

Análisis de grandes volúmenes de datos

TC4034.10

Proyecto | lectura, escritura, archivos de Big Data PySpark

Equipo 37:

Alejandro Díaz Villagómez	A01276769
Alonso Pedrero Martínez	A01769076
César Iván Pedrero Martínez	A01366501
Emiliano Saucedo Arriola	A01659258

Docentes:

Dr. Iván Olmos Pineda Mtra. Verónica Sandra Guzmán de Valle Mtro. Alberto Daniel Salinas Montemayor

Fecha de entrega: 27 de abril del 2025

Proyecto I lectura, escritura, archivos de Big Data PySpark - Implementación

Tema de interés y propuesta

El análisis de reseñas de productos publicadas en Amazon utilizando herramientas de Big Data, como PySpark, representa una oportunidad valiosa para estudiar el comportamiento del consumidor a gran escala. Amazon, como líder mundial en comercio electrónico y servicios tecnológicos a través de su plataforma AWS (Amazon Web Services), ofrece una base de datos masiva que permite explorar patrones de opinión y tendencias de mercado.

Se propone desarrollar una solución que procese y analice grandes volúmenes de reseñas con el objetivo general de identificar el sentimiento asociado a los productos, clasificándolo en positivo o negativo.

El interés en este tema surge de la necesidad profesional de fortalecer competencias en análisis de datos masivos y procesamiento de lenguaje natural (NLP), así como de la demanda creciente en el mercado por herramientas que faciliten la interpretación de la percepción del consumidor en entornos de venta en línea.

Aplicaciones potenciales:

• Mejora de Productos: Analizar las reseñas puede ayudar a identificar problemas recurrentes o aspectos destacados de los productos, ayudando a los fabricantes a mejorar los diseños o la experiencia del cliente.

• Sistemas de Recomendación: El conjunto de datos puede usarse para mejorar los sistemas de recomendación, correlacionando las calificaciones de los productos, los sentimientos y las preferencias de los clientes.

• Atención al Cliente y Soporte: Identificar rápidamente las reseñas negativas permite a las empresas abordar las preocupaciones de los clientes de manera eficiente, mejorando la satisfacción del cliente.

Dataset:

Selección

Nombre: Amazon Product Reviews

Origen: Este conjunto de datos proviene de Kaggle. Es una colección de reseñas de productos publicadas por clientes de Amazon. El dataset está disponible públicamente en Kaggle para su análisis y exploración.

Tamaño Global:

• Tamaño estimado: Aproximadamente 3.68 GB.

Descripción General

Este conjunto de datos contiene reseñas de productos publicadas en Amazon, con un enfoque principal en productos electrónicos. La información es detallada, incluyendo las calificaciones de los usuarios, comentarios, votos útiles, y otros aspectos que permiten realizar un análisis profundo del comportamiento del consumidor y el sentimiento de las reseñas.

Atributos del Conjunto de Datos

1. marketplace:

- Descripción: La región o país en el que se realizó la compra (por ejemplo, "US").
- Tipo: Texto.
- 2. customer_id:
 - Descripción: Un identificador único para el cliente que dejó la reseña.
- Tipo: Numérico.
- Descripción: Un identificador único para cada reseña.

3. review_id:

- Tipo: Texto.
- 4. product_id:
 - Descripción: Un identificador único para el producto reseñado. • Tipo: Texto.
- 5. product_parent:
 - Descripción: Un identificador de grupo o "padre" para productos similares o relacionados.
- Tipo: Numérico. 6. product title:
 - Descripción: El nombre o título del producto que fue reseñado.
- 7. product_category:
- Descripción: La categoría a la que pertenece el producto (por ejemplo, "PC", "Electronics"). • Tipo: Texto.
- 8. star rating:
- Descripción: La calificación otorgada por el cliente, generalmente en una escala de 1 a 5 estrellas.
- Tipo: Numérico. 9. helpful_votes:
- Descripción: El número de votos que otros usuarios han marcado como útiles para esta reseña.
- 10. total_votes:
 - Descripción: El número total de votos que ha recibido la reseña (útiles o no).
- Tipo: Numérico. 11. vine:
 - Descripción: Indica si la reseña fue parte del programa Vine de Amazon (Y para sí, N para no).
- 12. verified_purchase: • Descripción: Indica si el producto fue verificado como comprado por el cliente (Y para sí, N para no).
- Tipo: Texto.
- 13. review_headline:
- Descripción: El encabezado o título de la reseña proporcionado por el cliente. • Tipo: Texto.
- 14. review_body:
- Descripción: El cuerpo de la reseña, donde el cliente describe su experiencia con el producto. • Tipo: Texto. 15. review_date:
- Descripción: La fecha en que se publicó la reseña.
- Tipo: Fecha. 16. sentiment:
 - Descripción: Un valor binario que indica el sentimiento de la reseña (1 para positivo, 0 para negativo).
 - Tipo: Numérico.

