Social Media Sentiment Analysis

Hyewon Lee, Martha McQuillan

Purpose

- Analyze the Social Media users' sentiment based on:
  - Platforms such as Facebook, Instagram, and Twitter
  - Terms (months/years)
  - Countries

Notes

- Includes Instagram, Facebook, and Twitter
- Ranges from 2010-Nov-12-20:00 — 2023-Jan-15-12:00, with 75% after 2019-Apr-05
- Dataset has 191 different sentiments.

Display few rows

print(df.head)

Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Likes	Country	Year	Month	Day	Hour
Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	User123	Twitter	#Nature #Park	15.0	30.0	USA	2023	1	15	12
Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	CommuterX	Twitter	#Traffic #Morning	5.0	10.0	Canada	2023	1	15	8
Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	FitnessFan	Instagram	#Fitness #Workout	20.0	40.0	USA	2023	1	15	15
Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	AdventureX	Facebook	#Travel #Adventure	8.0	15.0	UK	2023	1	15	18

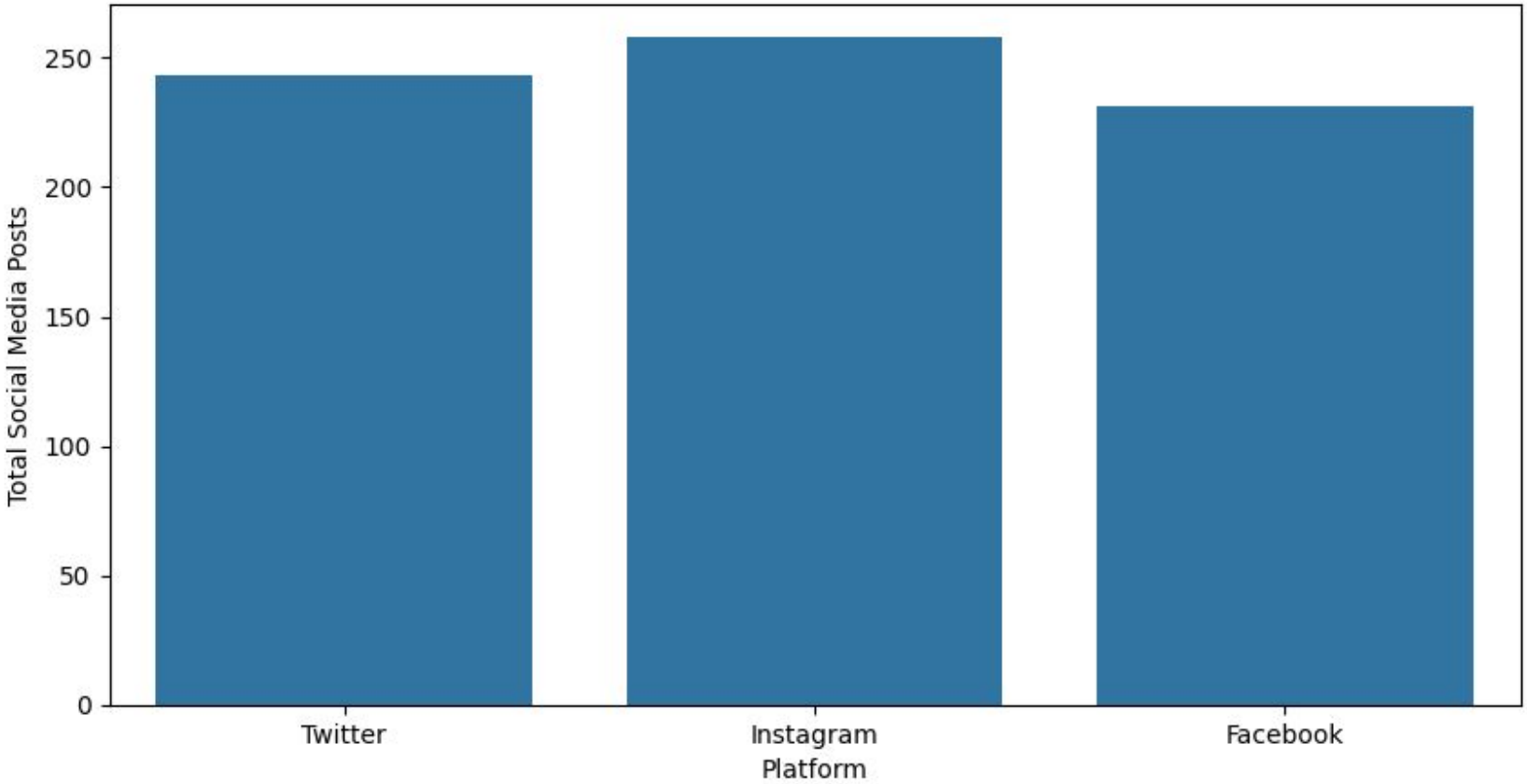
Number of Platform and data #posts Platform-wise

```
df['Platform'] = df['Platform'].str.strip().str.lower()
```

