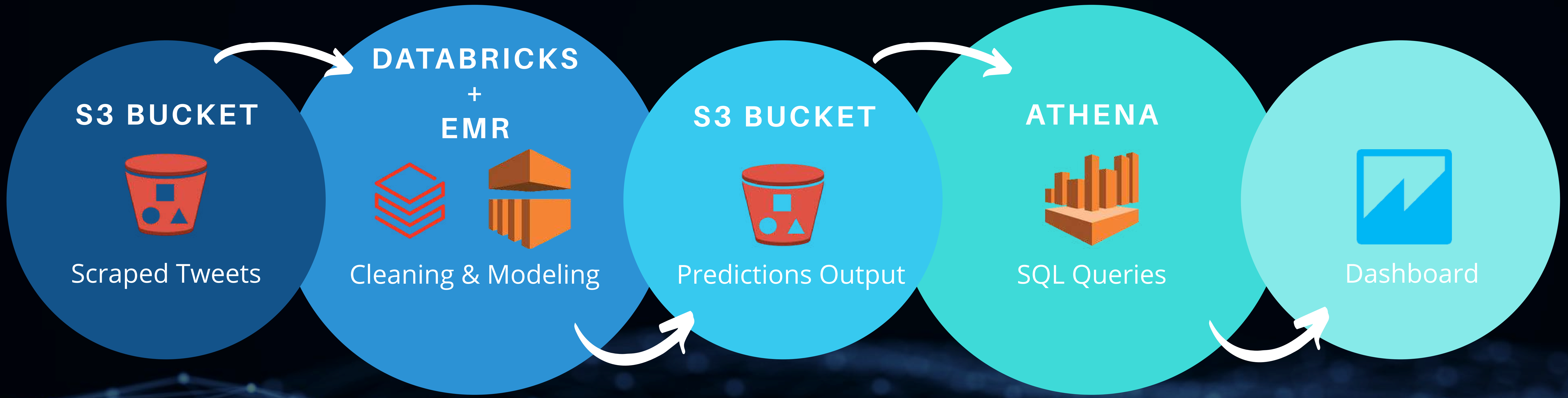
# SENTIMENT ANALYSIS PROJECT

## ELON MUSK DATASET 21-22.11.2022

KARINA OBORSKA

# DATA FLOW



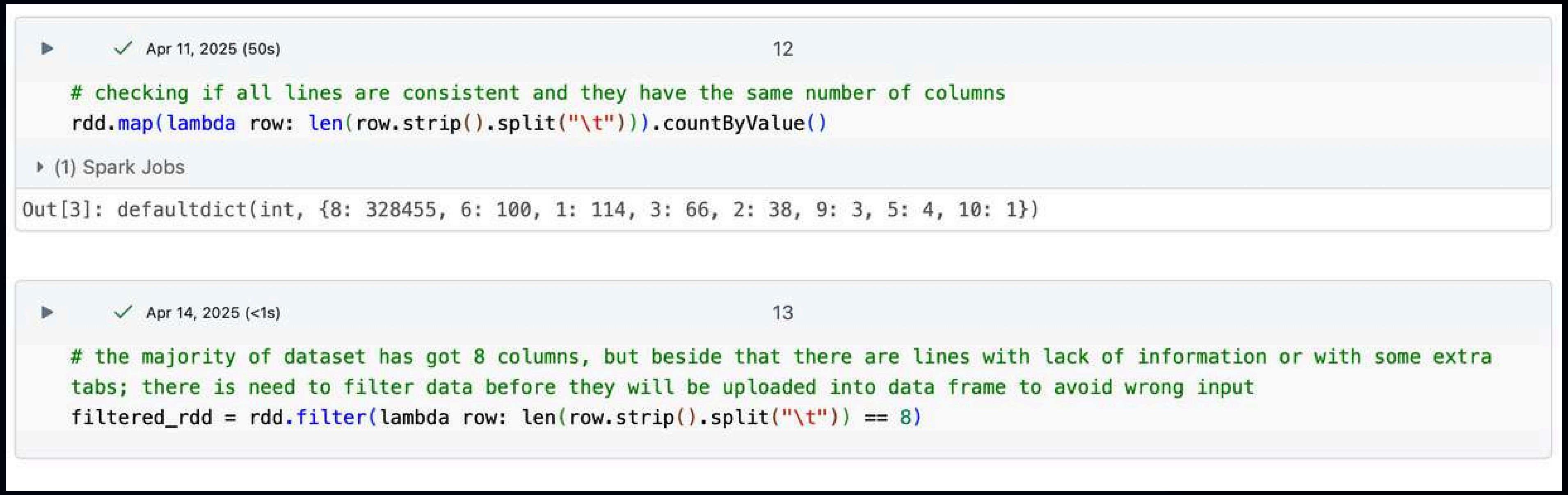
This project uses a cloud-based pipeline to collect, process, model, and visualize tweet data using AWS and Databricks.

