Twitter Sentiment

Thomas Moore

Data Overview

- The data was scraped from February of 2015 and contained information regarding the positive, negative, and neutral tweets about American, Delta, Southwest, United, US Airways, and the Virgin American airlines.
- The data was collected from Twitter for an 8 day span (2/16/15 - 2/24/15)
- There is a total of 15 parameters and 14640 tweets with some missing values as seen on the table to the right

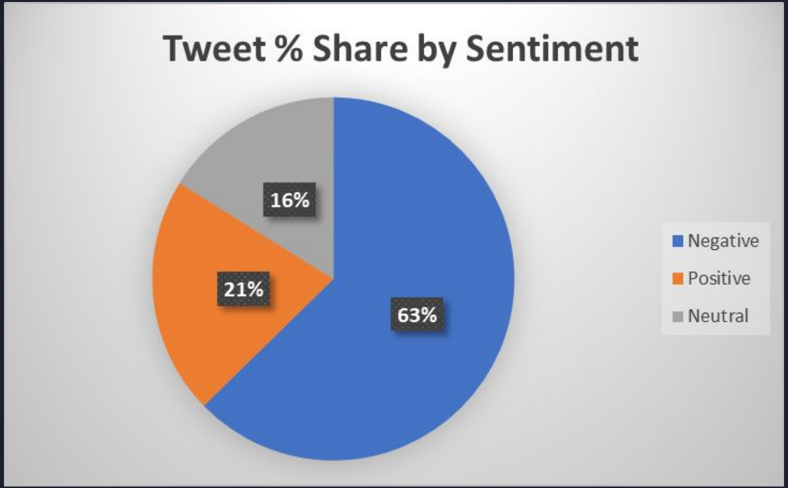
```
num_null = data.isnull().sum()
print(num_null)
```

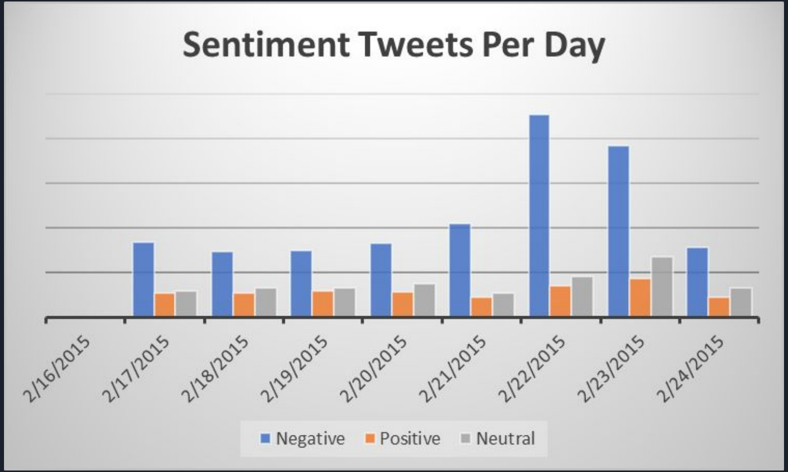
```
tweet_id                            0
airline_sentiment                   0
airline_sentiment_confidence        0
negativereason                   5462
negativereason_confidence        4118
airline                             0
airline_sentiment_gold          14600
name                                0
negativereason_gold             14608
retweet_count                       0
text                                0
tweet_coord                     13621
tweet_created                       0
tweet_location                   4733
user_timezone                    4820
dtype: int64
```

Data Overview

- The data showed a 63%, 21%, and 16% breakdown for negative, positive and neutral labeled tweets respectively