Análisis con PySpark

☐ From this point forward, we will use English for the remainder of the notebook to maintain consistency with code syntax and findings.

Note that to run this notebook, you'll need to install some dependencies. It's HIGHLY suggested to run this command to actually install it in a virtual environment by following this command:

python3 -m venv venv && source venv/bin/activate && pip install findspark pyspark matplotlib seaborn setuptools

```
import findspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, isnan, when, count
from pyspark.sql.types import StringType, DoubleType, FloatType
from os import path
import matplotlib.pyplot as plt
import seaborn as sns
```

spark = SparkSession.builder.master("local[*]").getOrCreate()

spark.conf.set("spark.sql.repl.eagerEval.enabled", True)

findspark.init() findspark.find()

In [1]:

In [2]:

Out[2]: '/Users/cesarivp/Documents/GitHub/TC4034.10-Equipo-37/venv/lib/python3.13/site-packages/pyspark'

In [3]:

spark

Setting default log level to "WARN". To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel). 25/04/27 13:58:06 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable 25/04/27 13:58:06 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

SparkSession - in-memory

SparkContext

Spark UI Version

Out[3]:

v3.5.5 Master

local[*] **AppName** pyspark-shell

In [4]:

PATH = "../files" FILE = "amazon_electronics.csv"

In [5]:

class FileManager(): @staticmethod

def open_csv_file(input_path : str, file_name : str): This method opens a csv file with pyspark