# DATASET INFORMATION

- Dataset contains 328,766 rows, but initial files had inconsistent formatting (different number of columns due to incorrect delimiters).

- After filtering, some rows still had nulls or merged columns, so I removed corrupted ones as it was just ~0.07%.

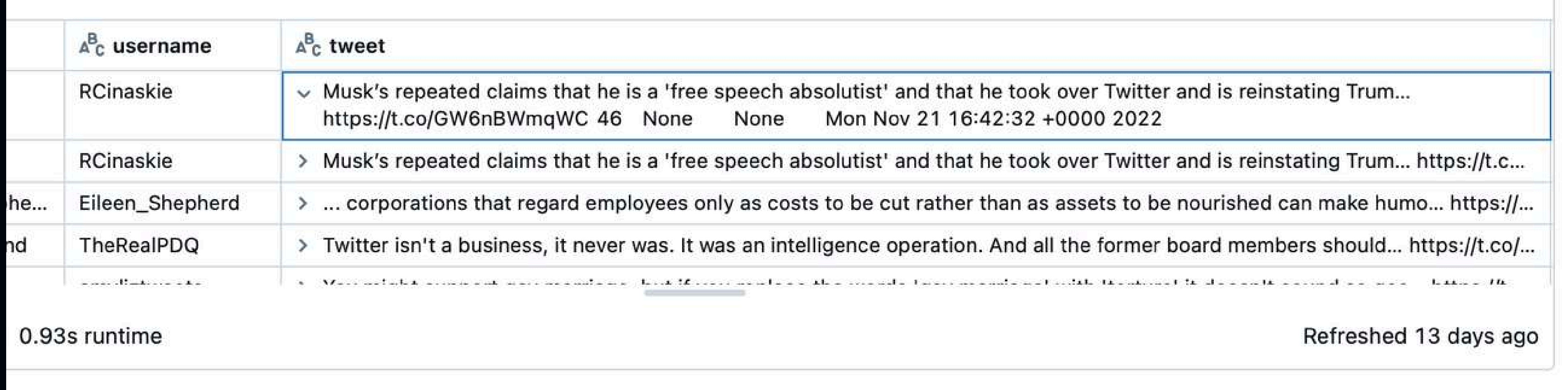
- Since sentiment was missing, I used VADER for labeling – it works well for social media, but struggles with sarcasm and context, which affected some predictions.



```
# checking if all lines are consistent and they have the same number of columns
rdd.map(lambda row: len(row.strip().split("\t"))).countByValue()
```

```
Out[3]: defaultdict(int, {8: 328455, 6: 100, 1: 114, 3: 66, 2: 38, 9: 3, 5: 4, 10: 1})
```