In order to identify “ Twitter”, “Twitter”, “twitter” as the same platform

```
print(df['Platform'].value_counts())
```

Platform	Count
Instagram	258
Facebook	231
Twitter	243



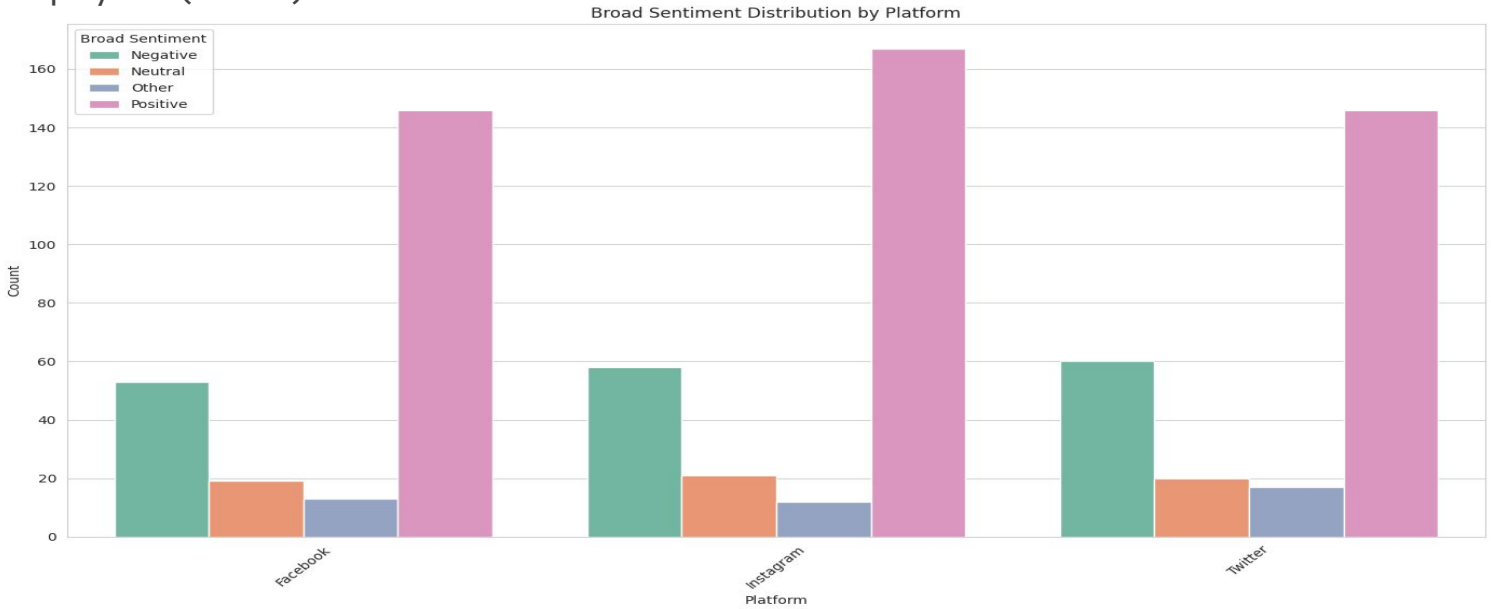
The Trend shows that Instagram is the most popular social media platform followed by Facebook and Twitter.

