Sentiment Analysis of Social Media Content

Final Project Presentation
Data Science Bootcamp Spring 2025

Team: Data Busters
Anushka Garg – ag10687
Nobodit
Debbika
Hrishik
Christopher

Project Overview

- **Goal:** Analyze social media posts to uncover emotional trends and user behavior.
- **Dataset includes** text, sentiments, timestamps, hashtags, engagement metrics, and geography.
- **Focus:** Sentiment patterns, hashtag trends, platform and regional engagement insights.

Problem Statement

1. Classify sentiment from user content.
2. Explore emotional trends over time and location.
3. Analyze user engagement (likes/retweets).
4. Compare platform-specific and hashtag-based behavior.

Dataset Exploration

- Cleaned unnecessary columns and standardized key features.
- Extracted time features: Hour, Day, Month.
- Parsed hashtags and cleaned platform/country fields.
- **Selected features:** Text, Sentiment, Time, Engagement, Platform, Country, Hashtags.

Sentiment Cleaning

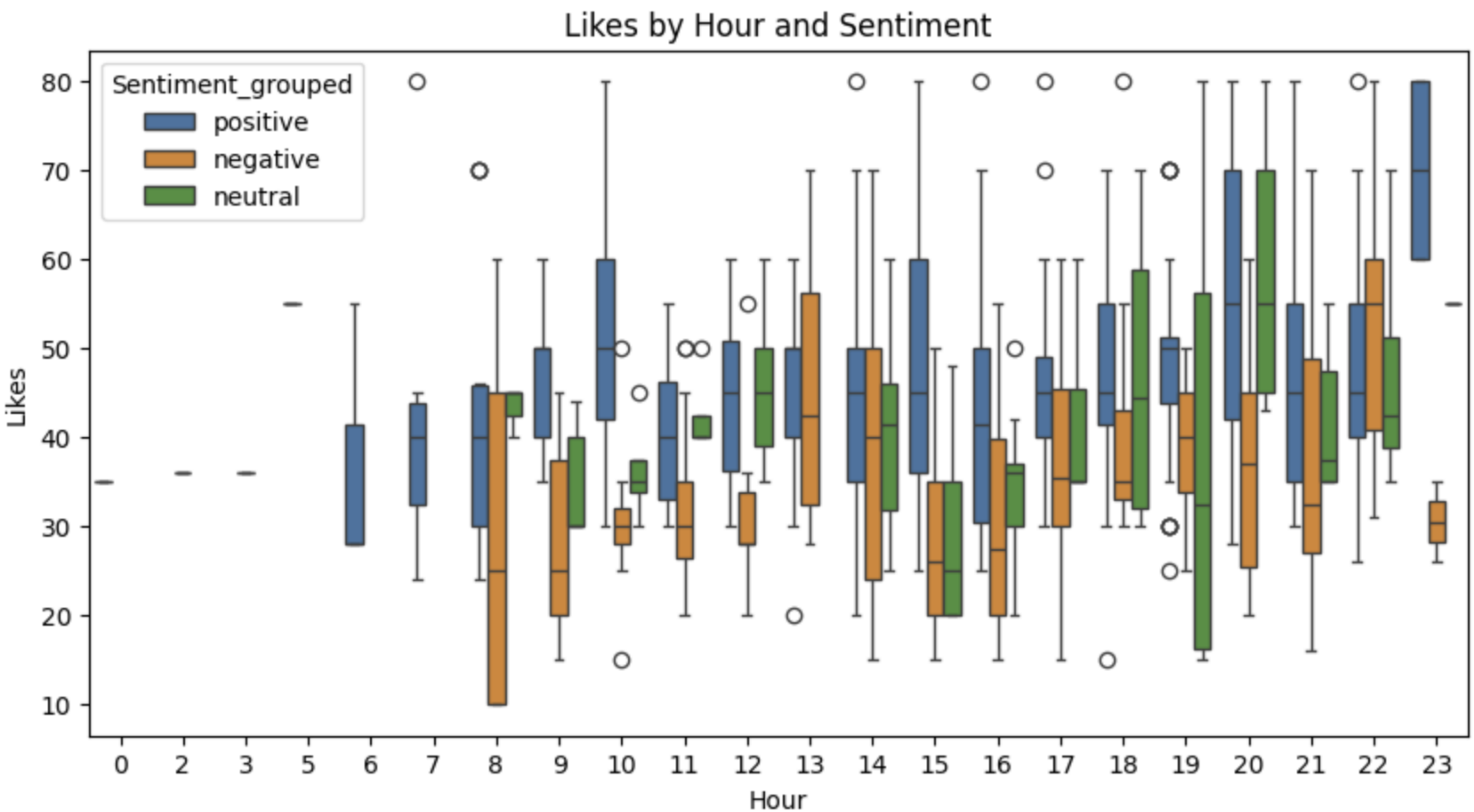
- Over 270 raw sentiment values grouped into: positive, neutral, negative.
- Used domain-based mapping to ensure 100% coverage.
- Prepared 'Sentiment_grouped' column for consistent analysis.

```
Remaining unmapped sentiments:
 Series([], Name: count, dtype: int64)
Remaining unmapped total: 0
```

