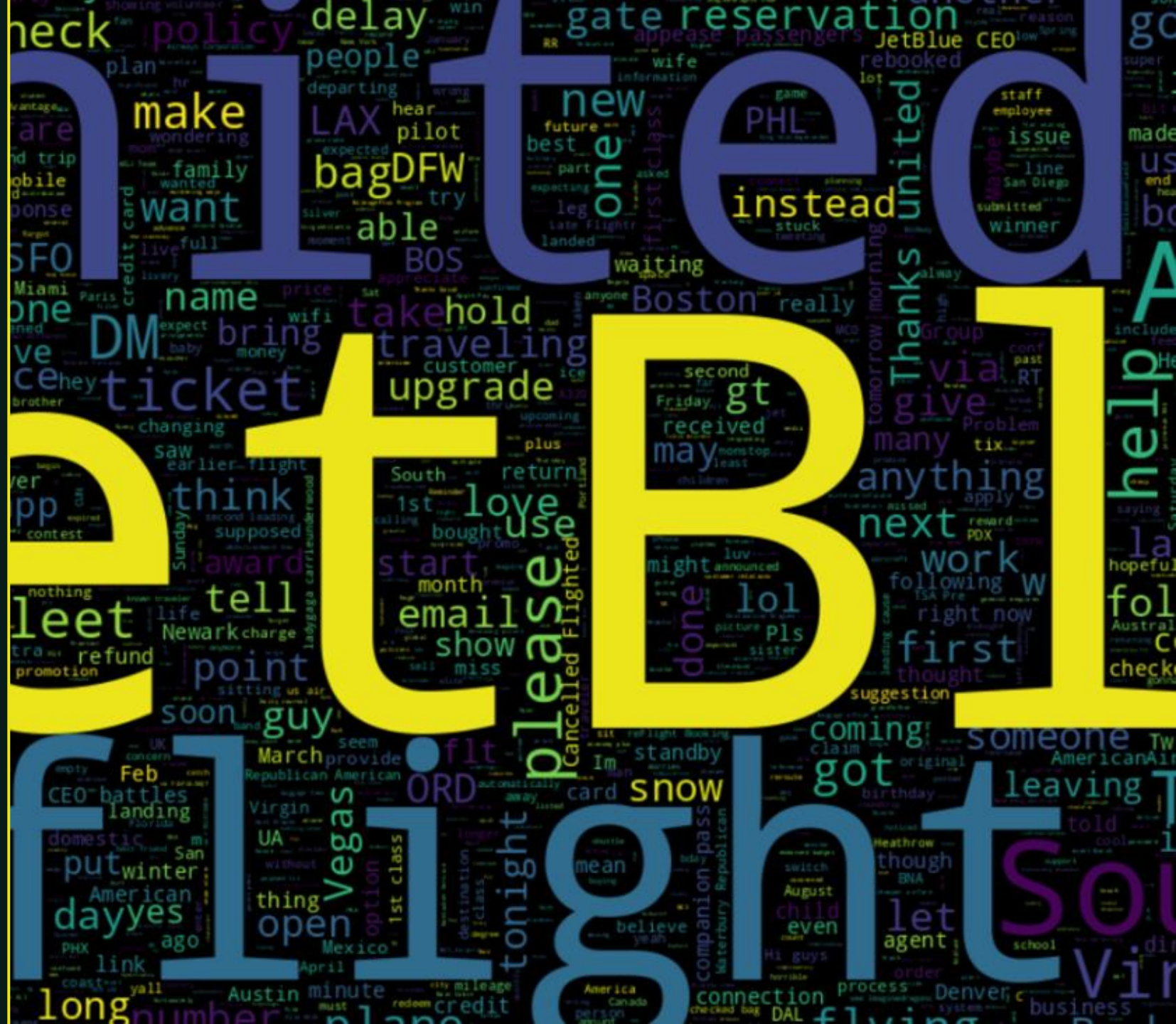
TWITTER FLIGHT SENTIMENT ANALYSIS

AMANDEEP SAROA



DATASET

WHERE? Sourced from **Kaggle.**

WHEN?: Data collected for the month of **February 2015.**

HOW? Approx. **14,000 tweets**, analyzing how travelers expressed their feelings about U.S. airlines on Twitter.

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & Canada)
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

Data Preprocessing

SELECTING RELVANT COLUMN: Focused on the airline sentiment and text columns.

Code Snippet: `data = df[["airline_sentiment", "text"]]`

Sample Output:

```
data.head()
```

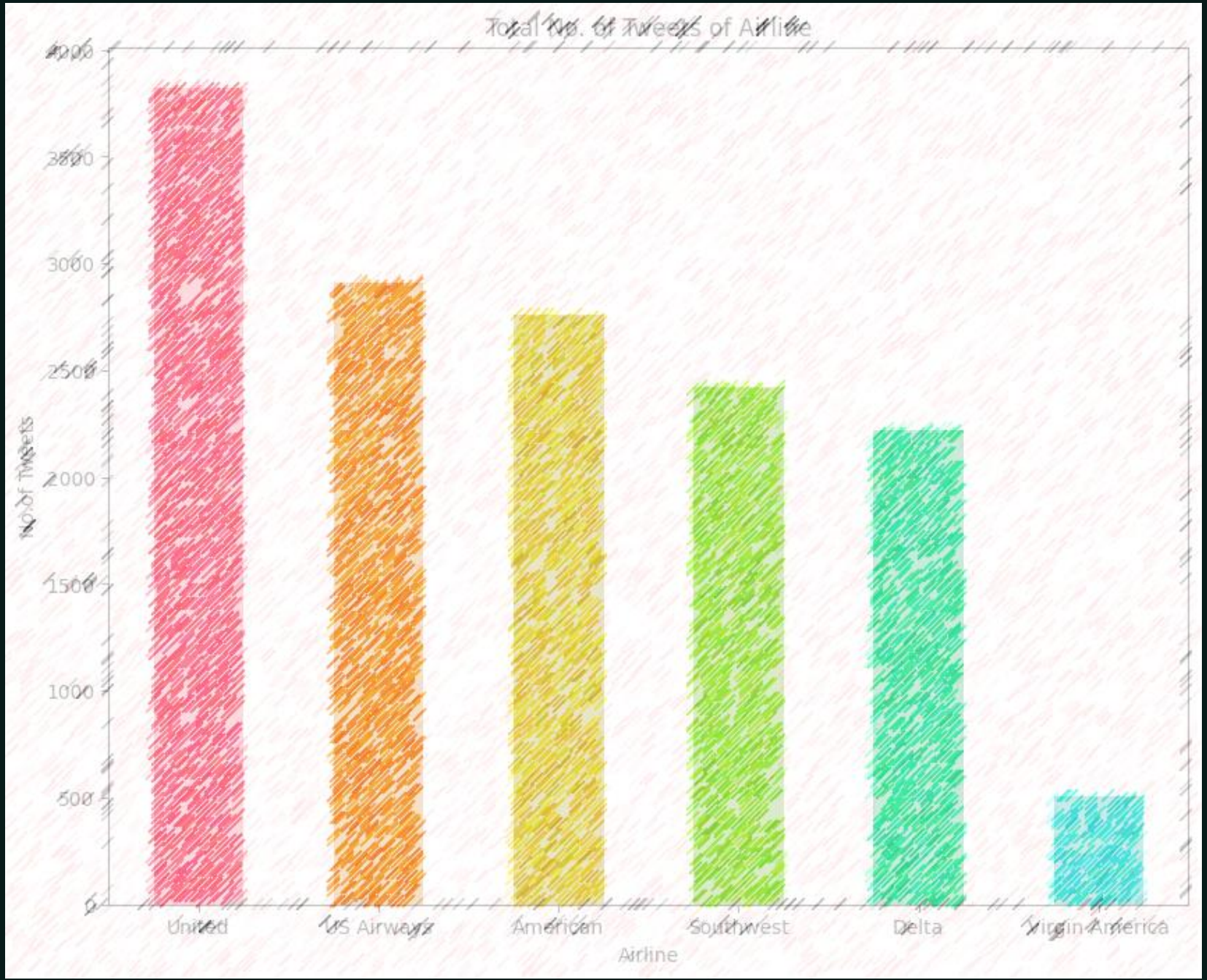
	airline_sentiment	text
0	neutral	@VirginAmerica What @dhepburn said.
1	positive	@VirginAmerica plus you've added commercials t...
2	neutral	@VirginAmerica I didn't today... Must mean I n...
3	negative	@VirginAmerica it's really aggressive to blast...
4	negative	@VirginAmerica and it's a really big bad thing...