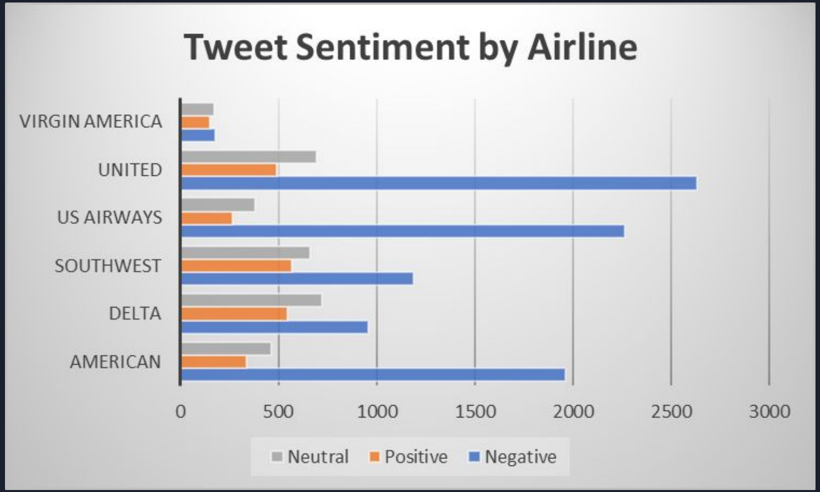
- There was a an average of 1,830 tweets per day with a fairly consistent tweet percentage share until there was a big spike in negative tweets during 2/22/15 and 2/23/15 (Sunday and Monday)





Sentiment Analysis per Airline

- The bar graph to the shows the total tweets by sentiment by each individual airline
- United and US Airlines have the most negative tweets and Southwest and Delta have the most positive.
- Virgin America has the least amount of total tweets



Sentiment Confidence

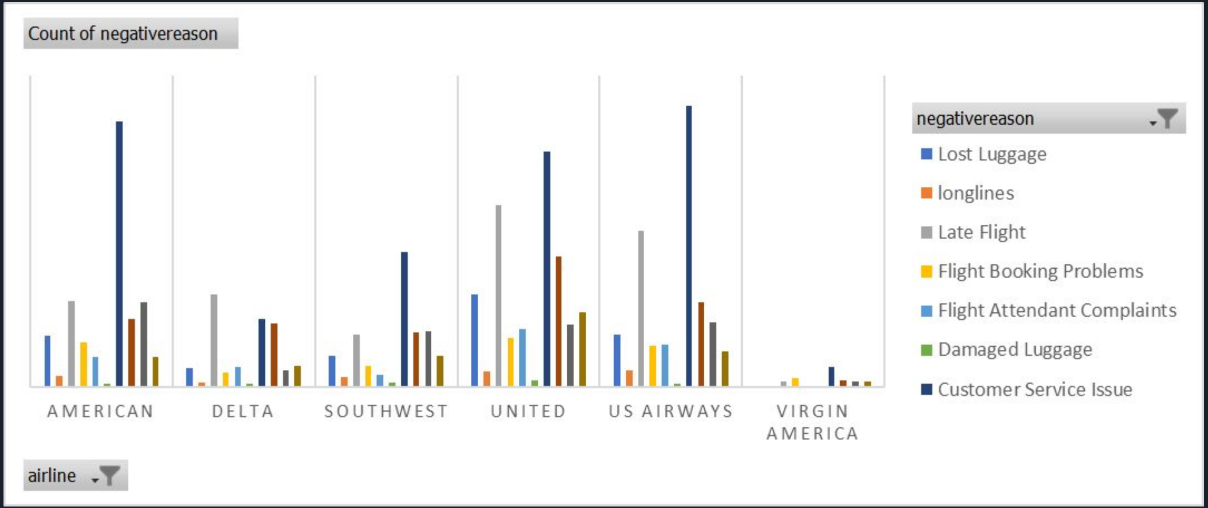
- The pivot table below shows the average sentiment analysis confidence related to each tweet breakdown from the previous chart
- The pivot table shows that US Airways has the highest sentiment negative confidence, United has the second highest negative confidence, and Delta has the worst average sentiment confidence.

- Across all airlines there was significantly better negative sentiment confidence than positive

Average of airline_sentiment_confidence	Column Labels			
Airlines	negative	neutral	positive	AVG
American	94.50%	82.59%	88.23%	91.74%
Delta	90.22%	82.93%	86.71%	86.99%
Southwest	92.05%	82.61%	88.61%	88.65%
United	93.34%	80.98%	85.60%	90.09%
US Airways	94.57%	82.19%	85.97%	92.16%
Virgin America	90.17%	83.84%	88.80%	87.61%

