In [87]:

```
from happytransformer import HappyTextClassification
```

In [88]:

```
happy_tc = HappyTextClassification(model_type="DISTILBERT", model_name="distilbert-base-uncased-finetuned-sst-2-english", num_
```

```
06/27/2023 19:10:30 - INFO - happytransformer.happy_transformer -   Using model: cpu
```

In [89]:

```
result = happy_tc.classify_text('''Estoy muy feliz hoy''')
result.label
```

Out[89]:

```
'POSITIVE'
```

In [90]:

```
import csv

# Path to the CSV file
csv_file = "reviews.csv"

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
data = pd.read_csv("reviews.csv")
```

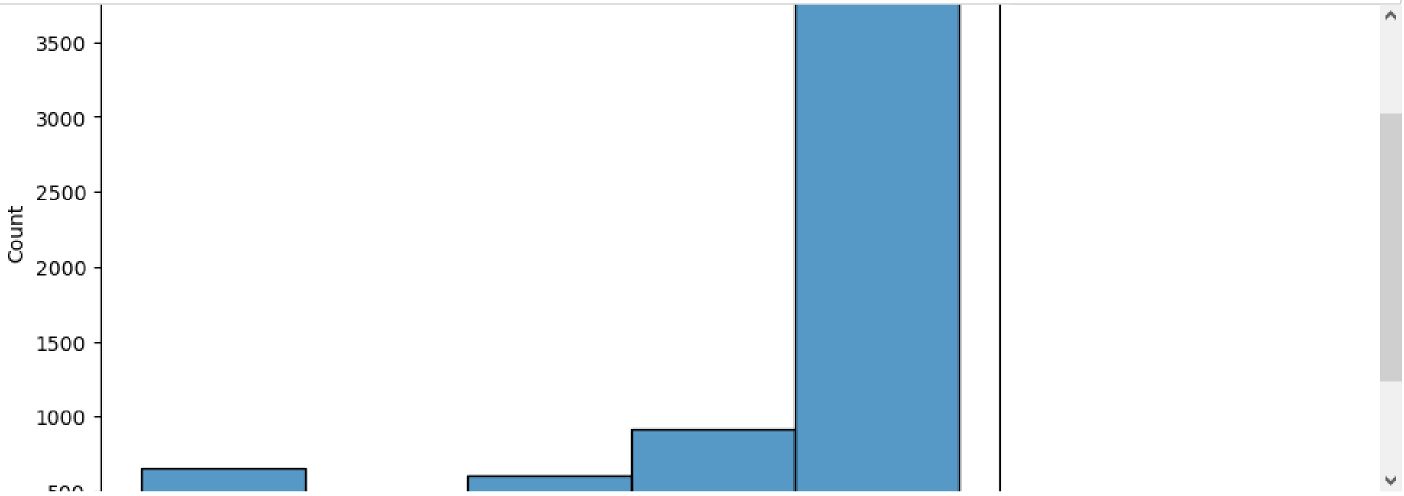
In [91]:

```


# Display the summary statistics
print(data.describe())
print(data.info())
"""
# Create a bar plot of product counts
plt.figure(figsize=(8, 6))
sns.countplot(data['product_name'])
plt.xlabel('Product Name')
plt.ylabel('Count')
plt.title('Bar Plot of Product Counts')
plt.xticks(rotation=90)
plt.show()
"""
# Create a histogram of review ratings
plt.figure(figsize=(8, 6))
sns.histplot(data['review_rating'], bins=5)
plt.xlabel('Review Rating')
plt.ylabel('Count')
plt.title('Histogram of Review Ratings')
plt.show()


```

```
       review_rating
count    6823.000000
mean        4.132493
std         1.336969
min         1.000000
25%         4.000000
50%         5.000000
75%         5.000000
max         5.000000
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6823 entries, 0 to 6822
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   url                6823 non-null   object
 1   product_name       6823 non-null   object
 2   reviewer_name      6823 non-null   object
 3   review_title       6822 non-null   object
 4   review_text        6814 non-null   object
 5   review_rating      6823 non-null   float64
```

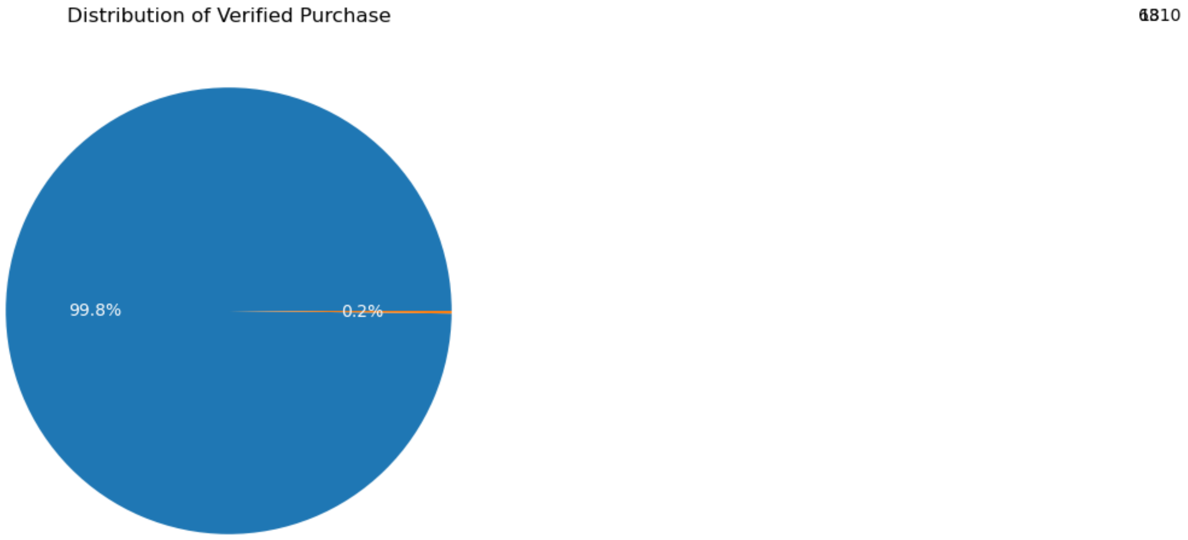
In [92]:

```
# Create a histogram of review ratings
plt.figure(figsize=(8, 6))
sns.histplot(data['review_rating'], bins=5)
plt.xlabel('Review Rating')
plt.ylabel('Count')
plt.title('Histogram of Review Ratings')
plt.show()
```