Grouped data by airline sentiment:

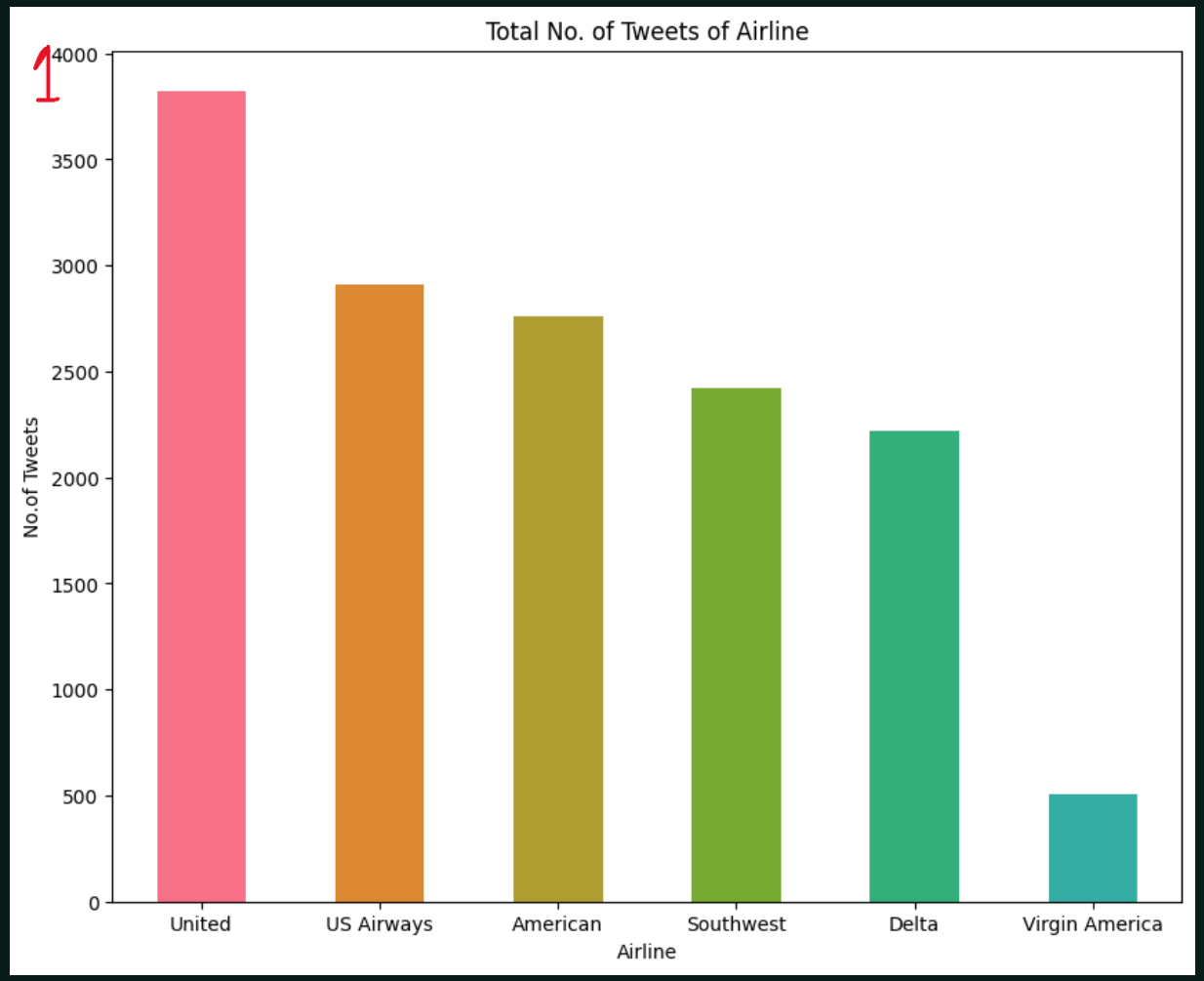
```
data.groupby('airline_sentiment').describe()
```

