Meta Project Final Presentation

Data Club of Notre Dame

Introduction

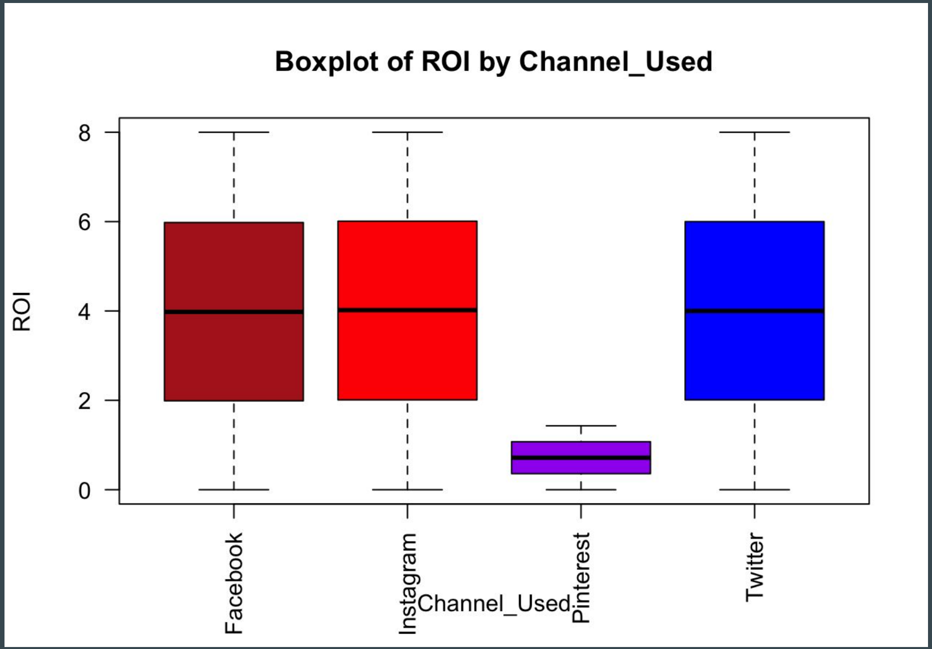
- Interested in exploring different ways to optimize ROI and user engagement for advertisers
- First approach uses data with personal and demographic information
- Second approach uses data about the content, utilizing sentiment analysis

Dataset 1 - Personal/Demographic Information

- Exploring how age, gender, platform, or location influence return on investment
- Models Discussed:
    - Regression
    - Random Forest
    - Gradient Boosting

- Summary of findings:
    - Platform used was the only significant predictor
    - The Kaggle dataset used was very synthetic, so there were not many meaningful trends, leading us to analyze another dataset

ROI vs. Channel Used




```
Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)    
(Intercept)           3.986930   0.007327  544.136   <2e-16 ***
Channel_UsedInstagram  0.021856   0.010364    2.109    0.035 *  
Channel_UsedPinterest -3.270498   0.010365 -315.536   <2e-16 ***
Channel_UsedTwitter    0.015307   0.010380    1.475    0.140    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.009 on 299996 degrees of freedom
Multiple R-squared:  0.3338,    Adjusted R-squared:  0.3338 
F-statistic: 5.012e+04 on 3 and 299996 DF,  p-value: < 2.2e-16
```

Only platform shows to generate significantly difference for ROI.

Pinterest: large negative beta
Instagram: small positive beta

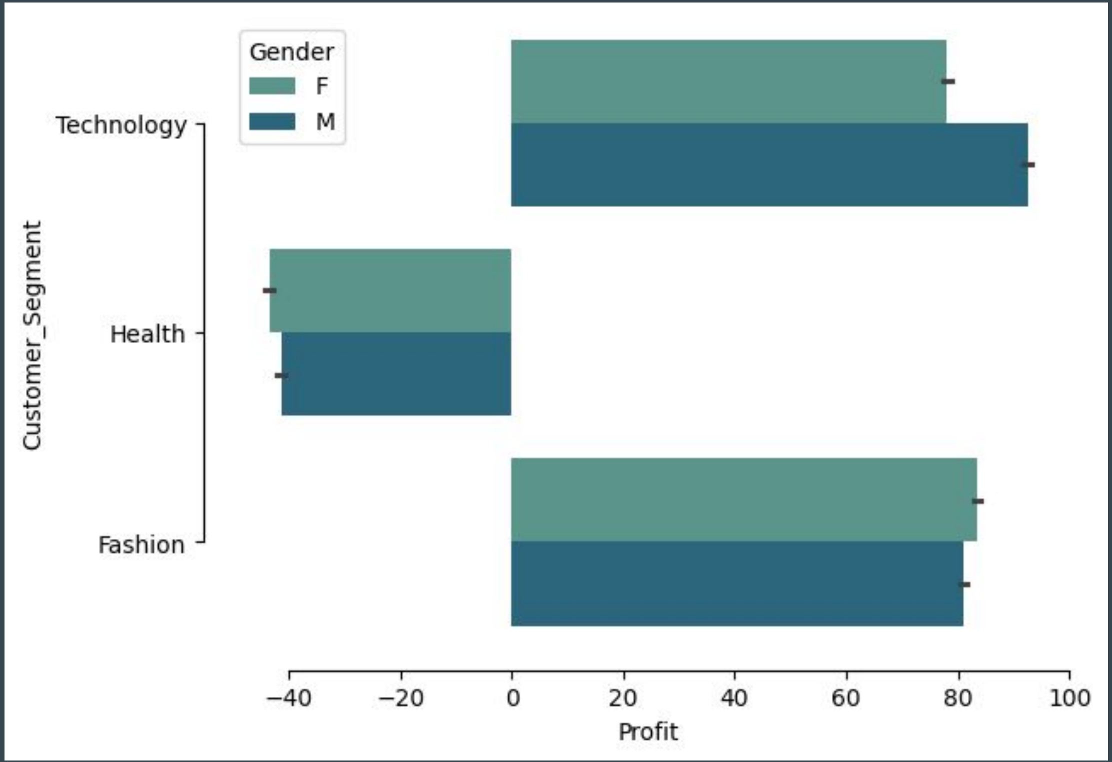
Changes to first model + (new dataset)

<table>
<tr><th colspan="8">ROI</th></tr>
<tr><td rowspan="2">Initial Social media data set</td><td>Location (City)</td><td colspan="2">Target_Audience</td><td>Conversion Rate</td><td>Clicks</td><td rowspan="2">Channel Used</td><td rowspan="2">Customer Segment</td></tr>
<tr><td>State</td><td>Age[]</td><td>Gender</td><td colspan="2">Customers</td></tr>
</table>

Merge

<table>
<tr><td rowspan="2">Additional data set</td><td rowspan="2">Gender</td><td rowspan="2">Age</td><td rowspan="2">State</td><td>Subcategory</td><td>Category</td><td>Profit.mean()</td><td>Quantity.sum()</td></tr>
<tr><td colspan="2">Customer Segment</td><td>Profit</td><td>Quantity</td></tr>
</table>

