
Name:- Madhav Kumar Rungta

Division:- CS5

Roll no:- 63

PRN:- 202401100097

```
# Simple US Twitter Airline Sentiment Analysis using NumPy and Pandas
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Read the CSV file
# Replace 'airline_tweets.csv' with your actual file path
df = pd.read_csv('/content/Tweets.csv')

# Display the first few rows of the dataset
print("First 5 rows of the dataset:")
print(df.head())

# 1. What is the total number of tweets in the dataset?
print("\n1. Total number of tweets in the dataset:")
total_tweets = len(df)
print(f"Total tweets: {total_tweets}")

# 2. How many tweets fall into each sentiment category?
print("\n2. Number of tweets by sentiment category:")
```

```
sentiment_counts = df['airline_sentiment'].value_counts()
print(sentiment_counts)

# 3. What is the percentage distribution of sentiments?
print("\n3. Sentiment distribution percentage:")
sentiment_percentage = (df['airline_sentiment'].value_counts() / total_tweets) * 100
print(sentiment_percentage)

# 4. Which airline has received the most tweets?
print("\n4. Airline with the most tweets:")
airline_counts = df['airline'].value_counts()
print(f"Most tweeted airline: {airline_counts.idxmax()} with {airline_counts.max()} tweets")
print(airline_counts)

# 5. What is the average sentiment confidence score?
print("\n5. Average sentiment confidence score:")
avg_sentiment_confidence = df['airline_sentiment_confidence'].mean()
print(f"Average sentiment confidence: {avg_sentiment_confidence:.4f}")

# 6. How many tweets are negative for each airline?
print("\n6. Number of negative tweets by airline:")
negative_by_airline = df[df['airline_sentiment'] == 'negative'].groupby('airline').size()
print(negative_by_airline)

# 7. What are the most common reasons for negative sentiment?
print("\n7. Most common reasons for negative sentiment:")
negative_reasons = df[df['airline_sentiment'] == 'negative']['negativereason'].value_counts()
```

```
print(negative_reasons.head(10))


# 8. What is the average confidence score for negative reasons?
print("\n8. Average confidence score for negative reasons:")
avg_negative_confidence = df['negativereason_confidence'].mean()
print(f"Average negative reason confidence: {avg_negative_confidence:.4f}")


# 9. How many tweets have been retweeted?
print("\n9. Number of tweets that have been retweeted:")
retweeted_count = df[df['retweet_count'] > 0].shape[0]
print(f"Tweets retweeted: {retweeted_count}")


# 10. What is the tweet with the highest retweet count?
print("\n10. Tweet with the highest retweet count:")
most_retweeted = df.loc[df['retweet_count'].idxmax()]
print(f"Most retweeted tweet has {most_retweeted['retweet_count']} retweets")
print(f"Text: {most_retweeted['text'][:100]}...")  # Show just first 100 chars


# 11. What is the relationship between airline and sentiment?
print("\n11. Relationship between airline and sentiment:")
airline_sentiment_matrix = pd.crosstab(df['airline'], df['airline_sentiment'])
print(airline_sentiment_matrix)


# 12. Which airline has the highest percentage of negative tweets?
print("\n12. Airline with the highest percentage of negative tweets:")
airline_sentiment_pct = airline_sentiment_matrix.div(airline_sentiment_matrix.sum(axis=1), axis=0)
* 100
print(f"Percentage of negative tweets by airline:")
```

```
print(airline_sentiment_pct['negative'].sort_values(ascending=False))


# 13. How many tweets include location information?
print("\n13. Tweets with location information:")
location_count = df['tweet_location'].notna().sum()
print(f"Tweets with location: {location_count}")


# 14. How many tweets were posted from each timezone?
print("\n14. Tweets count by timezone:")
timezone_counts = df['user_timezone'].value_counts().head(10)  # Top 10 timezones
print(timezone_counts)


# 15. What is the temporal distribution of tweets (by date)?
print("\n15. Temporal distribution of tweets:")
# Convert tweet_created to datetime if it's not already
if not pd.api.types.is_datetime64_any_dtype(df['tweet_created']):
    df['tweet_created'] = pd.to_datetime(df['tweet_created'])

tweets_by_date = df.groupby(df['tweet_created'].dt.date).size()
print(tweets_by_date)


# 16. Is there a correlation between retweet count and sentiment?
print("\n16. Average retweet count by sentiment:")
retweet_by_sentiment = df.groupby('airline_sentiment')['retweet_count'].mean()
print(retweet_by_sentiment)


# 17. For each airline, what is the most common negative reason?
```

```
print("\n17. Most common negative reason by airline:")
negative_df = df[df['airline_sentiment'] == 'negative']
for airline in df['airline'].unique():
    airline_negative = negative_df[negative_df['airline'] == airline]
    if len(airline_negative) > 0:
        most_common_reason = airline_negative['negativereason'].value_counts().idxmax()
        print(f"{airline}: {most_common_reason}")

# 18. What percentage of tweets have gold labels for sentiment?
print("\n18. Percentage of tweets with gold labels:")
gold_label_count = df['airline_sentiment_gold'].notna().sum()
gold_label_pct = (gold_label_count / total_tweets) * 100
print(f"Tweets with gold sentiment labels: {gold_label_pct:.2f}%")

# 19. How do tweet lengths vary across different sentiments?
print("\n19. Average tweet length by sentiment:")
df['tweet_length'] = df['text'].str.len()
avg_length_by_sentiment = df.groupby('airline_sentiment')['tweet_length'].mean()
print(avg_length_by_sentiment)

# 20. Which day of the week had the most negative tweets?
print("\n20. Negative tweets by day of week:")
if not pd.api.types.is_datetime64_any_dtype(df['tweet_created']):
    df['tweet_created'] = pd.to_datetime(df['tweet_created'])

df['day_of_week'] = df['tweet_created'].dt.day_name()
negative_by_day = df[df['airline_sentiment'] == 'negative'].groupby('day_of_week').size()
```

print(negative_by_day)

print("\nAnalysis complete! A sentiment distribution plot has been saved.")

```
First 5 rows of the dataset:
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000
```

