Tweets Sentiment Analysis using NLP, Machine Learning & Kafka

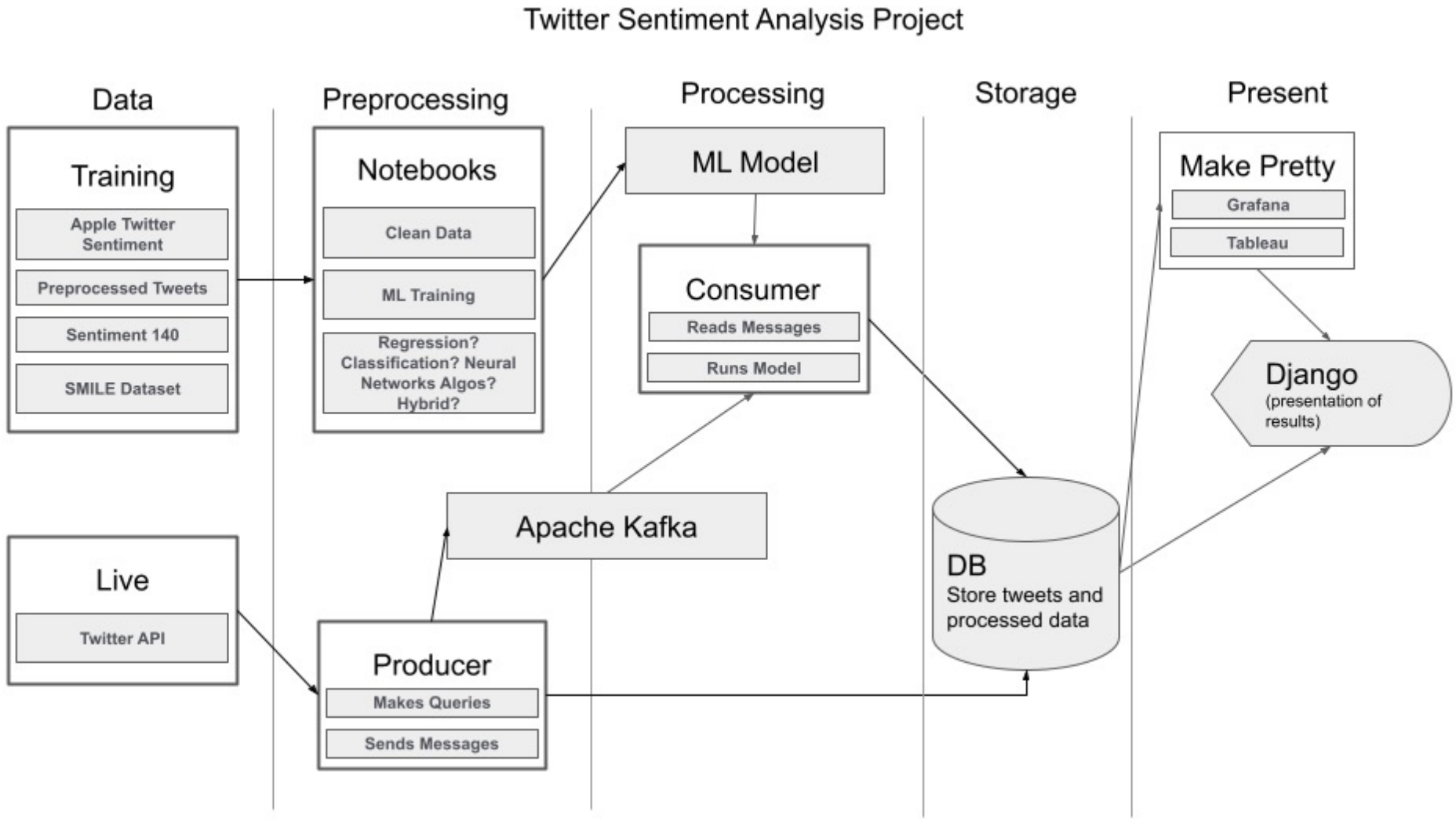
Topic: Mask Mandate

(Final Project Demo)