Platform wise sentiments trend

```
sns.countplot(x='Platform', hue='Sentiment', data=df1)

plt.xlabel('Platform')

plt.ylabel('Count')
```



It shows Sentiment wise Distribution for each platform. Positive sentiment is highest for each platform. Though it is similar, Twitter has a bit more negative sentiment

Useful Columns

```
specified_columns = ['Platform', 'Country', 'Year', 'Month']

for column in specified_columns:
    total_unique_values = df[column].nunique()
    print(f'Total unique values for {column}: {total_unique_values}')

    top_values = df[column].value_counts()

    for value, count in top_values.items():
        print(f'{value}: {count}')

    print('\n' + '=' * 30 + '\n')
```

Country wise number of social media response

df['Country'].value_counts()
df['Country'] = df['Country'].str.strip() #makes 'USA' = ' USA' = 'USA '

Previous code gives unique value for the Country

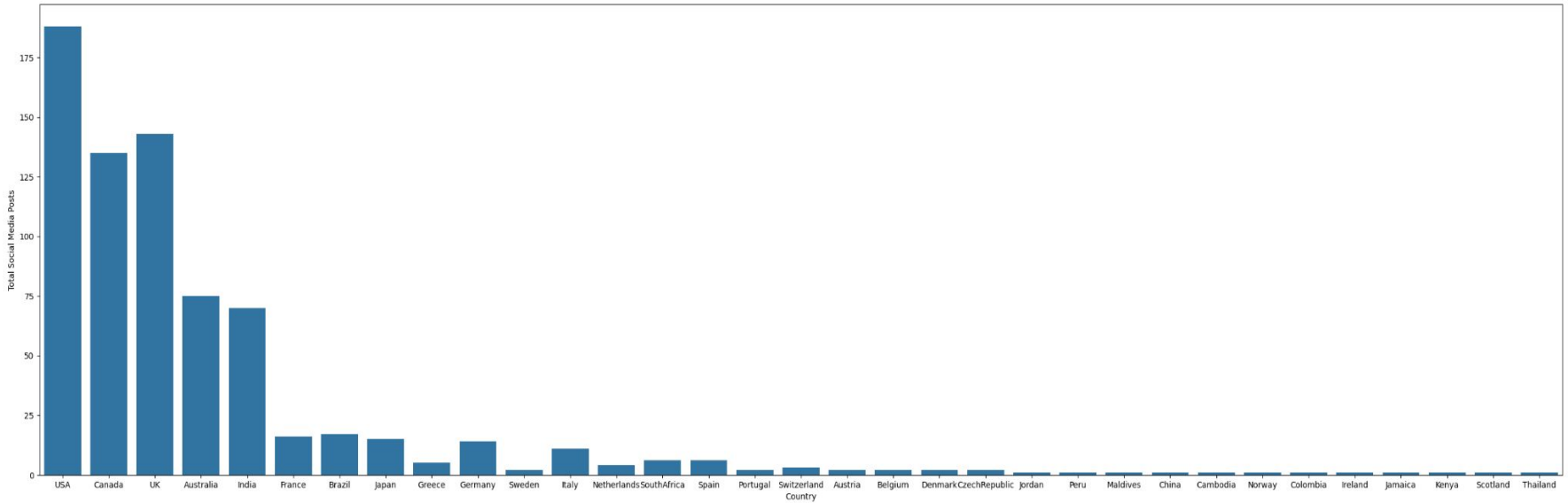
Since there are so many values,

df['Country'].value_counts().nlargest(10).plot(kind='bar')