Distribution of Profit per Customer Segment



Random Forest Model Adaptation

0.338 R^2

0.4330 R^2

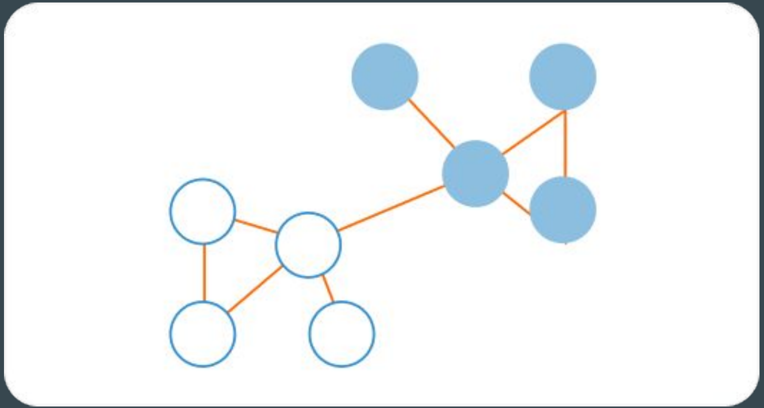
Age	Gender	State	Profit	Quantity	Customers	Customer Segment	Channel_Used

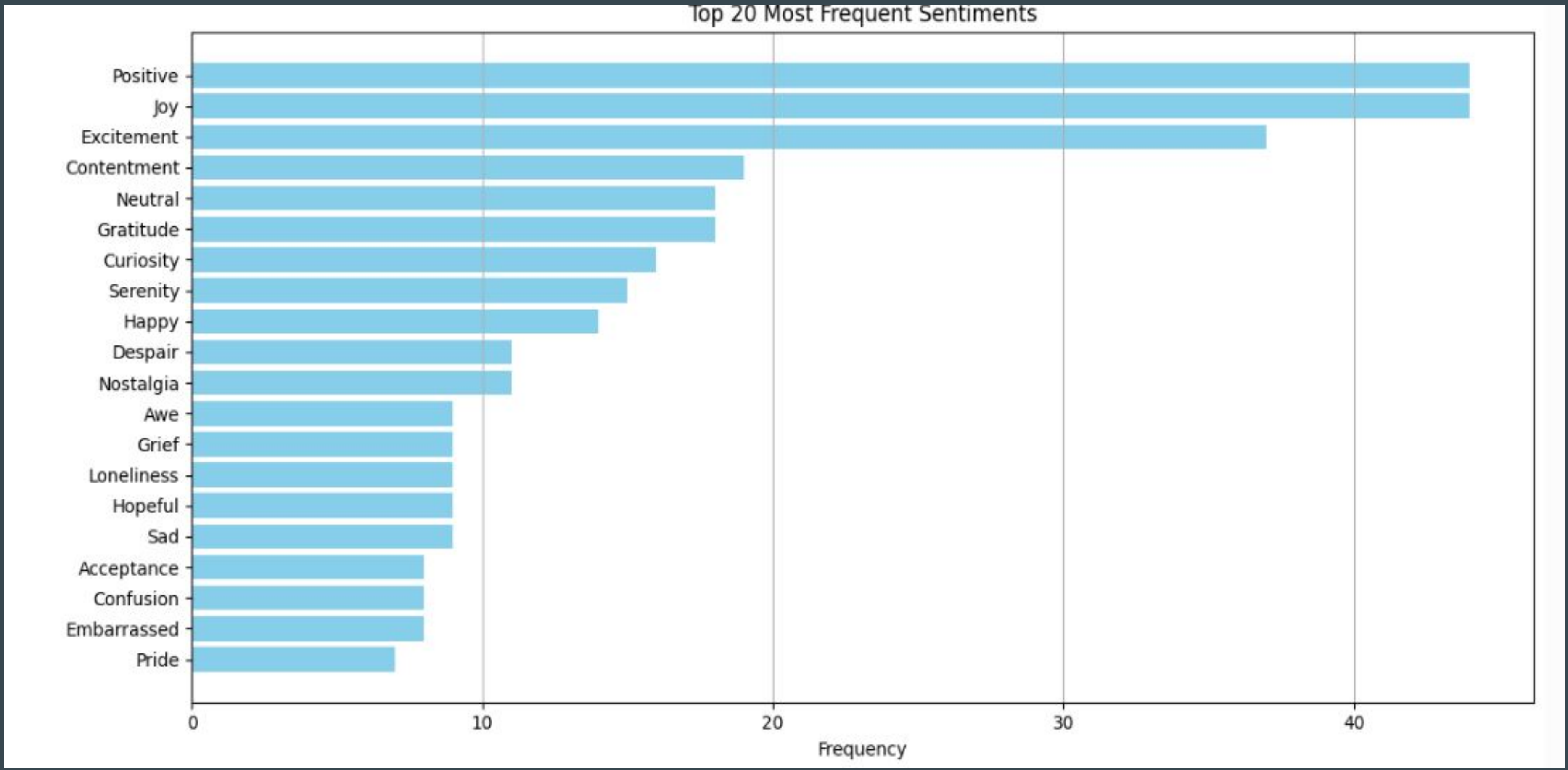
Age	Gender	Profit	Customers	Customer Segment	Channel Used

Dataset 2

- Rather than using personal, identifiable data, we wanted to tailor advertisements to users based on content
- Found a social media dataset of captions, hashtags, sentiment, and engagement (retweets, likes)

- Questions Explored:
    - Does sentiment affect engagement?
    - Is there correlation between the month or country with sentiment?