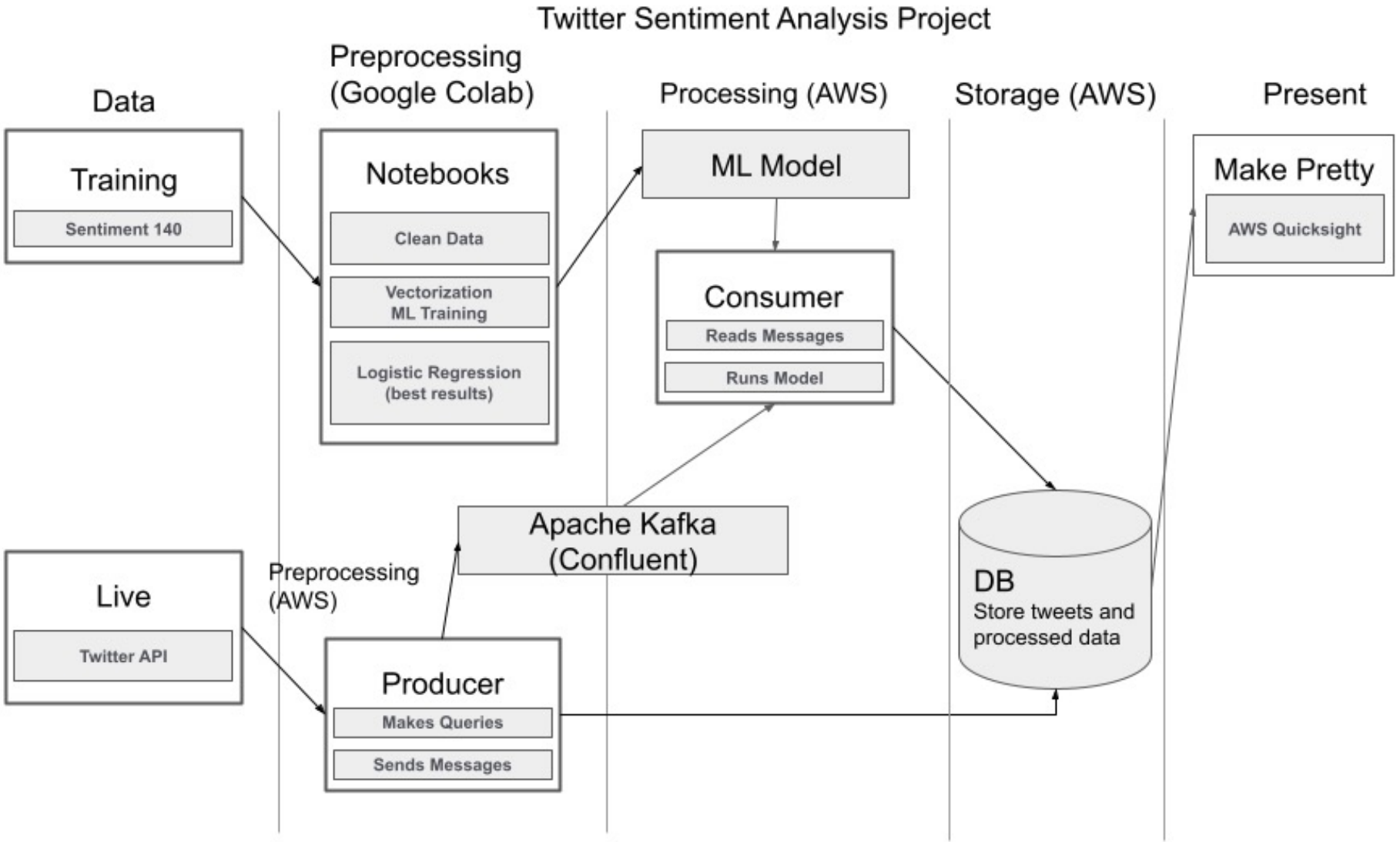
Creasen, Drake & Keerthi

Data 2.2, Zip Code Wilmington

Data Pipeline – The Initial Plan



Data Pipeline – The Final Version



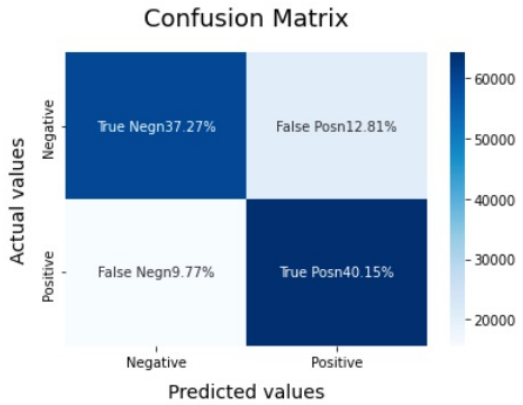
Developing the ML model

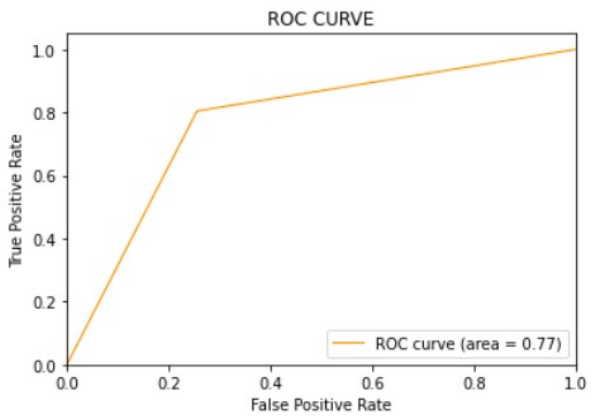
- Sentiment140 Dataset of 1,600,000 tweets (Twitter API)
- Salient Steps:
  - Data Pre-processing
    - Change to lower case for better generalization
    - Remove URLs and handles (@User), Stopwords, Punctuations, Numbers,
    - Tokenization of tweet text
    - Perform Stemming(reducing the words to their derived stems)
    - Perform Normalization - Lemmatization (reducing the derived words to their root form known as lemma)
    - Separate input feature and label
- Splitting into train and test subsets