Gives important 10 countries with a bar graph

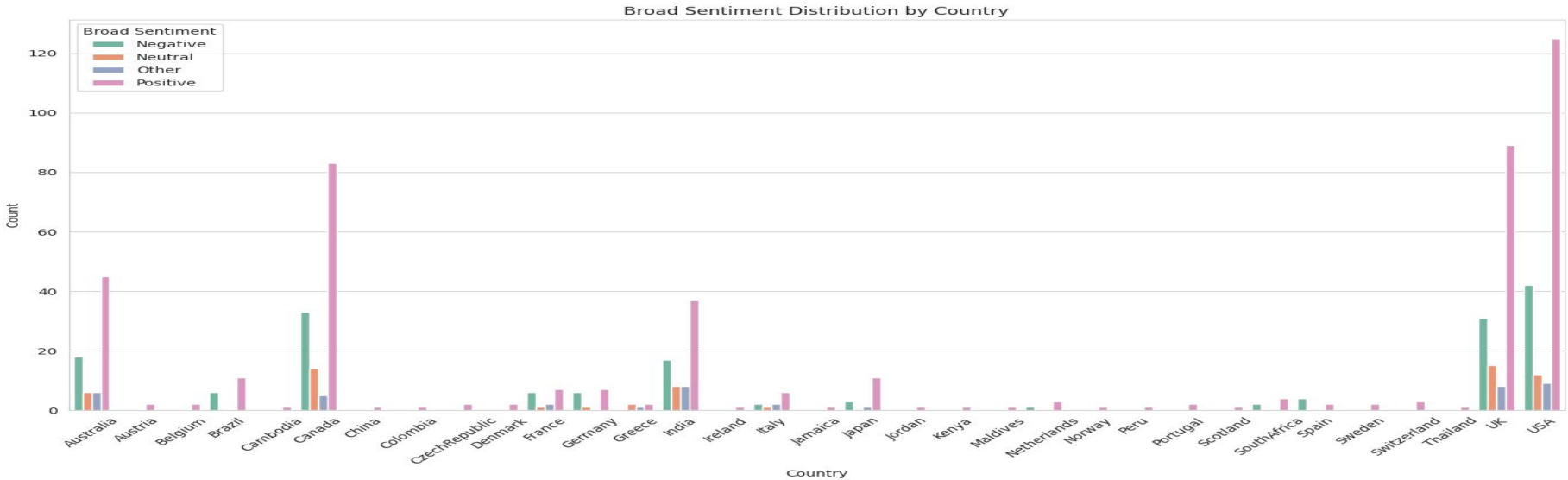
```
Total unique values for Country: 33
USA: 188
UK: 143
Canada: 135
Australia: 75
India: 70
Brazil: 17
France: 16
Japan: 15
Germany: 14
Italy: 11
Spain: 6
South Africa: 6
Greece: 5
Netherlands: 4
```

Country wise number of social media likes



This trend shows number of posts on social media country wise. Showing highest number of likes/tweets from USA followed by UK and Canada

Country wise sentiment trends



This bar chart displays the distribution of broad sentiment categories across various countries.

The USA, Canada, and UK show the highest counts of social media posts, predominantly positive, with some negative and neutral sentiments as well. Other countries have fewer posts and a lower distribution of sentiments.

Monthly Media response

- The useful columns code prints unique sentiment values for each month:
  - February, January, August, and September had approximately 80 posts showing that those months are when users actively posts