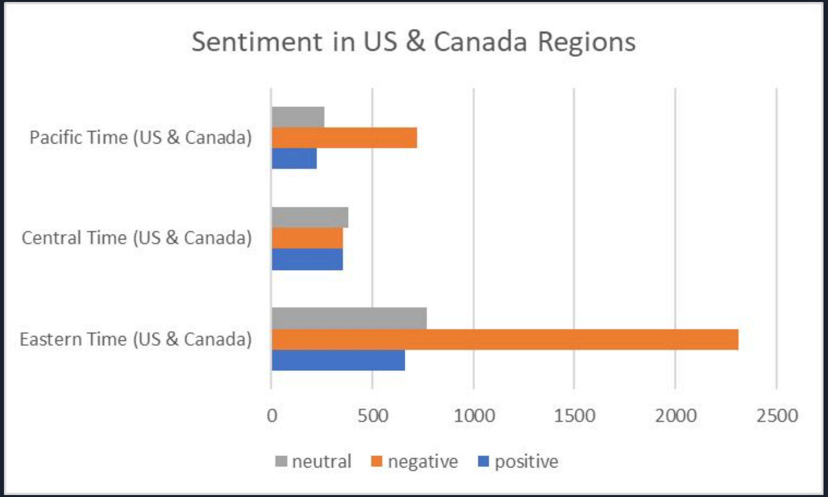
Reason for negative sentiments

- Each tweet with a negative sentiment label was given a reason why
- The most common issue was customer service across all airlines during this month
- The 5 most frequent issues in order was “customer service”, “late flight”, “cancelled flight”, “lost luggage”, and finally “bad flight”.



Tweet Sentiment by Region

- In the bar graph to the right, there are over double as many negatively labeled tweets in the states in the Eastern time zone compared to the Pacific and Central time zones over the month of February 2015



Summary

- The vast majority of tweets were labeled with a “negative” sentiment
- All of the negatively labeled tweets had higher confidence scores than the positive and neutral labels
- These tweets were labeled negative mainly due to bad customer service, late flights, and cancelled flights
- Sunday has the most negative tweets according to the “Sentiment Tweets per Day” bar graph showing the spike in negative tweets during 2/22/15 and 2/23/15
- The main focus for these airline companies would be to improve their customer service departments to decrease the amount of negatively labeled sentiment tweets

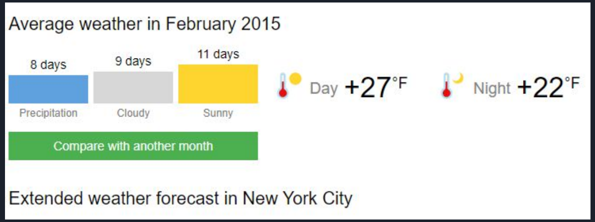
APPENDIX

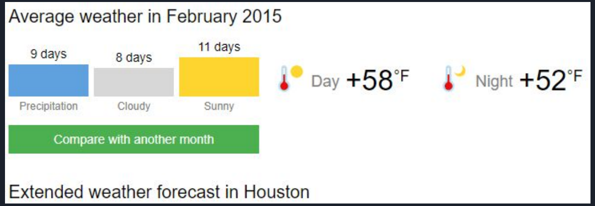
Action Steps

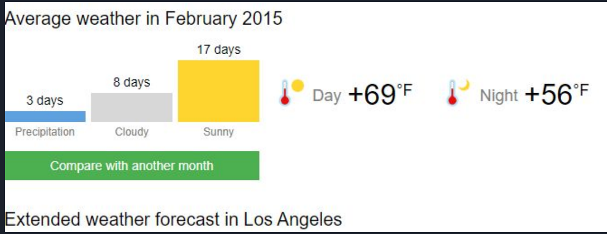
1. I first attempted opening google colab in attempts to create a model and remove null values but found it might be easier to operate in excel
2. I created a naive bayes model to see which words appeared most in the tweets and this gave me a very good idea of what words associated negative tweets and positive tweets
3. I decided not to use most of the data with missing/null values in my analysis because I figured it would not be easy to work with or resemble anything useful
4. From there I chose to look mainly at the sentiment results associated with each airline, confidence in the results, sentiment tweets per day, and which region held the most negative/positive/neutral tweets
5. I created visuals in excel with the data and conditionalized regions with associated sentiment results with a count of each
6. After I had my assumptions as to why the sentiment imbalance was the way it was and researched a little more on the weather during the specified dates