> csv_df = spark.read.csv(path.join(input_path, file_name), header=True,

> > inferSchema=True, multiLine=True,

```
csv df.show(truncate=20)
       return csv df
In [6]:
class StatisticalAnalysisHelper():
   @staticmethod
   def dataset_dimensions(df_input):
       print("columns in the dataset:", len(df input.columns))
       print("rows in the dataset:", df_input.count())
   @staticmethod
   def schema_information(df_input):
       This method shows the current schema of the data.
       df input.printSchema()
   @staticmethod
   def descriptive statistics(df input):
        This method shows the descriptive statistics of the data.
       df_input.summary().show(truncate=False)
   @staticmethod
   def missing_values_table(df_input):
       Displays a table with the count of missing values per column.
       missing exprs = []
        for c in df input.schema.fields:
           field name = c.name
           field_type = c.dataType
           if isinstance(field type, (DoubleType, FloatType)):
               missing exprs.append(
                   count(when(col(field name).isNull() | isnan(col(field name)), field name)).alias(field name)
           elif isinstance(field_type, StringType):
               missing_exprs.append(
                   count(when(col(field name).isNull() | (col(field name) == ""), field name)).alias(field name)
           else:
               missing_exprs.append(
                   count(when(col(field_name).isNull(), field_name)).alias(field_name)
       df missing values = df input.select(missing exprs)
        return df_missing_values
In [7]:
class VisualAnalysisHelper:
    @staticmethod
   def missing_values_plot(df_missing_values):
        This method shows the missing values and plots them.
       df_missing_values_pd = df_missing_values.toPandas()
       df_missing_values_pd = df_missing_values_pd.melt(var_name="column", value_name="missing_count")
       plt.figure(figsize=(10,6))
       sns.barplot(x="column", y="missing count", data=df missing values pd)
       plt.xticks(rotation=45, ha='right')
       plt.title("Missing Values per Column")
       plt.tight_layout()
       plt.show()
   def column distribution plot(df input, column name, x label, y label):
        This method plots the distribution of the 'star_rating' column.
       rating counts = (
           df_input.groupBy(column_name)
            .orderBy(column name)
       plot_data = rating_counts.toPandas()
       plt.figure(figsize=(8, 5))
       sns.barplot(x=column name, y="count", data=plot data)
       plt.xlabel(x label)
       plt.ylabel(y_label)
       plt.title("Distribution of Values")
       plt.xticks(rotation=0)
       plt.tight_layout()
        plt.show()
In [8]:
df_reviews = FileManager.open_csv_file(PATH, FILE)
25/04/27 13:58:20 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users should configure it(them) to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors
                                                                                                                                                                                              review_body|review date|sentiment|
                           review_id|product_id|product_parent|
                                                                       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|
                                                                                                                                                                     review headline|
|marketplace|customer_id|
                                                      335625766|Plemo 14-Inch Lap...|
         US| 22873041|R3ARRMDEGED8RD|B00KJWQIIC|
                                                                                                                                        0 | N |
                                                                                                                                                               Y|Pleasantly surprised|I was very surpri...| 2015-08-31|
               30088427| RQ28TSA020Y6J|B013ALA9LA|
                                                      671157305|TP-Link OnHub AC1...|
                                                                                                  PC|
                                                                                                                           24|
                                                                                                                                       31| N|
                                                                                                                                                               N|OnHub is a pretty...|I am a Google emp...| 2015-08-31|
                                                                                                                                                                                                                             1 |
         US| 20329786| RUXJRZCT6953M|B00PML2GQ8|
                                                      982036237|AmazonBasics USB ...|
                                                                                                  PC|
                                                                                                              1 |
                                                                                                                            2 |
                                                                                                                                        2| N|
                                                                                                                                                               N|None of them work...|Bought cables in ...| 2015-08-31|
                                                                                                                                                                                                                             0 |
         US| 14215710| R7E00U06BPB71|B001NS00Z4|
                                                      576587596|Transcend P8 15-i...|
                                                                                                  PC|
                                                                                                              1 |
                                                                                                                            0 |
                                                                                                                                        0| N|
                                                                                                                                                               Y|just keep searching.|nope, cheap and slow| 2015-08-31|
                                                                                                                                                                                                                             0 |
               38264512|R39NJY2YJ1JFSV|B00AQMTND2|
                                                      964759214|Aleratec SATA Dat...|
                                                                                                  PC|
                                                                                                              5|
                                                                                                                            0 |
                                                                                                                                        0| N|
                                                                                                                                                                          Five Stars|Excellent! Great ... | 2015-08-31|
                                                                                                                                                                                                                             1 |
                                                      170101802|Kingston Digital ...|
               30548466|R31SR7REWNX7CF|B00KX4TORI|
                                                                                                  PC |
                                                                                                              5|
                                                                                                                            0 |
                                                                                                                                        0| N|
                                                                                                                                                               Y|Good quality, wor...|Good quality,work...| 2015-08-31|
                                                                                                                                                                                                                             1|
                                                      206124740|White 9 Inch Unlo...|
                                                                                                              3|
                                                                                                                                        2 | N |
                                                                                                                                                               Y|in fact this is t...|This demn tablet ...| 2015-08-31|
                 589298| RVBP8I1R0CTZ8|B00P17WEMY|
                                                                                                  PC|
                                                                                                                            1|
                                                                                                                                                                                                                             0 |
               49329488|R1QF6RS1PDLU18|B00TR05L9Y|
                                                      778403103|Lenovo TAB2 A10 -...|
                                                                                                  PC |
                                                                                                               4 |
                                                                                                                            1|
                                                                                                                                        1 | N |
                                                                                                                                                                                Good|I am not sure I d...| 2015-08-31|
                                                                                                                                                                                                                             1 |
                                                                                                                                                               Υ|
               50728290|R23AICGEDAJQL1|B0098Y77OG|
                                                      177098042|
                                                                                                  PC|
                                                                                                              1 |
                                                                                                                            0 |
                                                                                                                                        0 |
                                                                                                                                            Ν|
                                                                                                                                                               Y|You get what you ...|After exactly 45 ...| 2015-08-31|
                                                                                                                                                                                                                             0 |
                                                      602496520|AzureWave Broadco...|