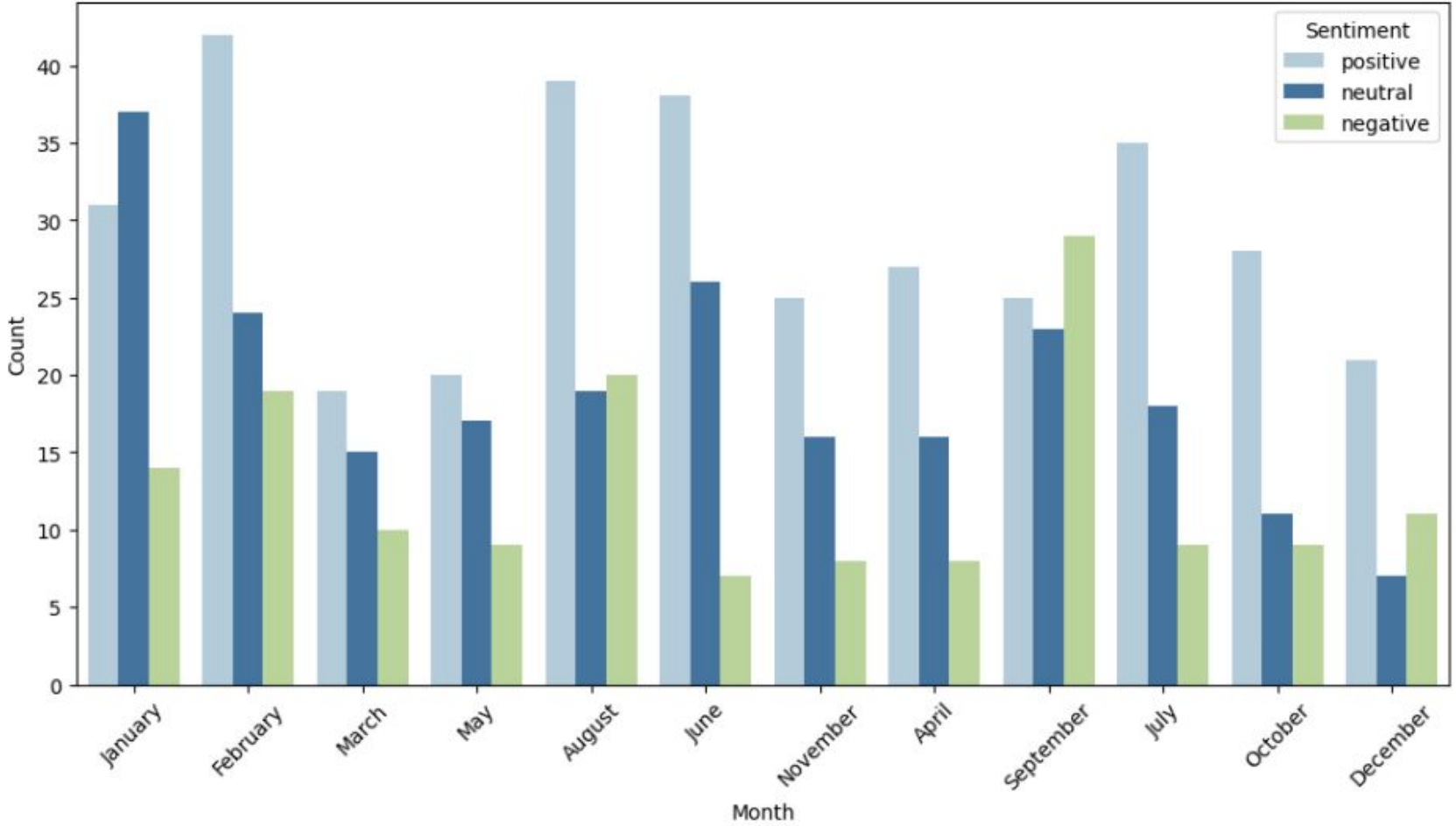
- Each month includes 3 sentiments: positive, negative, and neutral

```
sns.countplot(x='Month', hue ='Sentiment', data=df1, palette='Paired')

plt.xlabel('Month')

plt.ylabel('Count')
```

```
Total unique values for Month: 12
February: 85
January: 82
August: 78
September: 77
June: 71
July: 62
April: 51
November: 49
October: 48
May: 46
March: 44
December: 39
```