In [93]:

```


# Calculate the count of each value in the "verified_purchase" column
verified_counts = data['verified_purchase'].value_counts()

# Create a pie plot of "verified_purchase" values with count numbers
plt.figure(figsize=(8, 6))
patches, texts, autotexts = plt.pie(verified_counts, autopct='%1.1f%%', textprops={'color': 'white'})
plt.title('Distribution of Verified Purchase')

# Add count numbers to the pie plot
for i, count in enumerate(verified_counts):
    #angle = (verified_counts.index.get_loc(i) / len(verified_counts)) * 360
    x = 1.3 * 180 / 180 * 3.14  # Adjust the distance of count numbers
    y = 1.3
    plt.text(x, y, f"{count}")

plt.show()
verified_counts
```

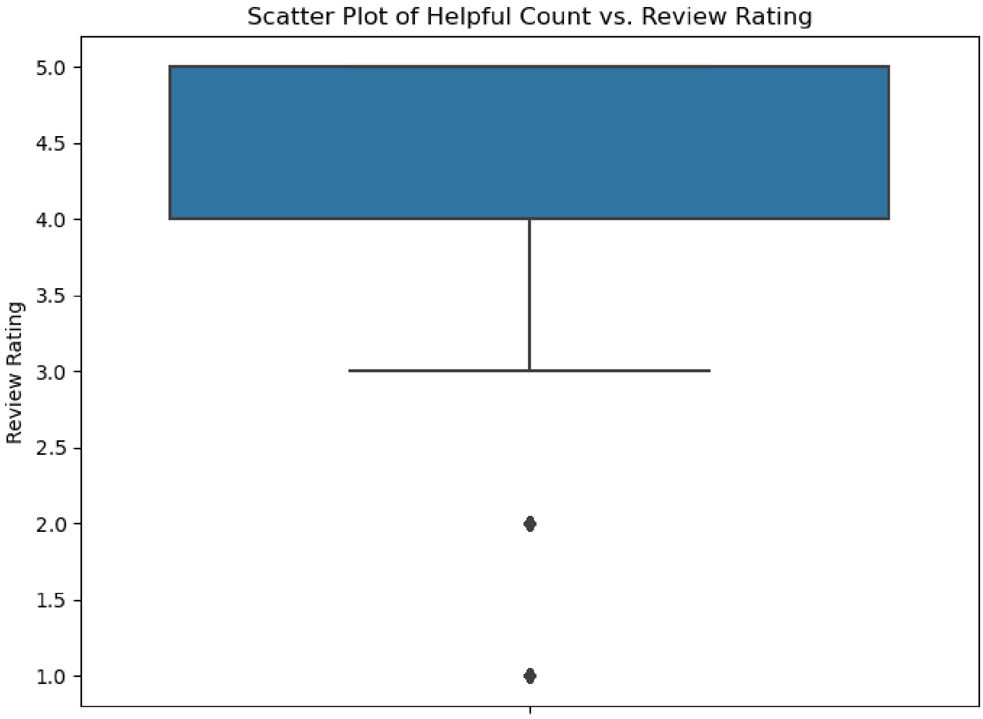


Out[93]:

```
True     6810
False      13
Name: verified_purchase, dtype: int64
```

In [94]:

```
# Create a scatter plot to visualize the relationship between helpful_count and review_rating
plt.figure(figsize=(8, 6))
sns.boxplot( y='review_rating', data=data)

plt.ylabel('Review Rating')
plt.title('Scatter Plot of Helpful Count vs. Review Rating')
plt.show()
```



In [95]:

```
data
```

Out[95]:

url	product_name	reviewer_name	review_title	review_text	review_rating	verified_purchase	review_date	helpful_count	u
/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Jocelyn McSayles	Love em	Love these. Was looking for converses and thes...	5.0	True	Reviewed in the United States on 2 June 2020	2 people found this helpful	36e 2894 d2b330
/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Kenia Rivera	The plastic ripped	The shoes are very cute, but after the 2nd day...	2.0	True	Reviewed in the United States on 28 October 2021	NaN	f47 307 ffce41
/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Chris Souza	Good quality	Good quality	5.0	True	Reviewed in the United States on 20 January 2021	NaN	db5 d40 df4f29
/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Amazon Customer	Good	Great	5.0	True	Reviewed in the United States on 22 April 2021	NaN	75a 646 27d336
dp/B08SW434MG	GUESS Women's Bradly Gymnastics Shoe, White, 7 UK	Graziella	PERFETTE!!	Ho scelto il modello bianco con rifinitura die...	5.0	True	Reviewed in Italy on 2 April 2021	2 people found this helpful	232 849 efb3f4
...	...	...	...	...	...	...	...	...	
dp/B07TPYWFVN	Clarks Vennor Wing, Men's Low-Top Sneakers, Bl...	mauti72	Schick und leicht	Schicker Schuh, läuft sich gut.	5.0	True	Reviewed in Germany on 15 October 2020	NaN	085 fa2 ad4915
dp/B07TPYWFVN	Clarks Vennor Wing, Men's Low-Top Sneakers, Bl...	Charles Lechesnier	EXCELLENT	Mieux que je ne l'imaginais. Très bonne taille...	5.0	True	Reviewed in France on 23 August 2020	NaN	4bf ea7 2aee3f
dp/B084WB2D93	Rohde Men's Tivoli-H Mule, 82 Anthracite, 12.5 UK	Rebecca Lützenkirchen	Einfach schöne Hausschuhe	Habe sie als Geschenk gekauft und sie sind seh...	5.0	True	Reviewed in Germany on 4 October 2021	NaN	5b1 a43 217a17
dp/B084WB2D93	Rohde Men's Tivoli-H Mule, 82 Anthracite, 12.5 UK	Sergej Friedel	Langlebig.	Trage diese Hausschuhe fast zwei Monate jeden ...	5.0	True	Reviewed in Germany on 31 January 2021	NaN	911 98d 16a253
dp/B084WB2D93	Rohde Men's Tivoli-H Mule, 82 Anthracite, 12.5 UK	Swidurski	Hausschuhe für lange kalte Winterzeiten.	Die Hausschuhe sind sehr warm und tolle Leder ...	5.0	True	Reviewed in Germany on 27 January 2021	NaN	5e1 fe1 c9941b