#

```
   negativereason  negativereason_confidence         airline  \
0             NaN                        NaN  Virgin America
1             NaN                     0.0000  Virgin America
2             NaN                        NaN  Virgin America
3      Bad Flight                     0.7033  Virgin America
4      Can't Tell                     1.0000  Virgin America

   airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                     NaN     cairdin                 NaN              0
1                     NaN    jnardino                 NaN              0
2                     NaN  yvonnalynn                 NaN              0
3                     NaN    jnardino                 NaN              0
4                     NaN    jnardino                 NaN              0

                                                text tweet_coord  \
0                @VirginAmerica What @dhepburn said.         NaN
1  @VirginAmerica plus you've added commercials t...         NaN
2  @VirginAmerica I didn't today... Must mean I n...         NaN
3  @VirginAmerica it's really aggressive to blast...         NaN
4  @VirginAmerica and it's a really big bad thing...         NaN

              tweet_created tweet_location               user_timezone
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)
```

```
1. Total number of tweets in the dataset:
Total tweets: 14640

2. Number of tweets by sentiment category:
airline_sentiment
negative    9178
neutral     3099
positive    2363
Name: count, dtype: int64

3. Sentiment distribution percentage:
airline_sentiment
negative    62.691257
neutral     21.168033
positive    16.140710
Name: count, dtype: float64

4. Airline with the most tweets:
Most tweeted airline: United with 3822 tweets
airline
United            3822
US Airways        2913
American          2759
Southwest         2420
Delta             2222
Virgin America     504
Name: count, dtype: int64

5. Average sentiment confidence score:
Average sentiment confidence: 0.9002

6. Number of negative tweets by airline:
airline
American          1960
Delta              955
Southwest         1186
US Airways        2263
United            2633
Virgin America     181
dtype: int64
```

```
7. Most common reasons for negative sentiment:
negativereason
Customer Service Issue        2910
Late Flight                   1665
Can't Tell                    1190
Cancelled Flight               847
Lost Luggage                   724
Bad Flight                     580
Flight Booking Problems        529
Flight Attendant Complaints    481
longlines                      178
Damaged Luggage                 74
Name: count, dtype: int64

8. Average confidence score for negative reasons:
Average negative reason confidence: 0.6383

9. Number of tweets that have been retweeted:
Tweets retweeted: 767

10. Tweet with the highest retweet count:
Most retweeted tweet has 44 retweets
Text: @USAirways 5 hr flight delay and a delay when we land . Is that even real life ? Get me off this pla...
```

```
11. Relationship between airline and sentiment:
airline_sentiment  negative  neutral  positive
airline
American               1960      463       336
Delta                   955      723       544
Southwest              1186      664       570
US Airways             2263      381       269
United                 2633      697       492
Virgin America          181      171       152

12. Airline with the highest percentage of negative tweets:
Percentage of negative tweets by airline:
airline
US Airways        77.686234
American          71.040232
United            68.890633
Southwest         49.008264
Delta             42.979298
Virgin America    35.912698
Name: negative, dtype: float64

13. Tweets with location information:
Tweets with location: 9907
```

```
14. Tweets count by timezone:
user_timezone
Eastern Time (US & Canada)     3744
Central Time (US & Canada)     1931
Pacific Time (US & Canada)     1208
Quito                           738
Atlantic Time (Canada)          497
Mountain Time (US & Canada)     369
Arizona                         229
London                          195
Alaska                          108
Sydney                          107
Name: count, dtype: int64

15. Temporal distribution of tweets:
tweet_created
2015-02-16       4
2015-02-17    1408
2015-02-18    1344
2015-02-19    1376
2015-02-20    1500
2015-02-21    1557
2015-02-22    3079
2015-02-23    3028
2015-02-24    1344
dtype: int64

16. Average retweet count by sentiment:
airline_sentiment
negative    0.093375
neutral     0.060987
positive    0.069403
Name: retweet_count, dtype: float64

17. Most common negative reason by airline:
Virgin America: Customer Service Issue
United: Customer Service Issue
Southwest: Customer Service Issue
Delta: Late Flight
US Airways: Customer Service Issue
American: Customer Service Issue

18. Percentage of tweets with gold labels:
Tweets with gold sentiment labels: 0.27%

19. Average tweet length by sentiment:
airline_sentiment
negative    113.947919
neutral      87.359471
positive     86.082945
Name: tweet_length, dtype: float64
```

```
20. Negative tweets by day of week:
day_of_week
Friday        835
Monday       1922
Saturday     1049
Sunday       2266
Thursday      751
Tuesday      1619
Wednesday     736
dtype: int64

Analysis complete! A sentiment distribution plot has been saved.
```