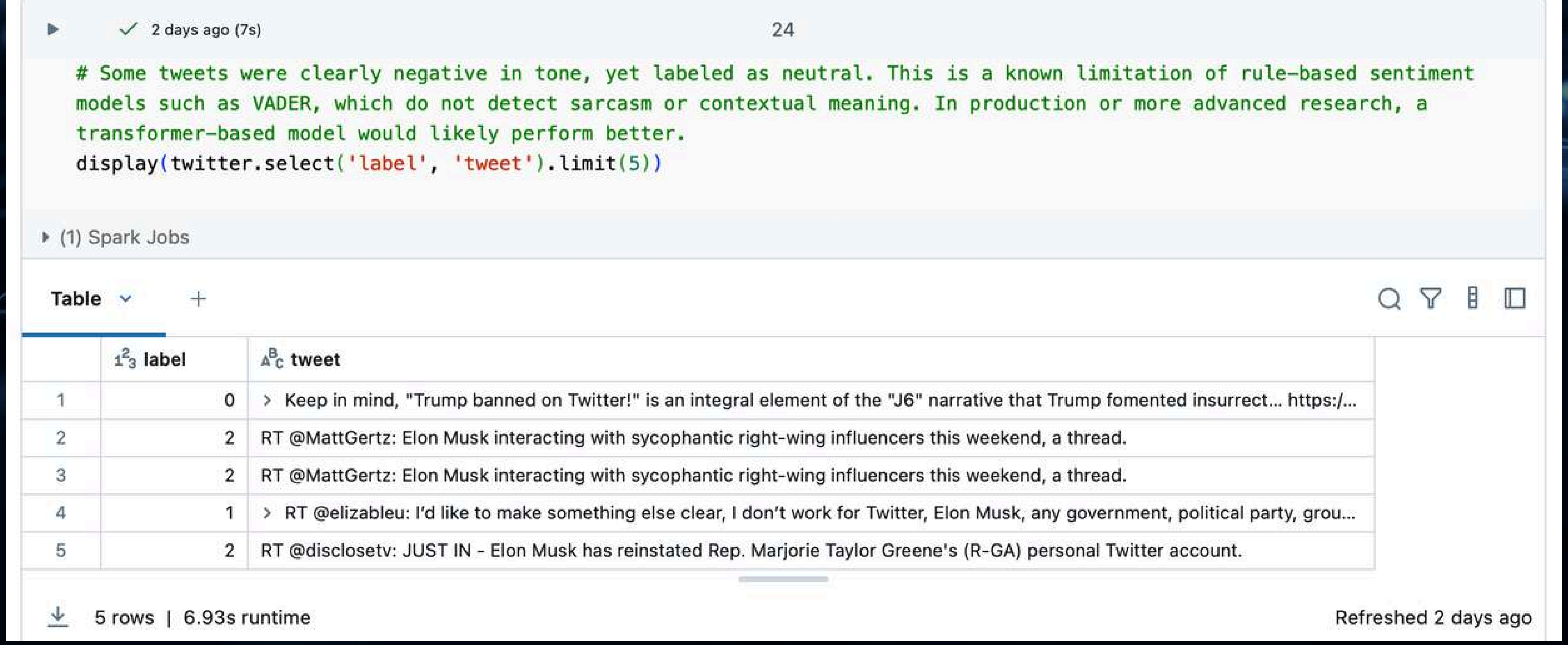
```
# the majority of dataset has got 8 columns, but beside that there are lines with lack of information or with some extra
tabs; there is need to filter data before they will be uploaded into data frame to avoid wrong input
filtered_rdd = rdd.filter(lambda row: len(row.strip().split("\t")) == 8)
```



| | username | tweet |
|---|---|---|
| | RCinaskie | ⌄ Musk's repeated claims that he is a 'free speech absolutist' and that he took over Twitter and is reinstating Trum... https://t.co/GW6nBWmqWC 46 None None Mon Nov 21 16:42:32 +0000 2022 |
| | RCinaskie | › Musk's repeated claims that he is a 'free speech absolutist' and that he took over Twitter and is reinstating Trum... https://t.c... |
| he... | Eileen_Shepherd | › ... corporations that regard employees only as costs to be cut rather than as assets to be nourished can make humo... https://... |
| nd | TheRealPDQ | › Twitter isn't a business, it never was. It was an intelligence operation. And all the former board members should... https://t.co/... |

0.93s runtime

Refreshed 13 days ago



```
# Some tweets were clearly negative in tone, yet labeled as neutral. This is a known limitation of rule-based sentiment
models such as VADER, which do not detect sarcasm or contextual meaning. In production or more advanced research, a
transformer-based model would likely perform better.
display(twitter.select('label', 'tweet').limit(5))
```