- Feature Scaling: TF-IDF Vectorizer
- Model Evaluation: Accuracy Score, Confusion Matrix with Plot, ROC-AUC Curve
- Models: (1) Bernoulli Naïve Bayes (2) SVM (3) Logistic Regression (4) XGBoost
- Best Model based on Evaluation (on next slide) –> Logistic Regression

Results:

(1) BNB

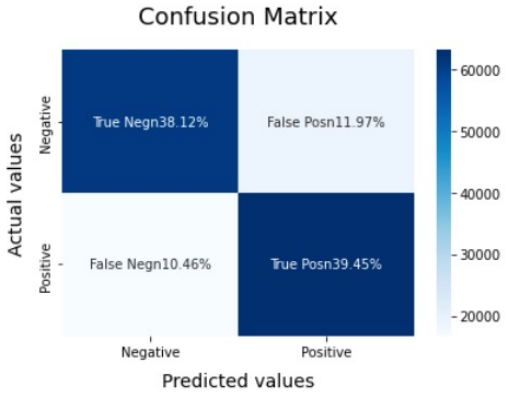
	precision	recall	f1-score	support
0	0.79	0.74	0.77	80139
1	0.76	0.80	0.78	79861
accuracy			0.77	160000
macro avg	0.78	0.77	0.77	160000
weighted avg	0.78	0.77	0.77	160000

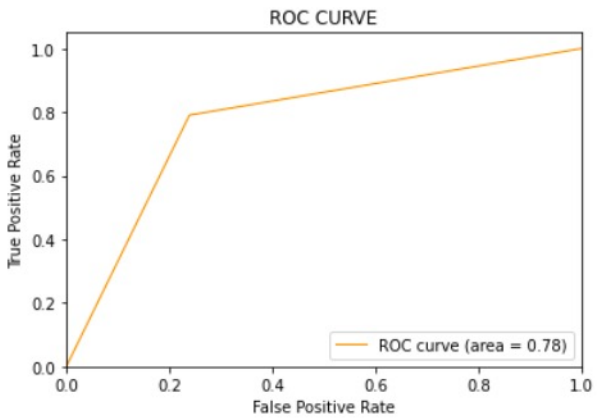




(2) SVM

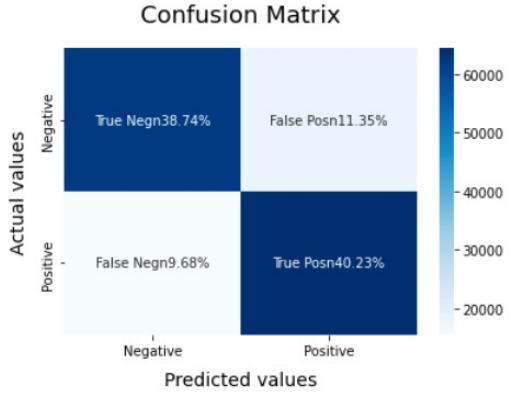
	precision	recall	f1-score	support
0	0.78	0.76	0.77	80139
1	0.77	0.79	0.78	79861
accuracy			0.78	160000
macro avg	0.78	0.78	0.78	160000
weighted avg	0.78	0.78	0.78	160000

