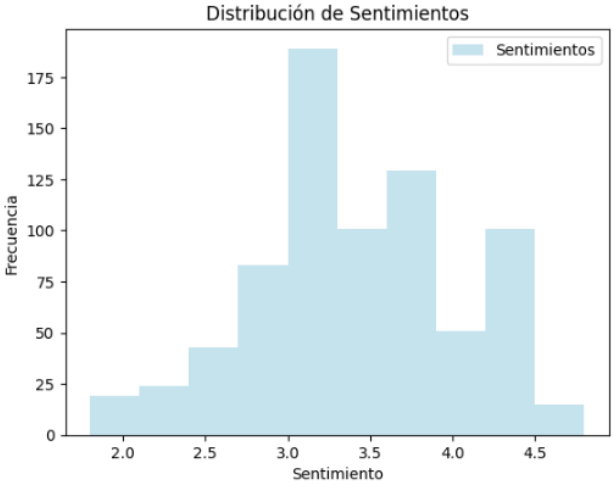

Figura 4.1: Histograma de sentimientos después de aplicar una transformación lineal estándar.

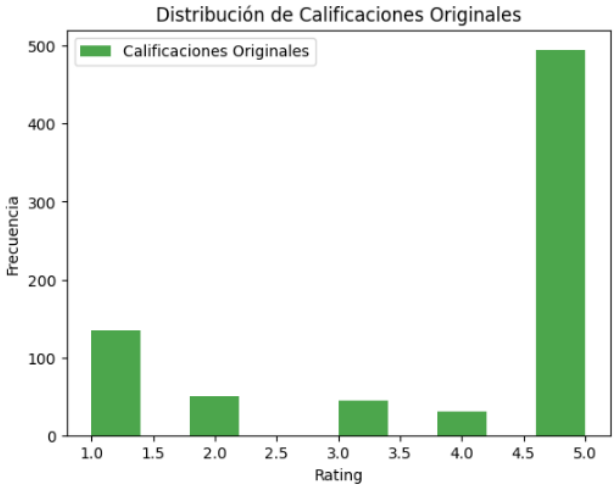

Figura 4.2: Histograma de calificaciones realizadas por los usuarios.

4.2. Correlación entre Calificación y Sentimiento

Para cuantificar la relación entre la puntuación objetiva (`rating`) y la medición subjetiva (`sentimiento_transformado`), se calculó el coeficiente de correlación de Pearson entre ambas variables. El valor resultante fue moderado y positivo (0.47), lo que indica que, en términos generales, a mayores calificaciones numéricas tiende a corresponder un sentimiento más positivo en el texto.

En la Figura 4.3 se presenta un diagrama de dispersión que permite visualizar esta relación. La disposición de los puntos sugiere una tendencia ascendente, aunque se pueden detectar algunos valores atípicos donde la calificación y el sentimiento no coinciden en su polaridad.

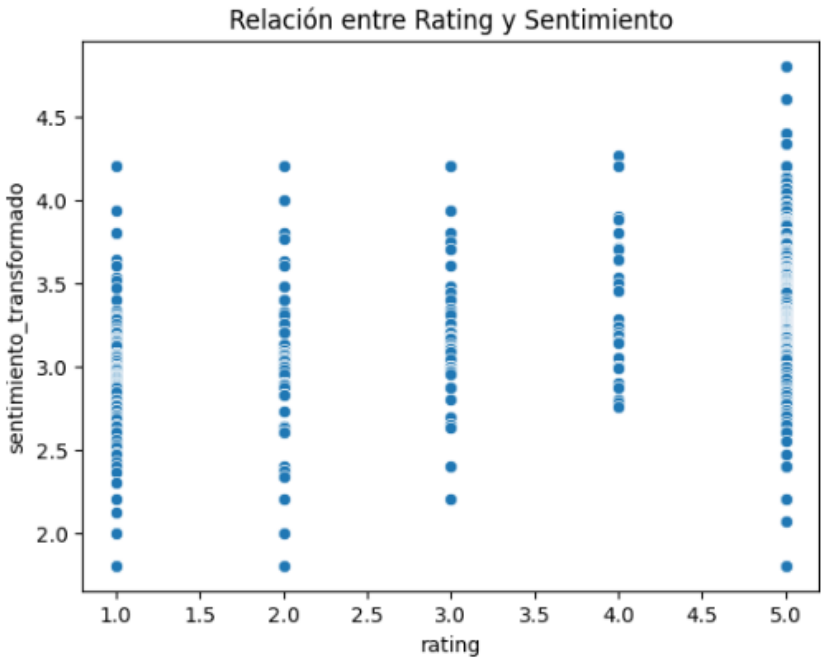

Figura 4.3: Diagrama de dispersión entre calificaciones reales y sentimiento generado.

4.3. Identificación de Outliers

Uno de los objetivos de este análisis era detectar reseñas cuyo valor de `rating` difiere notablemente del sentimiento expresado en su texto. Al filtrar aquellas con `rating` alto (por ejemplo, 4 o 5) pero `sentimiento_transformado` bajo (inferior a 2), se hallaron algunos casos que podrían indicar:

- Lenguaje sarcástico o irónico en la reseña.
- Errores de etiquetado (el usuario eligió una calificación distinta de lo que expresaba).
- Uso de expresiones negativas o de calificativos fuertes para describir aspectos específicos, pese a mantener una valoración global alta del juego.

Se procedió a revisar manualmente un subconjunto de estas reseñas atípicas, confirmando que, en varios casos, el texto sí presentaba un tono más negativo que el reflejado en la calificación.

