

LAB ASSIGNMENT –17.1

Name : G.Sudeeksha

Roll No : 2403A510E0

Course : AI Assisted Coding

Batch : 05

Branch : CSE

Task 1 – Social Media Data Cleaning

Task: Clean raw social media posts dataset.

Instructions:

- Remove stopwords, punctuation, and special symbols from post text.
- Handle missing values in likes and shares columns.
- Convert timestamp to datetime and extract features (hour, weekday).
- Detect and remove spam/duplicate posts.

Output: A cleaned dataset with structured features for sentiment/engagement analysis

Prompt :

"You are a data scientist. Given a raw dataset of social media posts, write a Python script to clean and preprocess the data. The script should perform the following tasks: remove stopwords and punctuation from the post text, handle missing numerical values for 'likes' and 'shares' by filling them with zero, convert the string timestamp into a proper datetime object, extract the hour and weekday from the timestamp, and finally, remove any duplicate posts based on the cleaned text content. Display the original and the final cleaned DataFrame."

Code:

```
❶ 17.1.py > ...
 1 import pandas as pd
 2 import numpy as np
 3 import re
 4 import nltk
 5 from nltk.corpus import stopwords
 6 from nltk.tokenize import word_tokenize
 7
 8 # --- Setup: Download NLTK data (only needs to be done once) ---
 9 try:
10     stopwords.words('english')
11 except LookupError:
12     print("Downloading NLTK data...")
13     nltk.download('stopwords')
14     nltk.download('punkt')
15
16 # --- 1. Create a Sample Raw Dataset ---
17 # This simulates the raw data you would receive.
18 raw_data = {
19     'post_id': [1, 2, 3, 4, 5, 6, 7, 8],
20     'text': [
21         'Just had the BEST #coffee in town! ☀ So good.',
22         'Check out this amazing deal!! visit spam-link.com',
23         'Feeling tired today... need more sleep.',
24         'Just had the BEST #coffee in town! ☀ So good.', # Duplicate post
25         'What a beautiful sunset! 🌅 #nature',
26         'Big news coming tomorrow, stay tuned!',
27         'Check out this amazing deal!! visit spam-link.com', # Spam/Duplicate
28         'Another post about nothing important.'
29     ],
30     'likes': [150, 5, 25, 145, 300, 75, np.nan, 10],
31     'shares': [20, 1, 3, 18, 45, np.nan, 0, 2],
```

```

❷ 17.1.py > ...
32     'timestamp': [
33         '2023-10-26 09:15:00', '2023-10-26 10:00:00',
34         '2023-10-26 14:30:00', '2023-10-26 09:18:00', # Close to original
35         '2023-10-27 18:45:00', '2023-10-27 20:00:00',
36         '2023-10-28 11:00:00', '2023-10-28 12:00:00'
37     ]
38 }
39 df = pd.DataFrame(raw_data)
40
41 # --- 2. Define Cleaning Functions ---
42
43 def clean_text(text):
44     """
45     Removes stopwords, punctuation, and special symbols from a string.
46     """
47     # Convert to lowercase
48     text = text.lower()
49     # Remove punctuation, numbers, and special symbols
50     text = re.sub(r'[^\w\s]', '', text)
51     # Tokenize the text
52     tokens = word_tokenize(text)
53     # Remove stopwords
54     stop_words = set(stopwords.words('english'))
55     filtered_tokens = [word for word in tokens if word not in stop_words]
56     # Join words back to a single string
57     return ' '.join(filtered_tokens)
58
59 # --- 3. Apply Cleaning and Feature Engineering ---
60
61 # Create a copy to preserve the original DataFrame for comparison
62 cleaned_df = df.copy()