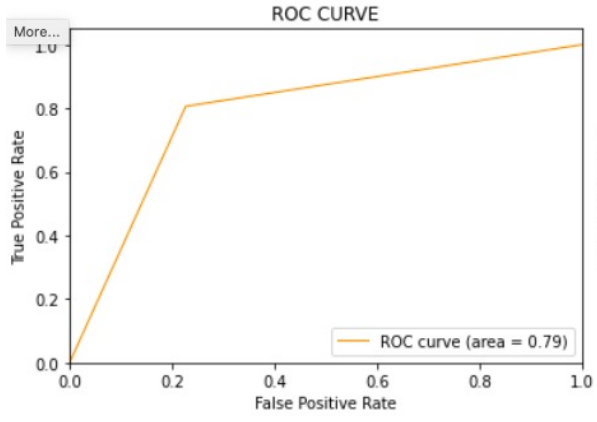




(3) LR

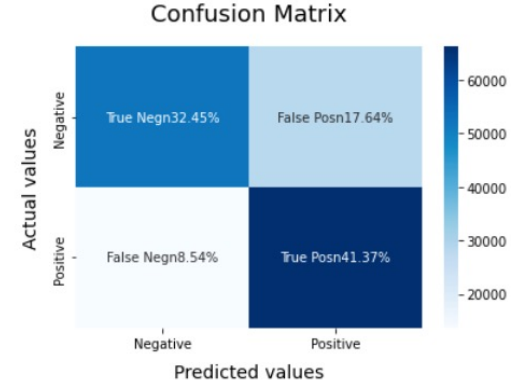
	precision	recall	f1-score	support
0	0.80	0.77	0.79	80139
1	0.78	0.81	0.79	79861
accuracy			0.79	160000
macro avg	0.79	0.79	0.79	160000
weighted avg	0.79	0.79	0.79	160000

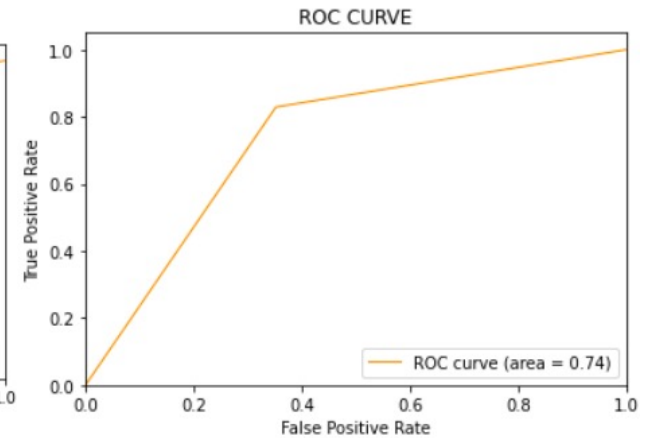




(4) XGBoost

	precision	recall	f1-score	support
0	0.79	0.65	0.71	80139
1	0.70	0.83	0.76	79861
accuracy			0.74	160000
macro avg	0.75	0.74	0.74	160000
weighted avg	0.75	0.74	0.74	160000