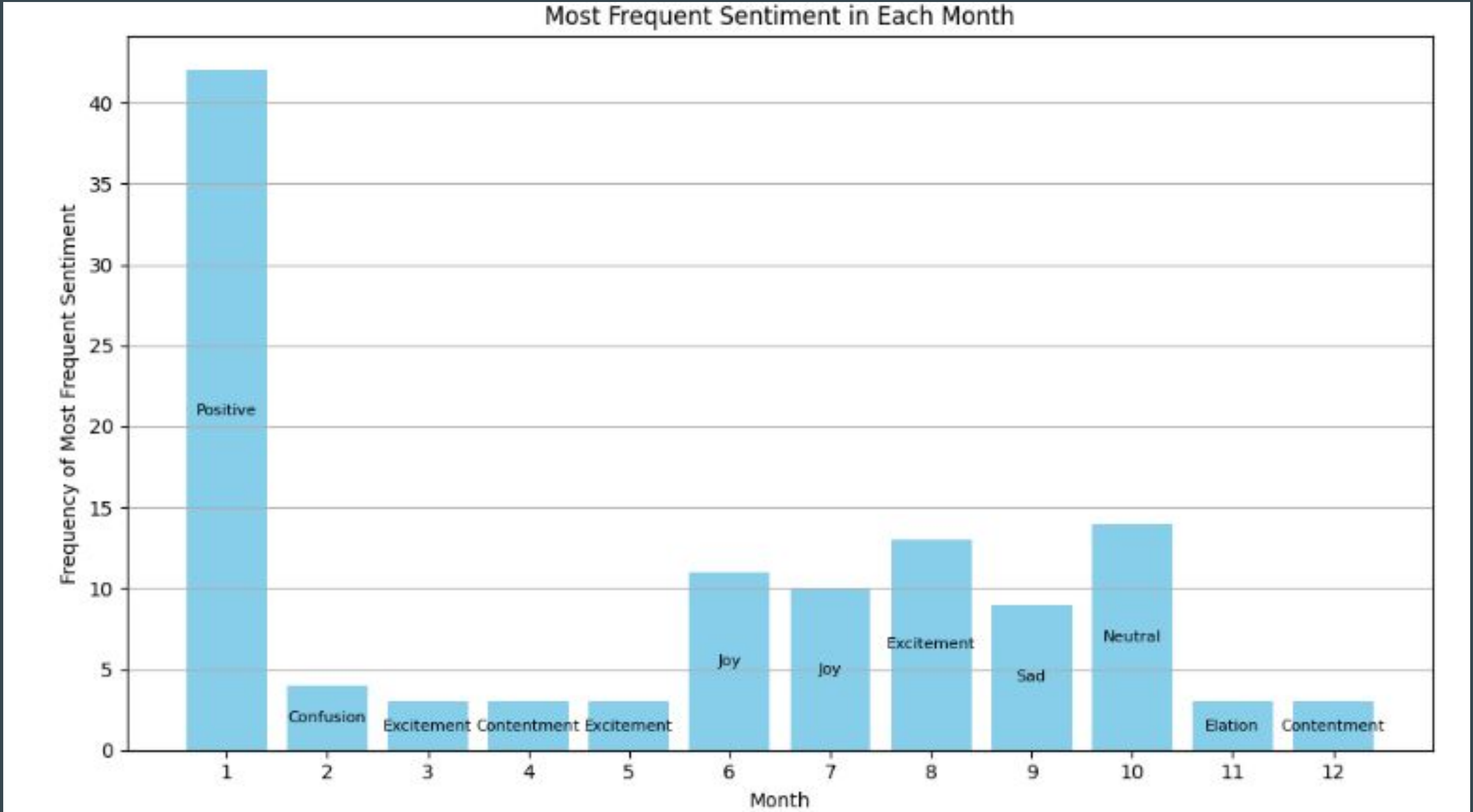




Are people more happy? More positive? More sad on social media? Could this contribute to something greater?

```python
from collections import defaultdict
import pandas as pd
from collections import Counter
import matplotlib.pyplot as plt
```

```python
top_20_sentiments = Counter(df['Sentiment']).most_common(20)
top_20_df = pd.DataFrame(top_20_sentiments, columns=['Sentiment', 'Count'])

# Plot
plt.figure(figsize=(12, 6))
plt.barh(top_20_df['Sentiment'], top_20_df['Count'], color='skyblue')
plt.xlabel('Frequency')
plt.title('Top 20 Most Frequent Sentiments')
plt.gca().invert_yaxis()  # Highest at top
plt.grid(axis='x')
plt.tight_layout()
plt.show()
```

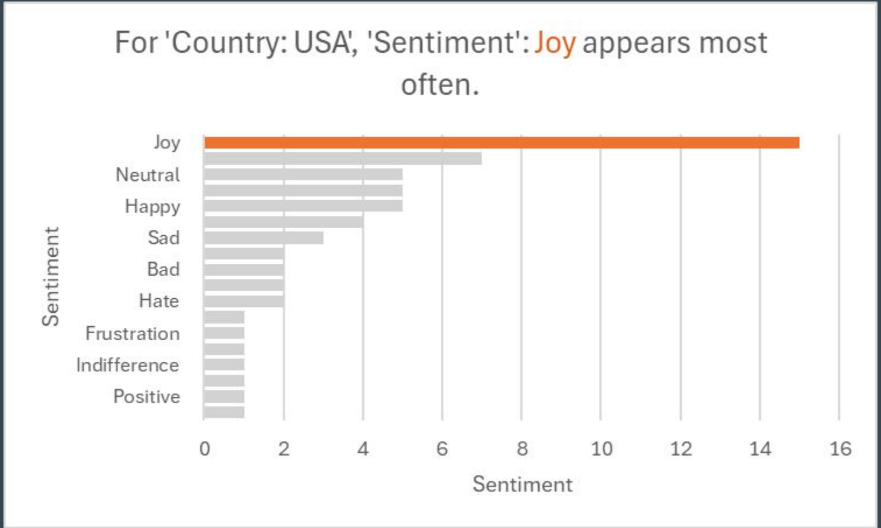


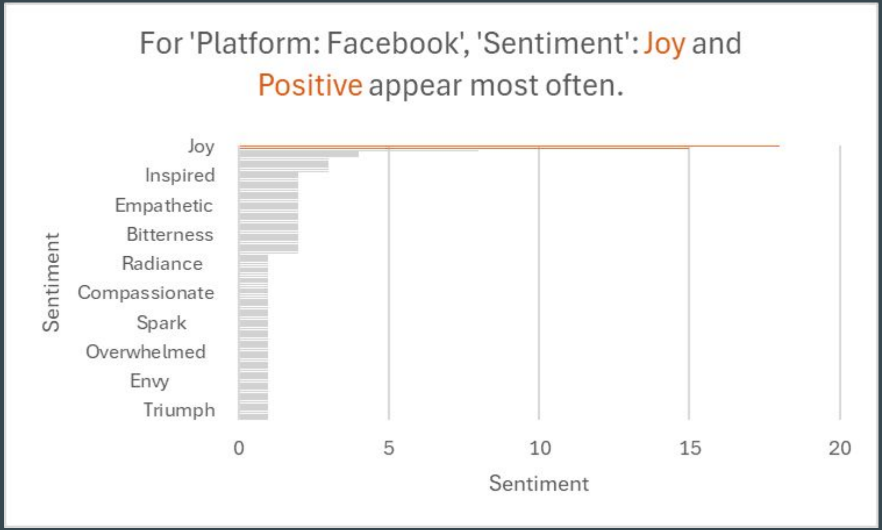
Most Frequent Sentiment in Each Month
Frequency of Most Frequent Sentiment
Month
0
5
10
15
20
25
30
35
40
1
2
3
4
5
6
7
8
9
10
11
12
Positive
Confusion
Excitement
Contentment
Excitement
Joy
Joy
Excitement
Sad
Neutral
Elation
Contentment

```python
from collections import defaultdict
import pandas as pd
from collections import Counter
import matplotlib.pyplot as plt
```

```python
most_frequent_sentiments = []
for month in range(1, 13):
    month_df = df[df['Month'] == month]

    most_frequent_sentiment = Counter(month_df['Sentiment']).most_common(1)[0]
    most_frequent_sentiments.append((month, most_frequent_sentiment[0], most_frequent_sentiment[1]))

most_frequent_df = pd.DataFrame(most_frequent_sentiments, columns=['Month', 'Most Frequent Sentiment', 'Count'])

plt.figure(figsize=(10, 6))
bars = plt.bar(most_frequent_df['Month'], most_frequent_df['Count'], color='skyblue')
plt.xlabel('Month')
plt.ylabel('Frequency of Most Frequent Sentiment')
plt.title('Most Frequent Sentiment in Each Month')
plt.xticks(range(1, 13))
plt.grid(axis='y')

for bar, sentiment in zip(bars, most_frequent_df['Most Frequent Sentiment']):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height() / 2, sentiment,
             ha='center', va='center', color='black', fontsize=8)

plt.tight_layout()
plt.show()
```