Temporal & Platform Trends

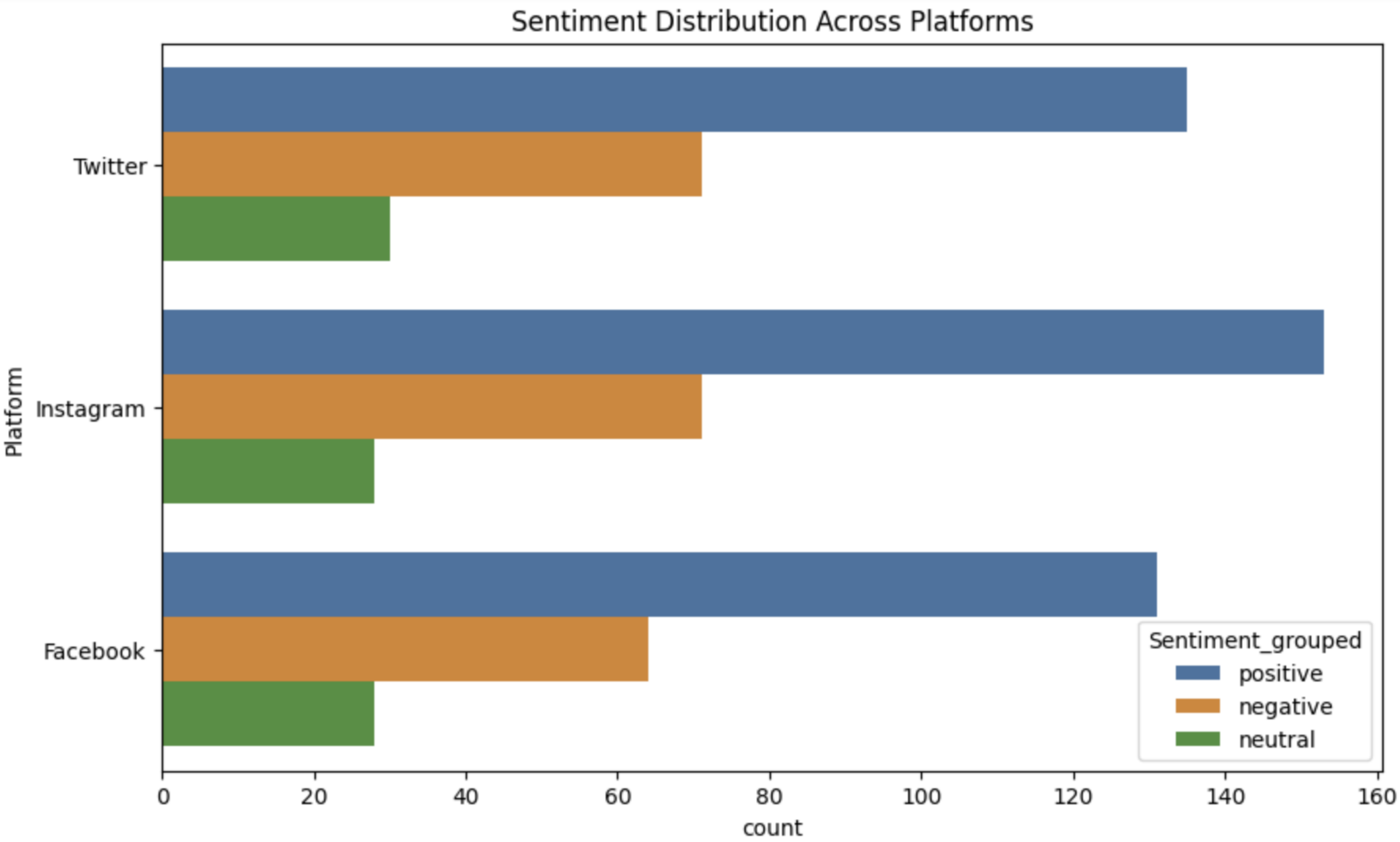
- Analyzed sentiment by hour of day.
- Detected engagement peaks during mid-day.
- Twitter: Neutral | Instagram: Positive | Facebook: Balanced

Results:



Likes by Hour and Sentiment
Sentiment_grouped
positive
negative
neutral
Likes
80
70
60
50
40
30
20
10
0
2
3
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

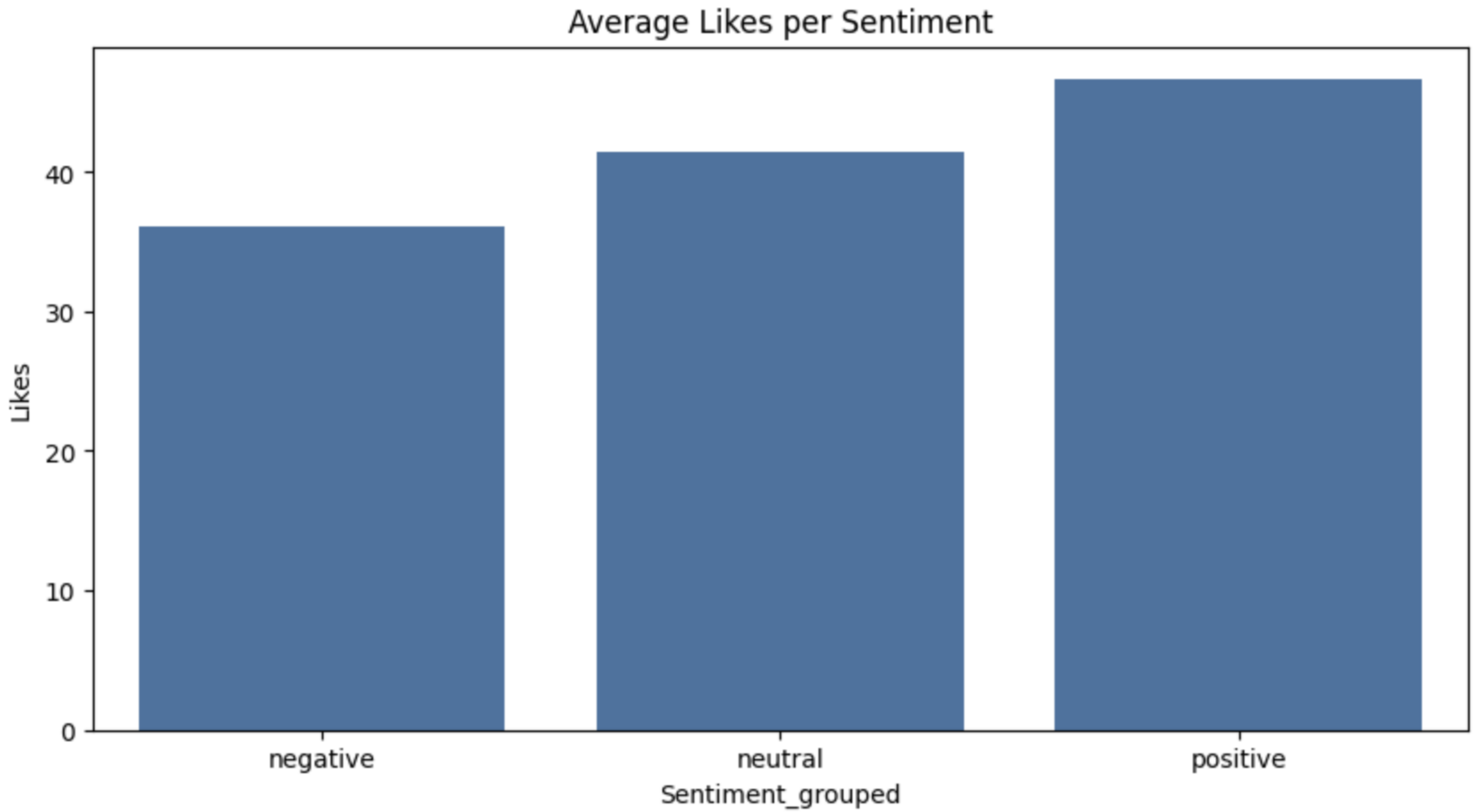
Results:


Sentiment Distribution Across Platforms
Twitter
Instagram
Facebook
Platform
0
20
40
60
80
100
120
140
160
count
Sentiment_grouped
positive
negative
neutral

User Engagement Insights

- Analyzed average Likes and Retweets by sentiment.
- Found positive sentiment yields higher engagement.
- Likes vs. Retweets Correlation: 0.998

Results:

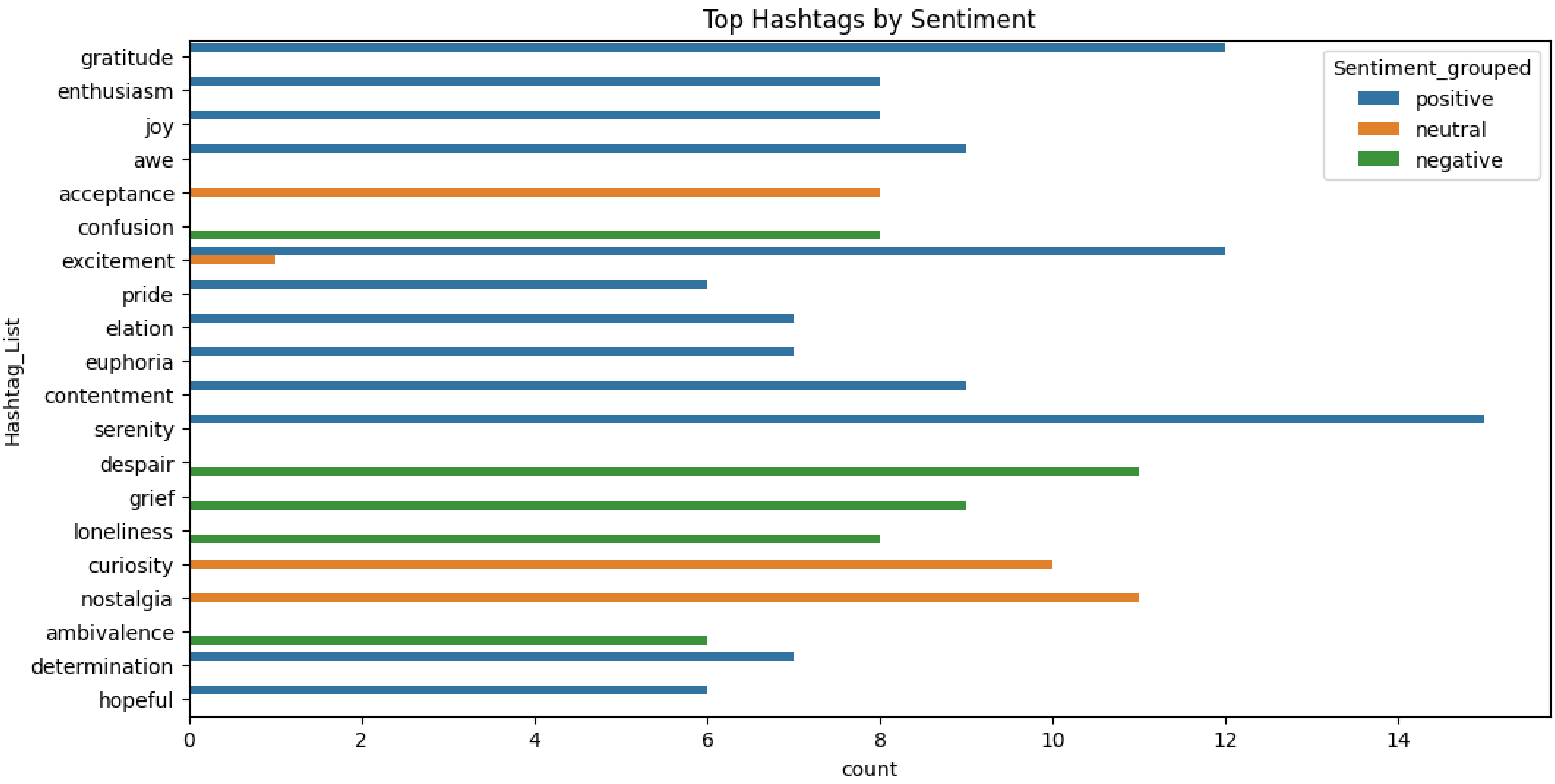


Average Likes per Sentiment
Likes
0
10
20
30
40
negative
neutral
positive
Sentiment_grouped