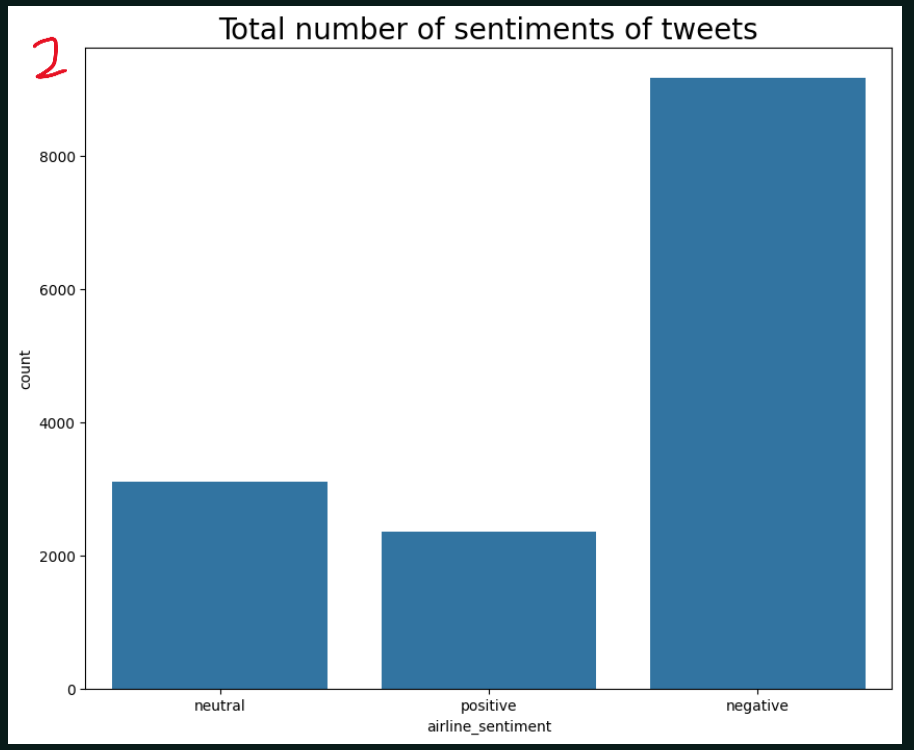
	text			
	count	unique	top	freq
airline_sentiment				
negative	9178	9087	@AmericanAir that's 16+ extra hours of travel ...	2
neutral	3099	3067	@SouthwestAir sent	5
positive	2363	2298	@JetBlue thanks!	5