Sentiment, Platform, and Country Comparison





Hashtag frequency analysis

Instagram

Emotional Depth
Both positive and negative emotions
(evolcative and reflective)

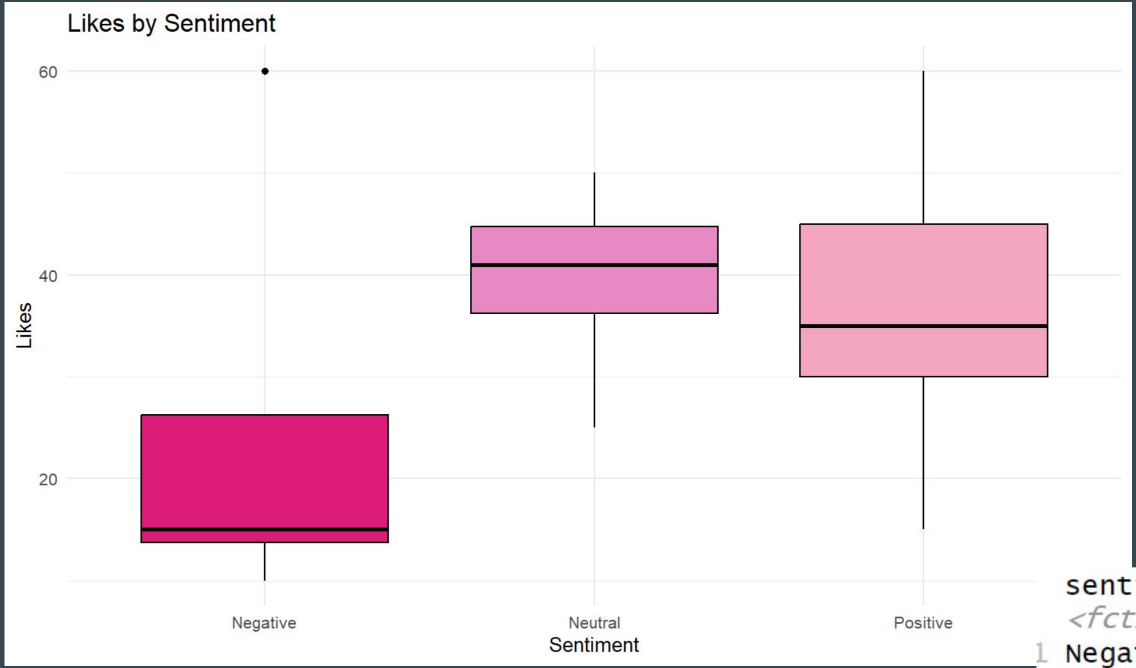
Facebook

Emotional Positive
Life storytelling/ Retrospective

Twitter

Emotional Extreme
Both extreme positive + negative

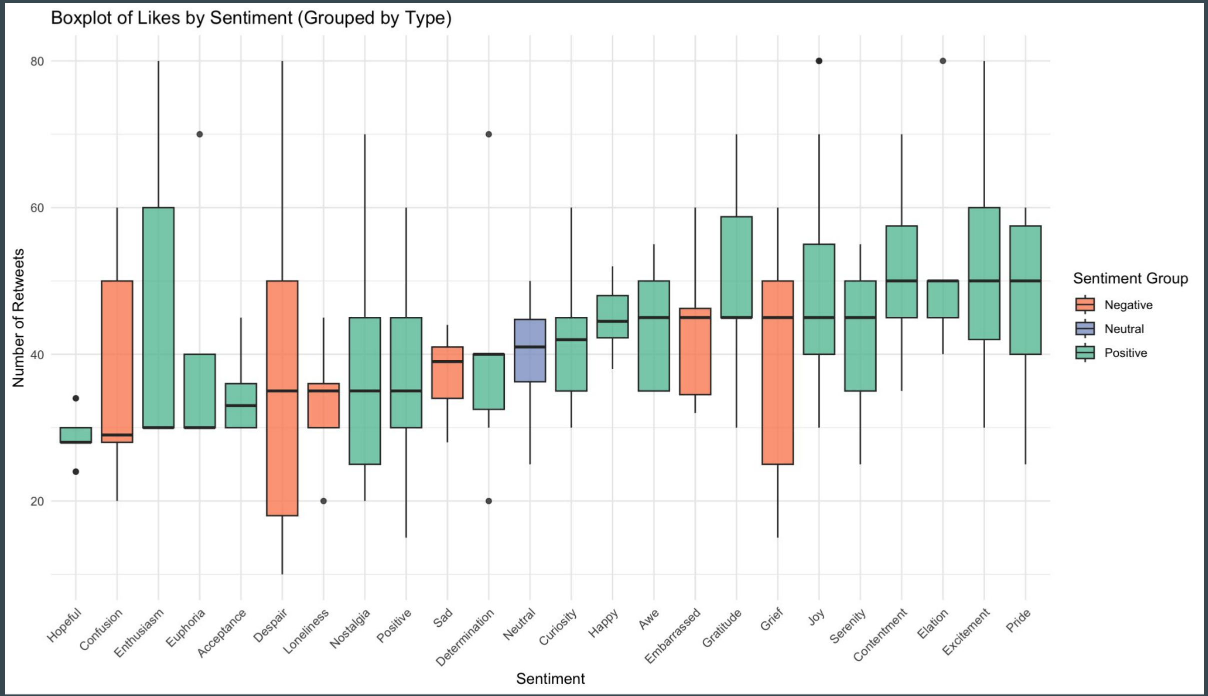
Is there a relationship between **likes** and **sentiment**?



	sentiment_cat	Mean	Median	Range	Mode
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	Negative	25	15	50	15
2	Neutral	40.5	41	25	35
3	Positive	37.8	35	45	30