Outside research

- I speculate the reason why there are more negative tweets associated with flight cancellations and delays is due to weather.
- In the winter (which is when this data was collected), the east coast had much harsher climate with higher amounts of precipitation and snow
- Flights could have been canceled due to this and that is why I believe that there are less negative tweets in the west regions







Code

P1-MGSC410.ipynb

File Edit View Insert Runtime Tools Help

Comment Share

+ Code + Text RAM Disk Editing

```
# processing text into just english

stop_words = stopwords.words('english')

data['text_without_stopwords'] = data['text'].apply(lambda x: ' '.join([word for word in x.split() if word r

countVect = CountVectorizer(min_df= 10)

binaryVector = countVect.fit_transform(data.text_without_stopwords)
```

```
y = data.airline_sentiment
X = binaryVector

train_X, test_X, train_y, test_y = train_test_split(X, y, random_state=123)

print([x.shape for x in [train_X, test_X, train_y, test_y]])
```

```
[(10980, 1872), (3660, 1872), (10980,), (3660,)]
```

```
# naive bayes model
MNB = MultinomialNB()
MNB.fit(train_X, train_y)

predicted = MNB.predict(test_X)
accuracy_score = metrics.accuracy_score(predicted, test_y)
confusion_count = metrics.confusion_matrix(predicted, test_y)


print('Accuracy: ',accuracy_score,'\n')
print('Confusion Matrix:\n',confusion_count)
```

```
Accuracy:  0.7808743169398907

Confusion Matrix:
 [[2015  304   76]
 [ 186  430   60]
 [ 111   65  413]]
```

```
# Most seen words associated with positive / negative

neg_class_prob_sorted = MNB.feature_log_prob_[0, :].argsort()[::-1]
```

0s completed at 8:12 AM


```
print('Accuracy: ',accuracy_score,'\n')
print('Confusion Matrix:\n',confusion_count)
```

```
Accuracy:  0.7808743169398907

Confusion Matrix:
 [[2015  304   76]
 [ 186  430   60]
 [ 111   65  413]]
```

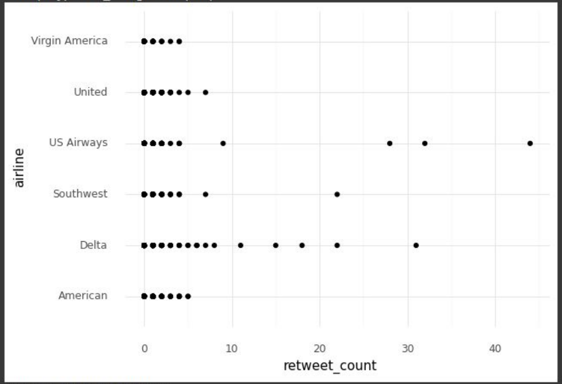

```
# Most seen words associated with positive / negative

neg_class_prob_sorted = MNB.feature_log_prob_[0, :].argsort()[::-1]
pos_class_prob_sorted = MNB.feature_log_prob_[1, :].argsort()[::-1]

print('Negative words:\n', np.take(countVect.get_feature_names(),
                                   neg_class_prob_sorted[:25]))
print('\nPositive words:\n', np.take(countVect.get_feature_names(),
                                     pos_class_prob_sorted[:25]))
```

```
Negative words:
 ['united' 'flight' 'usairways' 'americanair' 'southwestair' 'jetblue'
 'get' 'cancelled' 'service' 'hours' 'hold' 'can' 'customer' 'help' 'time'
 'plane' 'amp' 'delayed' 'still' 'you' 'us' 'co' 'one' 'call' 'http']

Positive words:
 ['jetblue' 'united' 'southwestair' 'flight' 'co' 'http' 'americanair'
 'usairways' 'get' 'please' 'flights' 'virginamerica' 'need' 'thanks'
 'help' 'can' 'dm' 'would' 'know' 'it' 'our' 'fleek' 'fleet' 'us' 'you']
```

Code

```
g1 = (ggplot(data, aes(x = "retweet_count", y = "airline"))
+ geom_point() + theme_minimal())

g1
```

```
/usr/local/lib/python3.7/dist-packages/plotnine/utils.py:1246: FutureWarning: is_categorical is deprecated ar
  if pdtypes.is_categorical(arr):
```




```
<ggplot: (8742021338913)>
```

References

- Day of the Week Calculator
- Twitter US Airline Sentiment | Kaggle