               37802374|R2EY3N4K9W19UP|B00IFYEYXC|
                                                                                                  PC|
                                                                                                              5|
                                                                                                                            3 |
                                                                                                                                        4 | N |
                                                                                                                                                               Y|Great for Windows...|Replaced my Intel...| 2015-08-31|
                                                                                                                                                                                                                             1 |
                                                      977217357|HDE Rotating iPad...|
                                                                                                                                                                            One Star|IT HAS ALREADY CR...| 2015-08-31|
         US| 52027882| RC9AW4HKJ016M|B0091ITP0S|
                                                                                                  PC|
                                                                                                              1 |
                                                                                                                            0 |
                                                                                                                                        0 | N |
                                                                                                                                                                                                                             0 |
                                                                                                                                                               Υ|
         US| 41770239|R2ALWJE9N6ZBXD|B008I21EA2|
                                                      295632907|Linksys AC1750 Wi...|
                                                                                                              1|
                                                                                                                            0 |
                                                                                                                                        0 | N |
                                                                                                                                                               N| Very Disappointed|Very disappointed...| 2015-08-31|
                                                                                                                                                                                                                             0 |
                                                                                                  PC|
         US| 42560427|R2G5FPA40X37GV|B00MRB7SB0|
                                                      922591915|iPad Pro 9.7, iPa...|
                                                                                                  PC|
                                                                                                              5|
                                                                                                                            1|
                                                                                                                                        1| N|
                                                                                                                                                                          Five Stars|Works well. I use...| 2015-08-31|
                                                                                                                                                                                                                             1 |
         US| 46345923|R1IKTSEVXSIMOD|B00LLER2CS|
                                                      997551273|SanDisk 16GB CZ43...|
                                                                                                  PC |
                                                                                                              5|
                                                                                                                            0 |
                                                                                                                                        0 | N |
                                                                                                                                                               Y|The encryption so...|The encryption so...| 2015-08-31|
                                                                                                                                                                                                                             1|
         US| 41751192|R2YA6G6SRFEWF6|B00B0CQCCC|
                                                      937999925|TRENDnet Wireless...|
                                                                                                  PC|
                                                                                                              1 |
                                                                                                                            0 |
                                                                                                                                        1 | N |
                                                                                                                                                               Y|Didn't last 2 years.|I have owned this...| 2015-08-31|
                                                                                                                                                                                                                             0 |
         US| 21176481| RS9H1N9I3Z1IA|B00GU8W5AE|
                                                       13865167|Redragon M901 PER...|
                                                                                                  PC|
                                                                                                              5|
                                                                                                                            0 |
                                                                                                                                        0 | N |
                                                                                                                                                               Y|Awesome gaming mouse|My first gaming m...| 2015-08-31|
                                                                                                                                                                                                                             1 |
                                                                                                                            0 |
                                                                                                  PC|
                                                                                                              1 |
                                                                                                                                        0 | N |
                                                                                                                                                                            One Star|I cannot get it t...| 2015-08-31|
         US| 10674058| RKKLBI76VTDNT|B00XHMXJQ0|
                                                      967483469|Mudder MHL Adapte...|
                                                                                                                                                                                                                             0 |
         US| 43341796|R2NJ3WFUS4E5G6|B00YGJJQ6U|
                                                      986548413|Fintie iPad Air 2...|
                                                                                                  PC|
                                                                                                              4 |
                                                                                                                            0 |
                                                                                                                                        0| N|
                                                                                                                                                               Y|Great choices of ...|Love that Finite ...| 2015-08-31|
                                                                                                                                                                                                                             1 |
         USI
               13232866|R21PTQNLGCBN0I|B00XMN20Y6|
                                                      873354048|Fintie iPad 2/3/4...|
                                                                                                  PC|
                                                                                                              5|
                                                                                                                            0 |
                                                                                                                                        0 | N |
                                                                                                                                                                          Five Stars|Nice color, I lov...| 2015-08-31|
                                                                                                                                                                                                                             1|
               29333557|R3G4RT3EQ9RSY7|B00MA40W9I|
                                                      535866197|Egoway® New Lapto...|
                                                                                                  PC |
                                                                                                              1|
                                                                                                                                        0 | N |
                                                                                                                                                               Y|Totally wasted $6...|Totally wasted $6...| 2015-08-31|
                                                                                                                                                                                                                             0 |
                                                                                                                                                              only showing top 20 rows
In [9]:
StatisticalAnalysisHelper.dataset dimensions(df reviews)
columns in the dataset: 16
                                                                  (0 + 1) / 1]
[Stage 3:>
rows in the dataset: 6906564
In [10]:
StatisticalAnalysisHelper.schema_information(df_reviews)
|-- marketplace: string (nullable = true)
|-- customer id: integer (nullable = true)
 |-- review id: string (nullable = true)
 |-- product id: string (nullable = true)
 |-- product parent: integer (nullable = true)
 |-- product_title: string (nullable = true)
 |-- product category: string (nullable = true)
 |-- star rating: integer (nullable = true)
|-- helpful votes: integer (nullable = true)
|-- total votes: integer (nullable = true)
|-- vine: string (nullable = true)
|-- verified purchase: string (nullable = true)
|-- review headline: string (nullable = true)
 |-- review body: string (nullable = true)
|-- review date: date (nullable = true)
```