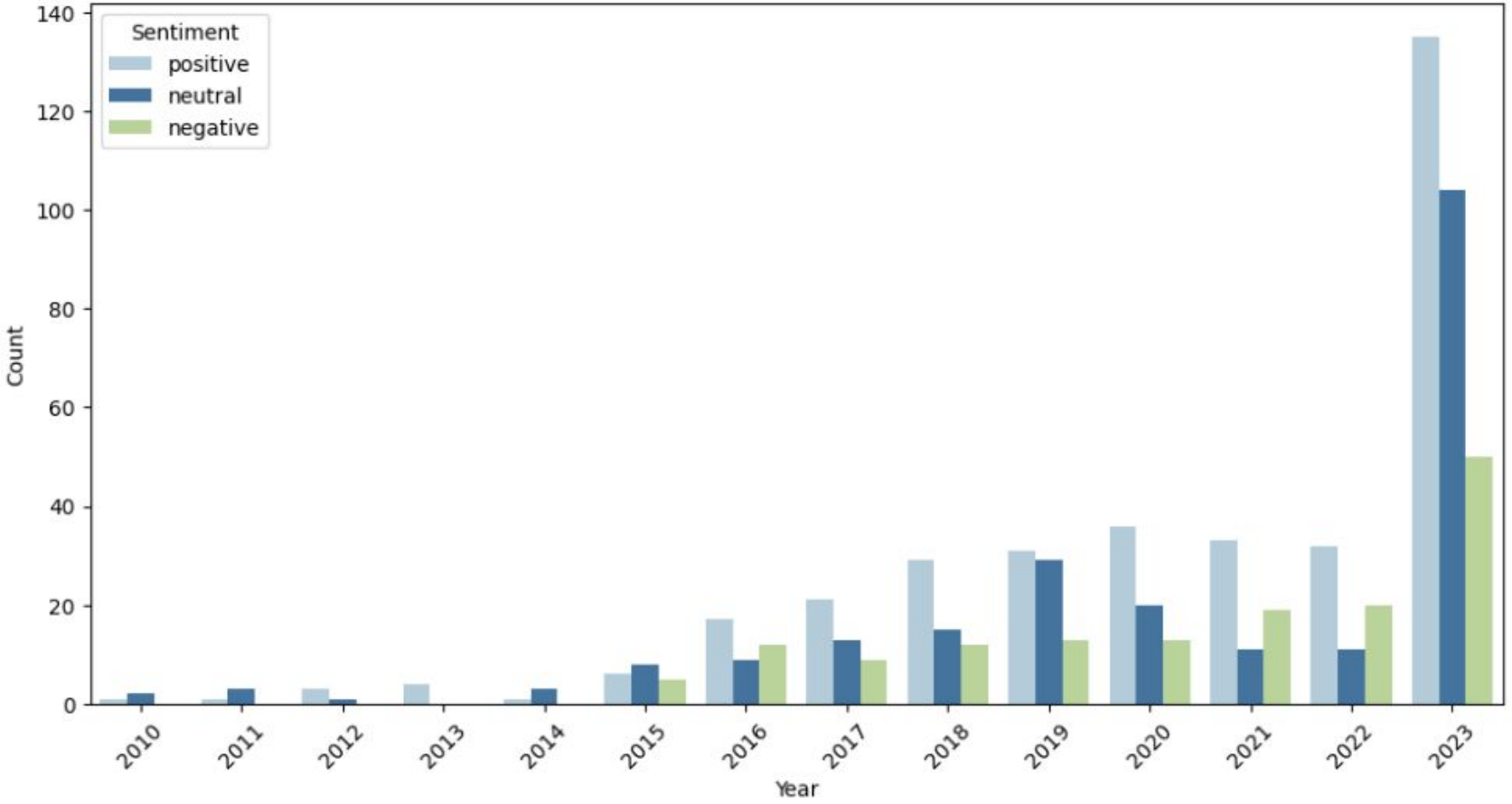
Can analyze how users responded each month and which sentiment were dominant:

Graph shows that overall, positive response were dominant except January and September.