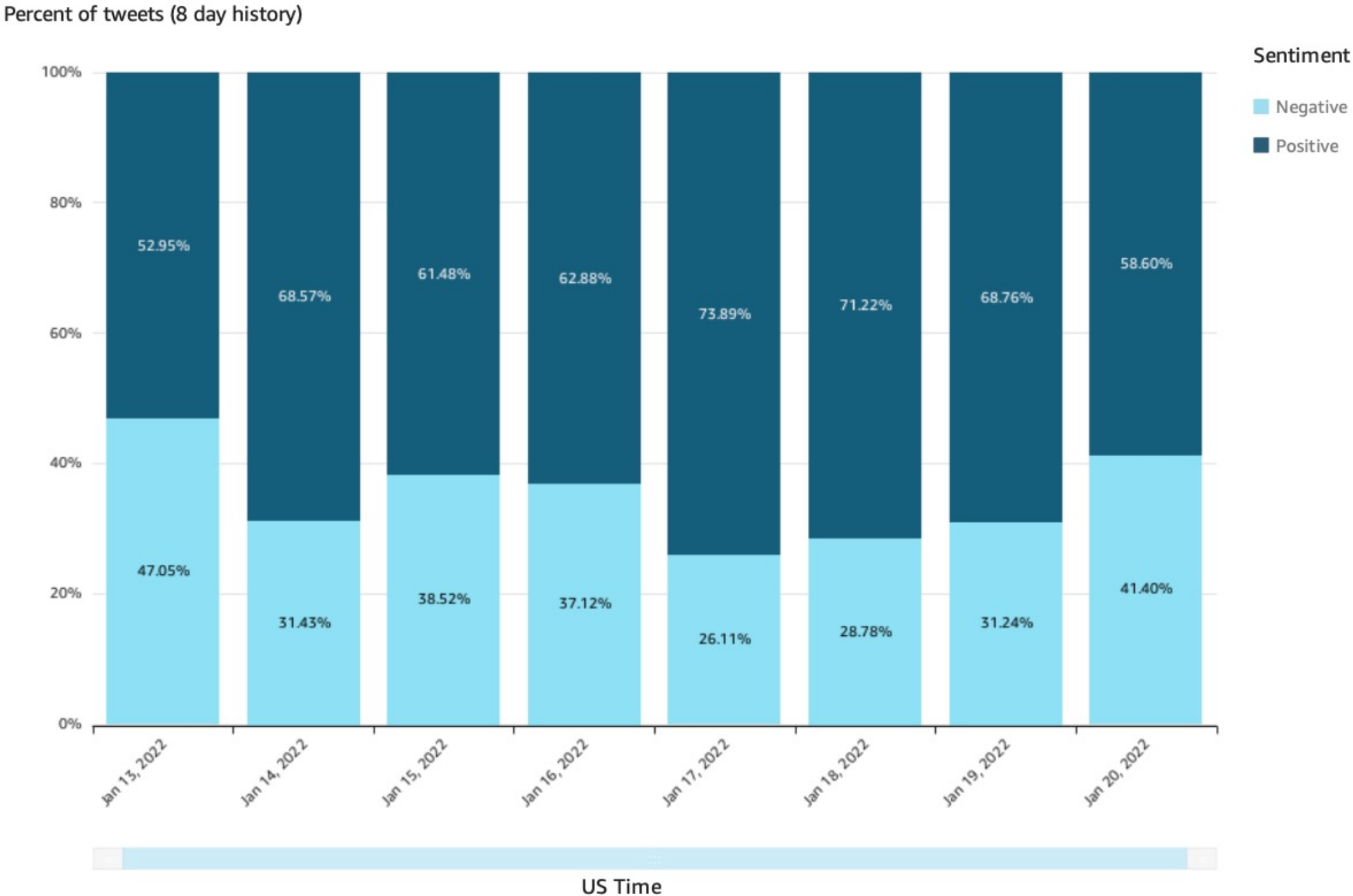


Twitter API, Kafka & Confluent, AWS, DB

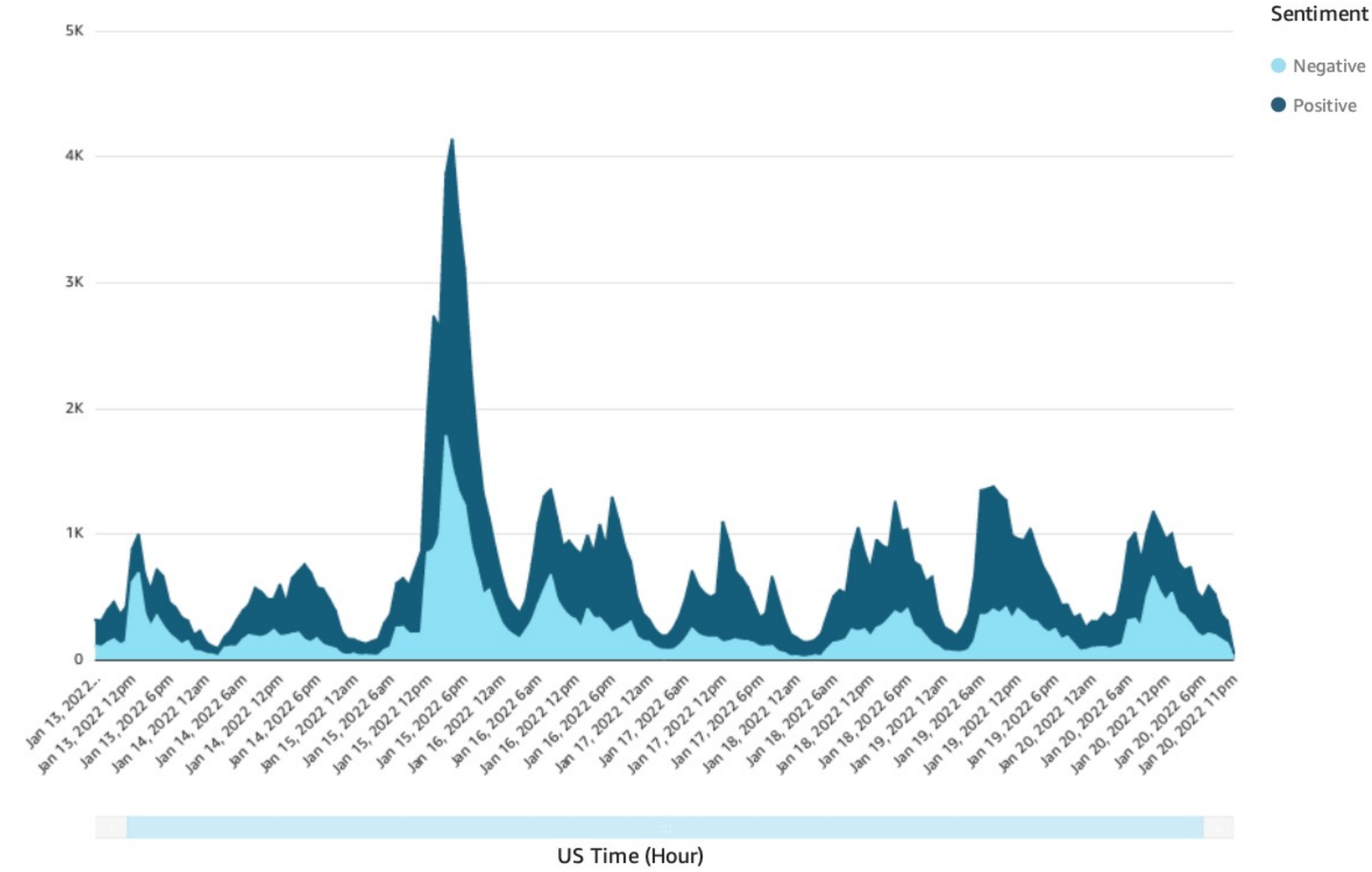
- Twitter API – tweets search based on ‘Mask Mandate’ keyword and English language
- Postgres database (DB) is used to store tweet data
- Kafka
  - Producer – extracts tweets using Twitter API and sends meta data to DB and tweet text with ID to topic
  - Consumer – reads tweet text with ID from topic and classifies the tweet (using LR model & vectorizer from ML model) for sentiment updating the DB
- Kafka Producer, Kafka Consumer and DB are hosted on AWS
- Kafka is hosted on Confluent