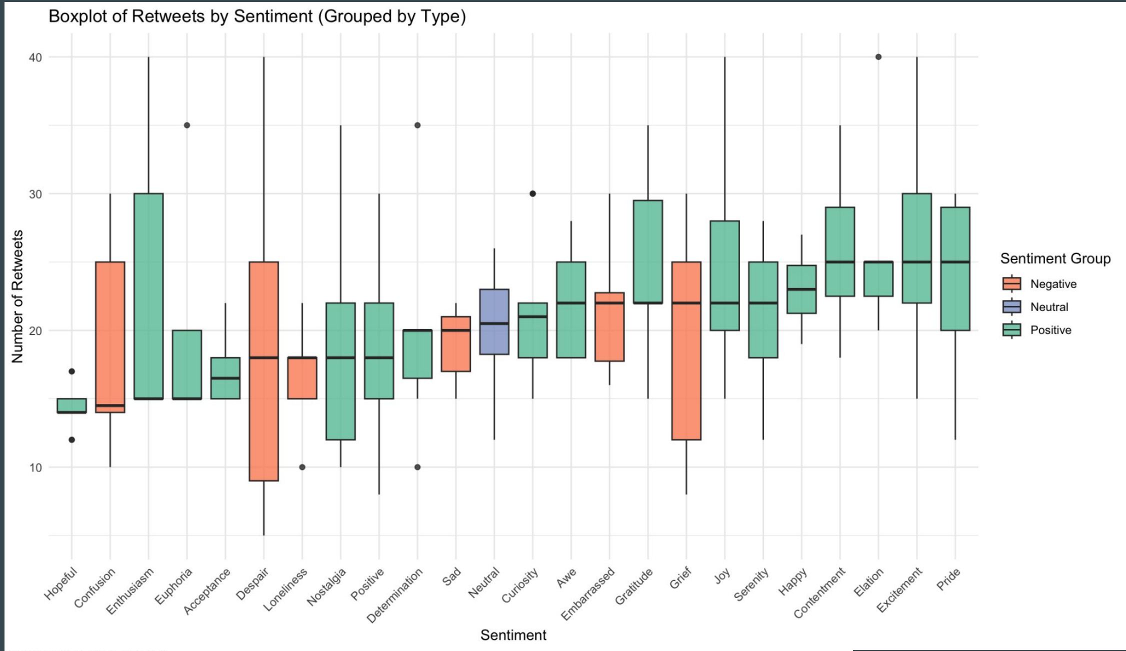
190+ Sentiments in total

Top 20 frequent Sentiment




```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                18.7593     0.8702  21.556  < 2e-16 ***
SentimentGroupNeutral       1.7963     1.7405   1.032 0.302757
SentimentGroupPositive      3.5193     0.9505   3.703 0.000248 ***
---
```

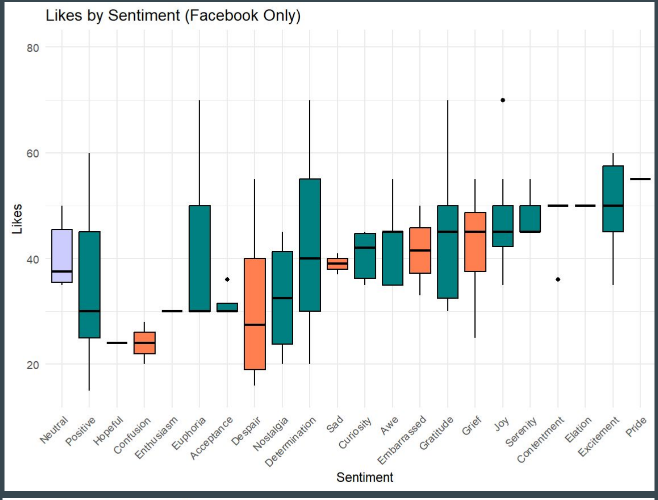
Is there a relationship between Reweets and sentiment?
Does Retweets and Likes show similar pattern for different sentiment?

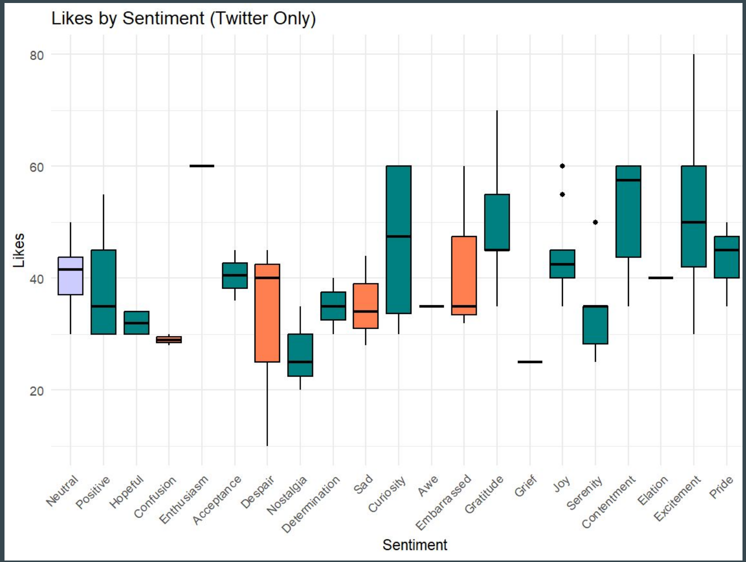


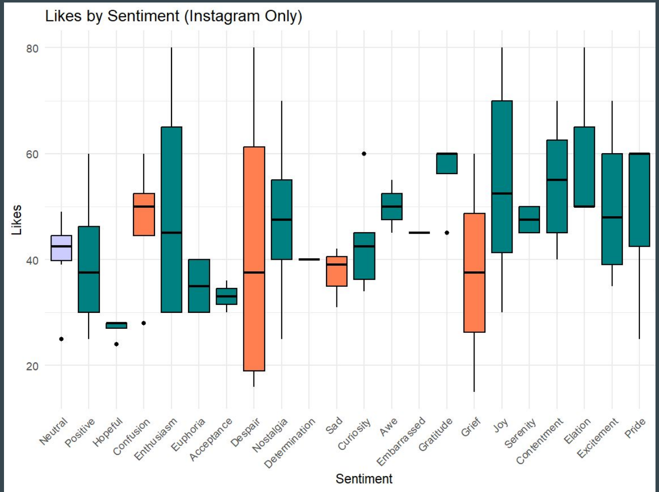
	Sentiment	median_retweets	median_likes	retweet_rank	like_rank	rank_diff
1	Happy	23	44.5	5	11	-6
2	Awe	22	45	6	5	1
3	Embarrassed	22	45	7	6	1
4	Gratitude	22	45	8	7	1
5	Grief	22	45	9	8	1
6	Joy	22	45	10	9	1
7	Serenity	22	45	11	10	1
8	Acceptance	16.5	33	20	20	0
9	Confusion	14.5	29	23	23	0
10	Contentment	25	50	1	1	0
11	Curiosity	21	42	12	12	0
12	Despair	18	35	16	16	0
13	Determination	20	40	14	14	0
14	Elation	25	50	2	2	0
15	Enthusiasm	15	30	21	21	0
16	Euphoria	15	30	22	22	0
17	Excitement	25	50	3	3	0
18	Hopeful	14	28	24	24	0
19	Loneliness	18	35	17	17	0
20	Neutral	20.5	41	13	13	0
21	Nostalgia	18	35	18	18	0
22	Positive	18	35	19	19	0
23	Pride	25	50	4	4	0
24	Sad	20	39	15	15	0

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	18.7593	0.8702	21.556	< 2e-16	***
SentimentGroupNeutral	1.7963	1.7405	1.032	0.302757	
SentimentGroupPositive	3.5193	0.9505	3.703	0.000248	***

Is there a relationship between **likes** and **sentiment** based on **Platform** ?