In [106]:

```
import pandas as pd
from datetime import datetime

date = row["review_date"]

# Convert the date strings in the "review_date" column to date format
for index, row in data.iterrows():


# ...

    date = row["review_date"]

# Convert the date to the desired format
    parsed_date = datetime.strptime(date, "%Y-%m-%d")

    formatted_date = parsed_date.strftime("%Y-%m-%d")

# ...


    # Convert the date to the desired format
    parsed_date = datetime.datetime.strptime(date, "%Y-%m-%d")

    formatted_date = parsed_date.strftime("%Y-%m-%d")

    # Update the "review_date" column with the formatted date
    data.at[index, "review_date"] = formatted_date

# Print the updated dataset
print(data.head())
```

```
File D:\run\anaconda\lib\_strptime.py:568, in _strptime_datetime(cls, data_string, format)
    565 def _strptime_datetime(cls, data_string, format="%a %b %d %H:%M:%S %Y"):
    566     """Return a class cls instance based on the input string and the
    567     format string."""
--> 568     tt, fraction, gmtoff_fraction = _strptime(data_string, format)
    569     tzname, gmtoff = tt[-2:]
    570     args = tt[:6] + (fraction,)

File D:\run\anaconda\lib\_strptime.py:349, in _strptime(data_string, format)
    347 found = format_regex.match(data_string)
    348 if not found:
--> 349     raise ValueError("time data %r does not match format %r" %
    350                      (data_string, format))
    351 if len(data_string) != found.end():
    352     raise ValueError("unconverted data remains: %s" %
    353                       data_string[found.end():])

ValueError: time data 'Reviewed in the United States on 2 June 2020' does not match format '%Y-%m-%d'
```

In [113]:

```
import pandas as pd
import datetime

# Assuming 'data' is your DataFrame containing the 'review_date' column

# Function to extract the date from a string with the format "Reviewed in the United States on {day} {month} {year}"
def extract_date(date_string):
    parts = date_string.split(' ')
    day = int(parts[-3])
    month = parts[-2]
    year = int(parts[-1])
    return datetime.datetime(year, datetime.datetime.strptime(month, "%B").month, day)

# Apply the function to extract the date from the strings in the 'review_date' column
data['review_date'] = data['review_date'].apply(extract_date)

# Convert the 'review_date' column to datetime
data['review_date'] = pd.to_datetime(data['review_date'])

# Extract the month from the 'review_date' column
data['review_month'] = data['review_date'].dt.to_period('M')
```

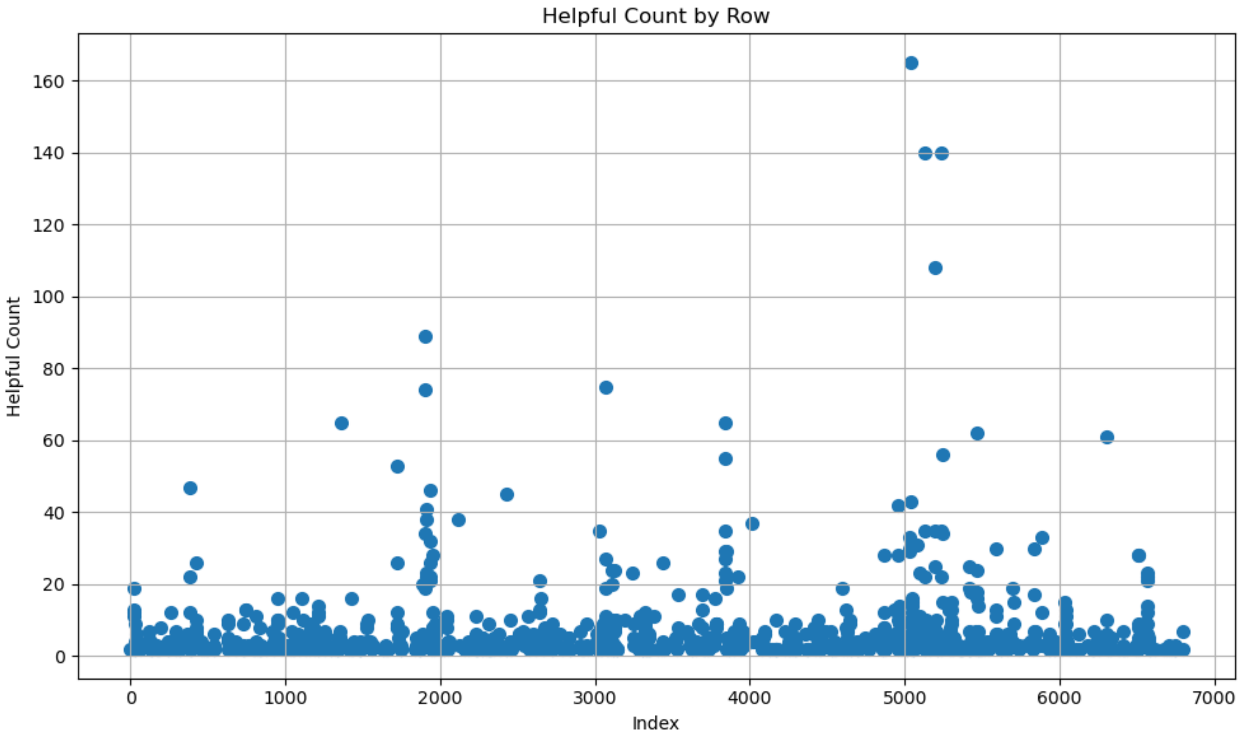
```
C:\Users\Afnan Qasim\AppData\Local\Temp\ipykernel_2964\1530078579.py:15: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vi
ew-versus-a-copy)
  data['review_date'] = data['review_date'].apply(extract_date)
C:\Users\Afnan Qasim\AppData\Local\Temp\ipykernel_2964\1530078579.py:18: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vi
ew-versus-a-copy)
  data['review_date'] = pd.to_datetime(data['review_date'])
C:\Users\Afnan Qasim\AppData\Local\Temp\ipykernel_2964\1530078579.py:21: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vi
ew-versus-a-copy)
  data['review_month'] = data['review_date'].dt.to_period('M')
```