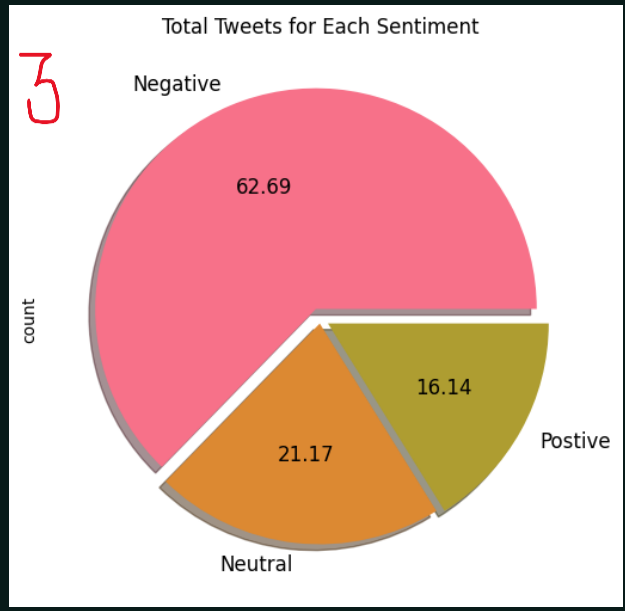
DATA VISUALIZATION

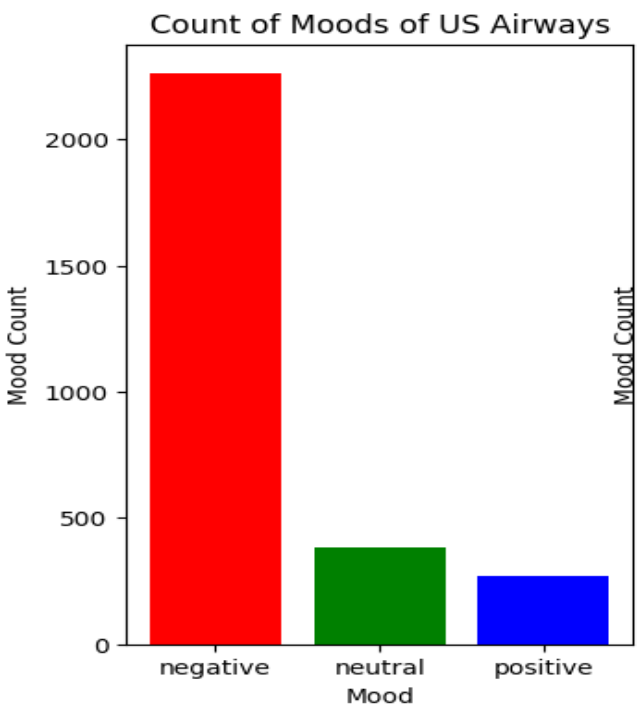


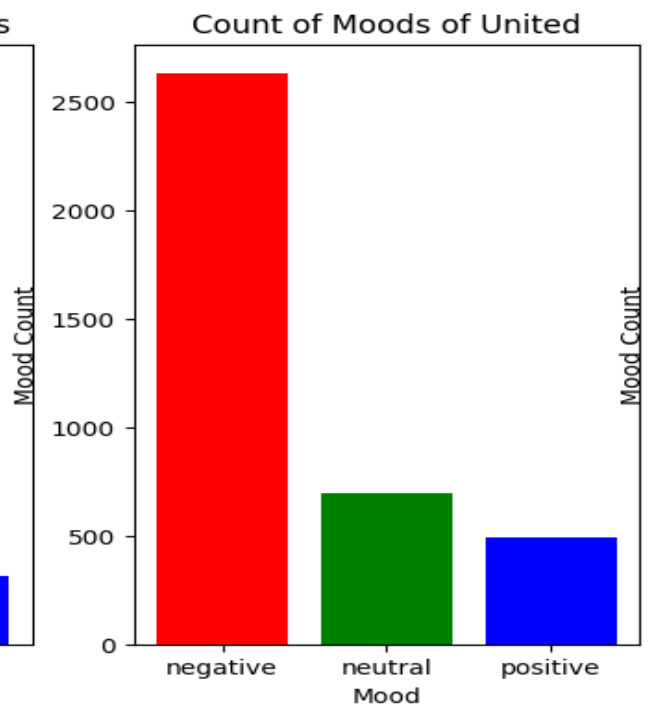
GRAPH's

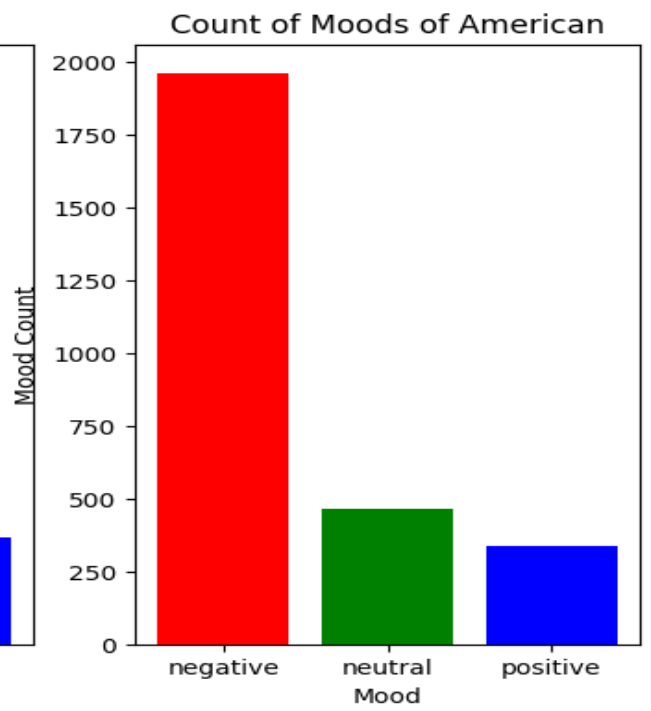

1
Total No. of Tweets of Airline
No.of Tweets
0
500
1000
1500
2000
2500
3000
3500
4000
United
US Airways
American
Southwest
Delta
Virgin America
Airline

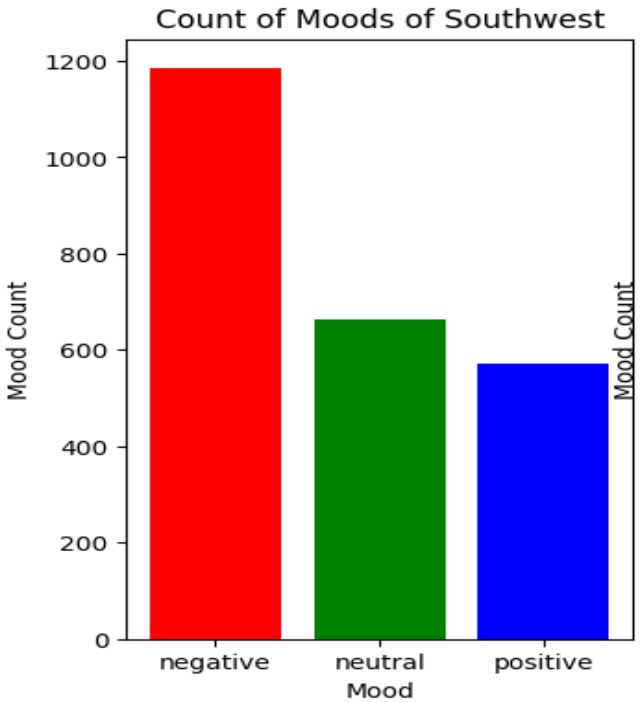

2
Total number of sentiments of tweets
count
0
2000
4000
6000
8000
neutral
positive
negative
airline_sentiment

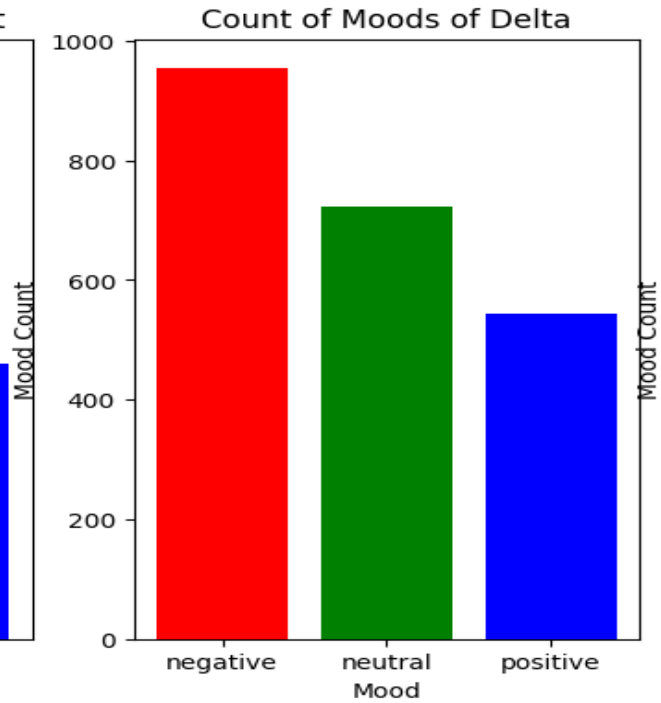

3
Total Tweets for Each Sentiment
Negative
62.69
16.14
Postive
21.17
Neutral
count

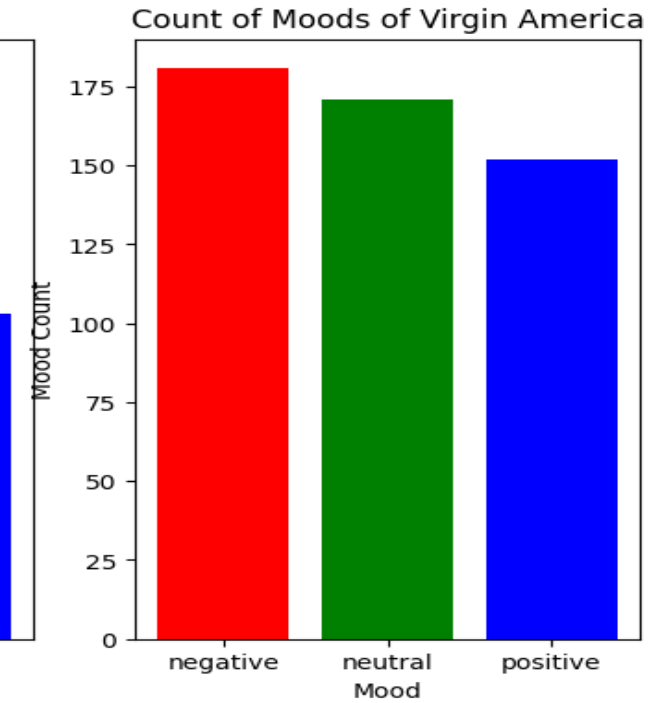

Count of Moods of US Airways
Mood Count
2000
1500
1000
500
0
negative
neutral
positive
Mood