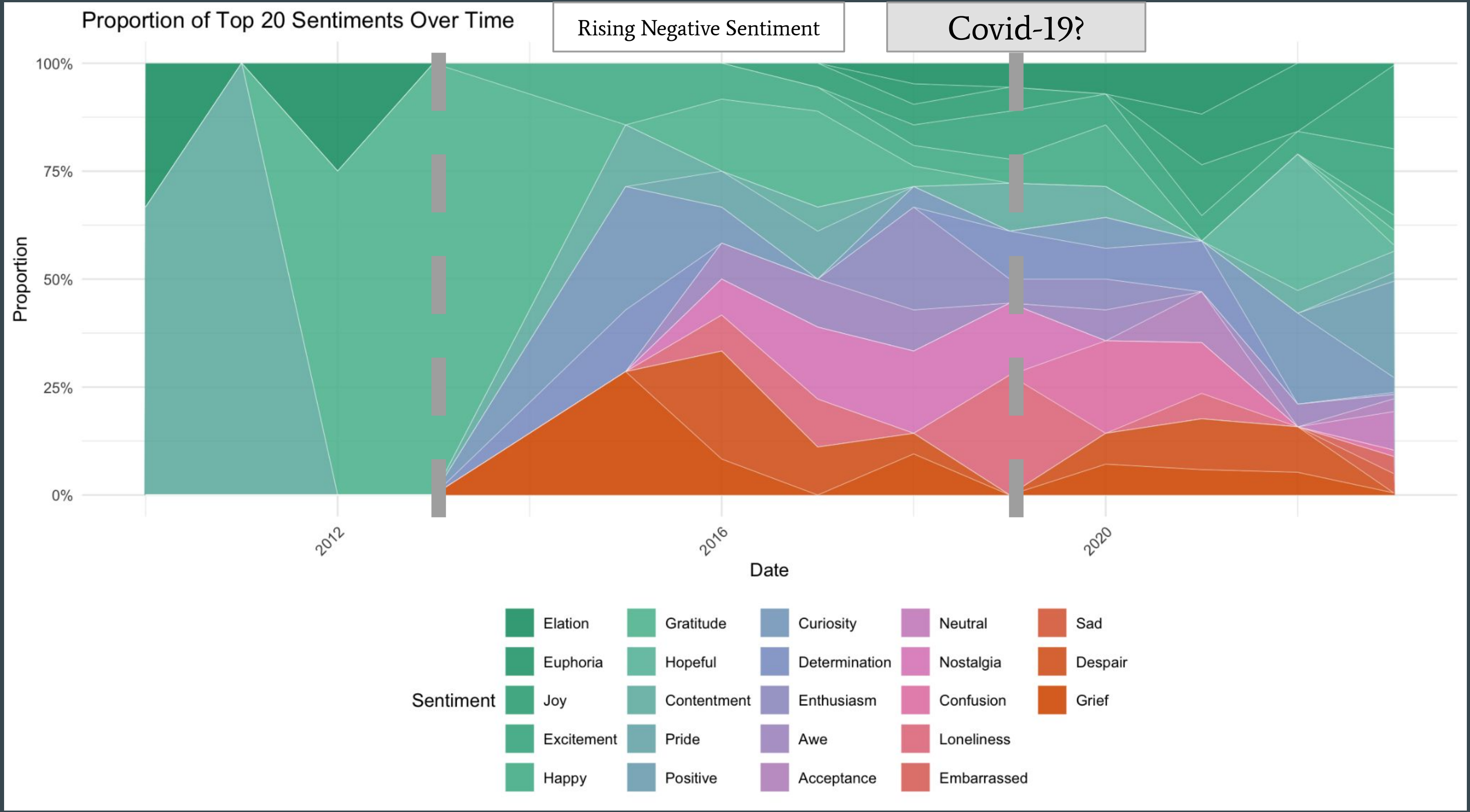




Conclusions:

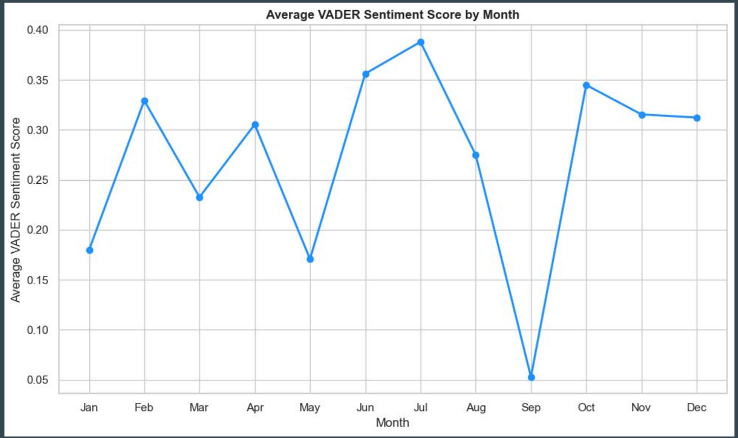
- The maximum number of likes for all three platforms came from positive sentiment captions.
- The minimum number of likes for all three platforms tended to come from negative sentiment captions.

Does sentiments **Changes** over times?

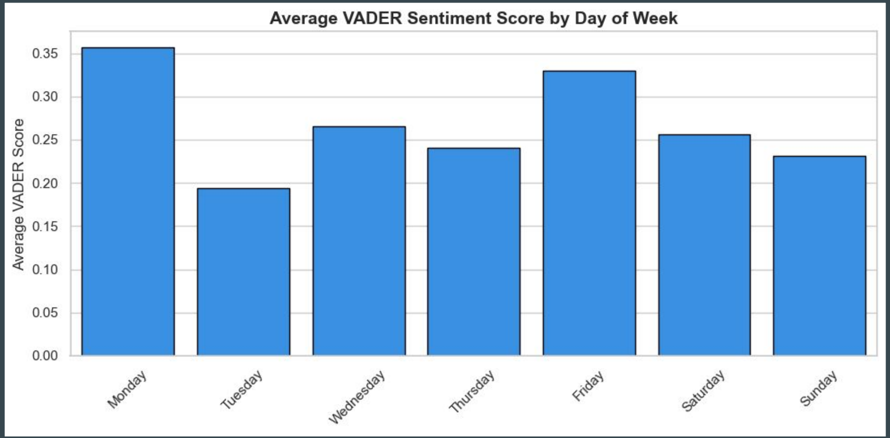


Is there a relationship between **month** or **day of the week** and **sentiment score**?