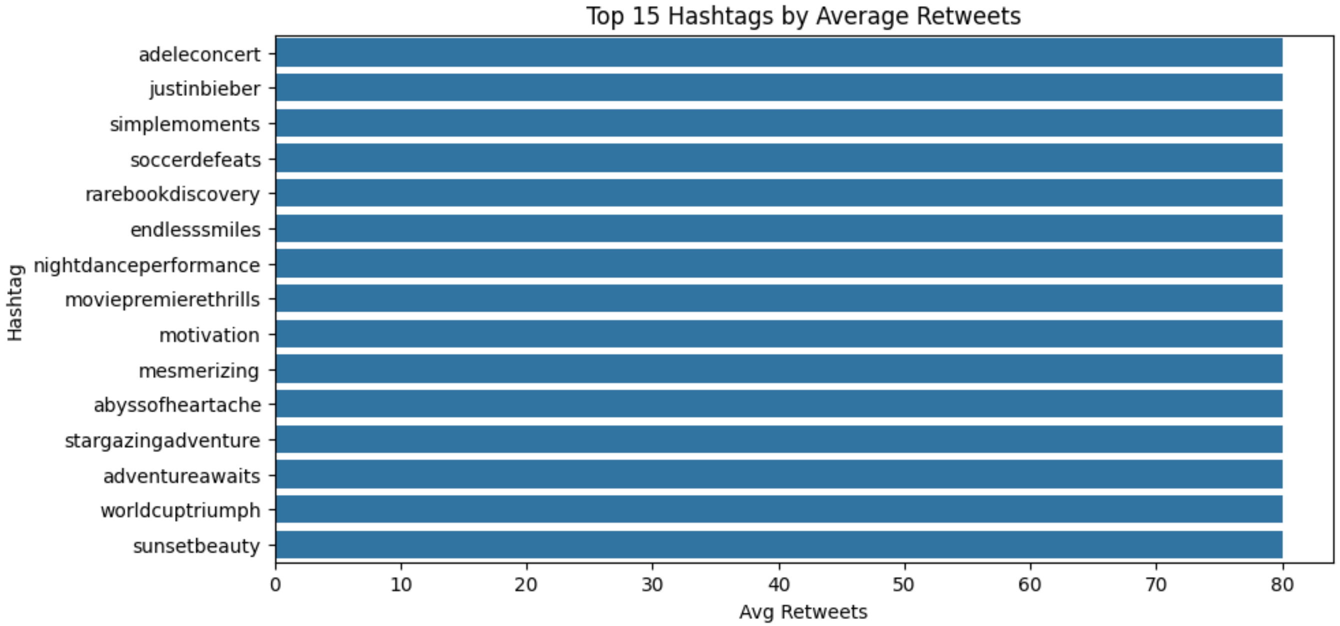
Hashtag Analysis

- Exploded hashtags for analysis.
- Plotted: top hashtags by count, sentiment, likes, retweets.
- High engagement examples: **#adeleconcert, #motivation, #sunsetbeauty**

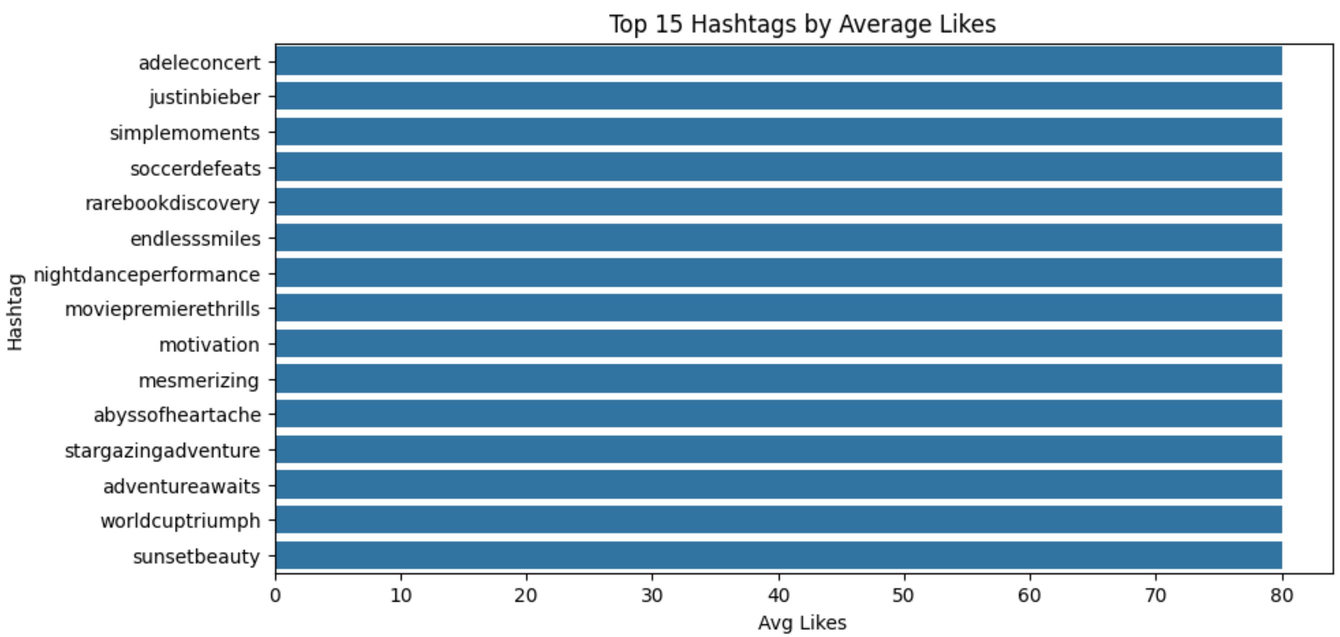
Results:



Top Hashtags by Sentiment
gratitude
enthusiasm
joy
awe
acceptance
confusion
excitement
pride
elation
euphoria
contentment
serenity
despair
grief
loneliness
curiosity
nostalgia
ambivalence
determination
hopeful
Hashtag_List
0
2
4
6
8
10
12
14
count
Sentiment_grouped
positive
neutral
negative