escape="\"", quote="\""

|-- sentiment: integer (nullable = true)

StatisticalAnalysisHelper.descriptive statistics(df reviews)

In [11]:

[Stage 6:>

|product id |product parent |product category|star rating |summary|marketplace|customer id |review id |product title |helpful votes |total votes |verified purchase|review headline |review body |count |6906564 |6906564 |6906564 |6906564 |6906564 |6906564 |6906564|6906564 |6906564 | 6906564 | 6906564 |6906564 |6906564 | 6906564 |2.799257247625172E7 |NULL |7.784060019910656E9 |4.9166335483220977E8|3063.133004032258 |mean |NULL |NULL |4.086460937739808|1.4844511684826203|1.9620770907212328|NULL |NULL |Infinity |63.88157894736842 |0.7636061 868101128| |stddev |NULL |1.5572197138978068E7|NULL |3.5410264135332127E9|2.8901115508420634E8|13807.905764291976 NULL |1.362854001234071|41.49142556564166 |43.07079218821245 |NULL |NULL |NaN |729.9122389794957 0.4248668 07848865 | |R100036YQULDQT|0012425362 |1602 |! NEW ! SDHC Memory Card (16gb) for NIKON SLR D40 D60|PC 10 |10016 |1 10 | N |min |US |\a\aThe only thing wrong with this monitor is the TN display which after calibration is okay. IPS panels are going to be more true in color and have better off viewing angles. This is a good overall value if you are looking for a multi purpose display. |0| |25% NULL |14419422 |7.303009426E9 |242889999 |700.0 NULL NULL NULL NULL 12.0 11.0 |1 |50% NULL |26791807 NULL |9.875849669E9 |488442804 1700.0 NULL |5 NULL |5.0 |NULL 11.0 |1 |75% |NULL 142659658 NULL |9.966633316E9 1742866380 |3911.0 NULL |5 |1 | 1 |NULL |NULL |350.0 11.0 | 1 1000006025 | 坩MTV 坩ェクロ 01 T―フコンバット ロ_150 ガルム1 117571 ITTC

25/04/27 13:58:34 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.

(0 + 1) / 1]

In [12]:

[Stage 9:>

missing_values = StatisticalAnalysisHelper.missing_values_table(df_reviews)

missing_values.show(truncate=False)

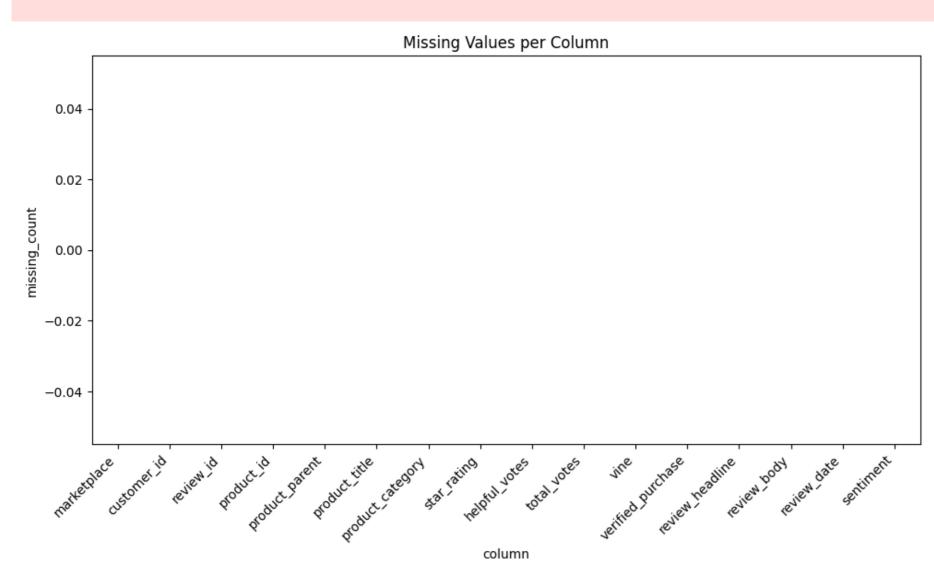
It's worth noting that this dataset seems somewhat suspicious. It doesn't have any missing values, which is great for our use case as it simplifies the analysis by eliminating the need to handle incomplete data.