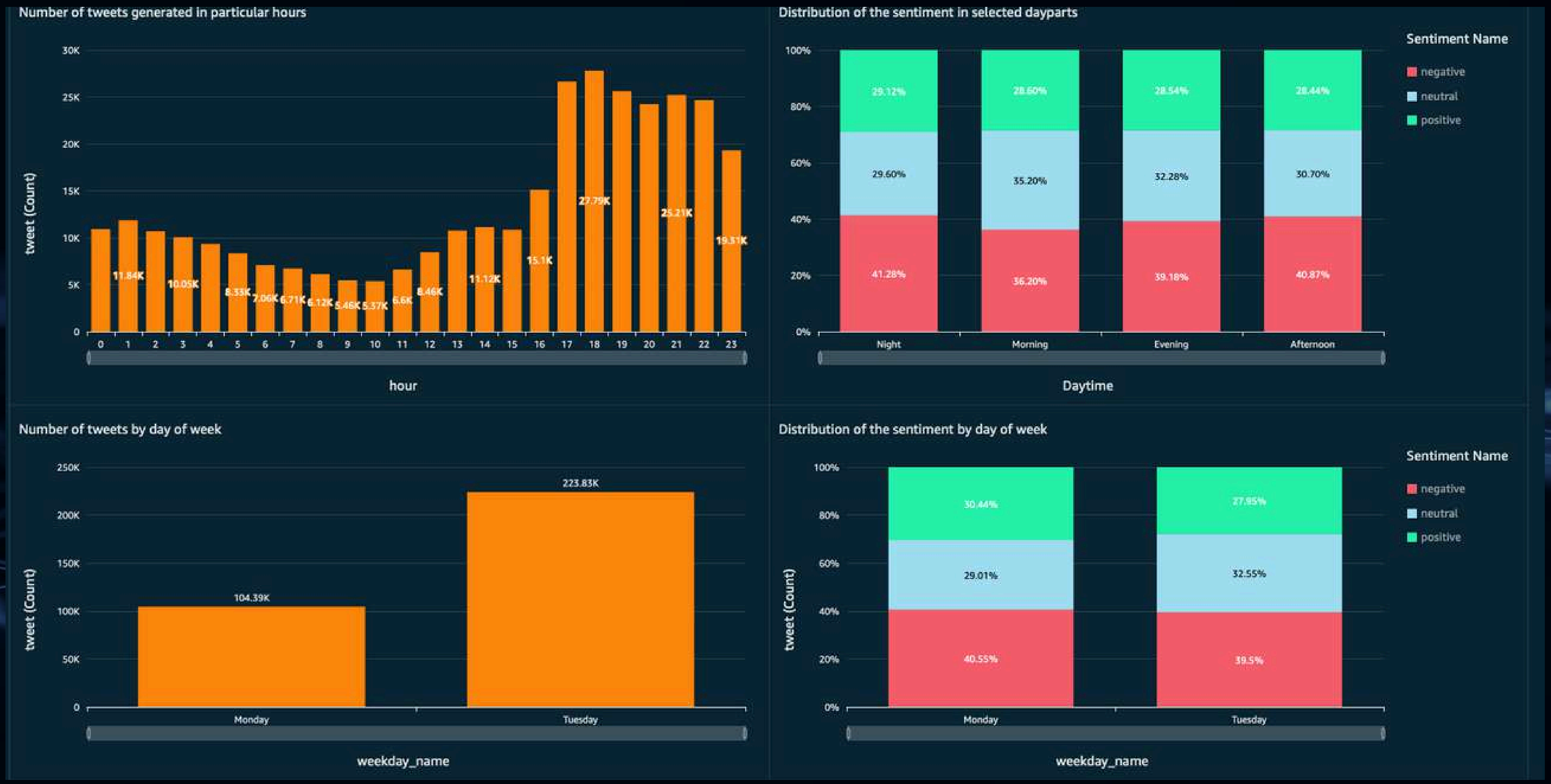
| | label | tweet |
|---|---|---|
| 1 | 0 | › Keep in mind, "Trump banned on Twitter!" is an integral element of the "J6" narrative that Trump fomented insurrect... https:/... |
| 2 | 2 | RT @MattGertz: Elon Musk interacting with sycophantic right-wing influencers this weekend, a thread. |
| 3 | 2 | RT @MattGertz: Elon Musk interacting with sycophantic right-wing influencers this weekend, a thread. |
| 4 | 1 | › RT @elizableu: I'd like to make something else clear, I don't work for Twitter, Elon Musk, any government, political party, grou... |
| 5 | 2 | RT @disclosetv: JUST IN - Elon Musk has reinstated Rep. Marjorie Taylor Greene's (R-GA) personal Twitter account. |