Month



Day of the Week



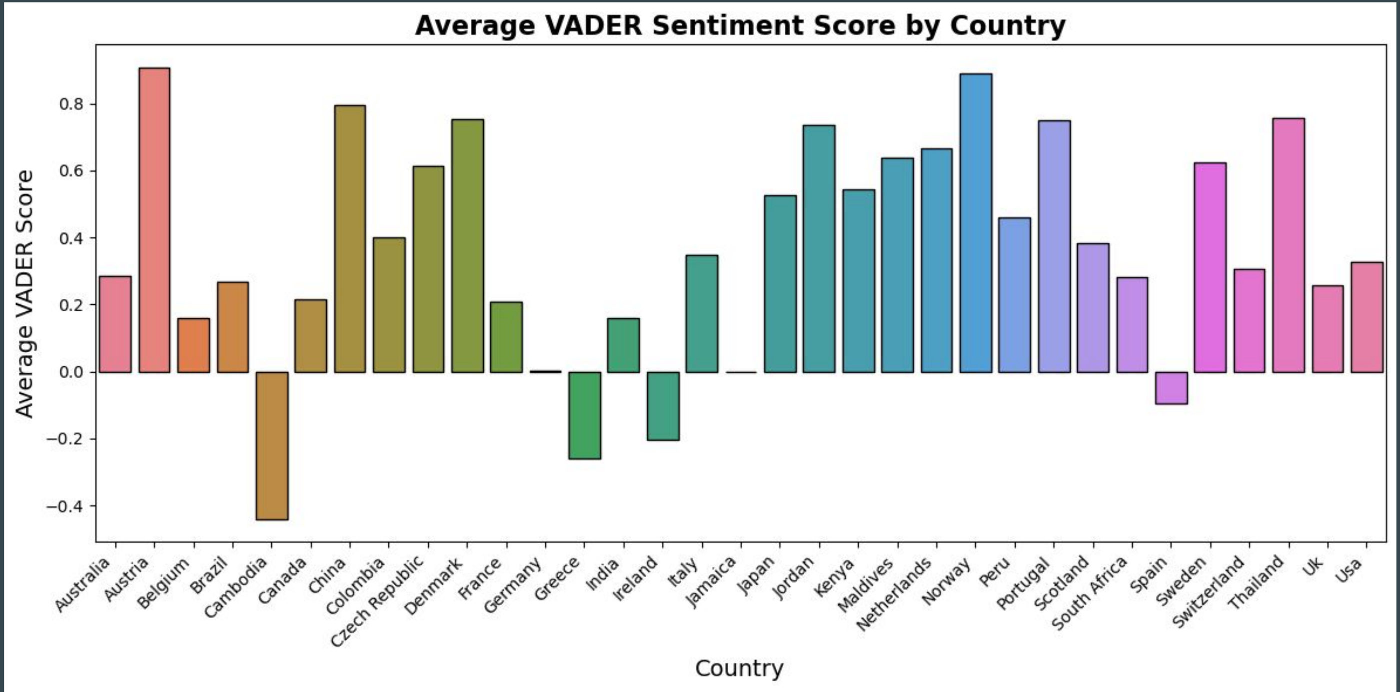
ANOVA Table — Month

	sum_sq	df	F	PR(>F)
C(Month)	6.996597	11.0	2.411058	0.006045
Residual	189.941166	720.0	NaN	NaN

ANOVA Table — Day of Week

	sum_sq	df	F	PR(>F)
C(Day_of_Week)	2.012336	6.0	1.247437	0.279827
Residual	194.925427	725.0	NaN	NaN

Is there a relationship between **country** and **sentiment score**?

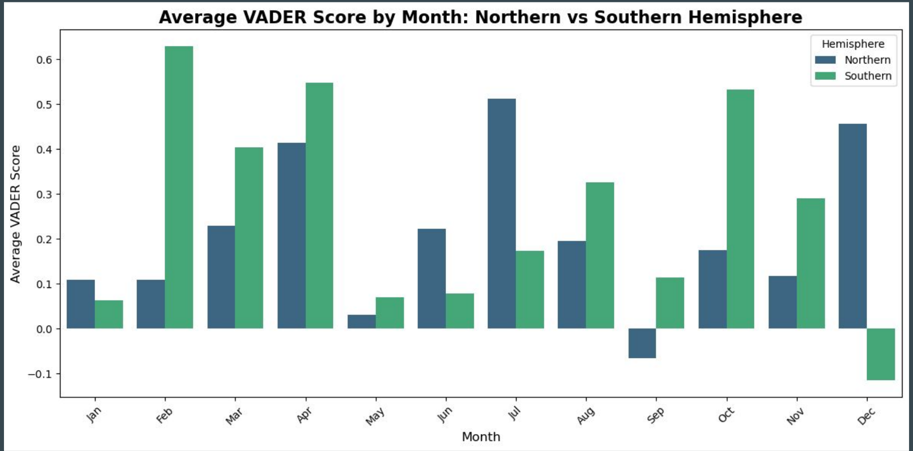


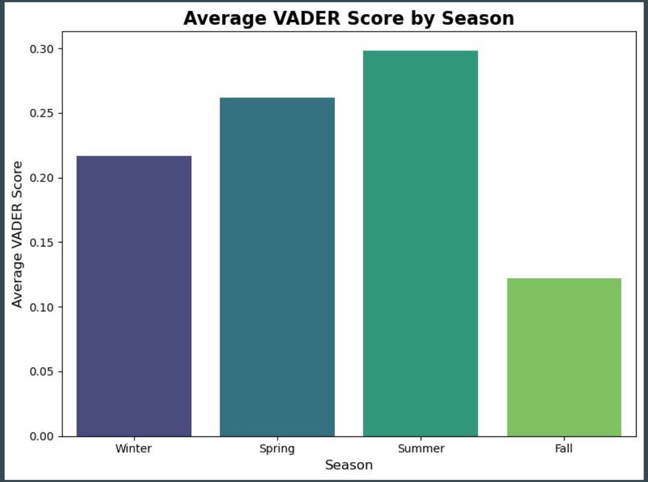
ANOVA Table — Country

	sum_sq	df	F	PR(>F)
C(Country)	11.276479	32.0	1.32672	0.109264
Residual	185.661284	699.0	NaN	NaN

How are **country** and **month** related to **sentiment score**?

Seasonal trends



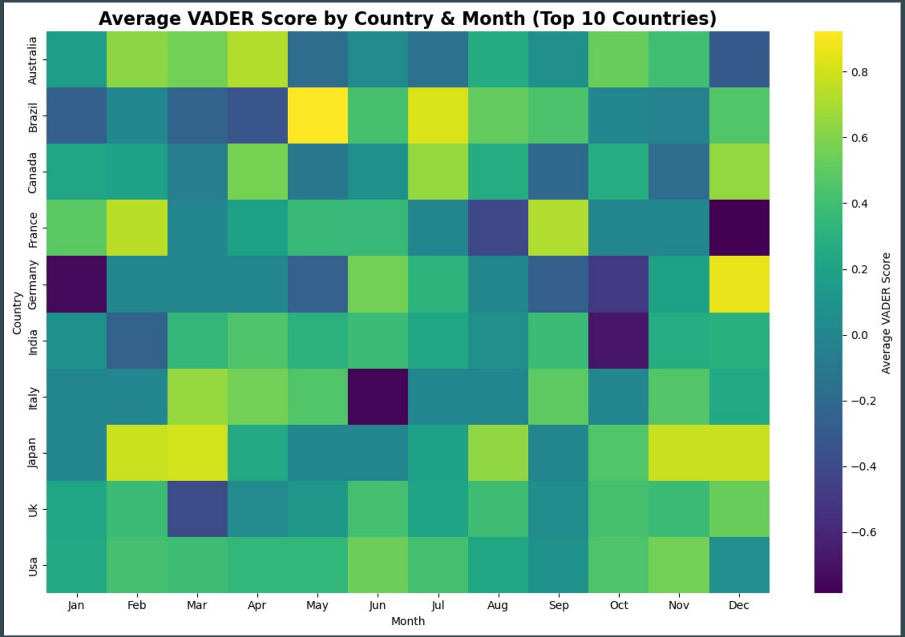


Northern = Blue
Southern = Green

Summer has the highest average VADER Score.