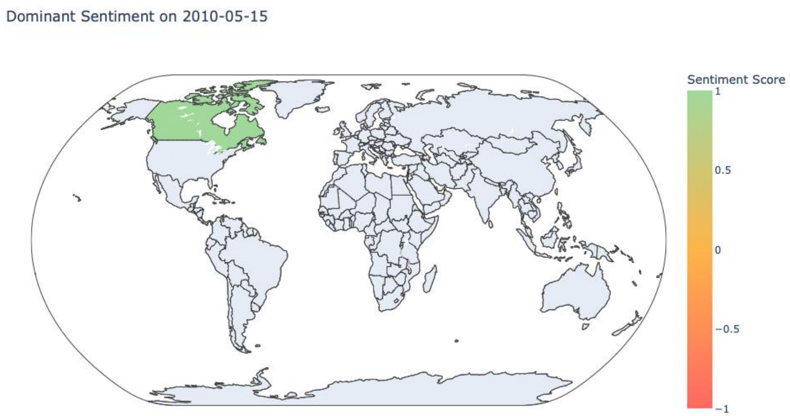
	Likes	Retweets
Likes	1.000000	0.998476
Retweets	0.998476	1.000000



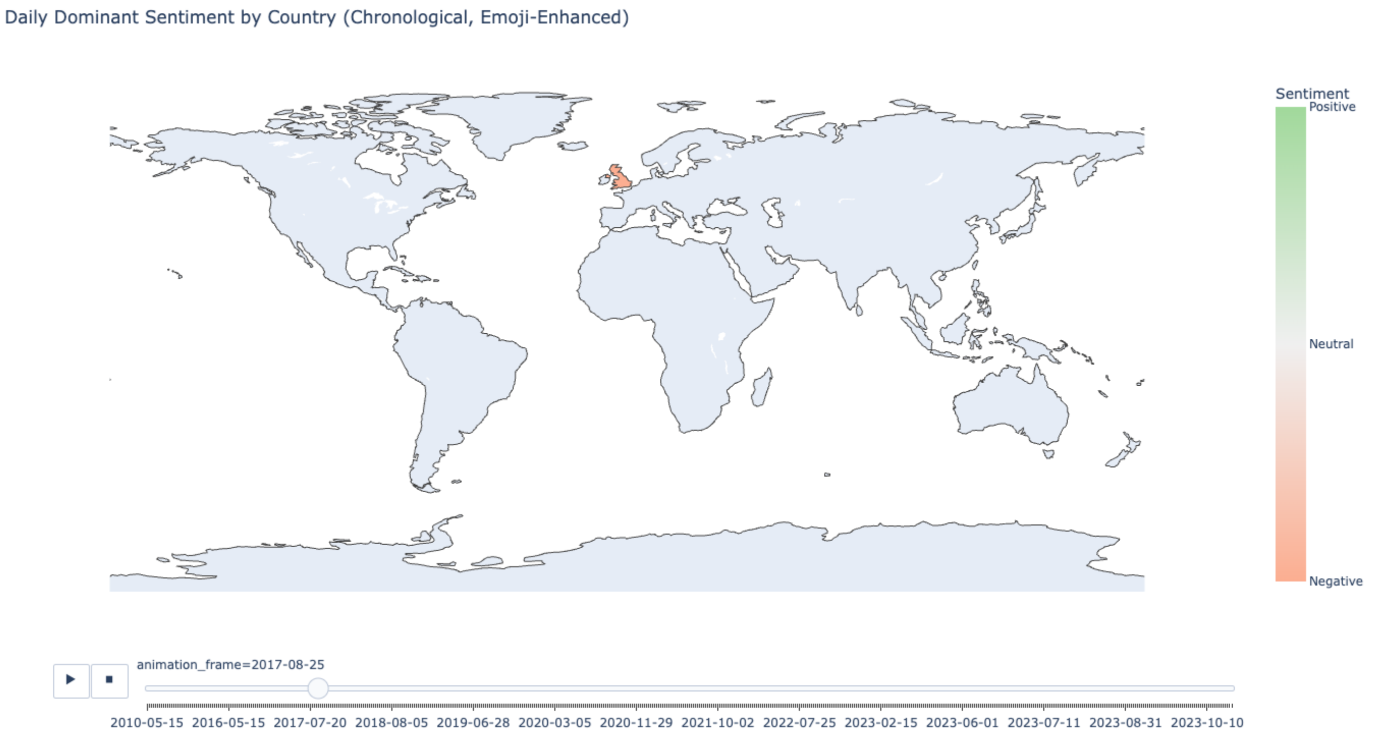
Geographical Trends

- Standardized country names.
- Top 10 countries plotted by sentiment group.
- USA, UK trend positive; some regions neutral.

Results:

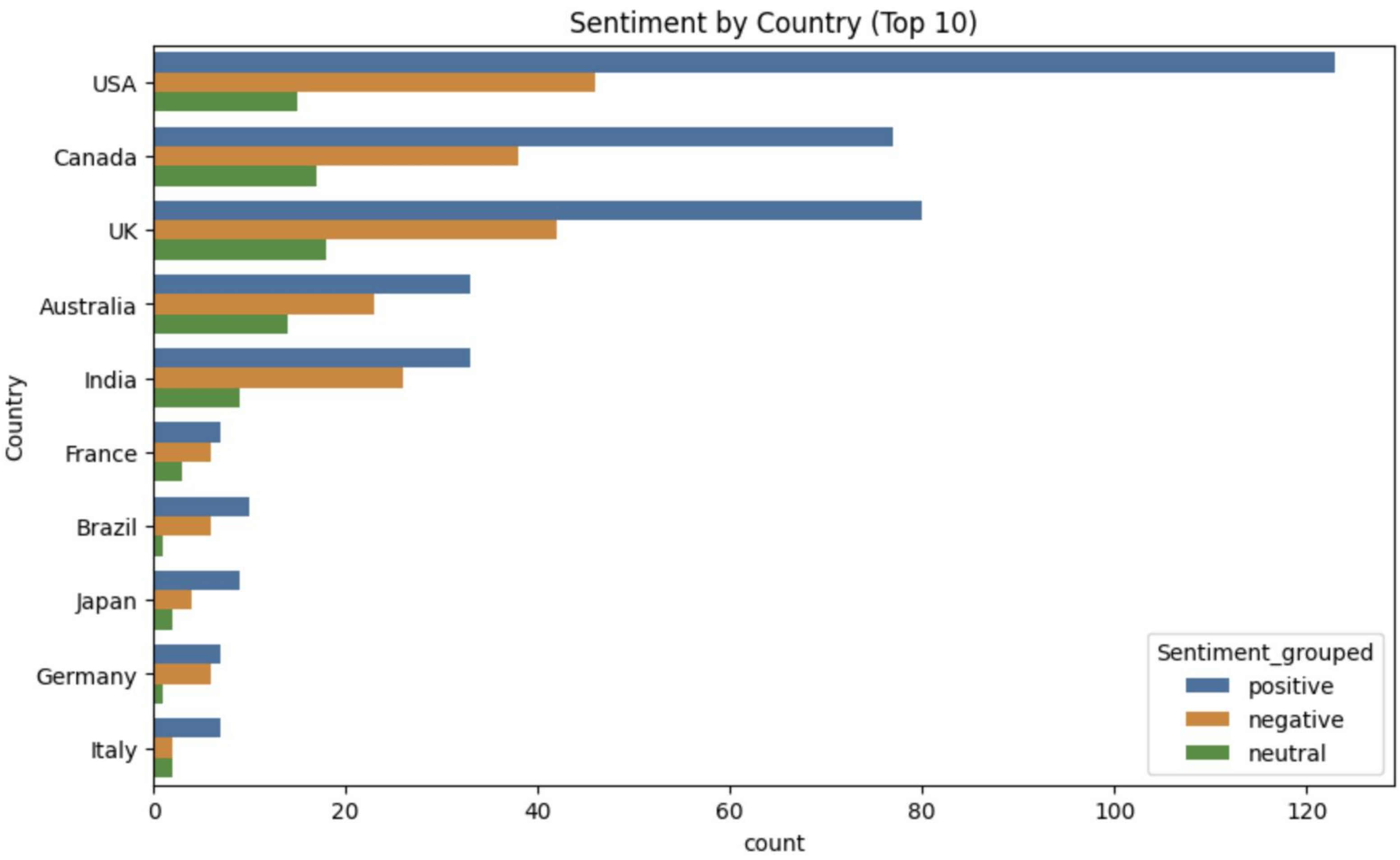


Dominant Sentiment on 2010-05-15
Sentiment Score
1
0.5
0
−0.5
−1



Daily Dominant Sentiment by Country (Chronological, Emoji-Enhanced)
Sentiment
Positive
Neutral
Negative
animation_frame=2017-08-25
2010-05-15
2016-05-15
2017-07-20
2018-08-05
2019-06-28
2020-03-05
2020-11-29
2021-10-02
2022-07-25
2023-02-15
2023-06-01
2023-07-11
2023-08-31
2023-10-10

Results:


Sentiment by Country (Top 10)
USA
Canada
UK
Australia
India
France
Brazil
Japan
Germany
Italy
Country
0
20
40
60
80
100
120
count
Sentiment_grouped
positive
negative
neutral

Timeline & Progress

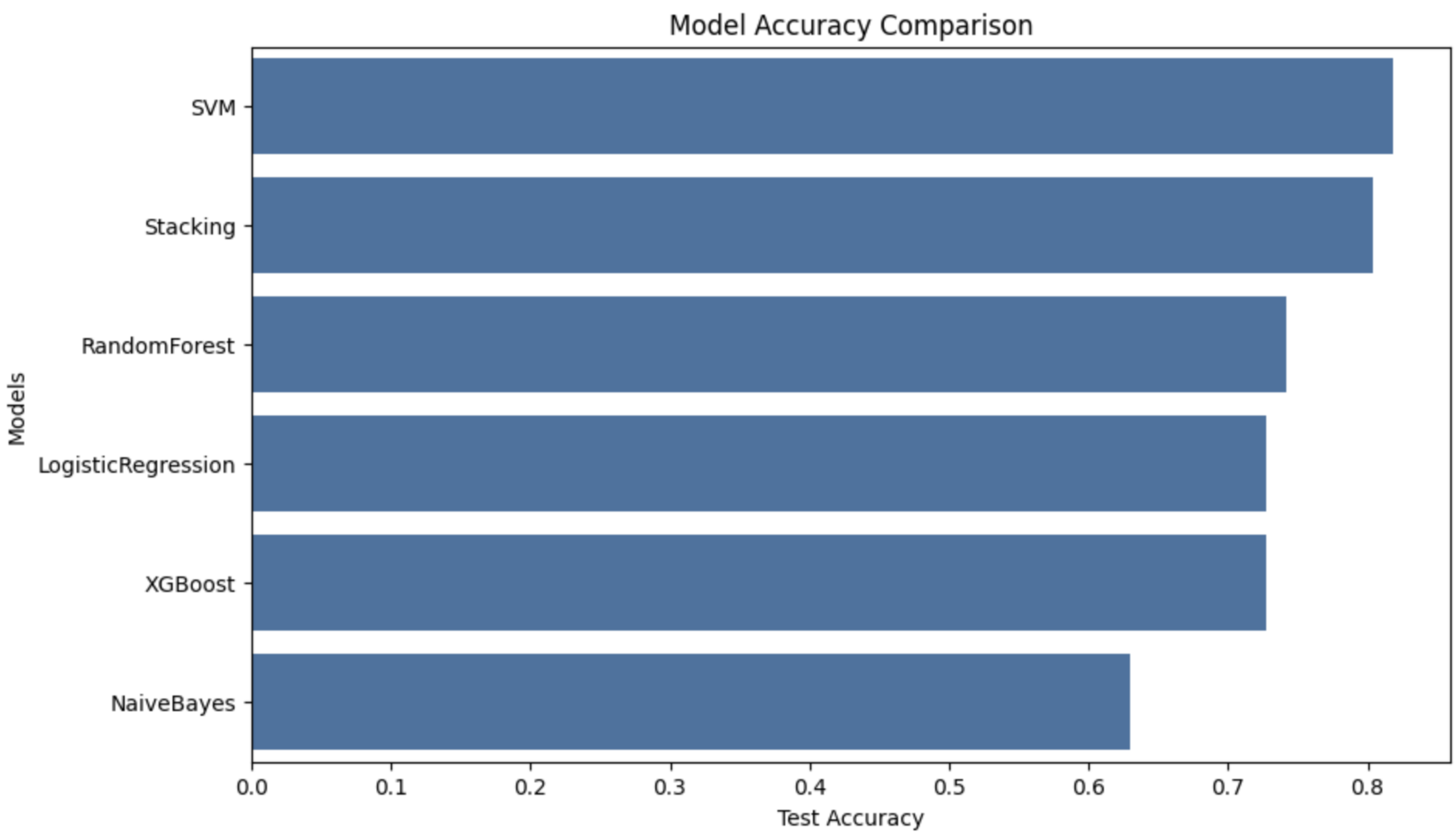
- ✅ Dataset Exploration
- ✅ Feature Engineering
- ✅ Sentiment Grouping
- ✅ EDA & Visualization
- ✅ Hashtag & Engagement Analysis
- 🔜 Predictive Modeling (Week 6)