5 rows | 6.93s runtime

Refreshed 2 days ago

# DATA ANALYSIS - TIME OF TWEETS

- Most tweets were posted between 5 PM and 11 PM.
- Tweet volume on Tuesday was twice as high as on Monday.
- Slight increase in negative sentiment at night.
- Sentiment distribution is stable across dayparts and weekdays.

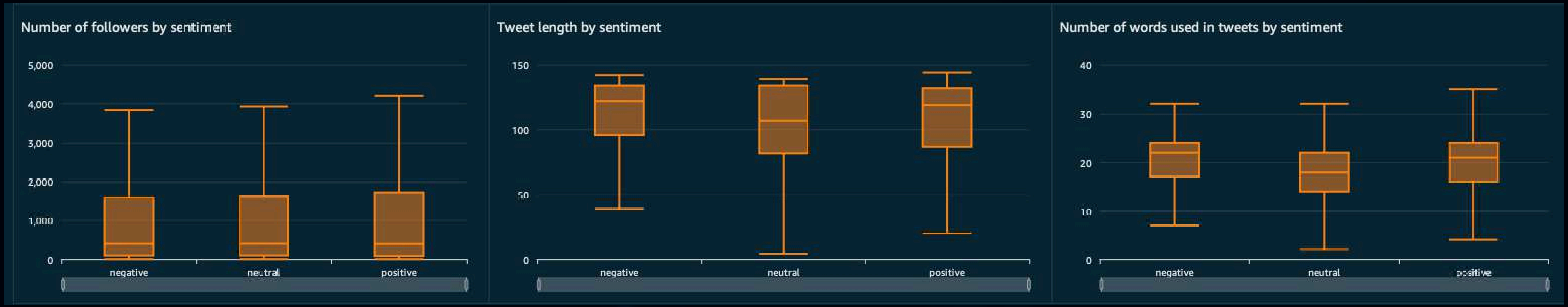

Number of tweets generated in particular hours

Distribution of the sentiment in selected dayparts

Number of tweets by day of week

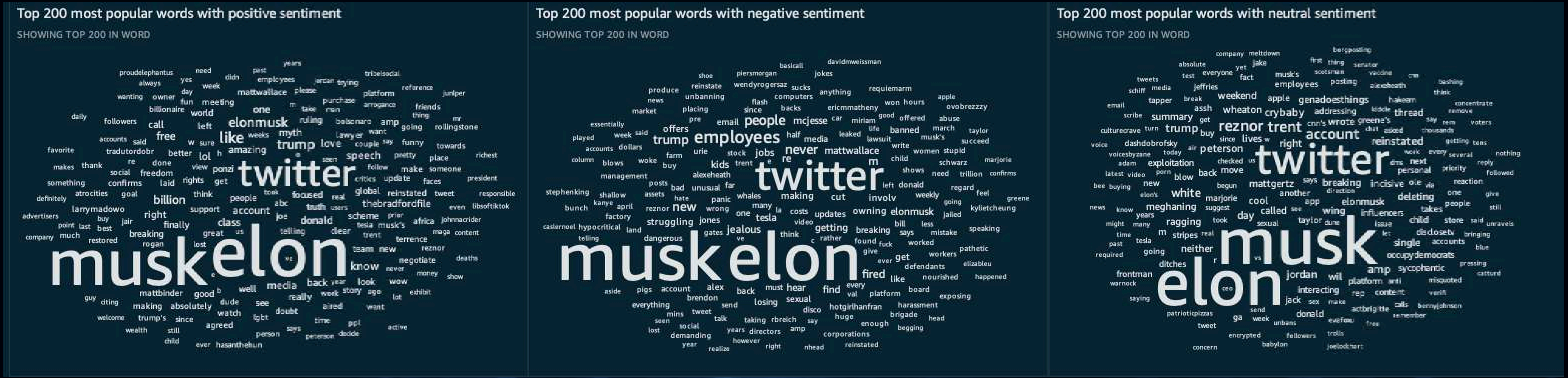
Distribution of the sentiment by day of week