In [117]:

```
data.to_csv('final_review.csv', index=False)
```

In [116]:

```

# Extract the count of helpful people
#data['helpful_count'] = data['helpful_count'].str.extract('(\d+)').astype(float)

# Filter rows with non-null helpful_count values
filtered_data = data[data['helpful_count'].notnull()]

# Create a line graph for helpful_count
plt.figure(figsize=(10, 6))
plt.scatter(filtered_data.index, filtered_data['helpful_count'], marker='o', linestyle='-', linewidth=2)
plt.xlabel('Index')
plt.ylabel('Helpful Count')
plt.title('Helpful Count by Row')
plt.grid(True)
plt.tight_layout()
plt.show()

# Find the highest helpful count
max_helpful_count = filtered_data['helpful_count'].max()

print(f"The highest number of 'found this helpful' is: {max_helpful_count}")
```



```
The highest number of 'found this helpful' is: 165.0
```

In [100]:

```
# Count the occurrences of each product name
product_counts = data['product_name'].value_counts()
z = pd.DataFrame(product_counts)
z
```

Out[100]:

```
                                                                                        product_name
                               Teva K Hurricane 3, Balboa Sodalite Blue, 12 UK Child            10
                                        New Balance Kids&#39; 574v1 Sport Sneaker               10
                  Reebok Women's Princess Sneaker, White/White/Collegiate Royal, 6 UK           10
                                 Propet Women's Ladybug Walking Shoe, Oyster, 11 W US           10
 MSMAX Black Patent Character Mary Jane Flexible Dance Tap Shoes Little Kid Size 11            10
                                                                                 ...            ...
                                      Dr. Brinkmann Women’s Flat Platform Size: 6 UK Blue        1
           adidas Originals Unisex VRX Low Skate Shoe, white/black/white, 4 M US Big Kid         1
                                    Aldo Women's RPPLFROST1B Sneaker, Light Pink, 6 UK           1
   Aigle Unisex Adults Brea Botte Iso Wellington Boots, Blue (Marine New 001), 10.5 UK           1
```

In [101]:

```
d = data
```

In [102]:

```
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
import string


# Define the function to clean the text
def clean_text(text):
    if pd.isnull(text):  # Check for missing values
        return ""

    # Remove punctuation
    text = text.translate(str.maketrans("", "", string.punctuation))

    # Convert to lowercase
    text = text.lower()

    # Tokenize the text
    tokens = word_tokenize(text)

    # Remove stop words
    stop_words = set(stopwords.words("english"))
    tokens = [token for token in tokens if token not in stop_words]

    # Lemmatization
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(token) for token in tokens]

    # Join the tokens back into a single string
    cleaned_text = " ".join(tokens)

    return cleaned_text

# Apply the clean_text function to the 'review_text' column
data['cleaned_text'] = data['review_text'].apply(clean_text)

# Print the updated dataframe
print(d)
```

```
                                         url  \
0     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
1     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
2     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
3     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
4     https://www.amazon.co.uk/dp/B08SW434MG (https://www.amazon.co.uk/dp/B08SW434MG)
...                                      ...
6818  https://www.amazon.co.uk/dp/B07TPYWFVN (https://www.amazon.co.uk/dp/B07TPYWFVN)
6819  https://www.amazon.co.uk/dp/B07TPYWFVN (https://www.amazon.co.uk/dp/B07TPYWFVN)
6820  https://www.amazon.co.uk/dp/B084WB2D93 (https://www.amazon.co.uk/dp/B084WB2D93)
6821  https://www.amazon.co.uk/dp/B084WB2D93 (https://www.amazon.co.uk/dp/B084WB2D93)
6822  https://www.amazon.co.uk/dp/B084WB2D93 (https://www.amazon.co.uk/dp/B084WB2D93)

                                           product_name  \
0     Klasified Women's Transparent Clear Sneaker Sh...
1     Klasified Women's Transparent Clear Sneaker Sh...
2     Klasified Women's Transparent Clear Sneaker Sh...
3     Klasified Women's Transparent Clear Sneaker Sh...
4     GUESS Women's Bradly Gymnastics Shoe, White, 7 UK
...                                                 ...
6818  Clarks Vennor Wing, Men’s Low-Top Sneakers, Bl...
6819  Clarks Vennor Wing, Men’s Low-Top Sneakers, Bl...
6820  Rohde Men's Tivoli-H Mule, 82 Anthracite, 12.5 UK
6821  Rohde Men's Tivoli-H Mule, 82 Anthracite, 12.5 UK
6822  Rohde Men's Tivoli-H Mule, 82 Anthracite, 12.5 UK

              reviewer_name                              review_title  \
0          Jocelyn McSayles                                   Love em
1              Kenia Rivera                        The plastic ripped
2               Chris Souza                              Good quality
3           Amazon Customer                                      Good
4                 Graziella                                PERFETTE!!
...                     ...                                       ...
6818                mauti72                         Schick und leicht
6819     Charles Lechesnier                                 EXCELLENT
6820  Rebecca Lützenkirchen                 Einfach schöne Hausschuhe
6821         Sergej Friedel                                Langlebig.
6822              Swidurski  Hausschuhe für lange kalte Winterzeiten.

                                            review_text  review_rating  \
0     Love these. Was looking for converses and thes...            5.0
1     The shoes are very cute, but after the 2nd day...            2.0
2                                          Good quality            5.0
3                                                 Great            5.0
4     Ho scelto il modello bianco con rifinitura die...            5.0
...                                                 ...            ...
6818                    Schicker Schuh, läuft sich gut.            5.0
6819  Mieux que je ne l'imaginais. Très bonne taille...            5.0
6820  Habe sie als Geschenk gekauft und sie sind seh...            5.0
6821  Trage diese Hausschuhe fast zwei Monate jeden ...            5.0
6822  Die Hausschuhe sind sehr warm und tolle Leder ...            5.0

      verified_purchase                                        review_date  \
0                  True      Reviewed in the United States on 2 June 2020
1                  True  Reviewed in the United States on 28 October 2021
2                  True  Reviewed in the United States on 20 January 2021
3                  True    Reviewed in the United States on 22 April 2021
4                  True                   Reviewed in Italy on 2 April 2021
...                 ...                                                ...
6818               True             Reviewed in Germany on 15 October 2020
6819               True               Reviewed in France on 23 August 2020
6820               True              Reviewed in Germany on 4 October 2021
6821               True             Reviewed in Germany on 31 January 2021
6822               True             Reviewed in Germany on 27 January 2021

      helpful_count                               uniq_id  \
0               2.0  36eae4e5-2894-5279-a0b7-d2b330e2b814
1               NaN  f4778bb8-3070-5cb1-b5aa-ffce41a97b57
2               NaN  db5a7525-d40b-5265-84d8-df4f29837a3b
3               NaN  75a42851-6462-54b5-988a-27d336221943
4               2.0  232dee43-849e-5d06-ba05-efb3f4814714
...             ...                                   ...
6818            NaN  0850eae1-fa2f-59e6-bf30-ad49151bfa20
6819            NaN  4bf117ed-ea7c-517c-967c-2aee3f80ed29
6820            NaN  5b129eb2-a438-5377-9c46-217a177615b2
6821            NaN  91144305-98db-5a55-8ec4-16a253beb811
6822            NaN  5e12b707-fe18-557e-96ba-c9941b4c7690

               scraped_at                                       cleaned_text
0     24/12/2021 02:26:25  love looking converse half price unique– ’ nev...
1     24/12/2021 02:26:25  shoe cute 2nd day wearing tongue started rippi...
2     24/12/2021 02:26:25                                       good quality
3     24/12/2021 02:26:25                                              great
4     24/12/2021 02:26:25  ho scelto il modello bianco con rifinitura die...
```

```
...                   ...                                                ...
6818  24/12/2021 02:29:39                      schicker schuh läuft sich gut
6819  24/12/2021 02:29:39  mieux que je ne limaginais très bonne taille b...
6820  24/12/2021 02:29:39  habe sie al geschenk gekauft und sie sind sehr...
6821  24/12/2021 02:29:39  trage diese hausschuhe fast zwei monate jeden ...
6822  24/12/2021 02:29:39  die hausschuhe sind sehr warm und tolle leder ...

[6823 rows x 12 columns]
```