Next Steps

1. Train models to predict sentiment and likes.
2. Conduct hypothesis testing on engagement trends.
3. Interpret model results and build final presentation.
4. Polish GitHub repo and create final report.

Predictive Modeling - Classification

- Used TF-IDF + metadata (platform, hashtags, time) as features
- Trained 5 models: Logistic Regression, Naive Bayes, Random Forest, XGBoost, SVM
- Compared test accuracies across all models
- **SVM achieved the highest performance**


Model Accuracy Comparison
SVM
Stacking
RandomForest
LogisticRegression
XGBoost
NaiveBayes
Models
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Test Accuracy

Model Evaluation (Before SMOTE)

- Best Parameters: C=1, kernel=`linear`
- Test Accuracy: **81.81%**
- Macro F1 Score: **0.74**
- Class imbalance impacted recall for Negative sentiment class (recall = 0.37)

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Best Parameters: {'C': 1, 'gamma': 'scale', 'kernel': 'linear'}
Best Cross-Validation Accuracy: 0.836283185840708

Test Accuracy: 0.8181818181818182

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.82      0.83        44
           1       1.00      0.37      0.54        19
           2       0.80      0.93      0.86        80

    accuracy                           0.82       143
   macro avg       0.88      0.70      0.74       143
weighted avg       0.84      0.82      0.80       143
```

SMOTE & Improvement

Performance After Balancing with SMOTE:

- Applied SMOTE to balance class distribution
- Recall for negative sentiment improved: 0.37 → 0.47
- Test Accuracy: **82.5%**
- Macro F1: 0.77 | Weighted F1: 0.82

```
Accuracy: 0.8251748251748252
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.84      0.84        44
           1       0.82      0.47      0.60        19
           2       0.82      0.90      0.86        80

    accuracy                           0.83       143
   macro avg       0.83      0.74      0.77       143
weighted avg       0.83      0.83      0.82       143
```