# DATA ANALYSIS - FOLLOWERS/TWEET STRUCTURE



- Positive tweets slightly more common among users with more followers.
- Negative tweets tend to be longer; neutral ones are the shortest.
- Word count follows a similar pattern to tweet length.

# DATA ANALYSIS - WORD CLOUDS BY SENTIMENT



Top 200 most popular words with positive sentiment

SHOWING TOP 200 IN WORD

Top 200 most popular words with negative sentiment

SHOWING TOP 200 IN WORD

Top 200 most popular words with neutral sentiment

SHOWING TOP 200 IN WORD

Word clouds reveal distinct differences in vocabulary across sentiments.

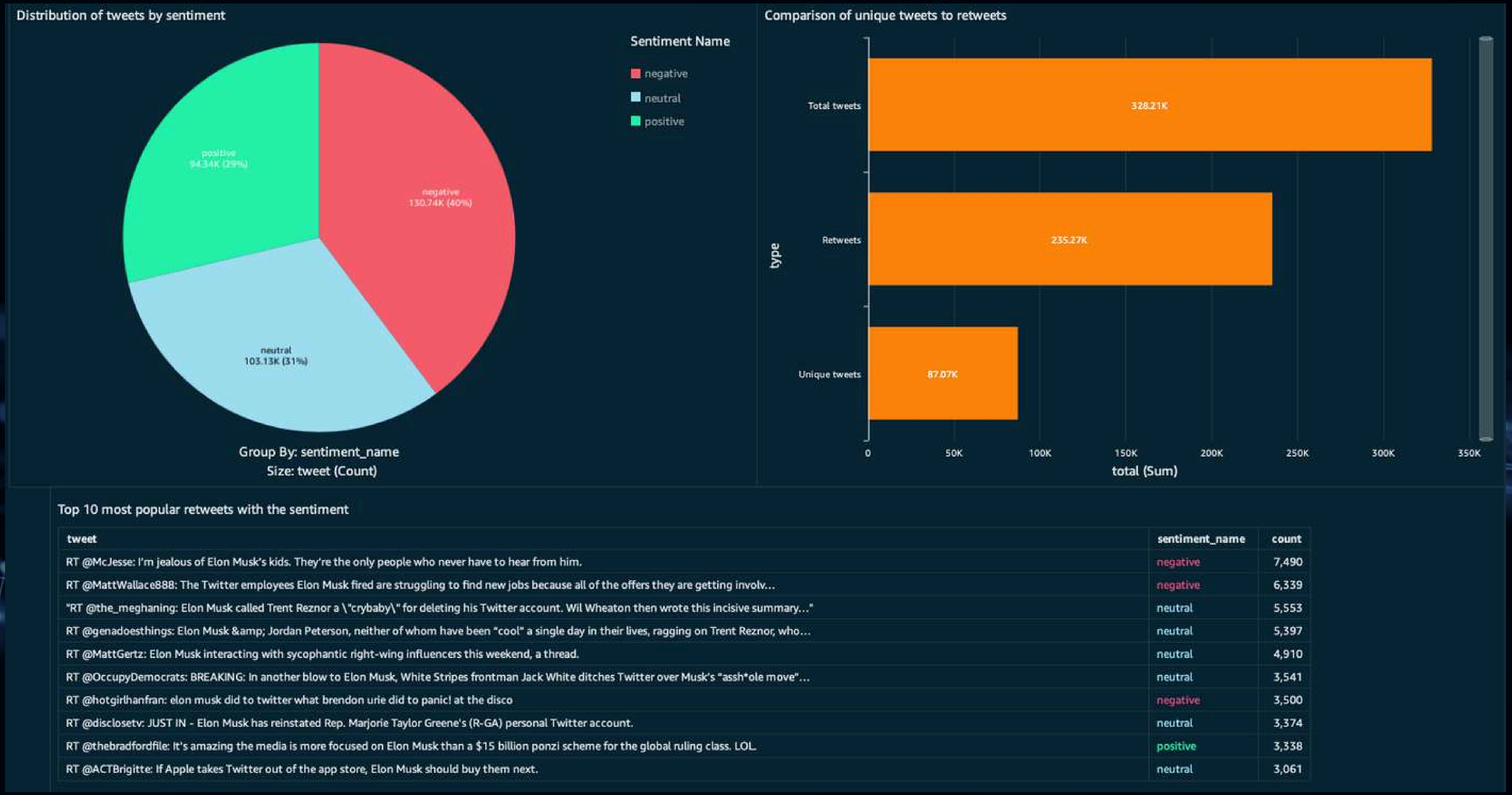
# DATA ANALYSIS - SENTIMENT DISTRIBUTION/RETWEETS