In [103]:

```
from langdetect import detect
import pandas as pd

# Load your data into a dataframe (assuming your data is already loaded)

# Function to check if the text is in English
def is_english(text):
    try:
        return detect(text) == 'en'
    except:
        return False

# Apply the language detection function to filter non-English rows
data = data[data['review_text'].apply(is_english)]

# Reset the index of the dataframe
d = data.reset_index(drop=True)

# Print the updated dataframe
print(d)
```

```
                                         url  \
0     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
1     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
2     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
3     https://www.amazon.co.uk/dp/B07SBX32T5 (https://www.amazon.co.uk/dp/B07SBX32T5)
4     https://www.amazon.co.uk/dp/B07S1XM3L7 (https://www.amazon.co.uk/dp/B07S1XM3L7)
...                                      ...
3827  https://www.amazon.co.uk/dp/B06XFT2G2F (https://www.amazon.co.uk/dp/B06XFT2G2F)
3828  https://www.amazon.co.uk/dp/B06XFT2G2F (https://www.amazon.co.uk/dp/B06XFT2G2F)
3829  https://www.amazon.co.uk/dp/B06XFT2G2F (https://www.amazon.co.uk/dp/B06XFT2G2F)
3830  https://www.amazon.co.uk/dp/B06XFT2G2F (https://www.amazon.co.uk/dp/B06XFT2G2F)
3831  https://www.amazon.co.uk/dp/B06XFT2G2F (https://www.amazon.co.uk/dp/B06XFT2G2F)

                                           product_name     reviewer_name  \
0     Klasified Women's Transparent Clear Sneaker Sh...  Jocelyn McSayles
1     Klasified Women's Transparent Clear Sneaker Sh...      Kenia Rivera
2     Klasified Women's Transparent Clear Sneaker Sh...       Chris Souza
3     Klasified Women's Transparent Clear Sneaker Sh...   Amazon Customer
4     adidas Women's Retrorun Shoes Running, Core Bl...           Lindsay
...                                                 ...               ...
3827  Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...         Shopper M
3828  Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...   Veronica Franco
3829  Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...   Kindle Customer
3830  Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...   Amazon Customer
3831  Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...               jen

                                         review_title  \
0                                             Love em
1                                  The plastic ripped
2                                        Good quality
3                                                Good
4                         Perfect right outta the box
...                                               ...
3827                          Great for early walkers
3828                                      Three Stars
3829                 Said they were very comfortable.
3830  They are smaller than other shoes the same size
3831              These shoes are great for the price

                                            review_text  review_rating  \
0     Love these. Was looking for converses and thes...            5.0
1     The shoes are very cute, but after the 2nd day...            2.0
2                                          Good quality            5.0
3                                                 Great            5.0
4     True to size. If between I'd probably go with ...            5.0
...                                                 ...            ...
3827  The only shoes (after many tries) that worked ...            5.0
3828            Too narrow hard to get on for a toddler            3.0
3829  My son loves them. Said they were very comfort...            5.0
3830  Size 8 but they are smaller than the size 7 my...            2.0
3831  These shoes are great for the price. Been lovi...            4.0

      verified_purchase                                        review_date  \
0                  True       Reviewed in the United States on 2 June 2020
1                  True   Reviewed in the United States on 28 October 2021
2                  True   Reviewed in the United States on 20 January 2021
3                  True     Reviewed in the United States on 22 April 2021
4                  True              Reviewed in Canada on 20 October 2021
...                 ...                                                ...
3827               True   Reviewed in the United States on 8 December 2017
3828               True      Reviewed in the United States on 23 June 2018
3829               True       Reviewed in the United States on 6 July 2018
3830               True  Reviewed in the United States on 27 September ...
3831               True   Reviewed in the United States on 17 October 2017

      helpful_count                               uniq_id  \
0               2.0  36eae4e5-2894-5279-a0b7-d2b330e2b814
1               NaN  f4778bb8-3070-5cb1-b5aa-ffce41a97b57
2               NaN  db5a7525-d40b-5265-84d8-df4f29837a3b
3               NaN  75a42851-6462-54b5-988a-27d336221943
4               NaN  b64632c5-6f24-51eb-9275-6614fed29f1a
...             ...                                   ...
3827            NaN  9b9e6d15-a4b1-57c0-bebd-d58b115b4ada
3828            NaN  66dc36c0-94b0-5aeb-a618-ce4d14f22740
3829            NaN  2d29c209-b745-5af9-bdde-0dc116387290
3830            NaN  c39f0eec-32e7-567b-a280-505e10a0a4fa
3831            NaN  c079d22a-0ad1-514f-9937-d650598f7c7d

               scraped_at                                       cleaned_text
0     24/12/2021 02:26:25  love looking converse half price unique— ’ nev...
1     24/12/2021 02:26:25  shoe cute 2nd day wearing tongue started rippi...
2     24/12/2021 02:26:25                                       good quality
3     24/12/2021 02:26:25                                              great
4     24/12/2021 02:26:25  true size id probably go lower end ie 885 go 8...
```

```
...                   ...                                         ...
3827  24/12/2021 02:29:38  shoe many try worked early walker bitty foot e...
3828  24/12/2021 02:29:38                       narrow hard get toddler
3829  24/12/2021 02:29:38                          son love said comfortable
3830  24/12/2021 02:29:38  size 8 smaller size 7 son outgrowing disappointed
3831  24/12/2021 02:29:38  shoe great price loving skechers shoe son two ...

[3832 rows x 12 columns]
```