(0 + 1) / 1]

However, the absence of missing values is unusual in real-world datasets, where incomplete records are often present. This raises the possibility that the dataset might not be entirely authentic or could have been autogenerated. While it's impossible to confirm, the dataset's structure and the perfect data cleanliness suggest that it may not come from a natural source. This could potentially affect the reliability of any analysis performed using this data.

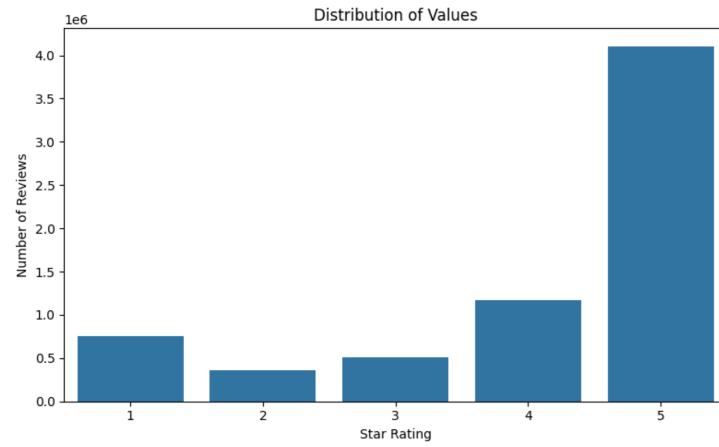
In [13]:

VisualAnalysisHelper.missing_values_plot(missing_values)



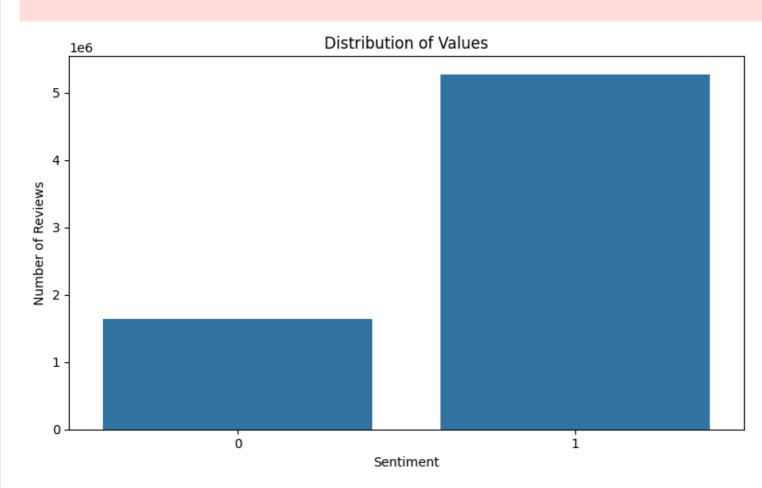
In [14]:

VisualAnalysisHelper.column_distribution_plot(df_reviews, "star_rating", "Star Rating", "Number of Reviews")



In [15]:

VisualAnalysisHelper.column_distribution_plot(df_reviews, "sentiment", "Sentiment", "Number of Reviews")



Despite the potential concerns regarding its authenticity, the dataset does show a certain level of consistency. For instance, as we can see above, the 5-star reviews align well with other attributes, such as the sentiment and the text provided in the review body. Positive reviews typically correlate with high ratings and sentiment scores, which makes sense and supports the integrity of the dataset, at least in this regard. This consistency is crucial for building reliable models, as it indicates that, despite the dataset's suspicious origin, it still follows logical patterns in terms of how reviews and ratings are structured.