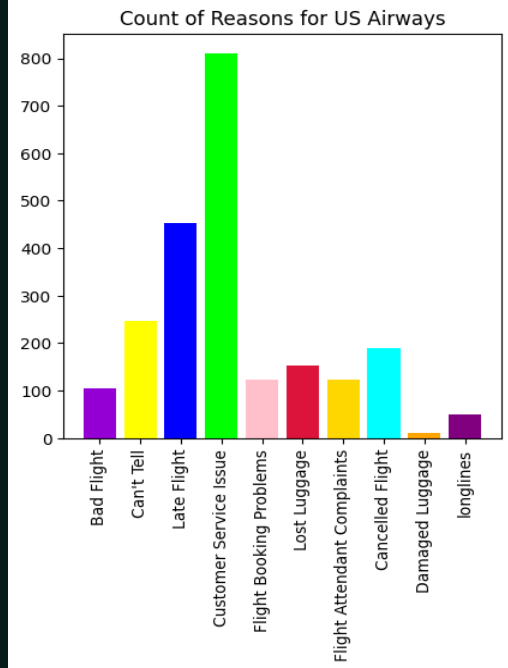

Count of Moods of United
Mood Count
2500
2000
1500
1000
500
0
negative
neutral
positive
Mood

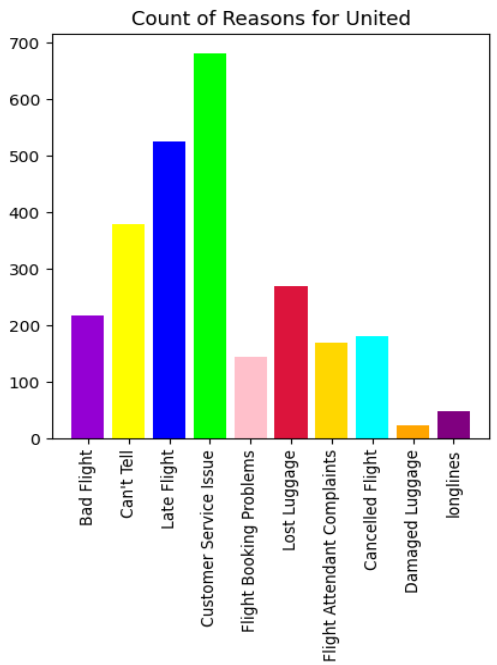

Count of Moods of American
Mood Count
2000
1750
1500
1250
1000
750
500
250
0
negative
neutral
positive
Mood

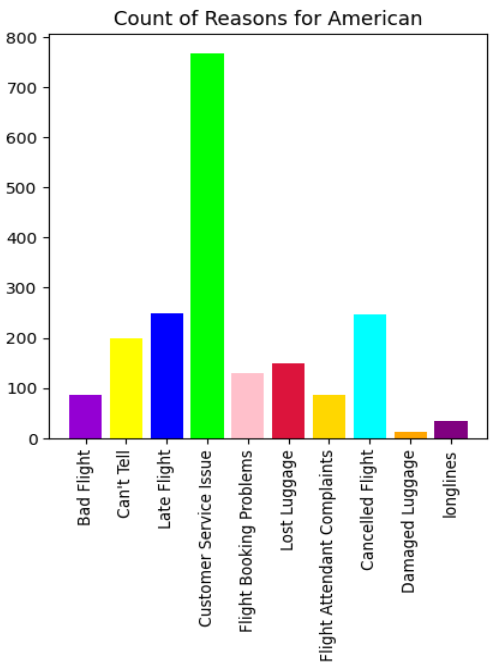

Count of Moods of Southwest
Mood Count
1200
1000
800
600
400
200
0
negative
neutral
positive
Mood

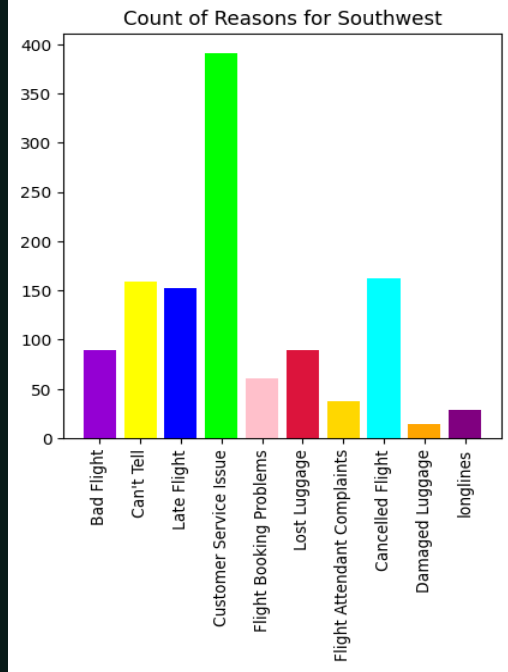

Count of Moods of Delta
Mood Count
1000
800
600
400
200
0
negative
neutral
positive
Mood

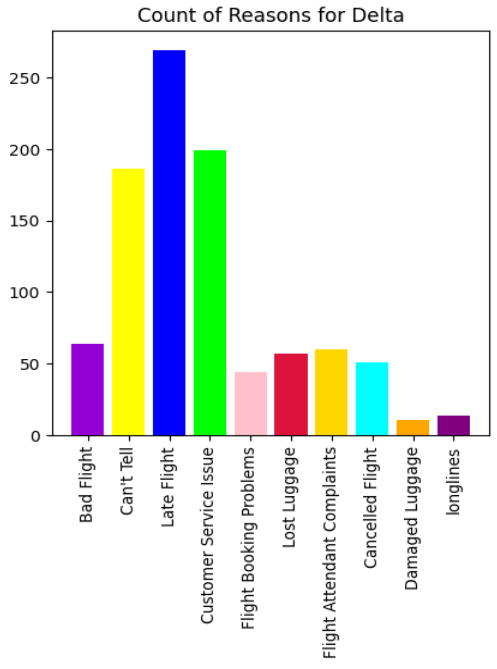

Count of Moods of Virgin America
Mood Count
175
150
125
100
75
50
25
0
negative
neutral
positive
Mood