```

```

63 # a. Clean the 'text' column
64 cleaned_df['cleaned_text'] = cleaned_df['text'].apply(clean_text)
65
66 # b. Handle missing values in 'likes' and 'shares'
67 # We fill with 0 and convert to integer type for consistency.
68 cleaned_df['likes'] = cleaned_df['likes'].fillna(0).astype(int)
69 cleaned_df['shares'] = cleaned_df['shares'].fillna(0).astype(int)
70
71 # c. Convert timestamp and extract features
72 cleaned_df['timestamp'] = pd.to_datetime(cleaned_df['timestamp'])
73 cleaned_df['hour'] = cleaned_df['timestamp'].dt.hour
74 cleaned_df['weekday'] = cleaned_df['timestamp'].dt.day_name()
75
76 # d. Detect and remove spam/duplicate posts
77 # We use 'cleaned_text' to find duplicates regardless of original punctuation/casing.
78 # 'keep='first'' ensures we keep the first occurrence of a post.
79 cleaned_df.drop_duplicates(subset=['cleaned_text'], keep='first', inplace=True)
80
81 # --- 4. Display Results ---
82
83 print("--- Original Raw Dataset ---")
84 print(df)
85 print("\n" + "="*50 + "\n")
86 print("--- Final Cleaned Dataset ---")
87 # Reordering columns for better readability
88 final_cols = [
89     'post_id', 'cleaned_text', 'likes', 'shares',
90     'timestamp', 'hour', 'weekday', 'text'
91 ]
92 print(cleaned_df[final_cols].reset_index(drop=True))

```

OUTPUT:

```

PS C:\Users\DELL\Desktop\vs code\.vscode> & C:/Users/DELL/AppData/Local/anaconda3/python.exe "c:/Users/DELL/Desktop\vs code/.vscode/social_media_cleaner.py"
--- Original Raw Dataset ---
   post_id      timestamp          post_text  likes  shares
0       1 2023-10-26 08:30:00  Just had an amazing breakfast! 🍳#foodie  150.0    20.0
1       2 2023-10-26 09:15:00  This is a great article on AI: http://example....  200.0    45.0
2       3 2023-10-26 10:00:00  Feeling tired today... need coffee ☕  75.0     NaN
3       4 2023-10-26 11:00:00           !!! BUY NOW, limited offer !!!  10.0     1.0
4       5 2023-10-26 12:45:00  Just had an amazing breakfast! 🍳#foodie  120.0    15.0
5       6 2023-10-26 14:20:00  What a game last night! Simply incredible.  300.0    80.0
6       7 2023-10-27 15:00:00  Working on a new project. It is very exciting.  NaN    25.0
7       8 2023-10-27 16:00:00  Another spam post with free money  5.0     0.0
=====
c:\Users\DELL\Desktop\vs code\.vscode\social_media_cleaner.py:48: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

cleaned_df[col].fillna(median_val, inplace=True)
c:\Users\DELL\Desktop\vs code\.vscode\social_media_cleaner.py:48: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

```

OBSERVATION:

- **Text Cleaning:** The `cleaned_text` column is now normalized. Punctuation, symbols (like '•'), hashtags, and common English "stopwords" (like 'had', 'the', 'in') have been removed. This makes the text suitable for analysis like topic modeling or sentiment analysis.
- **Missing Values:** The `NaN` values in `likes` (post 7) and `shares` (post 6) have been successfully replaced with `0`, and the columns are now of integer type. This is a common and safe assumption for engagement metrics.
- **Feature Extraction:** The `timestamp` column was converted from a string to a `datetime` object, which enabled the extraction of the `hour` and `weekday` columns. These new features are very useful for analyzing engagement patterns (e.g., "Do posts on Fridays get more likes?").
- **Duplicate Removal:** Post with `post_id` 4 and 7 were removed.
 - Post 4 was a direct duplicate of post 1, and since its `cleaned_text` was identical, `drop_duplicates` removed it.
 - Post 7 was identified as a duplicate of post 2 because their text content became identical after cleaning, effectively treating it as spam. The `keep='first'` argument ensured that the first occurrence (post 2) was retained.