Yearly Media Response

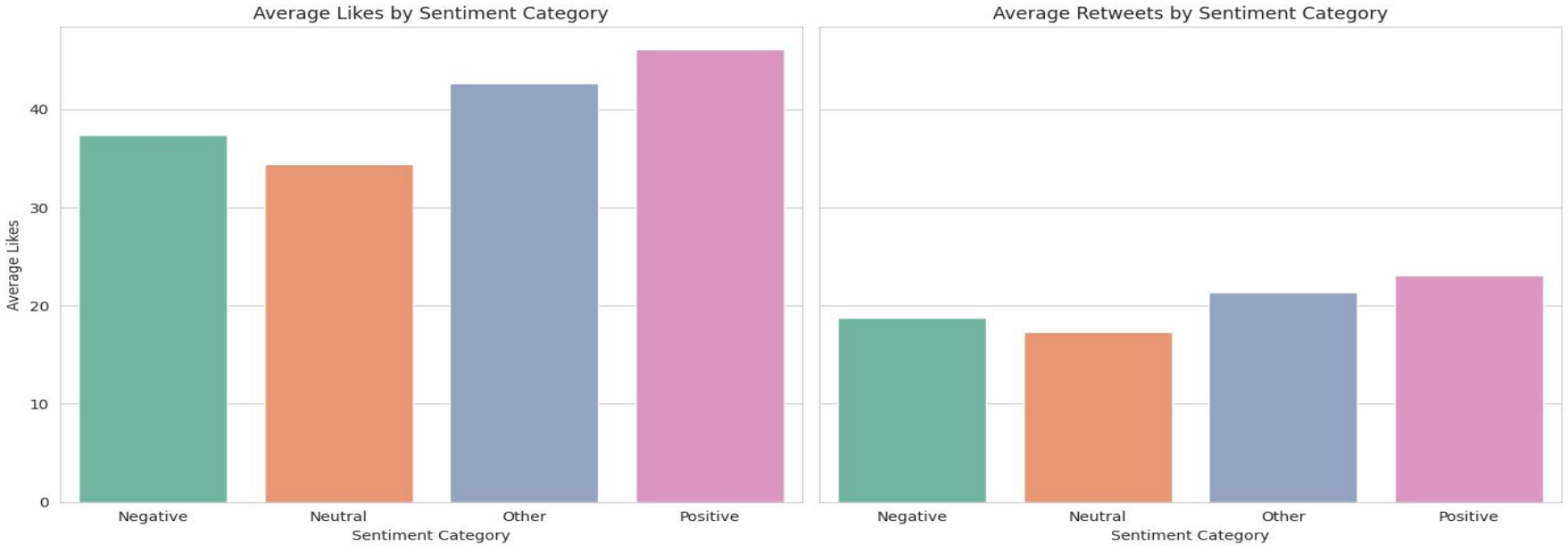
- From 2015 the posts had been increased less than 10 posts each year
- On 2023, posts has been increased greatly compared to the past

```
Total unique values for Year: 14
2023: 289
2019: 73
2020: 69
2021: 63
2022: 63
2018: 56
2017: 43
2016: 38
2015: 19
2011: 4
2013: 4
2012: 4
2014: 4
2010: 3
```



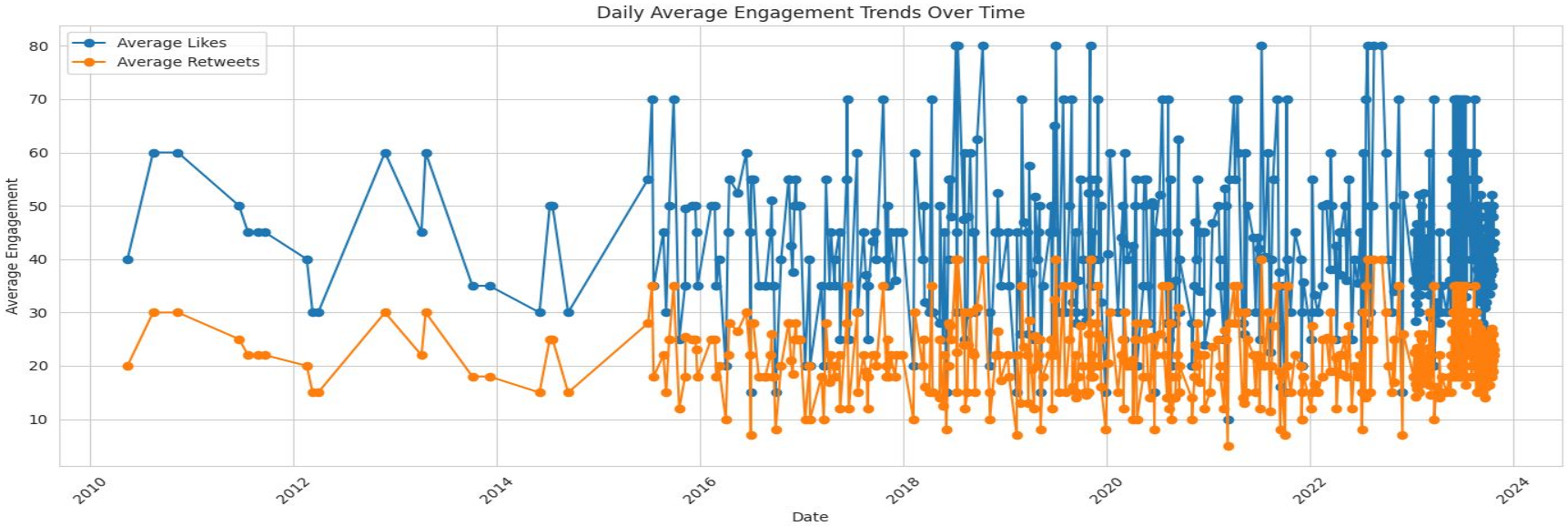
The bar graph with 3 different sentiments shows that most of the time positive sentiments are dominant

Sentiment wise average likes and retweets



Shows average Retweets for broad categories of sentiments. Indicates that Positive has highest number of retweets followed by ‘other’ sentiments. Similarly, this trend shows average Likes for broad categories of sentiments. Indicates that Positive has highest number of retweets followed by ‘other’ sentiments. This also shows positive trend between Likes and Retweets.