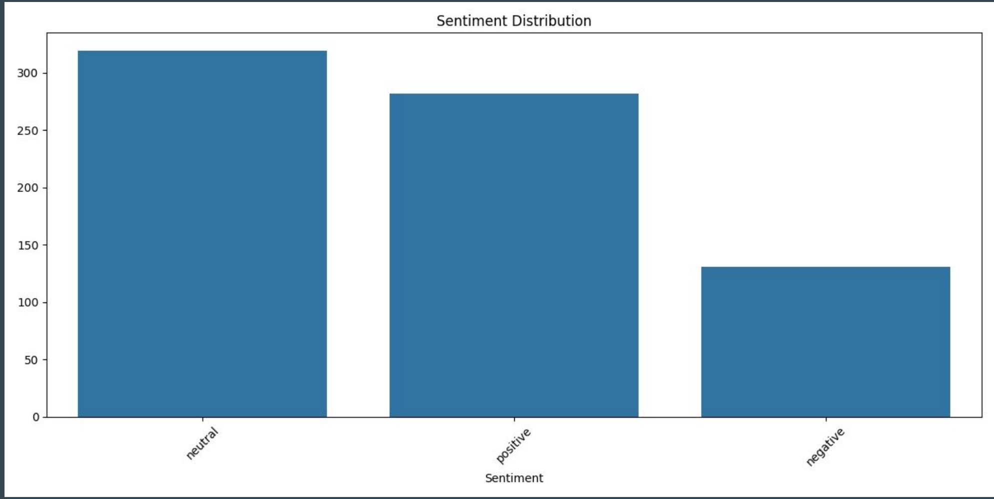
How are **country** and **month** related to **sentiment score**?
Heatmap

Yellow = High VADER Scores
Purple = Low VADER Scores



Sentiment Analysis from Captions / Data Cleaning

- The sentiment values were standardized to positive, neutral, and negative
- The text was cleaned with sentiment-aware preprocessing that preserved emotional signals like negations, punctuation, emojis, and boosted key sentiment words to enhance model accuracy



Model Performance Results

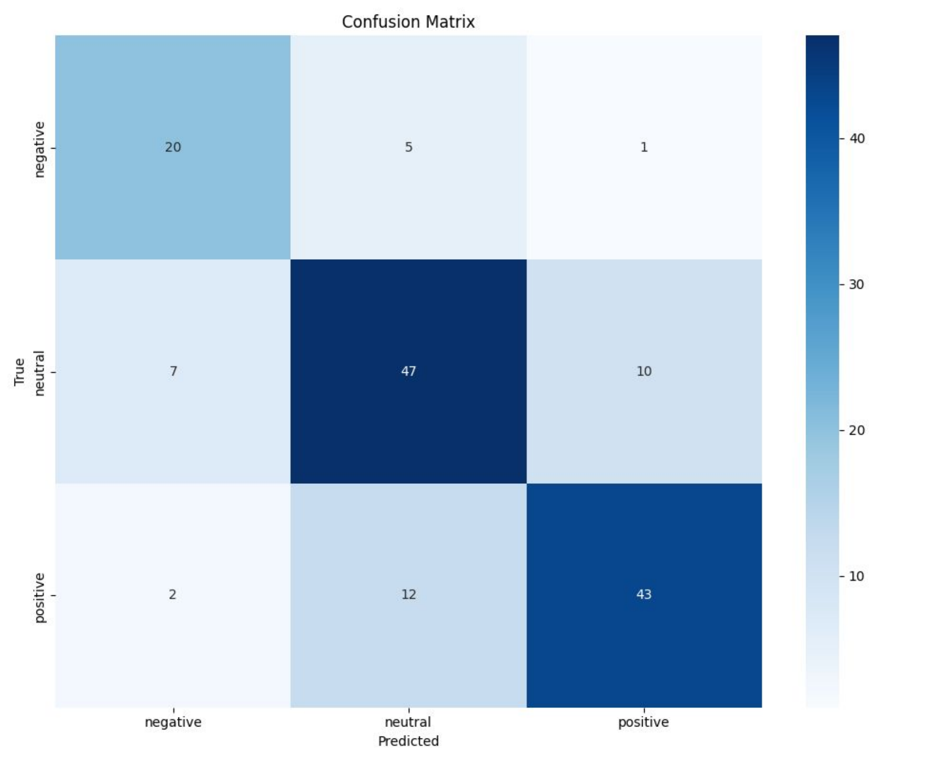
Logistic regression, random forest, naive bayes, support vector machine, and ensemble models were all used to learn off the dataset.

Out of the five, logistic regression, naive bayes, and SVM stood out to be the best predictors

LR - **72%** Accuracy

NB - **72%** Accuracy

SVM - **75%** Accuracy



Confusion matrix for the SVM model

Potential for Monetization

Prompt: "Today was absolutely magical — feeling so blessed and grateful for every moment! ✨❤️ #BestDayEver"

```
Predicted sentiment: positive
Top predictions:
  positive: 0.8777
  neutral: 0.0793
```

```
Product Recommendations:
--------------------------------------------------
1. Adventure Travel Package - $1299.99
   Turn your excitement into memories with our curated travel experiences
   Category: Travel
   Product ID: P005
--------------------------------------------------
2. Professional Camera Drone - $599.99
   Capture amazing adventures and create stunning aerial photography
   Category: Electronics
   Product ID: P003
--------------------------------------------------
3. Gourmet Chocolate Gift Box - $39.99
   Share happiness with our luxury chocolate assortment
   Category: Food
   Product ID: P004
```

Prompt: “Just another Tuesday… coffee, meetings, and more coffee. ☕ #DailyRoutine”

```
Predicted sentiment: neutral
Top predictions:
  neutral: 0.5413
  positive: 0.3174
```

```
Product Recommendations:
--------------------------------------------------
1. Stainless Steel Water Bottle - $29.99
   Durable, eco-friendly hydration for your daily routine
   Category: Lifestyle
   Product ID: U003
--------------------------------------------------
2. Monthly Planner - $19.99
   Organize your schedule with this practical and straightforward planner
   Category: Office
   Product ID: U005
--------------------------------------------------
3. Wireless Charging Pad - $39.99
   Convenient charging solution compatible with most modern devices
   Category: Electronics
   Product ID: U004
```

Prompt: “I hate when everything goes wrong at once. Seriously over today. 💔 #BadVibes”

```
Predicted sentiment: negative
Top predictions:
  negative: 0.7271
  neutral: 0.2296
```

```
Product Recommendations:
--------------------------------------------------
1. Guided Meditation App Subscription - $9.99
   Find peace and overcome negative emotions with expert-led meditation sessions
   Category: Digital
   Product ID: N003
--------------------------------------------------
2. Problem-Solving Strategy Book - $24.99
   Turn difficulties into opportunities with practical problem-solving techniques
   Category: Books
   Product ID: N005
--------------------------------------------------
3. Noise-Cancelling Headphones - $249.99
   Escape the frustrations of a noisy environment with our premium headphones
   Category: Electronics
   Product ID: N002
```