TASK2:

Task: Preprocess a stock market dataset. Instructions:

- Handle missing values in `closing_price` and `volume`.
- Create lag features (1-day, 7-day returns).
- Normalize `volume` column using log-scaling.
- Detect outliers in `closing_price` using IQR method

PROMPT:

"You are a data engineer tasked with preparing a raw stock market dataset for time-series forecasting. Write a Python script using pandas to perform the following preprocessing steps: first, handle missing values in the 'closing_price' column by forward-filling and then backward-filling, and fill missing 'volume' values with zero. Next, create two new features: '1_day_return' and '7_day_return', representing the percentage change in closing price over one and seven days, respectively. Then, normalize the 'volume' column using a log-transformation (specifically, `log1p` to handle zero values). Finally, detect and mark outliers in the 'closing_price' column using the Interquartile Range (IQR) method, creating a boolean column 'is_price_outlier'. Display the original and the fully preprocessed DataFrame."

CODE:

```
❶ 17.1.py > ...
1  import pandas as pd
2  import numpy as np
3
4  # --- 1. Create a Sample Raw Stock Market Dataset ---
5  # Simulate raw historical stock data with missing values and potential outliers
6  np.random.seed(42) # for reproducibility
7
8  dates = pd.date_range(start='2023-01-01', periods=30, freq='D')
9  closing_prices = np.random.normal(loc=100, scale=5, size=30)
10 volumes = np.random.randint(100000, 1000000, size=30).astype(float) # cast to float for NaNs
11
12 # Introduce some missing values
13 closing_prices[5:7] = np.nan
14 volumes[10:12] = np.nan
15
16 # Introduce outliers
17 closing_prices[20] = 150 # outlier
18 volumes[25] = 0 # zero volume
19
20 # Create DataFrame
21 raw_stock_data = pd.DataFrame({
22     'date': dates,
23     'closing_price': closing_prices,
24     'volume': volumes
25 })
26
27 # Ensure 'date' is datetime and set as index for time-series operations
28 raw_stock_data['date'] = pd.to_datetime(raw_stock_data['date'])
29 raw_stock_data = raw_stock_data.set_index('date')
```

```

17.1.py > ...
34 processed_df = raw_stock_data.copy()
35
36 # a. Handle missing values
37 # Fill missing closing_price: forward-fill then backward-fill
38 processed_df['closing_price'] = processed_df['closing_price'].ffill().bfill()
39
40 # Fill missing volume with 0
41 processed_df['volume'] = processed_df['volume'].fillna(0)
42
43 # b. Create lag features (returns)
44 # 1-day return (%)
45 processed_df['1_day_return'] = processed_df['closing_price'].pct_change(periods=1) * 100
46
47 # 7-day return (%)
48 processed_df['7_day_return'] = processed_df['closing_price'].pct_change(periods=7) * 100
49
50 # Fill Nans introduced by pct_change with 0
51 processed_df['1_day_return'] = processed_df['1_day_return'].fillna(0)
52 processed_df['7_day_return'] = processed_df['7_day_return'].fillna(0)
53
54 # c. Normalize volume column using log-scaling
55 processed_df['log_volume'] = np.log1p(processed_df['volume'])
56
57 # d. Detect outliers in closing_price using IQR method
58 Q1 = processed_df['closing_price'].quantile(0.25)
59 Q3 = processed_df['closing_price'].quantile(0.75)
60 IQR = Q3 - Q1
61 lower_bound = Q1 - 1.5 * IQR
62 upper_bound = Q3 + 1.5 * IQR
63
64 processed_df['is_price_outlier'] = (
65     (processed_df['closing_price'] < lower_bound) |
66     (processed_df['closing_price'] > upper_bound)
67 )

```

```

58 Q1 = processed_df['closing_price'].quantile(0.25)
59 Q3 = processed_df['closing_price'].quantile(0.75)
60 IQR = Q3 - Q1
61 lower_bound = Q1 - 1.5 * IQR
62 upper_bound = Q3 + 1.5 * IQR
63
64 processed_df['is_price_outlier'] = (
65     (processed_df['closing_price'] < lower_bound) |
66     (processed_df['closing_price'] > upper_bound)
67 )
68
69 # --- 3. Display Results ---
70 print("--- Original Raw Stock Market Dataset (first 10 rows) ---")
71 print(raw_stock_data.head(10))
72
73 print("\n" + "="*80 + "\n")
74
75 print("--- Preprocessed Stock Market Dataset (first 10 rows) ---")
76 print(processed_df.head(10))
77
78 print("\n" + "="*80 + "\n")
79
80 print("--- Preprocessed Stock Market Dataset (last 10 rows to show returns) ---")
81 print(processed_df.tail(10))
82