```
Best Params: {'C': 10, 'kernel': 'linear'}
Test Accuracy: 0.8111888111888111
```

Hyperparameter Tuning

- Re-ran GridSearch on SMOTE-balanced data
- Best Parameters: C=`10`, kernel=`linear`
- Slight drop in accuracy → stuck with original (C=1)
- Final model = SVM + SMOTE (C=1, linear)

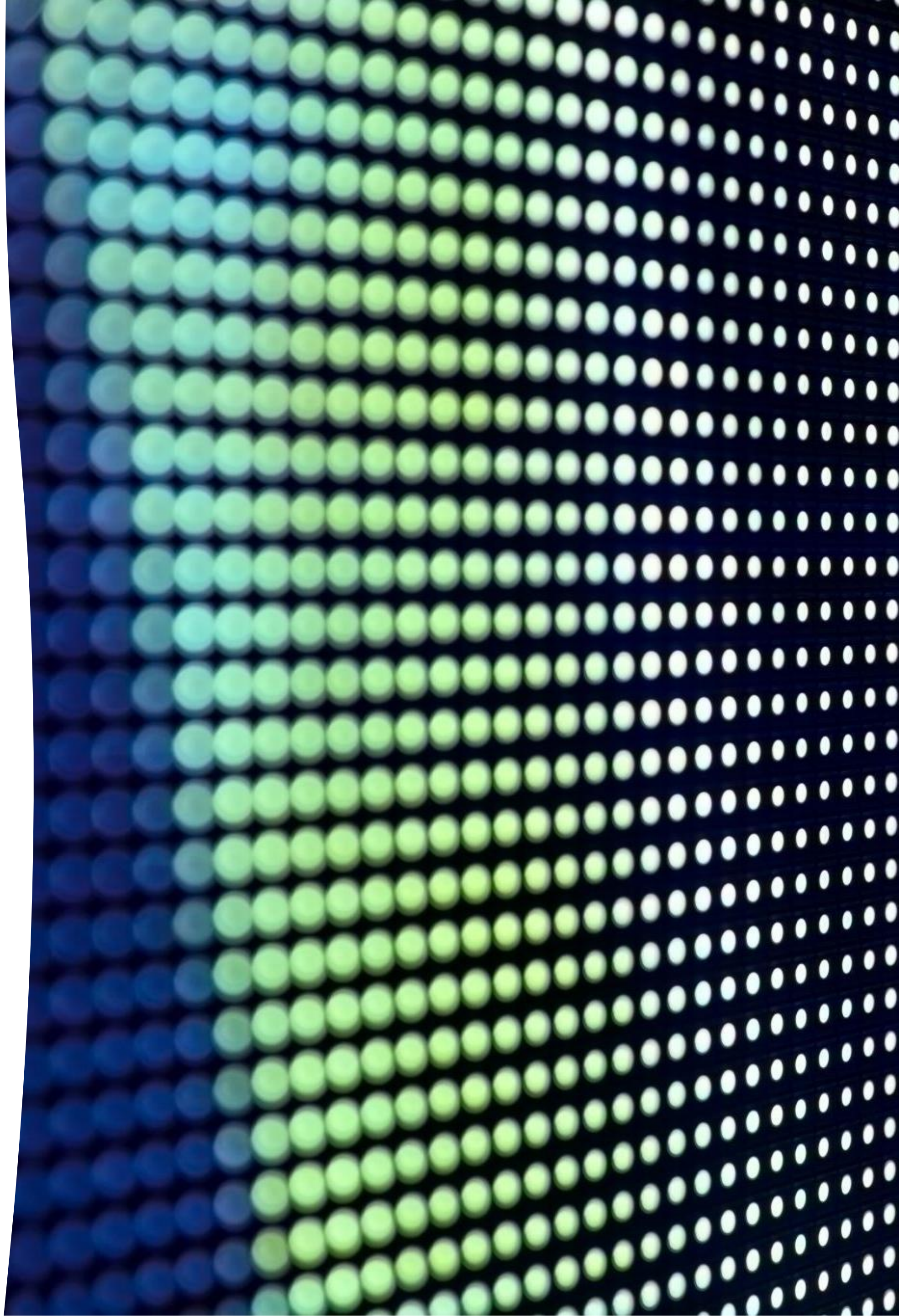
Engagement Prediction - Regression

Used Random Forest Regressor to predict post likes

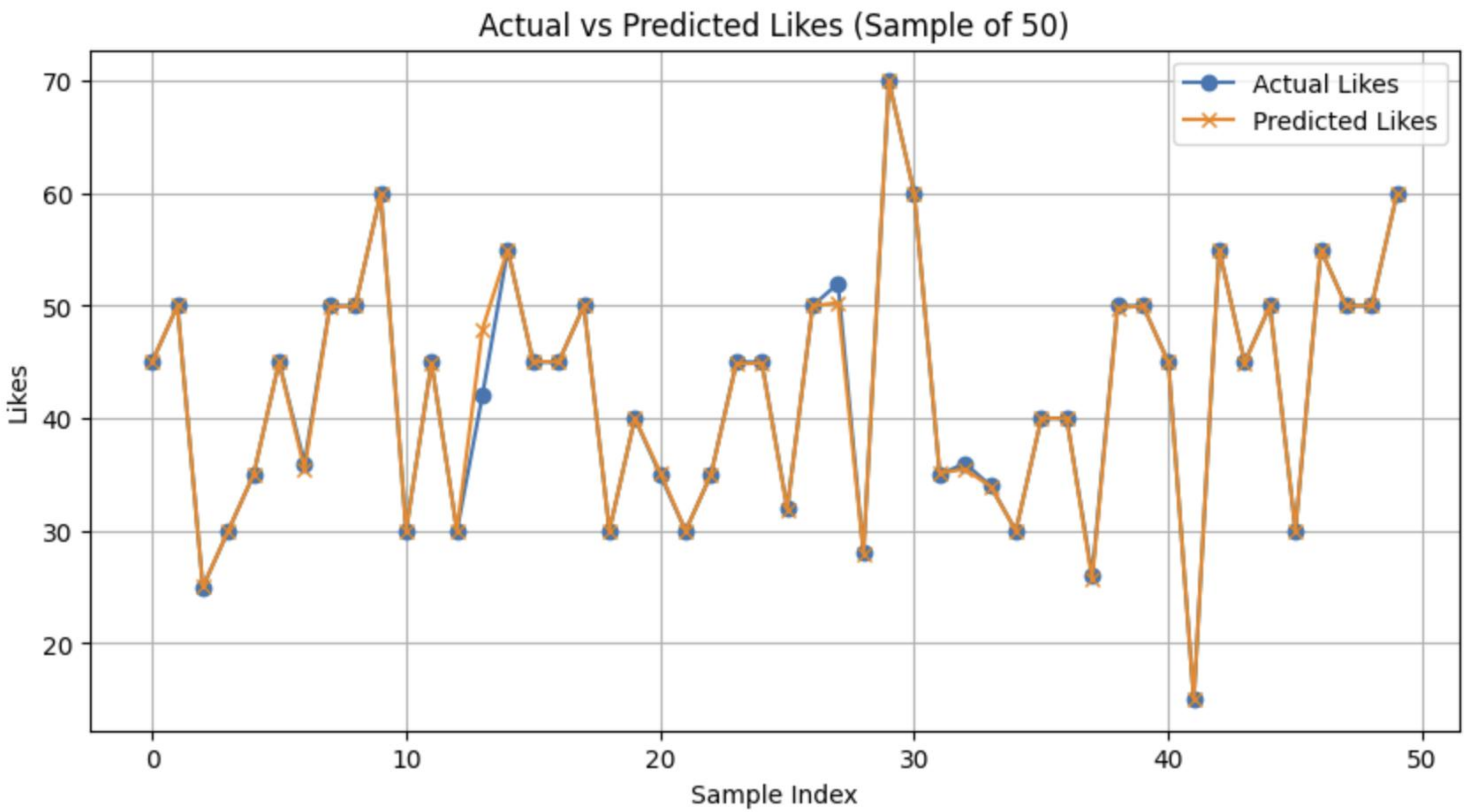
Features: TF-IDF + Time + Platform + Hashtags

RMSE: **0.63**

R^2 Score: **0.997** → Excellent fit



✅ Regression Results:
RMSE: 0.6300510579820776
R² Score: 0.9976485804042663



Final Insights

SVM + SMOTE = BEST SENTIMENT CLASSIFIER WITH BALANCED RECALL & PRECISION

POSITIVE POSTS LEAD TO HIGHER USER ENGAGEMENT

HASHTAGS LIKE #MOTIVATION, #ADELECONCERT DRIVE MOST LIKES

PREDICTIVE REGRESSION MODEL IS HIGHLY RELIABLE FOR ENGAGEMENT FORECASTING

Thank You

QUESTIONS?

TEAM: DATA BUSTERS

GITHUB:
GITHUB.COM/DOTBION/SENTIMENT-ANALYSIS-NYU-DSB