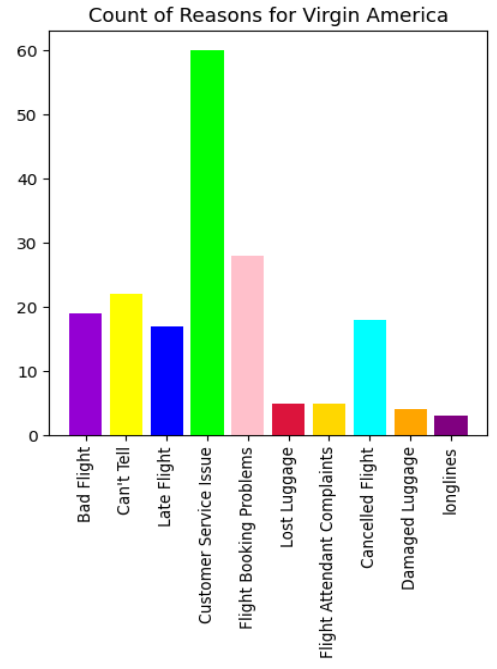

Count of Reasons for US Airways
0
100
200
300
400
500
600
700
800
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines


Count of Reasons for United
0
100
200
300
400
500
600
700
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines

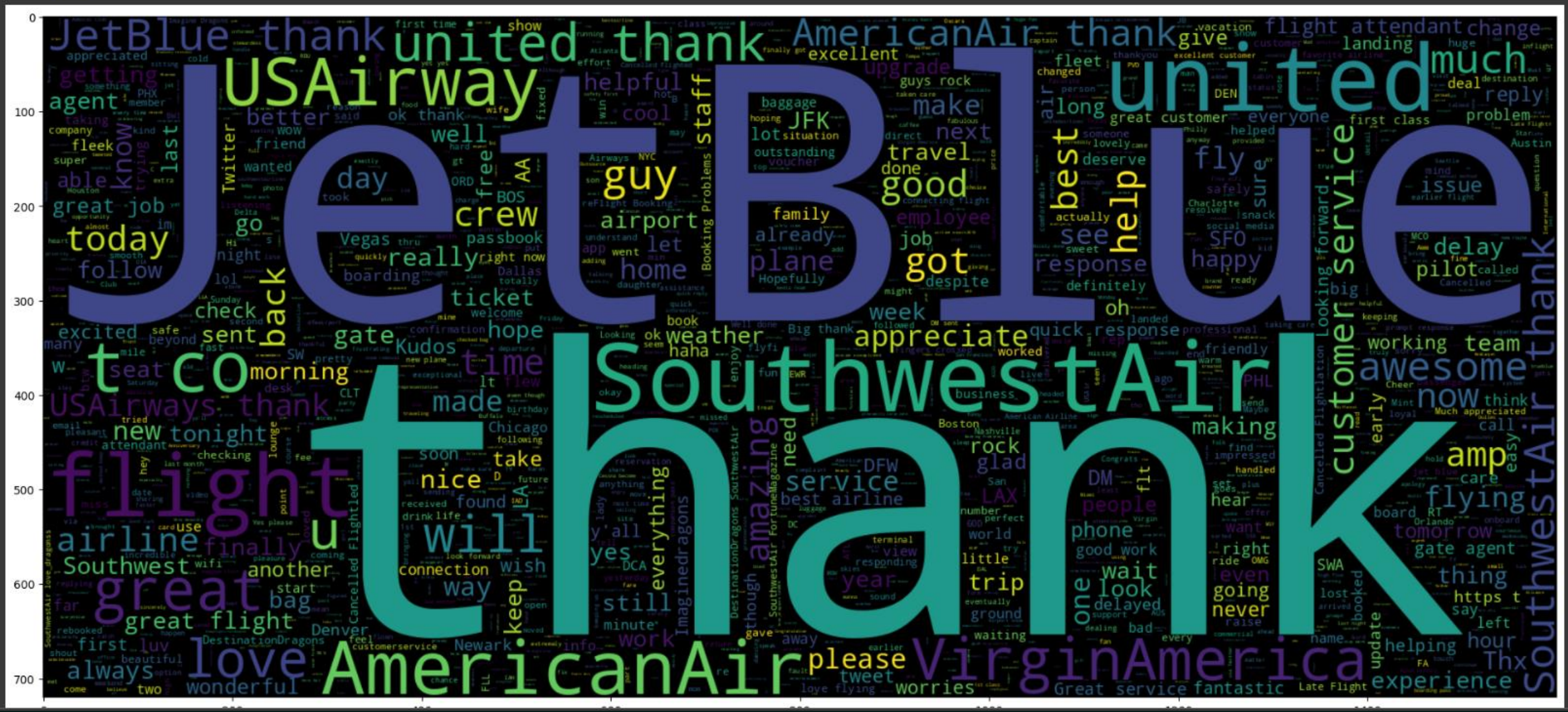

Count of Reasons for American
0
100
200
300
400
500
600
700
800
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines


Count of Reasons for Southwest
0
50
100
150
200
250
300
350
400
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines

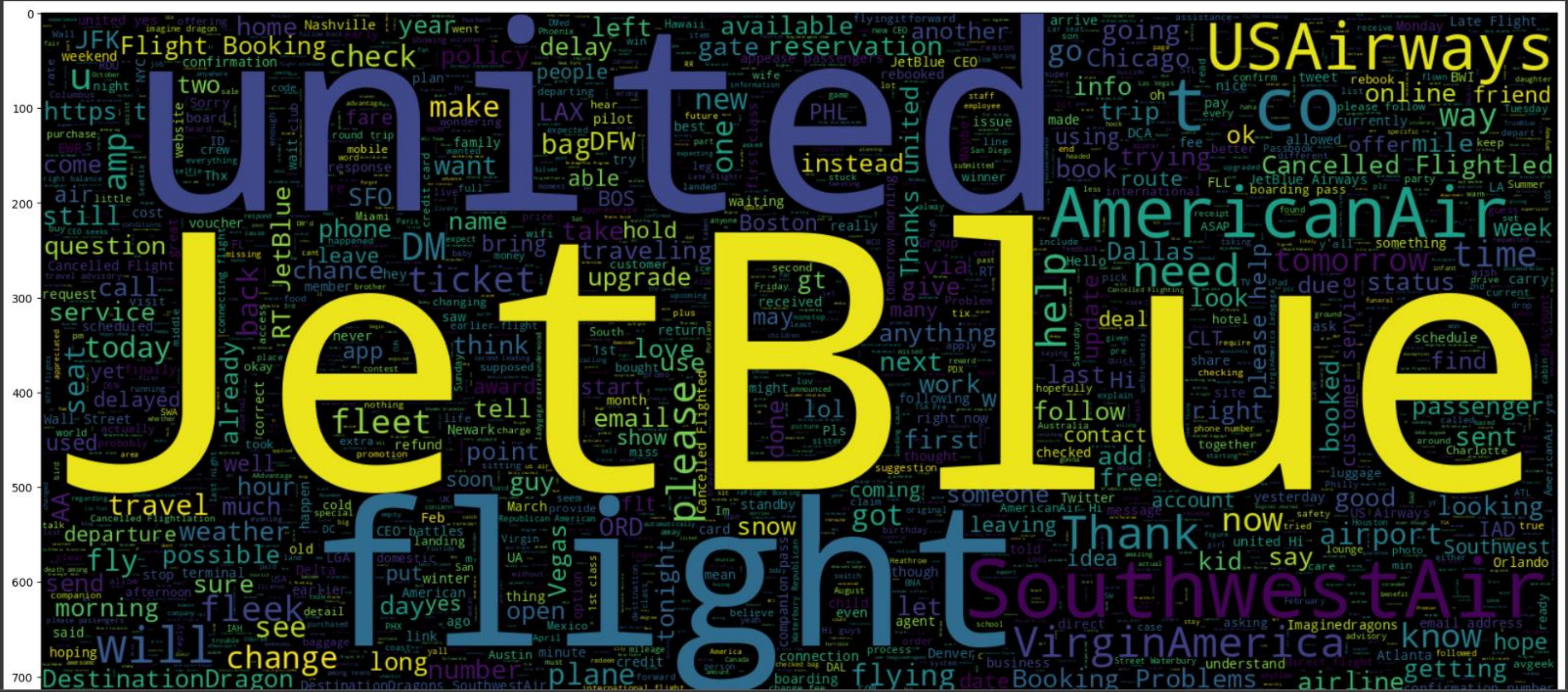

Count of Reasons for Delta
0
50
100
150
200
250
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines


Count of Reasons for Virgin America
0
10
20
30
40
50
60
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines

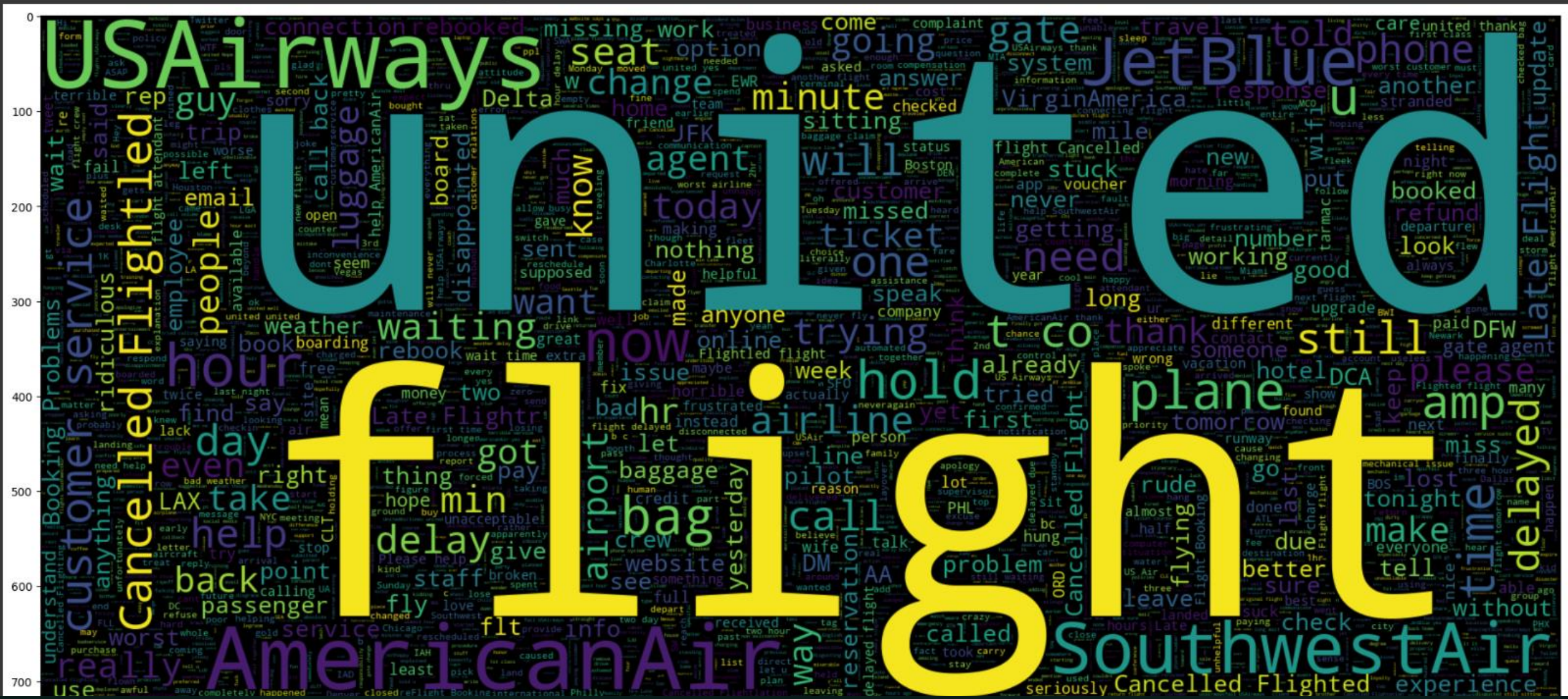
```
world_cloud_postive=WordCloud(min_font_size=3,max_words=3200,width=1600,height=720).generate("".join(positive))
plt.imshow(world_cloud_postive,interpolation='bilinear')
ax.grid(False)
```



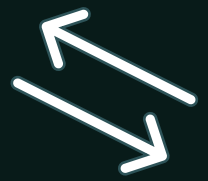
```
plt.figure(figsize=(24,12))
world_cloud_neutral=WordCloud(min_font_size=3,max_words=3200,width=1600,height=720).generate(" ".join(neutral))
plt.imshow(world_cloud_neutral,interpolation='bilinear')
ax.grid(False)
```



```
plt.figure(figsize = (24,12))
worldcould_neg = WordCloud(min_font_size = 3,  max_words = 3200 , width = 1600 , height = 720).generate(" ".join(negative))
plt.imshow(worldcould_neg,interpolation = 'bilinear')
ax.grid(False)
```



TEXT PROCESSING & CLEANING



Sentiment Conversion

- **Airline sentiment** to **numerical** values: 0,1,2

Remove Unnecessary text

- Stopwords Removal
- URL
- Punctuation Removal
- HTML Tag Removal
- Username Removal
- Emoji Removal

Apply RegEx

- Decontraction:
- Alphanumeric Separation
- Character Normalization
- Lowercasing

BEFORE:

```
df['text']
```

	text
0	@VirginAmerica What @dhepburn said.
1	@VirginAmerica plus you've added commercials t...
2	@VirginAmerica I didn't today... Must mean I n...
3	@VirginAmerica it's really aggressive to blast...
4	@VirginAmerica and it's a really big bad thing...
...	...
14635	@AmericanAir thank you we got on a different f...
14636	@AmericanAir leaving over 20 minutes Late Flig...
14637	@AmericanAir Please bring American Airlines to...
14638	@AmericanAir you have my money, you change my ...
14639	@AmericanAir we have 8 ppl so we need 2 know h...