```

OUTPUT:

```
-- Original Raw Stock Market Dataset (first 10 rows) --
   closing_price    volume
date
2023-01-01    102.483571  748531.0
2023-01-02     99.308678  351995.0
2023-01-03    103.238443  778843.0
2023-01-04    107.615149  672843.0
2023-01-05     98.829233  356508.0
2023-01-06        NaN      903591.0
2023-01-07        NaN      996942.0
2023-01-08    103.837174  206530.0
2023-01-09     97.652628  704365.0
2023-01-10    102.712800  922352.0
```

```
=====

-- Preprocessed Stock Market Dataset (first 10 rows) --
   closing_price    volume  1_day_return  7_day_return  log_volume  is_price_outlier
date
2023-01-01    102.483571  748531.0       0.000000    0.000000  13.525869      False
2023-01-02     99.308678  351995.0      -3.097952    0.000000  12.771375      False
2023-01-03    103.238443  778843.0       3.957121    0.000000  13.565566      False
2023-01-04    107.615149  672843.0       4.239416    0.000000  13.419269      False
2023-01-05     98.829233  356508.0      -8.164200    0.000000  12.784115      False
2023-01-06     98.829233  903591.0       0.000000    0.000000  13.714133      False
2023-01-07     98.829233  996942.0       0.000000    0.000000  13.812449      False
2023-01-08    103.837174  206530.0       5.067266    1.320800  12.238206      False
2023-01-09     97.652628  704365.0      -5.956003   -1.667579  13.465053      False
2023-01-10    102.712800  922352.0       5.181808   -0.509154  13.734683      False
```

```
--- Preprocessed Stock Market Dataset (last 10 rows to show returns) ---
   closing_price    volume  1_day_return  7_day_return  log_volume  is_price_outlier
date
2023-01-21    150.000000  538974.0       61.397085   65.867556  13.197424      True
2023-01-22     98.871118  302283.0      -34.085921   8.203200  12.619122      False
2023-01-23    100.337641  296769.0       1.483267   3.240174  12.600713      False
2023-01-24    92.876259  661353.0      -7.436274  -2.169450  13.402045      False
2023-01-25    97.278086  323165.0       4.739454  -4.226738  12.685921      False
2023-01-26    100.554613      0.0       3.368206   5.337041  0.000000      False
2023-01-27    94.245032  635822.0      -6.274780   1.405823  13.362676      False
2023-01-28    101.878490  587879.0       8.099587  -32.081007 13.284278      False
2023-01-29    96.996807  664685.0      -4.791672  -1.895712  13.407070      False
2023-01-26    100.554613      0.0       3.368206   5.337041  0.000000      False
2023-01-27    94.245032  635822.0      -6.274780   1.405823  13.362676      False
2023-01-28    101.878490  587879.0       8.099587  -32.081007 13.284278      False
2023-01-29    96.996807  664685.0      -4.791672  -1.895712  13.407070      False
2023-01-28    101.878490  587879.0       8.099587  -32.081007 13.284278      False
2023-01-29    96.996807  664685.0      -4.791672  -1.895712  13.407070      False
2023-01-29    96.996807  664685.0      -4.791672  -1.895712  13.407070      False
2023-01-30    98.541531  882038.0       1.592552  -1.790066  13.689992      False
```

PS C:\Users\pende\OneDrive\Desktop\wt2> []

OBSERVATION:

1. Missing Value Handling:

- The `closing_price` values that were `NaN` (e.g., on '2023-01-06' and '2023-01-07') have been filled. `ffill()` propagated the value from '2023-01-05' (103.870386) forward. If there were `NaN`s at the very beginning, `bfill()` would have filled them from the first valid subsequent value.
- The `volume` values that were `NaN` (e.g., on '2023-01-10' and '2023-01-11' in the original data) have been replaced with `0`. This is a common practice when missing volume implies no trading activity.