- Sentiment classes are slightly imbalanced: most tweets are negative (~40%).

- Only 27% of tweets are unique — the remaining 73% are retweets.

- Some retweets appear over 7.5k times, so random splitting could cause data leakage.

- Data was split chronologically to minimize this risk — the test set contains the latest 20% of tweets.



Distribution of tweets by sentiment

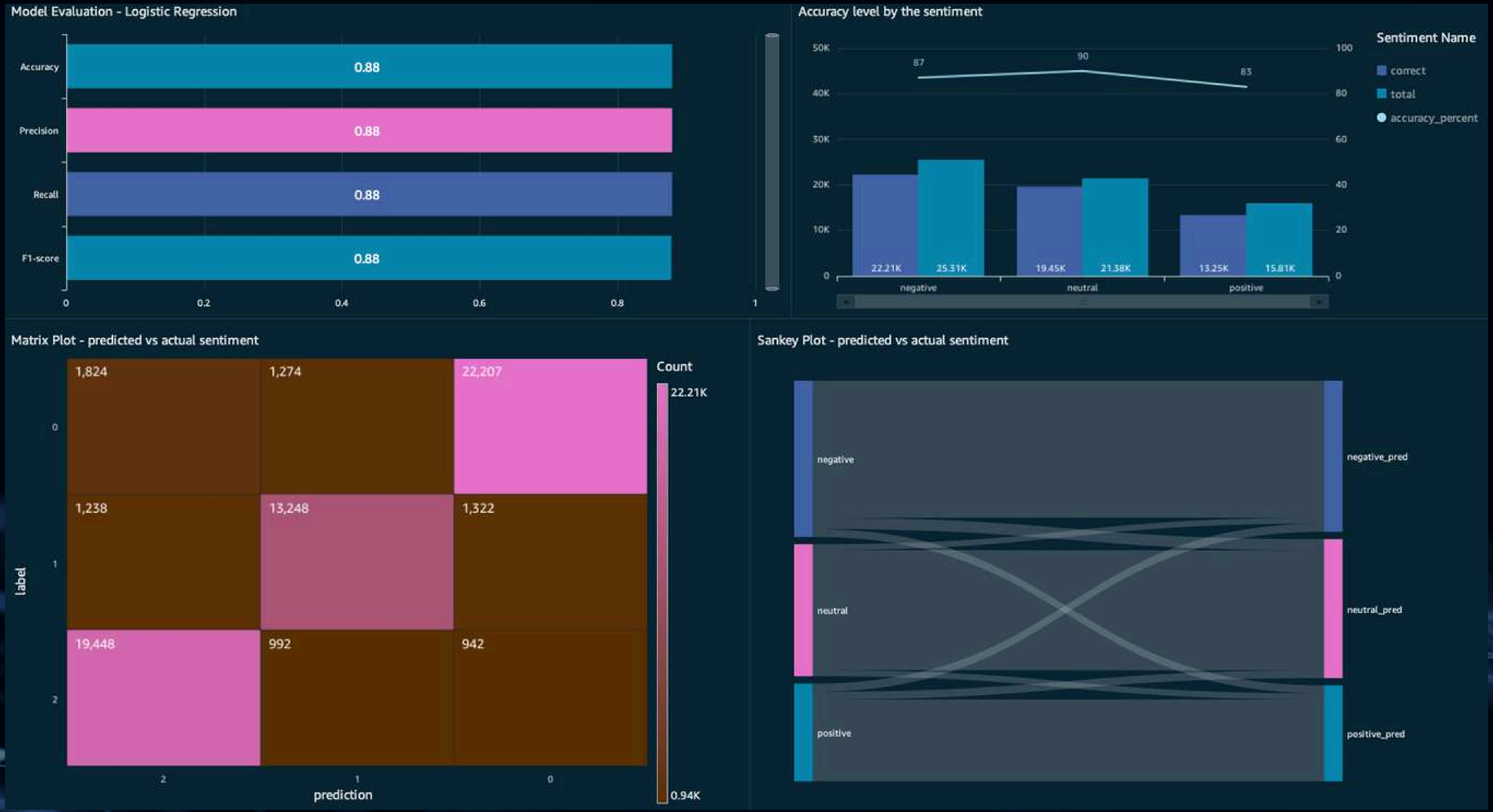
Comparison of unique tweets to retweets

Top 10 most popular retweets with the sentiment

| tweet | sentiment_name | count |
|---|---|---|
| RT @McJesse: I'm jealous of Elon Musk's kids. They're the only people who never have to hear from him. | negative | 7,490 |
| RT @MattWallace888: The Twitter employees Elon Musk fired are struggling to find new jobs because all of the offers they are getting involv… | negative | 6,339 |
| "RT @the_meghaning: Elon Musk called Trent Reznor a \"crybaby\" for deleting his Twitter account. Wil Wheaton then wrote this incisive summary…" | neutral | 5,553 |
| RT @genadoesthings: Elon Musk &amp; Jordan Peterson, neither of whom have been "cool" a single day in their lives, ragging on Trent Reznor, who… | neutral | 5,397 |
| RT @MattGertz: Elon Musk interacting with sycophantic right-wing influencers this weekend, a thread. | neutral | 4,910 |
| RT @OccupyDemocrats: BREAKING: In another blow to Elon Musk, White Stripes frontman Jack White ditches Twitter over Musk's "assh*ole move"… | neutral | 3,541 |
| RT @hotgirlhanfran: elon musk did to twitter what brendon urie did to panic! at the disco | negative | 3,500 |