Data Visualization

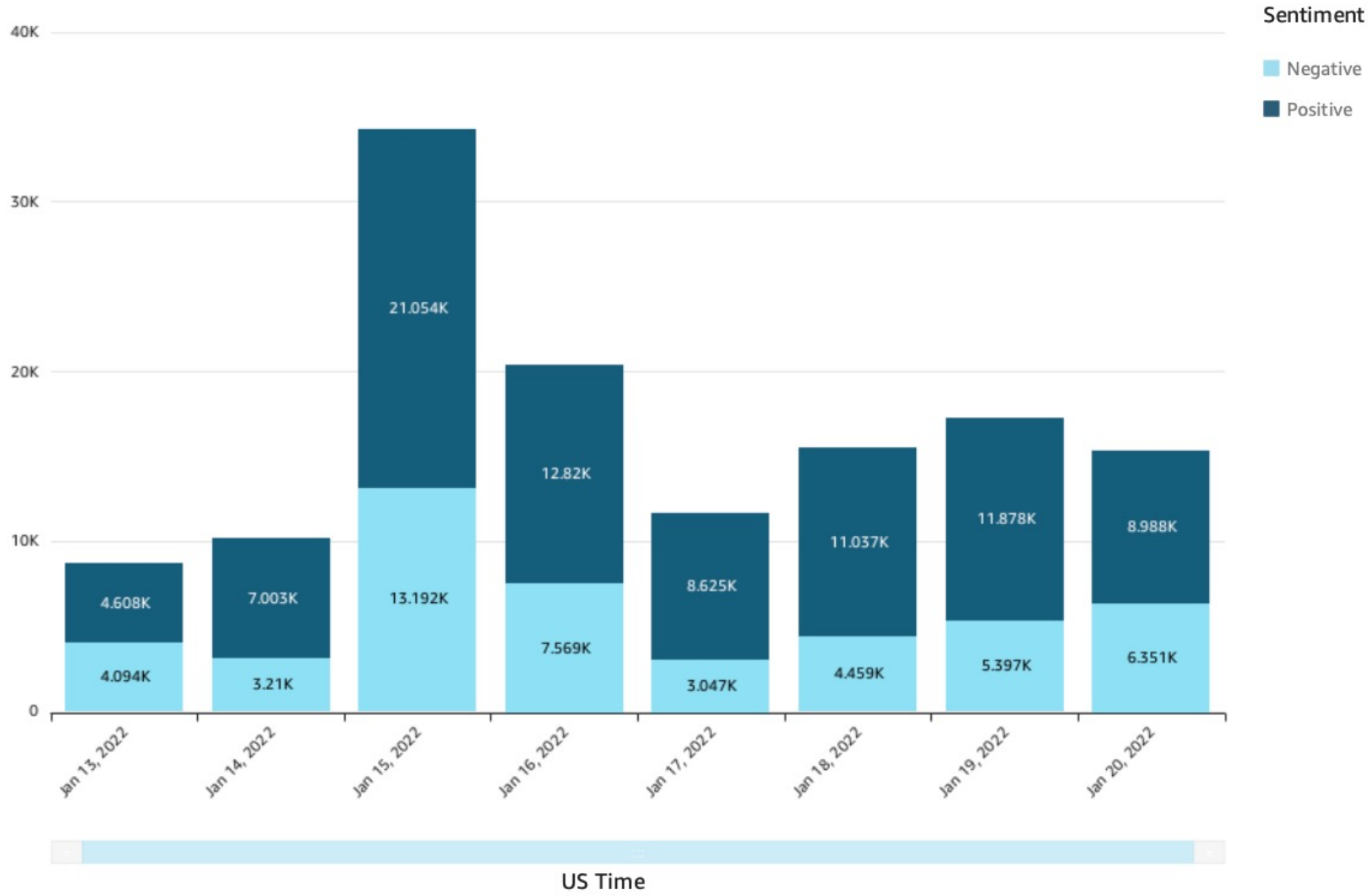
on AWS Quicksight

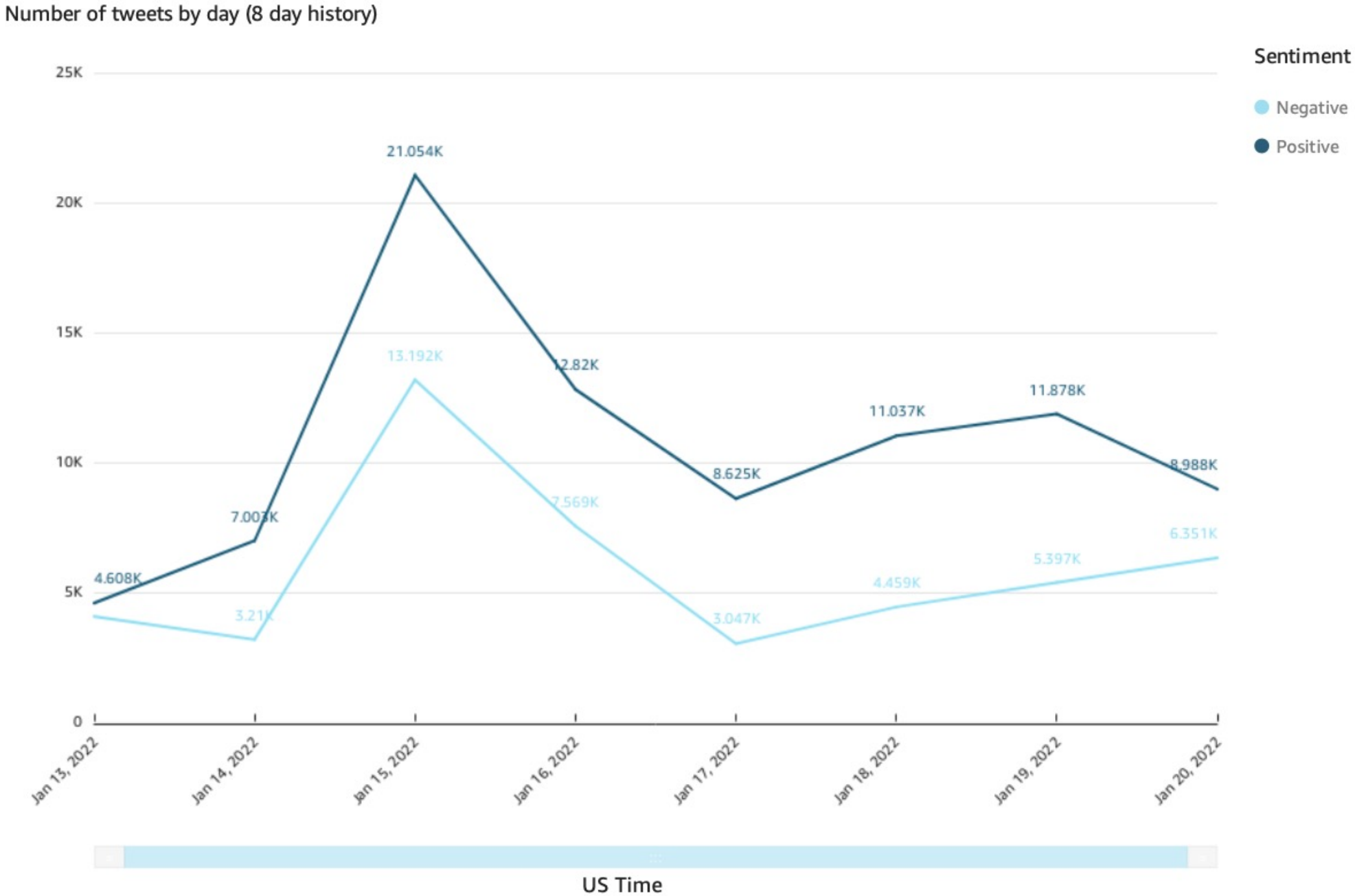

Percent of tweets (8 day history)
Sentiment
Negative
Positive
100%
80%
60%
40%
20%
0%
52.95%
47.05%
68.57%
31.43%
61.48%
38.52%
62.88%
37.12%
73.89%
26.11%
71.22%
28.78%
68.76%
31.24%
58.60%
41.40%
Jan 13, 2022
Jan 14, 2022
Jan 15, 2022
Jan 16, 2022
Jan 17, 2022
Jan 18, 2022
Jan 19, 2022
Jan 20, 2022
US Time