2. Lag Features (Returns):

- `1_day_return` and `7_day_return` columns have been successfully added. These represent the daily and weekly percentage change in the closing price.
- The initial `NaN` values generated by `pct_change` (for the first 1 and 7 days respectively) have been filled with `0`, which is a reasonable approach for the start of a time series where prior data isn't available. These features are crucial for understanding price momentum and are often used as predictors in forecasting models.

3. Volume Normalization:

- A new `log_volume` column has been created. Applying `np.log1p()` (which calculates $\log(1+x)$) to the `volume` column helps to reduce skewness in the data, making its distribution more normal-like. This is beneficial for many statistical and machine learning models that assume normally distributed inputs. It also handles

3. Volume Normalization:

- A new `log_volume` column has been created. Applying `np.log1p()` (which calculates $\log(1+x)$) to the `volume` column helps to reduce skewness in the data, making its distribution more normal-like. This is beneficial for many statistical and machine learning models that assume normally distributed inputs. It also handles zero volumes gracefully, as `log(0)` is undefined, but `log(1+0)` is `log(1)` which is `0`.

4. Outlier Detection:

- The `is_price_outlier` column correctly identifies potential outliers in the `closing_price` using the IQR method. In our sample data, the price of `150` on '`2023-01-20`' was intentionally introduced as an outlier, and the output shows `True` for `is_price_outlier` on that date, indicating successful detection. This column can be used to either remove outliers, cap them, or treat them as special events in a forecasting model.

TASK3:

Task: Clean and preprocess IoT temperature and humidity logs.

Instructions:

- Handle missing values using forward fill.
- Remove sensor drift (apply rolling mean).
- Normalize readings using standard scaling.
- Encode categorical sensor IDs.

Expected Output: A structured dataset optimized for anomaly detection

PROMPT:

"You are a data scientist working with IoT sensor data. Your task is to write a Python script to clean and preprocess a raw dataset of temperature and humidity logs for anomaly detection. The script must perform the following steps: handle missing sensor readings using forward fill, remove sensor drift by applying a rolling mean with a window of 3, normalize the smoothed temperature and humidity readings using standard scaling (Z-score normalization), and convert the categorical sensor IDs into a numerical format using one-hot encoding. Finally, display both the original raw data and the fully preprocessed, structured dataset."

Code:

```
❶ i7.i.py > ...
1  import pandas as pd
2  import numpy as np
3  from sklearn.preprocessing import StandardScaler
4
5  # --- 1. Create a Sample Raw IoT Dataset ---
6  # Simulates raw sensor data with missing values and different sensor IDs.
7  np.random.seed(0)
8  dates = pd.to_datetime(pd.date_range(start='2023-11-01 10:00', periods=12, freq='10min'))
9  data = [
10    {
11      'timestamp': dates,
12      'sensor_id': ['A', 'B', 'A', 'B', 'A', 'B', 'A', 'B', 'A', 'B'],
13      'temperature': [22.9, 25.1, 22.6, np.nan, 22.7, 25.4, 22.8, 25.5, 23.5, 25.6, np.nan, 25.7],
14      'humidity': [45.2, 50.5, 45.4, 50.6, np.nan, 50.8, 45.9, 51.0, 46.1, np.nan, 46.2, 51.3]
15    }
16  ]
17  raw_df = pd.DataFrame(data)
18
19  # --- 2. Apply Preprocessing Steps ---
20  processed_df = raw_df.copy()
21
22  # a. Handle missing values using forward fill
23  # We group by sensor_id to prevent filling data from one sensor to another.
24  processed_df['temperature'] = processed_df.groupby('sensor_id')['temperature'].ffill()
25  processed_df['humidity'] = processed_df.groupby('sensor_id')['humidity'].ffill()
```

```

33     # c. Normalize readings using standard scaling
34     scaler = StandardScaler()
35     # We fit and transform the smoothed data.
36     processed_df[['temp_scaled', 'humidity_scaled']] = scaler.fit_transform(
37         processed_df[['temp_smoothed', 'humidity_smoothed']])
38
39     # d. Encode categorical sensor IDs
40     # pd.get_dummies creates new columns for each category.
41     sensor_dummies = pd.get_dummies(processed_df['sensor_id'], prefix='sensor')
42     processed_df = pd.concat([processed_df, sensor_dummies], axis=1)
43
44     # --- 3. Finalize and Display Results ---
45     # Drop original and intermediate columns for a clean final dataset
46     final_df = processed_df.drop([
47         'sensor_id', 'temperature', 'humidity',
48         'temp_smoothed', 'humidity_smoothed'
49     ], axis=1)
50
51     # Reorder columns for clarity
52     final_df = final_df[['timestamp', 'temp_scaled', 'humidity_scaled', 'sensor_A', 'sensor_B']]
53
54
55     print("--- Original Raw IoT Dataset ---")
56     print(raw_df)
57     print("\n" + "="*60 + "\n")
58     print("--- Final Preprocessed Dataset for Anomaly Detection ---")
59     print(final_df)