4.4. Boxplots

Además de los histogramas antes mencionados, se elaboraron *boxplots* (Figura 4.4) para examinar las distribuciones de `sentimiento_transformado` en función de cada categoría de `rating`. En algunos casos, se observaron valores extremos que brindan información sobre la heterogeneidad de las reseñas mencionadas anteriormente:

- **Valoraciones bajas (1 y 2)**: Concentración de valores de sentimiento transformado por debajo de la media, lo que confirma la presencia de palabras con connotaciones negativas en el texto.
- **Valoraciones medias (3)**: Variedad de sentimientos, desde neutros hasta positivos o negativos, reflejando reseñas ambiguas o con opiniones mezcladas.
- **Valoraciones altas (4 y 5)**: La mayoría de las puntuaciones de sentimiento se concentran en la franja positiva; sin embargo, en la zona de valores bajos se hallaron los outliers ya mencionados.

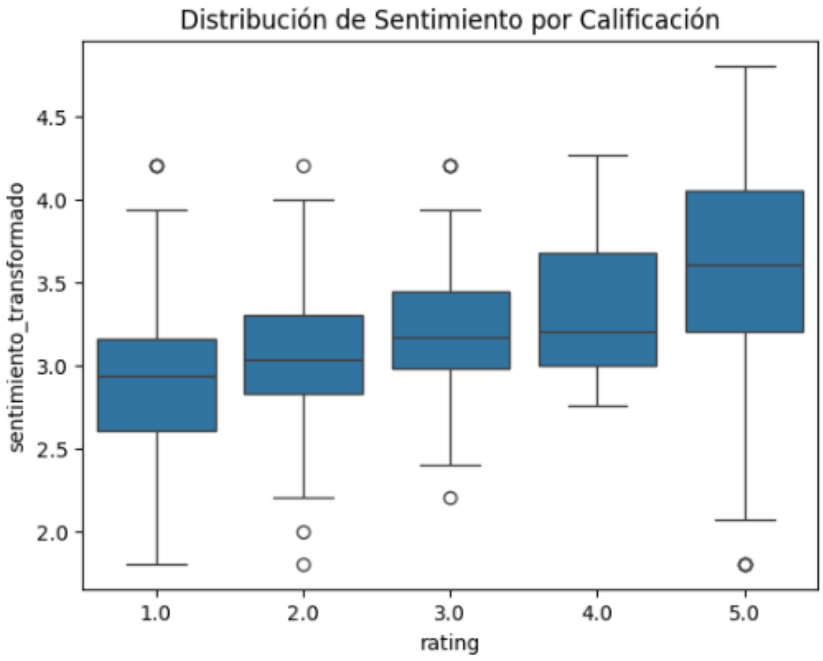

Figura 4.4: Boxplot de sentimiento por calificación

4.4.1. Nubes de Palabras

Con la intención de profundizar en la interpretación cualitativa de los comentarios, se generaron nubes de palabras separadas según:

1. **Sentimiento (positivo vs. negativo):** En las reseñas positivas (`sentimiento` ≥ 4), destacan términos alusivos a la calidad, la diversión y la experiencia inmersiva (palabras como "great","love", "fun"). Por el contrario, en los textos más negativos (`sentimiento` < 3), son frecuentes conceptos vinculados a la frustración o dificultad excesiva (palabras como "level", "time", "hard", çombat").

2. **Calificaciones numéricas (altas vs. bajas):** De manera similar, las palabras presentes en reseñas de `rating` alto se centran en elogios y descripciones favorables, mientras que en las reseñas con `rating` bajo aparecen palabras no tan relacionadas con alguna frustración o dificultad.

En la Figura 4.5 pueden apreciarse visualmente los términos relevantes, proporcionando una aproximación rápida a los aspectos más valorados y los principales focos de crítica.

Figura 4.5: Nubes de palabras basadas en sentimientos y calificaciones reales.

4.5. Hallazgos Relevantes

A modo de síntesis, se destacan los siguientes puntos:

- Existe una correlación positiva moderada entre `rating` y `sentimiento_transformado`, lo que sugiere coherencia general entre la calificación y el tono del comentario.
- Se presentan casos de discrepancias notables, ya sea por posible sarcasmo, reseñas mal etiquetadas o apreciaciones específicas.
- El análisis de la frecuencia de palabras mediante nubes de términos permite identificar rápidamente los factores más influyentes (tanto positivos como negativos) que los usuarios mencionan en sus reseñas.

Capítulo 5

Conclusiones

A manera de cierre, el análisis de sentimientos implementado permitió corroborar que las calificaciones numéricas (`rating`) y la métrica de sentimiento calculada (`sentimiento_transformado`) presentan una correlación positiva moderada. Esto sugiere una coherencia general entre la opinión textual y la valoración otorgada. Sin embargo, también se observaron casos de discrepancias significativas, atribuibles a factores como la ironía, errores de etiquetado o puntos de vista muy específicos. Estos hallazgos revelan la necesidad de implementar modelos más sofisticados para procesar adecuadamente el lenguaje natural en reseñas de videojuegos.

Asimismo, el uso de diversas visualizaciones (histogramas, boxplots, diagramas de dispersión y nubes de palabras) permitió identificar de forma gráfica las tendencias dominantes en las reseñas, destacando términos relevantes tanto en opiniones positivas como negativas. De esta forma, se logró una comprensión más amplia de los factores que inciden en la satisfacción o insatisfacción del usuario.

Finalmente, este estudio sienta las bases para futuras mejoras, como la incorporación de técnicas de aprendizaje profundo o la ampliación del alcance a una mayor variedad de juegos y contextos. Con ello, se busca profundizar en la comprensión de la experiencia de los usuarios y potenciar el valor de la retroalimentación que brindan mediante sus reseñas.