| RT @disclosetv: JUST IN - Elon Musk has reinstated Rep. Marjorie Taylor Greene's (R-GA) personal Twitter account. | neutral | 3,374 |
| RT @thebradfordfile: It's amazing the media is more focused on Elon Musk than a $15 billion ponzi scheme for the global ruling class. LOL. | positive | 3,338 |
| RT @ACTBrigitte: If Apple takes Twitter out of the app store, Elon Musk should buy them next. | neutral | 3,061 |

# DATA ANALYSIS - MODEL METRICS

- Logistic Regression based on tweet text achieved the best results.
- Overall F1-score: 88%.
- Best performance on negative tweets – ~90% accuracy.
- Most confusion between negative and neutral classes.



Model Evaluation - Logistic Regression

Accuracy level by the sentiment

Matrix Plot - predicted vs actual sentiment

| label \ prediction | 2 | 1 | 0 |
|---|---|---|---|
| 0 | 1,824 | 1,274 | 22,207 |
| 1 | 1,238 | 13,248 | 1,322 |
| 2 | 19,448 | 992 | 942 |

Sankey Plot - predicted vs actual sentiment

# SUMMARY & REFLECTIONS

- The dataset covers only two days of tweets, which limits the ability to observe longer-term trends or draw more robust conclusions.
- Project work was constrained by the limitations of the Databricks Community version — more complex operations (like GBT with OneVsRest) were inefficient or failed to complete.
- Sentiment was labeled using VADER, which works well for informal text, but using a transformer-based model could potentially improve results (at the cost of time and compute).
- Despite these constraints, the pipeline worked well, and the model achieved solid performance.

# THANK YOU