In [104]:

```
d = pd.DataFrame(d)
d
```

3828	https://www.amazon.co.uk/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Veronica Franco	Three Stars	Too narrow hard to get on for a toddler	3.0	True	Reviewed in the United States on 23 June 2018
3829	https://www.amazon.co.uk/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Kindle Customer	Said they were very comfortable.	My son loves them. Said they were very comfort...	5.0	True	Reviewed in the United States on 6 July 2018
3830	https://www.amazon.co.uk/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Amazon Customer	They are smaller than other shoes the same size	Size 8 but they are smaller than the size 7 my...	2.0	True	Reviewed in the United States on 27 September ...
3831	https://www.amazon.co.uk/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	jen	These shoes are great for the price	These shoes are great for the price. Been lovi...	4.0	True	Reviewed in the United States on 17 October 2017

In [50]:

```
a = d.to_csv('final_review.csv', index=False)
```

In [52]:

```
d
```

Out[52]:

url	product_name	reviewer_name	review_title	review_text	review_rating	verified_purchase	review_date	helpful_count	
k/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Jocelyn McSayles	Love em	Love these. Was looking for converses and thes...	5.0	True	2020-06-02	2.0	36 289 d2b330
k/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Kenia Rivera	The plastic ripped	The shoes are very cute, but after the 2nd day...	2.0	True	2021-10-28	NaN	f4 307 ffce41
k/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Chris Souza	Good quality	Good quality	5.0	True	2021-01-20	NaN	db d40 df4f29
k/dp/B07SBX32T5	Klasified Women's Transparent Clear Sneaker Sh...	Amazon Customer	Good	Great	5.0	True	2021-04-22	NaN	75 646 27d336
k/dp/B07S1XM3L7	adidas Women's Retrorun Shoes Running, Core Bl...	Lindsay	Perfect right outta the box	True to size. If between I'd probably go with ...	5.0	True	2021-10-20	NaN	b6 6f2 6614f
...	...	...	...	...	...	...	...	...	
k/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Shopper M	Great for early walkers	The only shoes (after many tries) that worked ...	5.0	True	2017-12-08	NaN	9b a4b d58b11
k/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Veronica Franco	Three Stars	Too narrow hard to get on for a toddler	3.0	True	2018-06-23	NaN	66 94b ce4d1
k/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Kindle Customer	Said they were very comfortable.	My son loves them. Said they were very comfort...	5.0	True	2018-07-06	NaN	2d b7 0dc116
k/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Amazon Customer	They are smaller than other shoes the same size	Size 8 but they are smaller than the size 7 my...	2.0	True	2017-09-27	NaN	c3 32e 505e1
k/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	jen	These shoes are great for the price	These shoes are great for the price. Been lovi...	4.0	True	2017-10-17	NaN	c0 0a d6505

In [47]:

```
d1 = d['review_text']
d1
```

Out[47]:

```
0       Love these. Was looking for converses and thes...
1       The shoes are very cute, but after the 2nd day...
2                                            Good quality
3                                                   Great
4       True to size. If between I'd probably go with ...
                              ...                        
3832    The only shoes (after many tries) that worked ...
3833              Too narrow hard to get on for a toddler
3834    My son loves them. Said they were very comfort...
3835    Size 8 but they are smaller than the size 7 my...
3836    These shoes are great for the price. Been lovi...
Name: review_text, Length: 3837, dtype: object
```

grouping and saving data for chunks according to review star

In [59]:

```
import pandas as pd

# Assuming you have a DataFrame named 'data' with a column 'review_star'

# Group the DataFrame by 'review_star'
grouped_data = data.groupby('review_rating')

# Iterate over the groups and save each group in separate CSV files
for group_name, group_data in grouped_data:
    filename = f"group_{group_name}.csv"  # Generate a unique filename for each group
    group_data.to_csv(filename, index=False)
```

In [ ]:

```
from happytransformer import HappyTextClassification
import pandas as pd
```

In [64]:

```



# Load the DistilBERT tokenizer
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')




# Initialize the maximum sequence length variable
max_seq_length = 0

# Iterate over the 'review_text' column
for review in data['review_text']:
    # Tokenize the review text
    tokens = tokenizer.encode(review, add_special_tokens=True)
    # Update the maximum sequence length if necessary
    max_seq_length = max(max_seq_length, len(tokens))

# Print the maximum sequence length
print("Maximum sequence length:", max_seq_length)
```

```
Token indices sequence length is longer than the specified maximum sequence length for this model (655 > 512). R
unning this sequence through the model will result in indexing errors

Maximum sequence length: 655
```

In [ ]:

```
from happytransformer import HappyTextClassification
import pandas as pd

# Assuming you have a list of filenames for each group
file_names = ["group_1.0.csv", "group_2.0.csv", "group_3.0.csv", "group_4.0.csv", "group_5.0.csv"]

# Initialize the HappyTextClassification model
happy_tc = HappyTextClassification(model_type="DISTILBERT", model_name="distilbert-base-uncased-finetuned-sst-2-english", num_

# Set the maximum sequence length for the model
max_seq_length = happy_tc.tokenizer.model_max_length

# Iterate over the file names and classify the sentiment for each file
for file_name in file_names:
    # Load the CSV file into a DataFrame
    data = pd.read_csv(file_name)

    # Truncate or limit the length of the input text to the maximum sequence length
    data['review_text'] = data['review_text'].apply(lambda x: x[:max_seq_length])

    # Apply sentiment classification to the 'review_text' column using the HappyTextClassification model
    data['label'] = data['review_text'].apply(lambda x: happy_tc.classify_text(x).label)

    # Save the updated DataFrame with the new 'label' column to a new CSV file
    new_file_name = file_name.replace(".csv", "_classified.csv")
    data.to_csv(new_file_name, index=False)
```