Engagement Trends over years



This line plot displays the average number of likes and retweets over time, highlighting daily trends.

There is high variability in engagement, with spikes indicating increased engagement on certain days. Likes generally trend higher than retweets, indicating that content may be more "liked" than "retweeted."

Logistic regression

```
from sklearn.linear_model import LogisticRegression
logistic_classifier = LogisticRegression(max_iter=50, random_state=42)
logistic_classifier.fit(x_train, y_train)

LogisticRegression(max_iter=50, random_state=42)
y_pred_logistic = logistic_classifier.predict(x_train)
accuracy_logistic = accuracy_score(y_test, y_pred_logistic)
classification_rep_logistic = classification_report(y_test, y_pred_logistic)
print("Logistic Regression Results:")
print(f"Accuracy: {accuracy_logistic}")
print("Classification Report:\n", classification_rep_logistic)
```

Logistic Regression Result

```
Logistic Regression Results:
Accuracy: 0.6326530612244898
Classification Report:
              precision    recall  f1-score   support

    negative       0.89      0.50      0.64        32
     neutral       0.80      0.36      0.50        55
    positive       0.55      0.95      0.70        60

    accuracy                           0.63       147
   macro avg       0.75      0.60      0.61       147
weighted avg       0.72      0.63      0.61       147
```

It shows that the accuracy is approximately 0.63 meaning that the results are approximately 63.27% reliable

Conclusions - Logistic Regression

- **Precision** shows the accuracy of predictions
- **Recall** shows how many actual positive/negative/neutral instances were correctly identified
- **F1-score** shows harmonic mean between precision and recall
- **Precision and recall for positive class**: The model is very good at identifying positive instances (high recall of 0.95), but it doesn't do as well in terms of precision (only 0.55). This suggests the model may classify some neutral instances as positive.
- **Precision and recall for neutral and negative classes**: The model has a high precision for negative instances (0.89) but a relatively low recall (0.50), indicating it misses many negative instances. For neutral instances, precision (0.80) is better, but recall (0.36) is lower.

In conclusion, the logistic regression model seems to favor predicting positive instances correctly but struggles with accurately predicting negative and neutral classes, especially when it comes to recall.

Conclusion

The data reveals several key insights:

1. **Positive Sentiment Dominance**: Positive sentiments dominate across platforms, particularly on Instagram and Twitter, which suggests that users on these platforms are more inclined to share positive content.
2. **Higher Engagement for Positive Sentiments**: Positive posts attract more likes and retweets on average, indicating that uplifting content garners higher engagement across social media.
3. **Geographic Distribution**: The USA, Canada, and the UK are the most active countries in this dataset, with positive sentiments being the most common type of engagement.
4. **Platform Preference**: Instagram and Twitter have higher engagement metrics and post counts compared to Facebook, suggesting these platforms may encourage more user interaction.

Conclusion

- These findings suggest that positive sentiment content drives engagement on social media, especially on Instagram and Twitter, and that engagement varies considerably by platform and geographic region.
- We can conclude that if we analyzing the response of a product
    - Users mostly liked the product especially on USA, UK, and Canada
    - Due to the rise in the use of social media platforms, it is likely to promote next products in these platforms, especially on Instagram and Twitter which leads more response on users.
    - The logistic regression gives higher reliability on Positive sentiments. Thus, focusing on Positive sentiments would be more accurate such as January, June, and September since they had the highest Positive sentiment
    - Additionally, gathering the users age, gender, or other aspects of their identity and analyze the sentiments of each generations could help us understand the data better and how to interpret and use it

THANK YOU