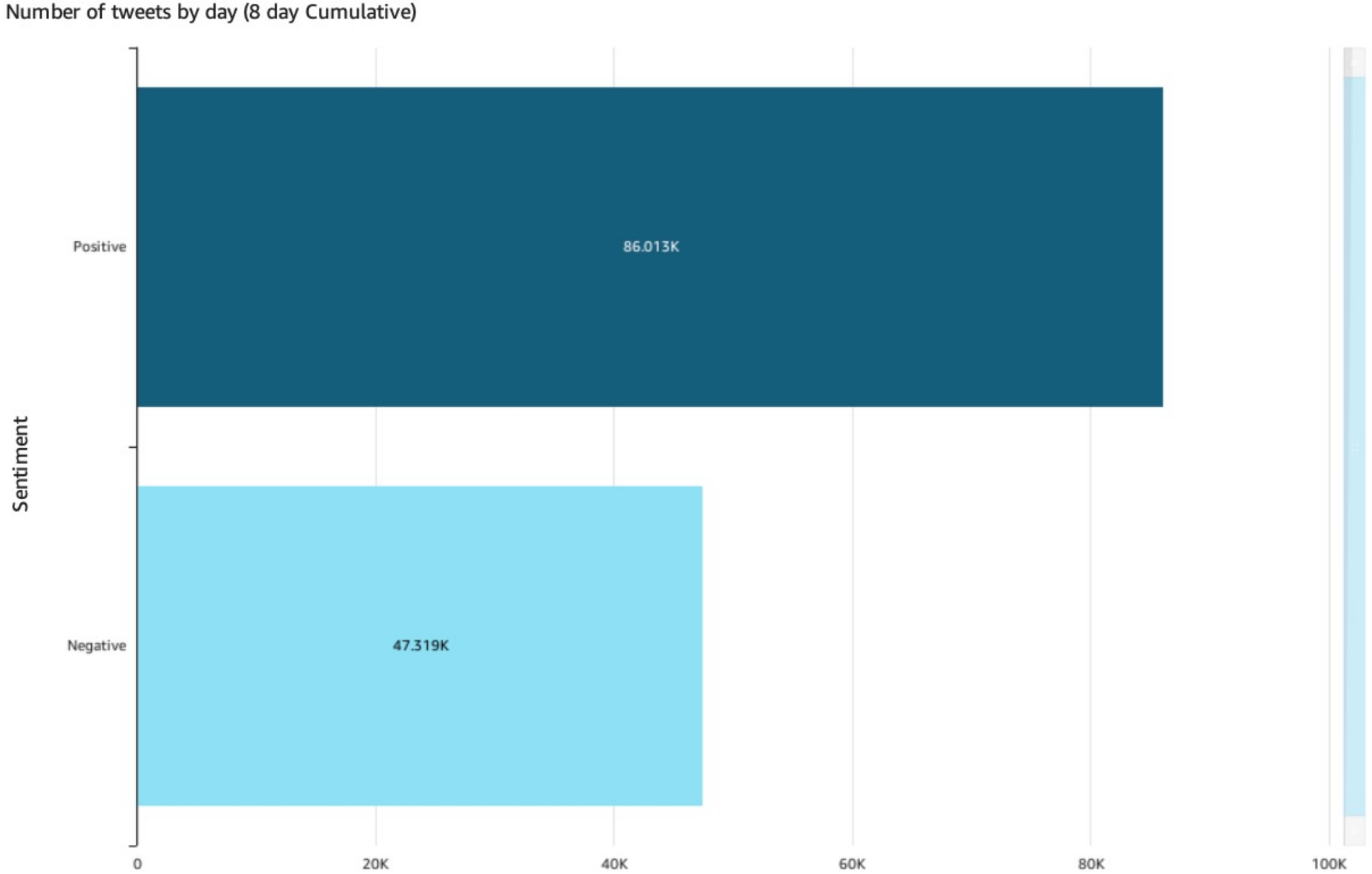
Number of tweets by hour (8 day history)



Number of tweets by day (8 day history)




Number of tweets by day (8 day history)
Sentiment
Negative
Positive
25K
20K
15K
10K
5K
0
21.054K
13.192K
12.82K
11.878K
11.037K
8.988K
8.625K
7.569K
7.003K
6.351K
5.397K
4.608K
4.459K
3.21K
3.047K
Jan 13, 2022
Jan 14, 2022
Jan 15, 2022
Jan 16, 2022
Jan 17, 2022
Jan 18, 2022
Jan 19, 2022
Jan 20, 2022
US Time


Number of tweets by day (8 day Cumulative)
Sentiment
Positive
86.013K
Negative
47.319K
0
20K
40K
60K
80K
100K

Questions?

Thank you!