14640 rows × 1 columns

dtype: object

AFTER:

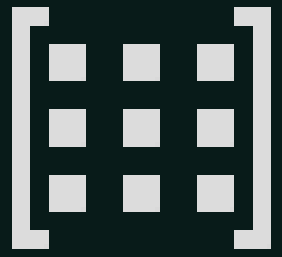
```
# result
df['final_text']
```

	final_text
0	said
1	plus added commercials experience tacky
2	today must mean need take another trip
3	bad flight really aggressive blast obnoxious e...
4	ca tell really big bad thing
...	...
14635	thank got different flight chicago
14636	customer service issue leaving minutes late fl...
14637	please bring american airlines blackberry
14638	customer service issue money change flight ans...
14639	ppl need know many seats next flight plz put u...

14640 rows × 1 columns

dtype: object

TRAINING MACHINE LEARNING MODEL



TF-IDF Vectorizer

- Convert the processed text into numerical features.
- capture the importance of words in each tweet.
- Giving more weight to unique words within each tweet.

SMOTE

- **Synthetic Minority Oversampling Technique.**
- Balance the dataset by generating synthetic samples.
- How? ensuring that each sentiment class (positive, neutral, negative) has an equal representation in the training data, which improves model performance.

Train-Test Split

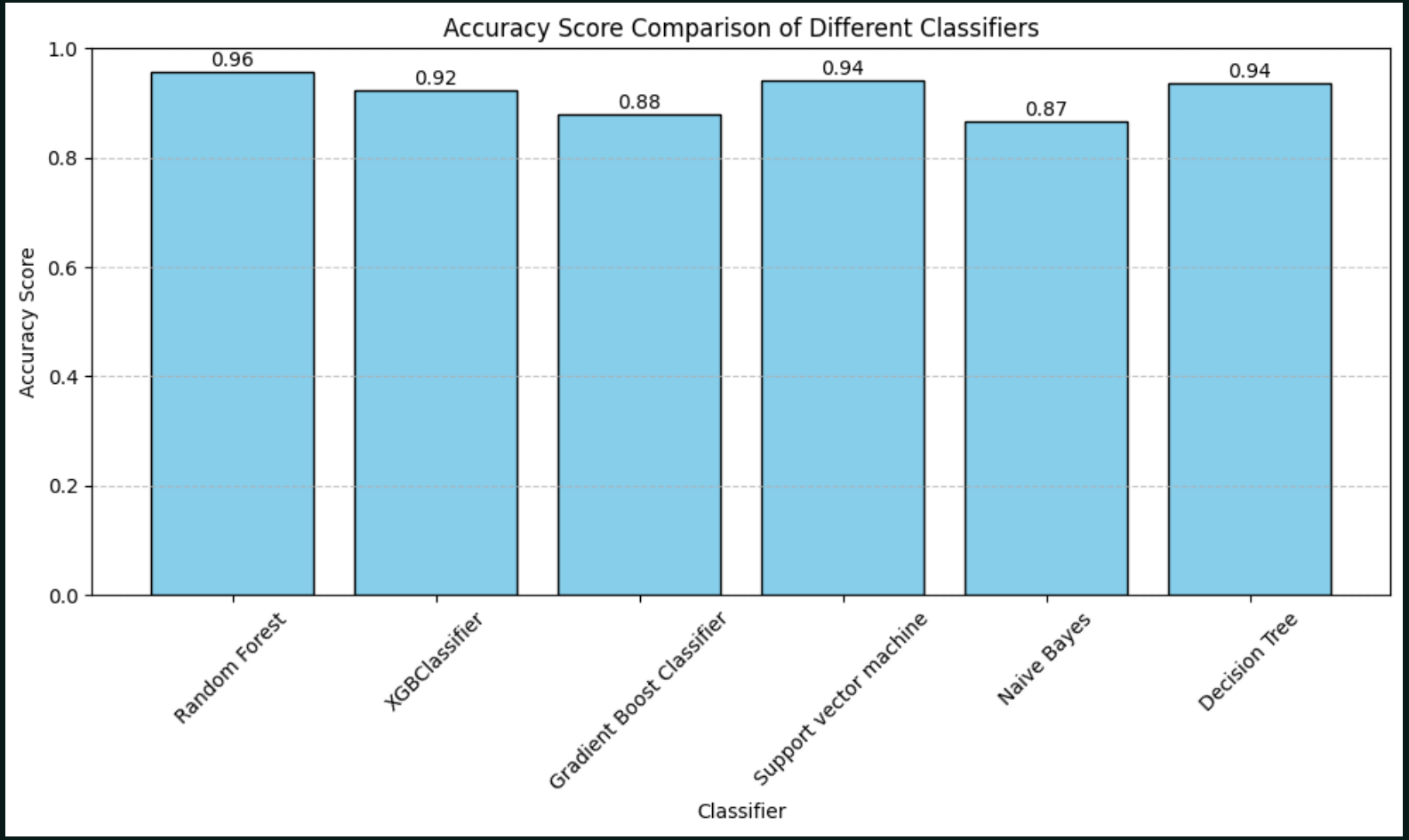
- **Training and Testing sets (75% training, 25% testing).**
- Train the model and evaluate its performance on unseen data, ensuring generalizability

ALGORITHMS APPLIED

1. XGBOOST CLASSIFIER
2. RANDOM FOREST
3. GRADIENT BOOSTING CLASSIFIER
4. SUPPORT VECTOR MACHINE (SVM)
5. NAÏVE BAYES
6. DECISION TREE

Each algorithm uses a different approach to learn from the data.
Helps in finding the most effective model for our task.

ACCURACY SCORE ACHIEVED



- The **Random Forest Algo.** model achieved the highest accuracy score of **0.96.**
- **Suggesting:** Learned patterns effectively from the training data and applied them accurately on new, unseen data.

Thank you

AMANDEEP SAROA

amandeep.saroa@ontariotechu.net

https://amandeepsaroaportfolio.netlify.app/