In [ ]:

```
import pandas as pd

# List of file names
file_names = ["group_1.0_classified.csv", "group_2.0_classified.csv", "group_3.0_classified.csv", "group_4.0_classified.csv", "

# Create an empty DataFrame to store the combined data
combined_data = pd.DataFrame()

# Iterate over the file names
for file in file_names:
    # Read each file into a DataFrame
    data = pd.read_csv(file)
    # Concatenate the data to the combined DataFrame
    combined_data = pd.concat([combined_data, data])

# Write the combined data to a new CSV file
combined_data.to_csv("combined_data.csv", index=False)
```

In [65]:

```
import pandas as pd

# Read the CSV file
data = pd.read_csv('combined_data.csv')

# Access and manipulate the data as needed
# For example, you can print the first few rows of the DataFrame
data
```

Out[65]:

_name	review_title	review_text	review_rating	verified_purchase	review_date	helpful_count	uniq_id	scraped_at	cleaned_text	review_
. Slate	NO SUPPORT! NOT FOR RUNNING!	I would NOT recommend these for running. They ...	1.0	True	2020-07-12	19.0	1bd3f6f9-6e70-50a8-a913-6c9af4f8c7c7	24/12/2021 02:26:25	would recommend running zero support could fee...	2
orge-y	Not as supportive as I had hoped for.	These shoes are cute online but in person...no...	1.0	True	2016-05-10	12.0	51b14655-18d6-556d-bf0b-1cbd9536d9a2	24/12/2021 02:26:25	shoe cute online personnot muchthe shoe super ...	2
Anna	Fell apart after one week	The laces broke after slightly over one week o...	1.0	True	2019-12-22	NaN	c111f9c3-e95b-57c1-8a7f-879f9ec476df	24/12/2021 02:26:25	lace broke slightly one week use horrible qual...	2
chel L.	Too Hard For My Son To Get On	It's a tight fit! The top part doesn't open so...	1.0	True	2021-06-14	NaN	d8a7aabc-a96a-5a59-9ba6-4d2bf0228407	24/12/2021 02:26:26	' tight fit top part ' open made super hard so...	2
Morris	wrong shoes size came	I order a size 11 kids size shoes received a...	1.0	True	2021-02-15	NaN	a3d47e92-821c-51e7-9907-0bc4fbcf2b5c	24/12/2021 02:26:26	order size 11 kid size shoe received 10 tight ...	2
...	...	...	...	...	...	...	...	...	...	
Tracy	Comfortable, like walking on a cloud.	There are no dislikes, every step is with cush...	5.0	True	2021-11-18	NaN	9eb0cc48-0aa1-536c-9ea4-d4df3a8ed326	24/12/2021 02:29:38	dislike every step cushion like feel become fa...	2
Biggs	These shoes are the bomb	I never had shoes this comfy. Definitely buy a...	5.0	True	2021-08-16	NaN	0f1f0492-1e18-5c7e-ad1e-066135767977	24/12/2021 02:29:38	never shoe comfy definitely buy another pair	2
mazon stomer	Comfortable	Walking on clouds.	5.0	True	2021-06-13	NaN	ac960483-75eb-5a56-ad95-1d17bf575620	24/12/2021 02:29:38	walking cloud	2
per M	Great for early walkers	The only shoes (after many tries) that worked ...	5.0	True	2017-12-08	NaN	9b9e6d15-a4b1-57c0-bebd-d58b115b4ada	24/12/2021 02:29:38	shoe many try worked early walker bitty foot e...	2
Kindle stomer	Said they were very comfortable.	My son loves them. Said they were very comfort...	5.0	True	2018-07-06	NaN	2d29c209-b745-5af9-bdde-0dc116387290	24/12/2021 02:29:38	son love said comfortable	2

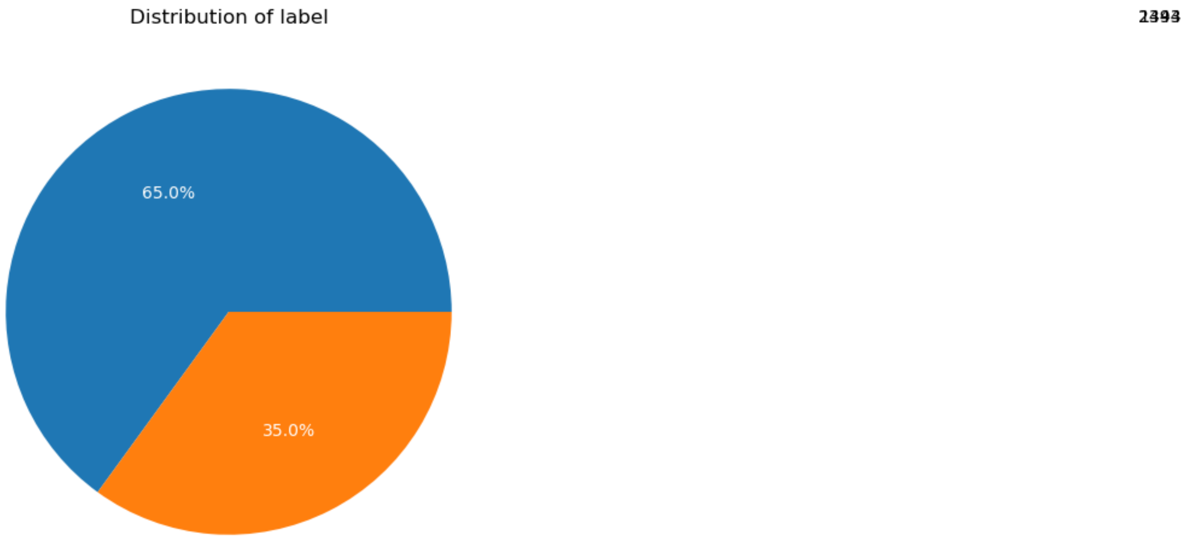
In [68]:

```
a = data['review_text']
b = data['label']
print(a[6])
print(b[6])
```

```
Ordered a size 6. Looked like size 3. Was very small. Not very good quality.
NEGATIVE
```