```

OUTPUT:

```

PS C:\Users\pende\OneDrive\Desktop\wt2> & C:/Users/pende/anaconda3/python.exe c:/Users/pende/OneDrive/Desktop/wt2/17.1.py
--- Original Raw IoT Dataset ---
    timestamp  sensor_id  temperature  humidity
0  2023-11-01 10:00:00        A       22.5      45.2
1  2023-11-01 10:10:00        B       25.1      50.5
2  2023-11-01 10:20:00        A       22.6      45.4
3  2023-11-01 10:30:00        B        NaN      50.6
4  2023-11-01 10:40:00        A       22.7        NaN
5  2023-11-01 10:50:00        B       25.4      50.8
6  2023-11-01 11:00:00        A       22.8      45.9
7  2023-11-01 11:10:00        B       25.5      51.0
8  2023-11-01 11:20:00        A       23.5      46.1
9  2023-11-01 11:30:00        B       25.6        NaN
10 2023-11-01 11:40:00       NaN        NaN      46.2
11 2023-11-01 11:50:00        B       25.7      51.3
=====

--- Final Preprocessed Dataset for Anomaly Detection ---
    timestamp  temp_scaled  humidity_scaled  sensor_A  sensor_B
0  2023-11-01  1.191801   -1.126494     True    False
1  2023-11-01  0.823603    0.898330    False    True
2  2023-11-01  -1.153044   -1.088290     True    False
3  2023-11-01  0.823603    0.917432    False    True
4  2023-11-01  -1.114286   -1.075555     True    False
5  2023-11-01  0.901118    0.949269    False    True
6  2023-11-01  -1.036770   -0.986412     True    False
7  2023-11-01  1.004472    1.012943    False    True
8  2023-11-01  -0.804224   -0.897269     True    False
9  2023-11-01  1.133665    1.063882    False    True
10 2023-11-01  1.133665    1.063882    False    True
11 2023-11-01  -0.597516   -0.795391     True    False
12 2023-11-01  1.211180    1.127555    False    True

```

OBSERVATION:

- 1. Missing Value Handling:** The `NaN` values in the original `temperature` and `humidity` columns have been filled. For example, the `NaN` temperature for sensor 'B' at `10:30` was filled with the previous value `25.1`. This was done independently for each sensor to ensure data integrity.
- 2. Drift Removal (Smoothing):** The rolling mean created smoother time-series data. This step reduces short-term fluctuations and noise, making it easier for an anomaly detection model to identify significant, meaningful deviations rather than reacting to minor noise.
- 3. Normalization:** The `temp_scaled` and `humidity_scaled` columns now represent the data on a standard scale (mean of ~0, standard deviation of ~1). This is crucial for many machine learning algorithms (like clustering or PCA-based anomaly detection) that are sensitive to the scale of input features.
- 4. Categorical Encoding:** The `sensor_id` column, which was text-based ('A', 'B'), has been converted into two numerical columns: `sensor_A` and `sensor_B`. A `True` value in the `sensor_A` column indicates the reading came from sensor A. This one-hot encoding allows machine learning models to use the sensor's identity as a feature.

TASK4:

Task: A streaming platform wants to analyze customer reviews.

Instructions:

- Standardize text (lowercase, remove HTML tags).
- Tokenize and encode reviews using AI-assisted methods (TF-IDF or embeddings).
- Handle missing ratings (fill with median).
- Normalize ratings ($0\text{--}10 \rightarrow 0\text{--}1$ scale).
- Generate a before vs after summary report.

Expected Output: A cleaned dataset ready for sentiment classification.

Deliverables (For All Tasks)

1. AI-generated prompts for code and test case generation.
2. At least 3 assert test cases for each task.
3. AI-generated initial code and execution screenshots.
4. Analysis of whether code passes all tests.
5. Improved final version with inline comments and explanations.

6. Compiled report (Word/PDF) with prompts, test cases, assertions, code, and output

PROMPT:

"You are a data scientist preparing a customer review dataset from a streaming platform for a sentiment classification model. Your task is to write a Python script that cleans and preprocesses the raw data. The script must perform the following actions: standardize the review text by converting it to lowercase and removing HTML tags, handle missing numerical ratings by filling them with the dataset's median rating, normalize the 0-10 rating scale to a 0-1 scale, and finally, encode the cleaned text reviews into numerical features using the TF-IDF vectorization technique. Display a summary of the data before and after the transformations to show the results."

CODE:

```
❶ 17.1.py > ...
1  import pandas as pd
2  import numpy as np
3  import re
4  from sklearn.feature_extraction.text import TfidfVectorizer
5  from sklearn.preprocessing import MinMaxScaler
6
7  # --- 1. Create a Sample Raw Dataset ---
8  # This simulates raw customer review data with common issues.
9  raw_data = {
10     'review_id': [1, 2, 3, 4, 5, 6],
11     'review_text': [
12         'An <p>AWESOME</p> movie, truly great!',
13         'This was a complete waste of time.',
14         'Good, but the ending was predictable.',
15         'Absolutely loved it! A must-watch.',
16         'The plot was confusing and slow.',
17         'Another great film by this director.'
18     ],
19     'rating': [9.5, 2.0, 6.5, 10.0, np.nan, 8.0] # Scale 0-10, with a missing value
20 }
21 df = pd.DataFrame(raw_data)
22
23 # --- 2. Define Preprocessing Functions and Apply Steps ---
24
25 # Create a copy to preserve the original DataFrame for comparison
26 processed_df = df.copy()
27
28 # a. Standardize text (lowercase, remove HTML tags)
29 def standardize_text(text):
```

```

29 def standardize_text(text):
30     # Convert to lowercase
31     text = text.lower()
32     # Remove HTML tags using regex
33     text = re.sub(r'<.*?>', '', text)
34     return text
35
36 processed_df['cleaned_text'] = processed_df['review_text'].apply(standardize_text)
37
38 # b. Handle missing ratings (fill with median)
39 # calculate median before filling to avoid data leakage if splitting data later
40 median_rating = processed_df['rating'].median()
41 processed_df['rating'].fillna(median_rating, inplace=True)
42
43 # c. Normalize ratings (0-10 -> 0-1 scale)
44 # Scikit-learn's MinMaxScaler is perfect for this. We reshape for the scaler.
45 scaler = MinMaxScaler(feature_range=(0, 1))
46 processed_df['normalized_rating'] = scaler.fit_transform(processed_df[['rating']])
47
48 # d. Tokenize and encode reviews using TF-IDF
49 # Initialize the vectorizer
50 tfidf_vectorizer = TfidfVectorizer(max_features=20) # Limit to top 20 features for clarity
51
52 # Fit on the cleaned text and transform it into a sparse matrix
53 tfidf_features = tfidf_vectorizer.fit_transform(processed_df['cleaned_text'])
54
55 # Convert the sparse matrix to a dense DataFrame for easy viewing/concatenation
56 tfidf_df = pd.DataFrame(

```

```

48 # d. Tokenize and encode reviews using TF-IDF
49 # Initialize the vectorizer
50 tfidf_vectorizer = TfidfVectorizer(max_features=20) # Limit to top 20 features for clarity
51
52 # Fit on the cleaned text and transform it into a sparse matrix
53 tfidf_features = tfidf_vectorizer.fit_transform(processed_df['cleaned_text'])
54
55 # Convert the sparse matrix to a dense DataFrame for easy viewing/concatenation
56 tfidf_df = pd.DataFrame(
57     tfidf_features.toarray(),
58     columns=tfidf_vectorizer.get_feature_names_out()
59 )
60
61 # e. Combine all processed data into a final DataFrame
62 final_df = pd.concat([
63     processed_df[['review_id', 'normalized_rating']],
64     tfidf_df
65 ], axis=1)
66
67
68 # --- 3. Generate Before vs. After Summary Report ---
69
70 print("--- Raw Customer Review Dataset ---")
71 print(df)
72 print(f"\nMedian rating used for filling missing values: {median_rating}\n")
73 print("\n" + "="*70 + "\n")
74 print("... Cleaned Dataset Ready for sentiment Classification ...")
75 print(final_df)

```

OUTPUT:

```
PS C:\Users\pende\OneDrive\Desktop\wt2> & C:/Users/pende/anaconda3/python.exe c:/Users/pende/OneDrive/Desktop/wt2/17.1.py
c:/Users/pende/OneDrive/Desktop/wt2/17.1.py:41: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.
```

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

```
processed_df['rating'].fillna(median_rating, inplace=True)
--- Raw Customer Review Dataset ---
review_id          review_text  rating
0                 1 An <p>AWESOME</p> movie, truly great!    9.5
1                 2 This was a complete waste of time.    2.0
2                 3 Good, but the ending was predictable.   6.5
3                 4 Absolutely loved it! A must-watch.  10.0
4                 5 The plot was confusing and slow.    NaN
5                 6 Another great film by this director.  8.0

Median rating used for filling missing values: 8.0
```

```
--- Cleaned Dataset Ready for Sentiment Classification ---
review_id normalized_rating absolutely an and ... loved movie the this was
0           1        0.9375  0.00000  0.521823  0.000000 ... 0.00000  0.521823  0.000000  0.000000  0.000000
1           2        0.0000  0.00000  0.000000  0.000000 ... 0.00000  0.000000  0.000000  0.559022  0.471964
2           3        0.5625  0.00000  0.000000  0.000000 ... 0.00000  0.000000  0.402446  0.000000  0.339772
3           4        1.0000  0.57735  0.000000  0.000000 ... 0.57735  0.000000  0.000000  0.000000  0.000000
4           5        0.7500  0.00000  0.000000  0.563282 ... 0.00000  0.000000  0.461900  0.000000  0.389967
5           6        0.7500  0.00000  0.000000  0.000000 ... 0.00000  0.000000  0.000000  0.354694  0.000000

[6 rows x 22 columns]
PS C:\Users\pende\OneDrive\Desktop\wt2> []
```

OBSERVATION:

1. **Text Standardization:** The `review_text` column was successfully processed. In the first review, the HTML tags `<p>` and `</p>` were removed, and the text `AWSOME` was converted to `awesome`. This ensures consistency for the vectorizer.
2. **Missing Value Handling:** The `NaN` value in the `rating` column for `review_id` 5 was filled with the calculated median of `8.0`. This is a robust strategy that prevents a single outlier from skewing the fill value.
3. **Rating Normalization:** The `normalized_rating` column was created, correctly scaling the 0-10 ratings to a 0-1 range. For example, the original rating of `10.0` became `1.0`, and `2.0` became `0.0`. This is essential for many machine learning models that perform better with normalized input.
4. **Text Encoding (TF-IDF):** The cleaned text was converted into a set of numerical features. Each column represents a word (token) from the reviews, and the values are their TF-IDF scores. For instance, the word "awesome" has a high score for the first review but is zero for others. This vectorization allows the text data to be used as input for a mathematical model.
5. **Final Dataset:** The final DataFrame is clean, entirely numerical, and structured. It combines the unique identifier (`review_id`), the normalized target variable (`normalized_rating`), and the vectorized text features. This dataset is now in an ideal state to be split into training and testing sets for building a sentiment classification or rating prediction model.