In [69]:

```
# Calculate the count of each value in the "verified_purchase" column
verified_counts = data['label'].value_counts()

# Create a pie plot of "verified_purchase" values with count numbers
plt.figure(figsize=(8, 6))
patches, texts, autotexts = plt.pie(verified_counts, autopct='%1.1f%%', textprops={'color': 'white'})
plt.title('Distribution of label')

# Add count numbers to the pie plot
for i, count in enumerate(verified_counts):
    #angle = (verified_counts.index.get_loc(i) / len(verified_counts)) * 360
    x = 1.3 * 180 / 180 * 3.14  # Adjust the distance of count numbers
    y = 1.3
    plt.text(x, y, f"{count}")

plt.show()
verified_counts
```



Out[69]:

```
POSITIVE    2494
NEGATIVE    1343
Name: label, dtype: int64
```

In [138]:

```
import pandas as pd
from transformers import pipeline
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
import string
from datetime import date

# Instantiate the HappyTextClassification model
happy_tc = HappyTextClassification(model_type="DISTILBERT", model_name="distilbert-base-uncased-finetuned-sst-2-english", num_

# Read the CSV file
data = pd.read_csv('combined_data.csv')

# Function to process user input and store it in a CSV file
def process_user_input():
    # Get user input
    user_input = input("Enter your text: ")
    #user_input1 = float(input("Enter your rating: "))
    user_input2 = input("Enter your name: ")



    user_rating = None

    while True:
        try:
            user_rating = float(input("Enter your rating (1-5): "))
            if 1 <= user_rating <= 5:
                user_rating = user_rating
                break  # Valid input, exit the loop
            else:
                print("Invalid rating. Please enter a number between 1 and 5.")
        except ValueError:
            print("Invalid input. Please enter a number.")











    # Remove punctuation
    text = user_input.translate(str.maketrans("", "", string.punctuation))

    # Convert to lowercase
    text = text.lower()

    # Tokenize the text
    tokens = word_tokenize(text)

    # Remove stop words
    stop_words = set(stopwords.words("english"))
    tokens = [token for token in tokens if token not in stop_words]

    # Lemmatization
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(token) for token in tokens]

    # Join the tokens back into a single string
    cleaned_text = " ".join(tokens)

    # Get the current date
    current_date = date.today().strftime("%Y-%m-%d")

    # Apply sentiment classification to user input using the HappyTextClassification model
    result = happy_tc.classify_text(cleaned_text)

    # Create a DataFrame with user input, label, and date
    user_data = pd.DataFrame({'reviewer_name': [user_input2],'review_rating': [user_rating],'review_text': [user_input],'cleane

    # Append the user data to the existing DataFrame
    updated_data = pd.concat([data, user_data], ignore_index=True)

    # Save the updated DataFrame to the CSV file
    updated_data.to_csv("combined_data.csv", index=False)
```

```
# Run the pipeline
process_user_input()
```

```
06/27/2023 20:11:35 - INFO - happytransformer.happy_transformer -   Using model: cpu

Enter your text: example
Enter your name: example
Enter your rating (1-5): example
Invalid input. Please enter a number.
Enter your rating (1-5): 4.5
```

In [135]:

```
'''
import pandas as pd

# Read the CSV file
data = pd.read_csv('combined_data.csv')

# Delete the row at index 3838
data = data.drop(index=3837)

# Save the updated DataFrame to the CSV file
data.to_csv('combined_data.csv', index=False)
'''
```

Out[135]:

```
"\nimport pandas as pd\n\n# Read the CSV file\ndata = pd.read_csv('combined_data.csv')\n\n# Delete the row at in
dex 3838\ndata = data.drop(index=3837)\n\n# Save the updated DataFrame to the CSV file\ndata.to_csv('combined_da
ta.csv', index=False)\n"
```

In [139]:

```
import pandas as pd

# Read the CSV file
data = pd.read_csv('combined_data.csv')

# Access and manipulate the data as needed
# For example, you can print the first few rows of the DataFrame
data
```

Out[139]:

	url	product_name	reviewer_name	review_title	review_text	review_rating	verified_purchase	revi
0	https://www.amazon.co.uk/dp/B07S1XM3L7	adidas Women's Retrorun Shoes Running, Core Bl...	B. Slate	NO SUPPORT! NOT FOR RUNNING!	I would NOT recommend these for running. They ...	1.0	True	20
1	https://www.amazon.co.uk/dp/B0125TMZGK	Aravon Women's Betty-AR Oxfords, Stone, 5.5 UK	George-y	Not as supportive as I had hoped for.	These shoes are cute online but in person...no...	1.0	True	20
2	https://www.amazon.co.uk/dp/B077TC44GZ	Merrell Boys' Burnt Rock Low Sneaker, Black, 1...	Anna	Fell apart after one week	The laces broke after slightly over one week o...	1.0	True	20
3	https://www.amazon.co.uk/dp/B08KRS9T98	PUMA EL Rey 2 Slip ON Sneaker, Desert Sage-Gra...	Rachel L.	Too Hard For My Son To Get On	It's a tight fit! The top part doesn't open so...	1.0	True	20
4	https://www.amazon.co.uk/dp/B08KRS9T98	PUMA EL Rey 2 Slip ON Sneaker, Desert Sage-Gra...	Tiffany Morris	wrong shoes size came	I order a size 11 kids size shoes received a...	1.0	True	20
...	...	...	...	...	...	...	...	
3833	https://www.amazon.co.uk/dp/B08GDV6DN6	adidas Zx 2k Boost Mens Fv9993 Size 11	Lyle Biggs	These shoes are the bomb	I never had shoes this comfy. Definitely buy a...	5.0	True	20
3834	https://www.amazon.co.uk/dp/B08GDV6DN6	adidas Zx 2k Boost Mens Fv9993 Size 11	Amazon Customer	Comfortable	Walking on clouds.	5.0	True	20
3835	https://www.amazon.co.uk/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Shopper M	Great for early walkers	The only shoes (after many tries) that worked ...	5.0	True	20
3836	https://www.amazon.co.uk/dp/B06XFT2G2F	Skechers Kids Boys' Nitrate-95358N Sneaker, Bl...	Kindle Customer	Said they were very comfortable.	My son loves them. Said they were very comfort...	5.0	True	20
3837	NaN	NaN	example	NaN	example	4.5	NaN	20

3838 rows × 14 columns

In [ ]: