Airline Sentiment Analysis

By Victoria Porter, Will Holbrook, Laura Cope, Lauren Cooper and Joel Wolinsky

Agenda

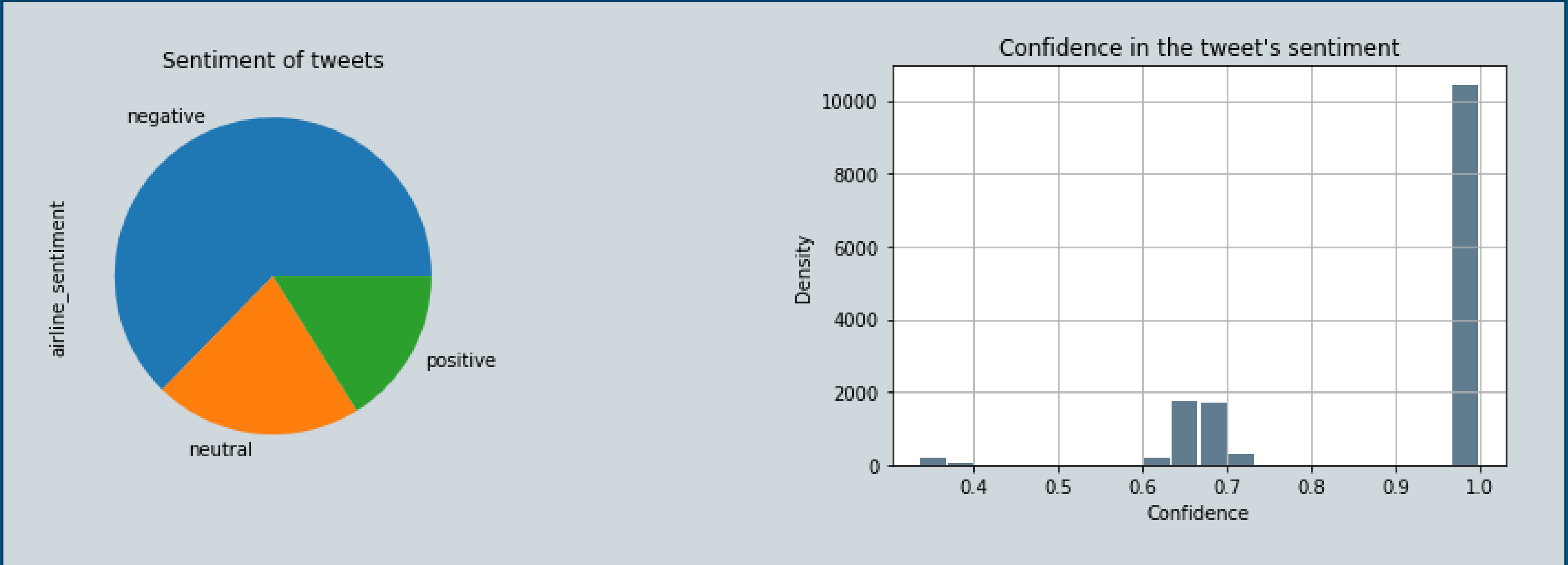
- Exploratory Analysis
- Preparing the Dataset
- Modelling Experiments
- Performance Analysis
- The Final Model
- Final Model Insights
- Teamwork
- Learning Outcomes
- Questions

Exploratory Analysis

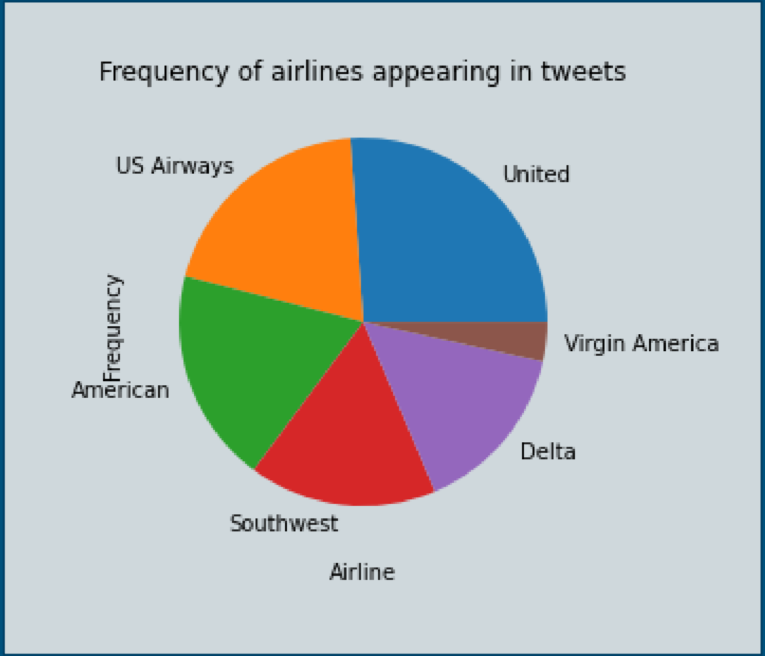
Variable identification

Type of variable	*Data Type*	*Variable Category*
Predictor Variable - airline - name - retweet_count - text - tweet_coord - tweet_created - tweet_location - user_timezone **Target Variable** - Sentiment **Other -** - airline_sentiment - airline_sentiment_gold - airline_sentiment_confidence - negativereason - airline - negativereason_gold - negativereason_confidence	**Float** - airline_sentiment_confidence - negativereason_confidence **Integer** - tweet_id - retweet_count **String** - tweet_coord *(co-ordinate)* - tweet_created *(date)* - airline_sentiment - negativereason - airline - airline_sentiment_gold - name - negativereason_gold - text - tweet_location - user_timezone	**Categorical** - airline_sentiment - negativereason - airline - airline_sentiment_gold - negativereason_gold - User_timezone **Numerical** - tweet_id - airline_sentiment_confidence - negativereason_confidence - retweet_count **Other** - name - text - Tweet_coord - Tweet_location - tweet_created

Sentiment

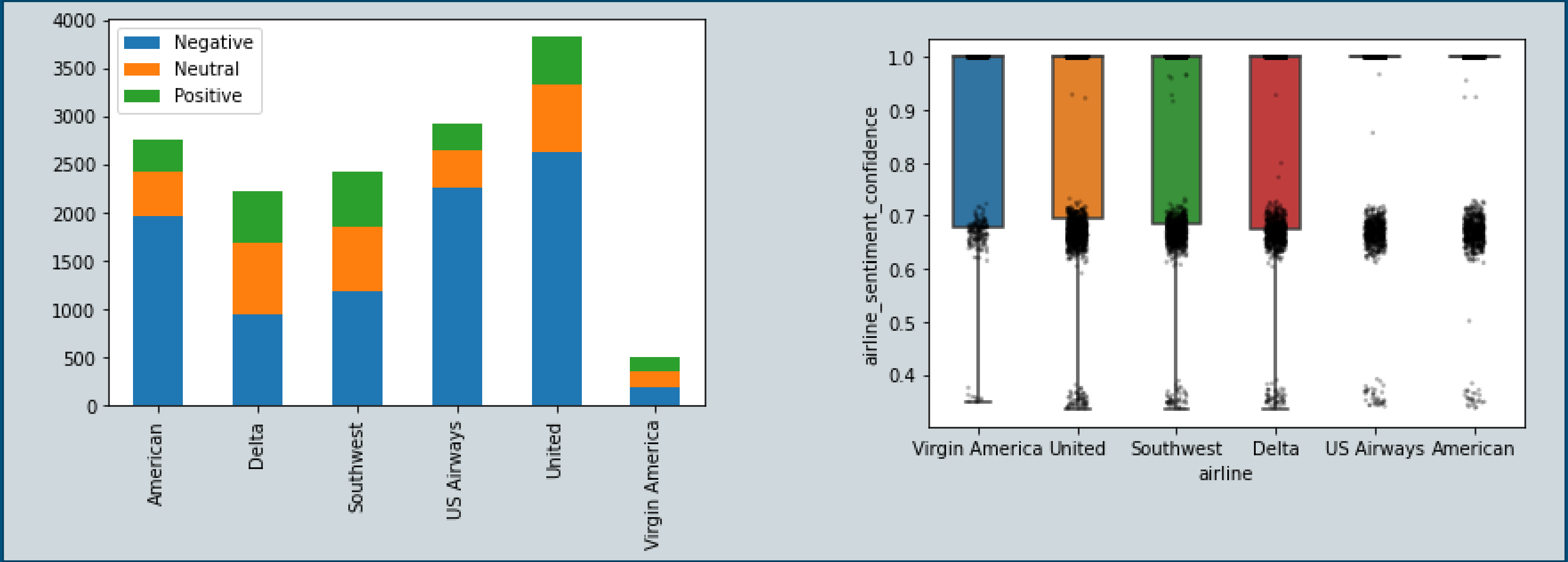

Sentiment of tweets
airline_sentiment
negative
neutral
positive
Confidence in the tweet's sentiment
Density
10000
8000
6000
4000
2000
0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Confidence

Airlines

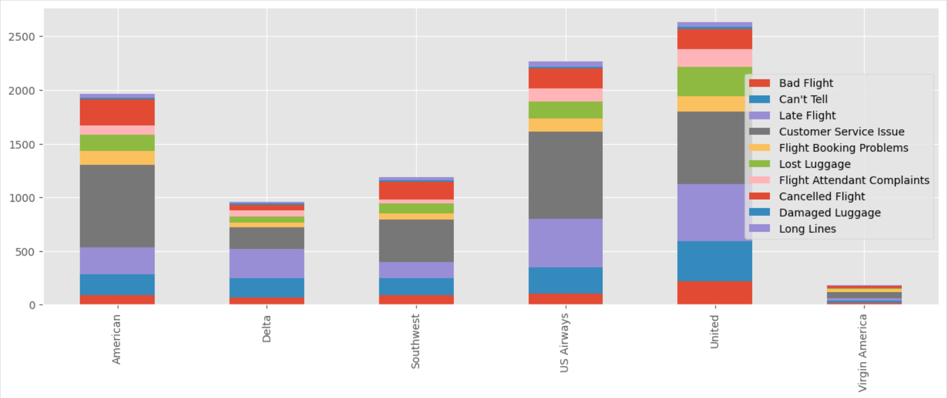


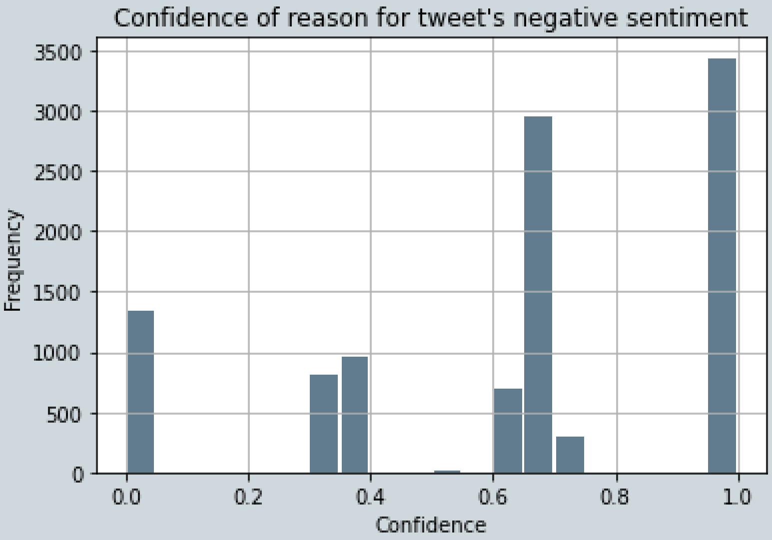
Frequency of airlines appearing in tweets
US Airways
United
Frequency
Virgin America
American
Delta
Southwest
Airline

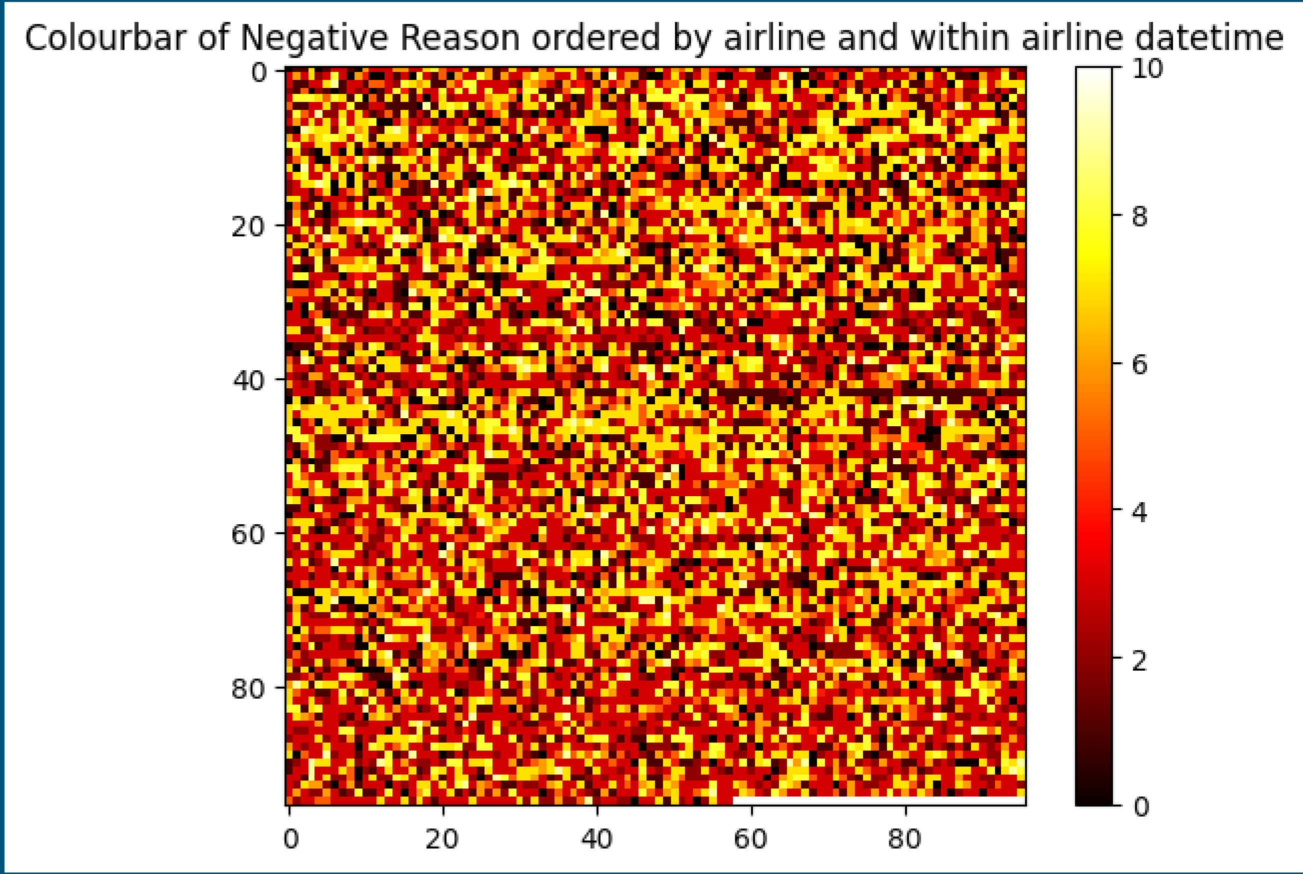
→ Airlines and Sentiment, and respective Confidence



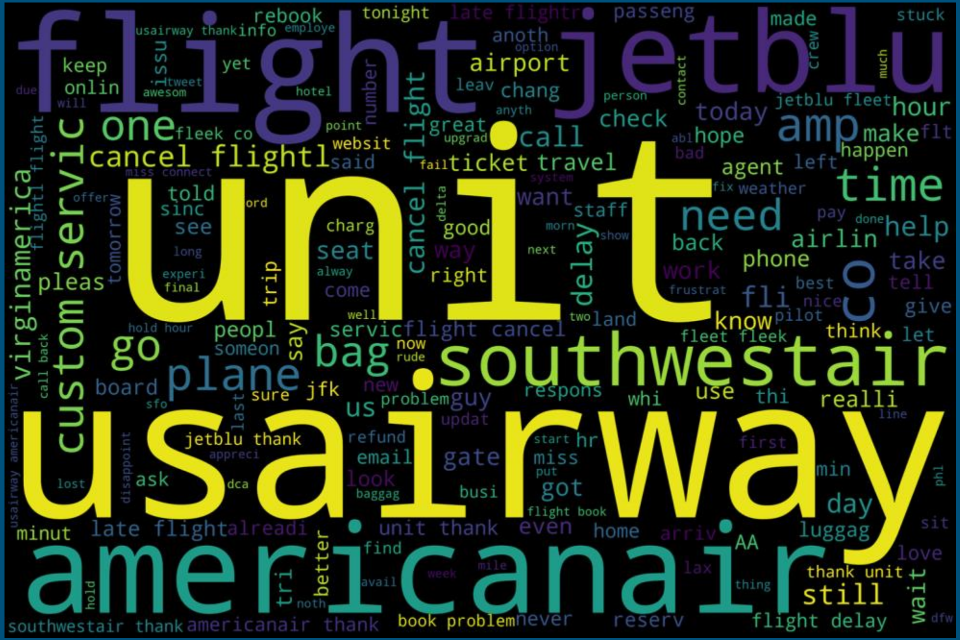
Reason for negative sentiment



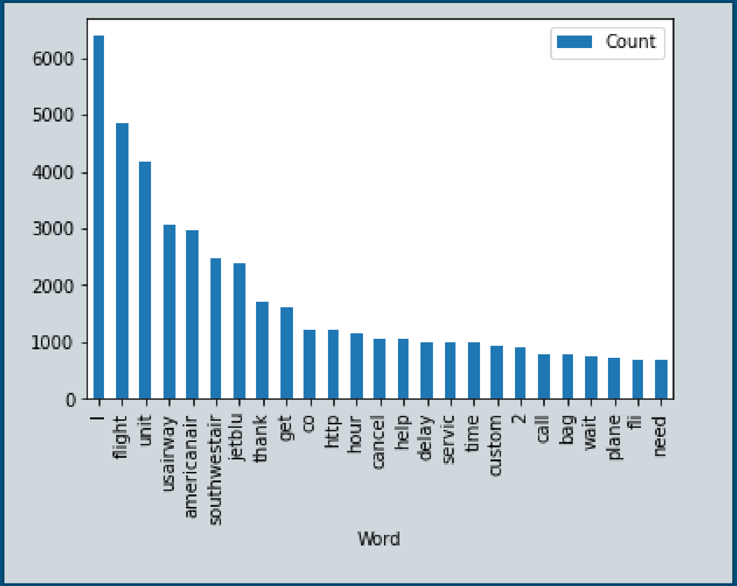



Colourbar of Negative Reason ordered by airline and within airline datetime
0
20
40
60
80
0
20
40
60
80
10
8
6
4
2
0

Text



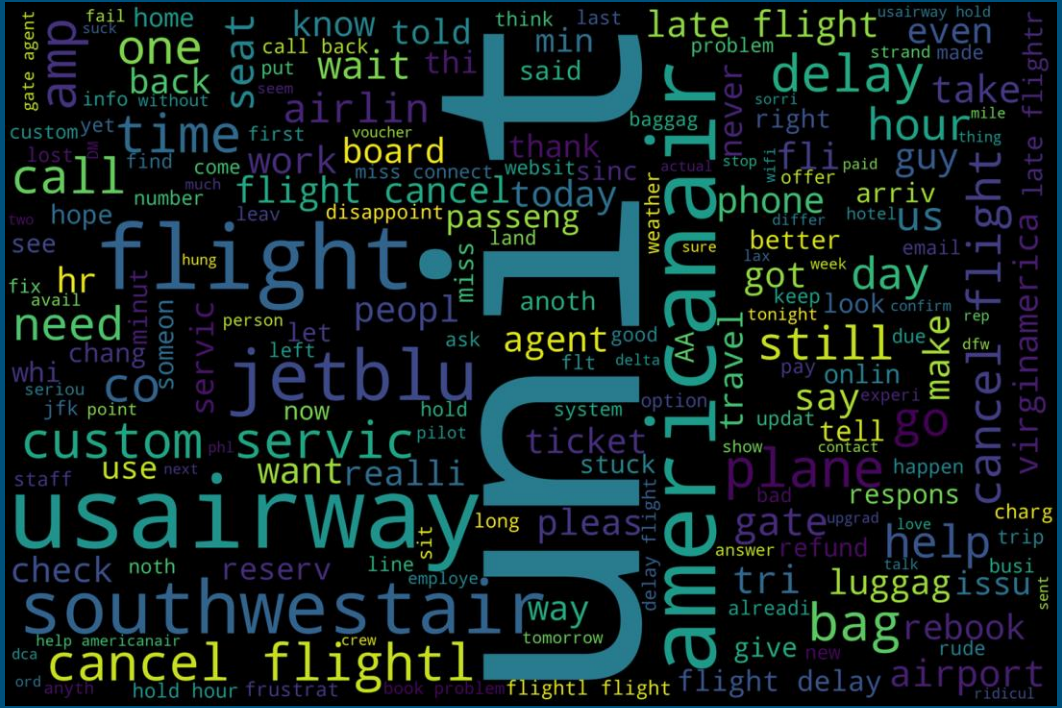
flight
jetblu
unit
usairway
americanair
southwestair
custom servic
virginamerica
cancel flightl
one
go
plane
bag
amp
time
need
co
airport
call
ticket
travel
check
today
hope
hour
make
help
airlin
back
phone
delay
cancel flight
seat
good
right
pleas
know
still
day
luggag
flight delay
jetblu thank
unit thank
thank unit
late flight
southwestair thank
americanair thank
book problem



Count
6000
5000
4000
3000
2000
1000
0
I
flight
unit
usairway
americanair
southwestair
jetblu
thank
get
co
http
hour
cancel
help
delay
servic
time
custom
2
call
bag
wait
plane
fli
need
Word

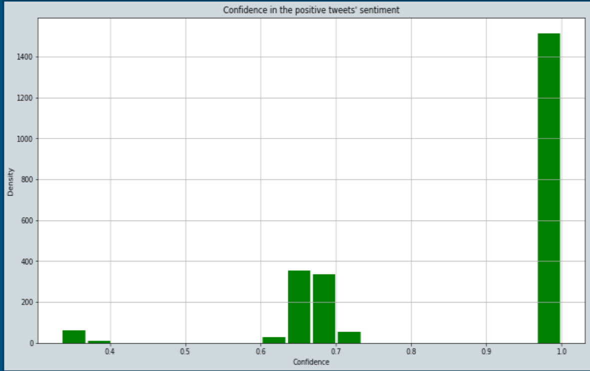
→ Text and Sentiment

Negative tweets

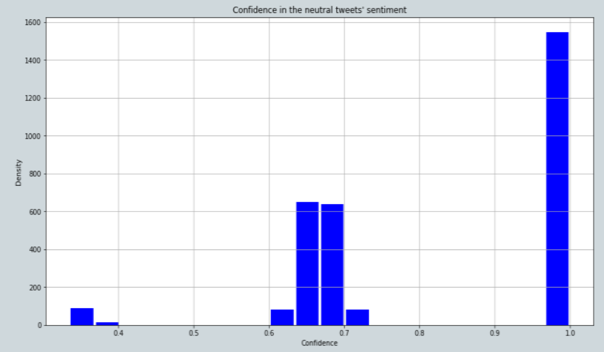


Sentiment Confidence

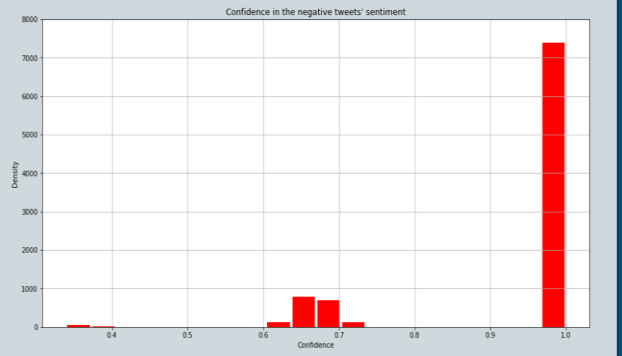
Positive



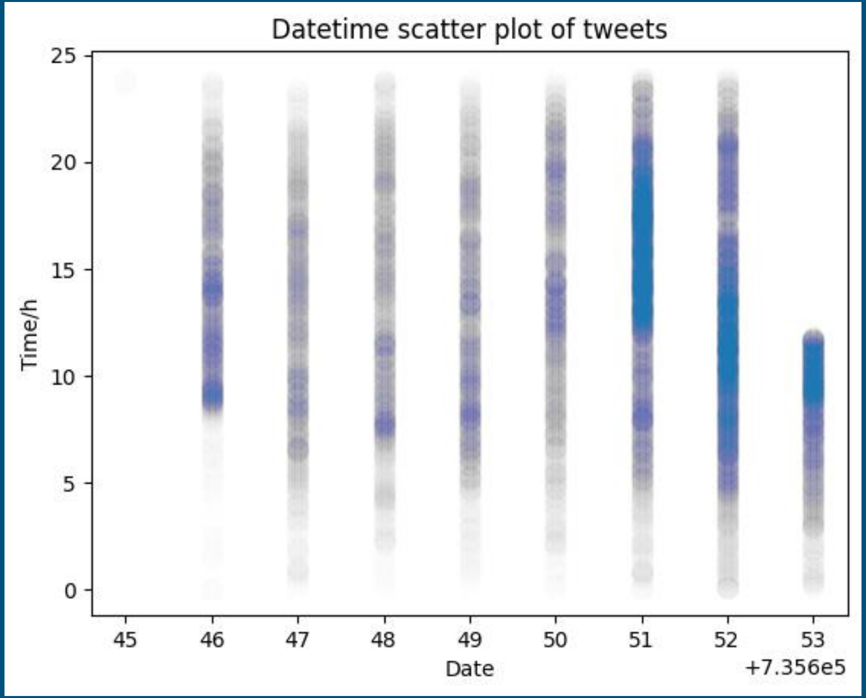
Neutral

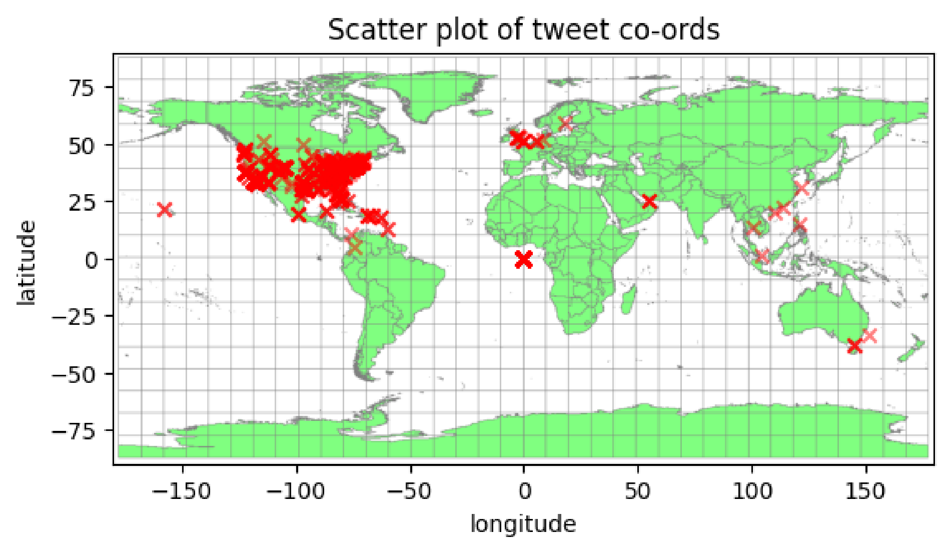


Negative



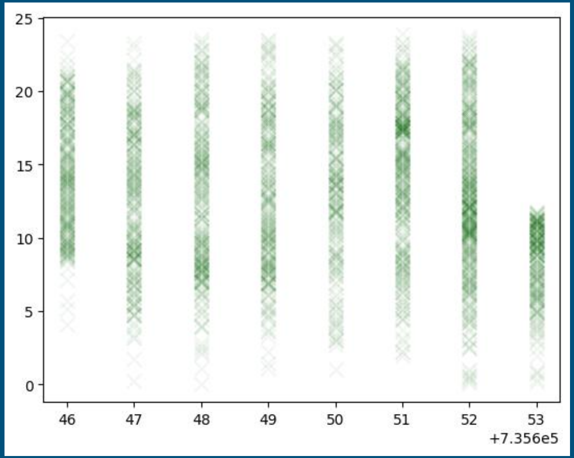
Creation datetime, Co-ords



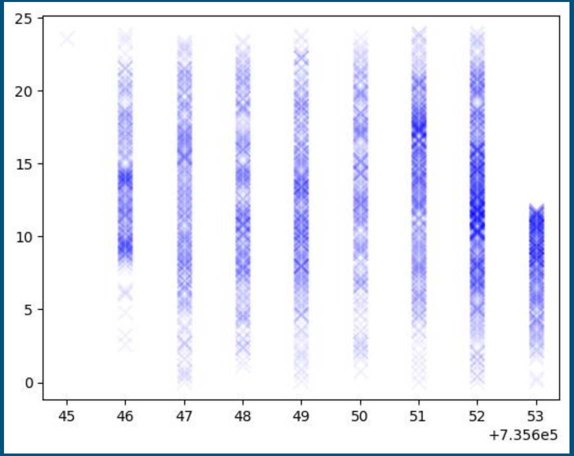


→ Creation datetime and Sentiment

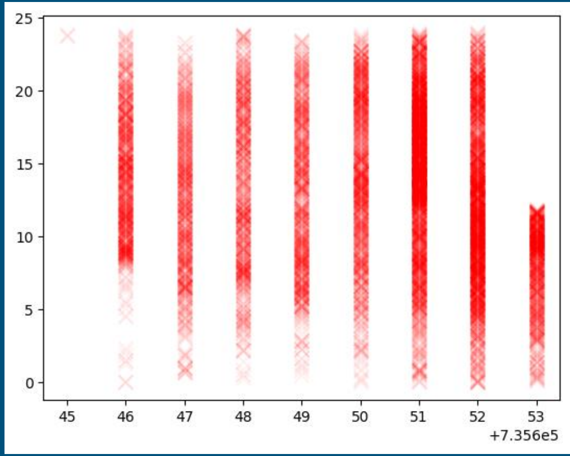
Positive



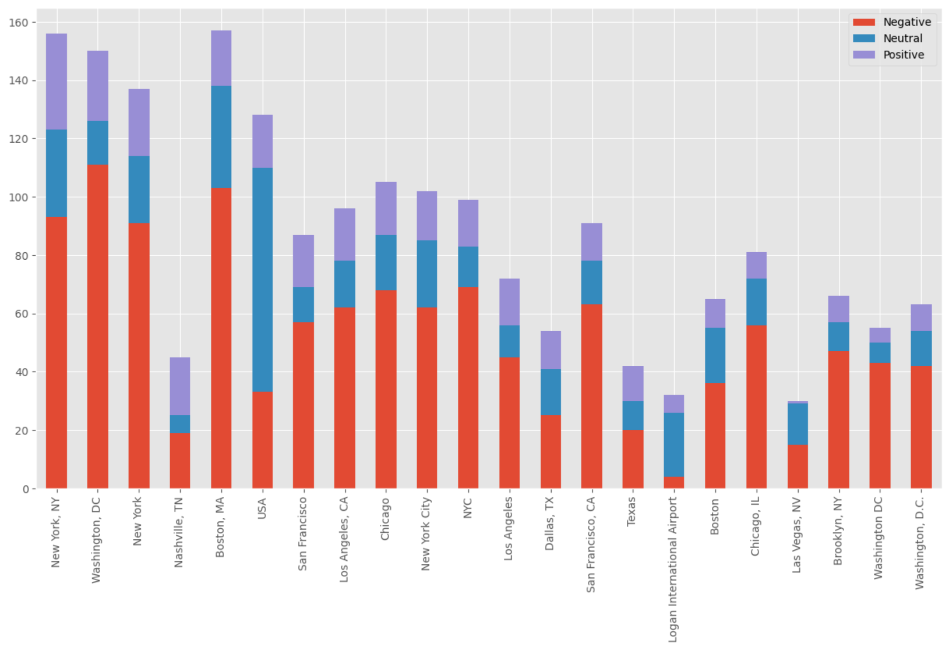
Neutral

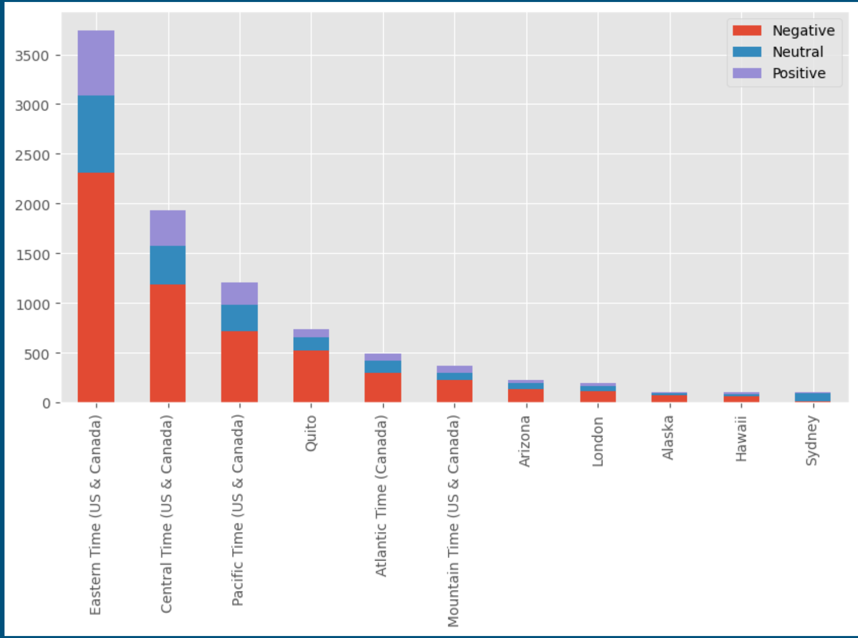


Negative

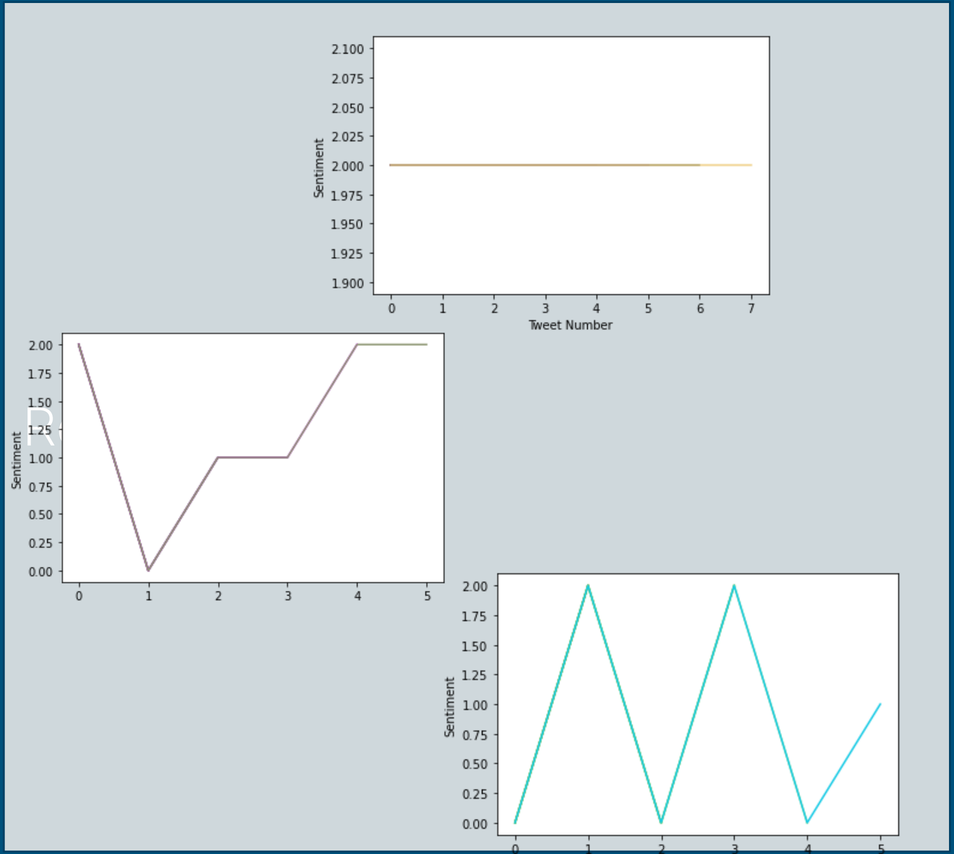


→ Location and Sentiment , Timezone and Sentiment

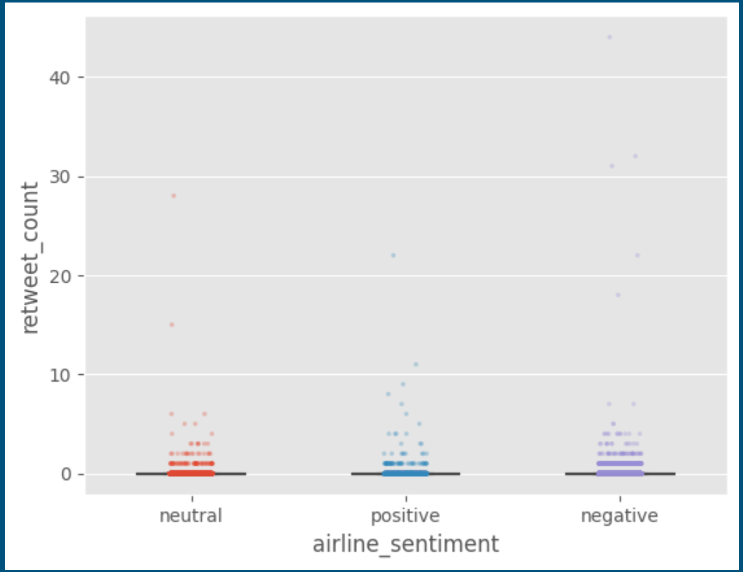




Name



Retweet Count



Preparing the Dataset

Techniques used:

- Tokenization
- Removing stops words
- Stemming & Lemmatizing
- Removing Airline_sentiment_confidence < 0.6

Before - '@VirginAmerica it was amazing, and arrived an hour early. You're too good to me.'

After - 'virginamerica amaz arriv hour earli you good'

Baseline Model: Naïve Bayes

Model Type	Model Name	Training Time/s	Testing Time/s	Accuracy	RAM Used/GB
Baseline	Naïve Bayes	2.186	45.665	0.7668	0.16

Experiment

-ve and +ve

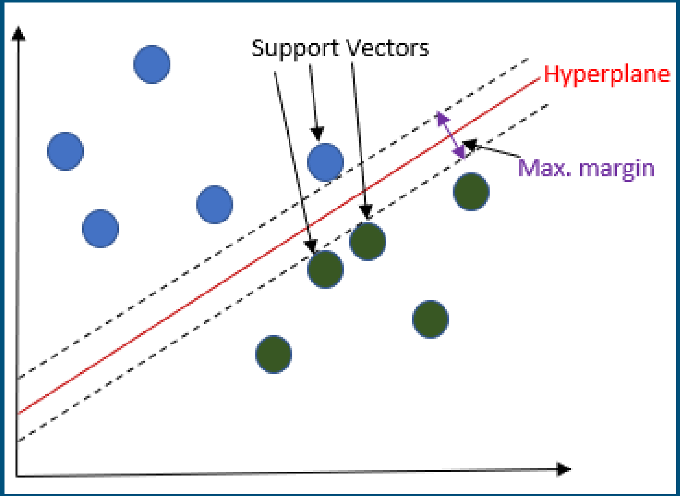
```
Confusion Matrix:
[[866  60]
 [ 72 157]]
```

-ve, neu and +ve

```
Confusion Matrix:
[[866  42  12]
 [143 130  23]
 [ 79  35 134]]
```

Non-Neural Model: Support Vector Machine (SVM)

Model Type	Model Name	Training Time/s	Testing Time/s	Accuracy	RAM Used/GB
Non-Neural	SVM	234.618	9.553	0.7766	0.47 GB



Pretrained Model: MonkeyLearn Sentiment Analysis

- MonkeyLearn was one of the few fully off-the-shelf pretrained sentiment analysis models available for free.
- Benefits include a simple implementation, a quick runtime with no training needed.
- Drawbacks include little available information on the model itself: it is very much a black box. No details available on the model or training data. This could limit the model's ability to understand tweets as opposed for formal language
- Only 300 free queries allowed per month limited how much we could test the model

MonkeyLearn Results

- Mid-level accuracy, possibly overfit to training data or trained or formal English language writing. This teaches us that the training data used is very important, and some pre-trained sentiment analysis models may not be suitable for informal language.
- The model worked significantly better on the raw tweets as opposed to the cleaned data set. The confidence score on the raw data was also much higher.
- Next steps would involve a paid subscription which would likely come with more detail on the model and training data itself, and the ability to test more data.

Test Data Used (300 tweets each)	Accuracy	Average Confidence
Cleaned	0.563	0.7092
Raw	0.780	0.7740

The Final Model

Model Type	Model Name	Training Time/s	Testing Time/s	Accuracy
Baseline	Naïve Bayes	2.186	45.665	0.7668
Non-Neural	SVM	234.618	9.553	0.7766
Pre-trained	MonkeyLearn Sentiment Analysis	N/A	150.650	0.7800

Key Predictors

```
Accuracy:  0.7766393442622951

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1826
           1       0.56      0.68      0.62       611
           2       0.72      0.70      0.71       491

    accuracy                           0.78      2928
   macro avg       0.72      0.74      0.73      2928
weighted avg       0.79      0.78      0.78      2928


Confusion Matrix:
[[1512  245   69]
 [ 130  418   63]
 [  65   82  344]]
```

Just Tweet Text

```
Accuracy:  0.7776639344262295

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1826
           1       0.56      0.68      0.62       611
           2       0.72      0.70      0.71       491

    accuracy                           0.78      2928
   macro avg       0.72      0.74      0.73      2928
weighted avg       0.79      0.78      0.78      2928


Confusion Matrix:
[[1515  243   68]
 [ 130  418   63]
 [  65   82  344]]
```

With TimeZone

```
Accuracy:  0.7766393442622951

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1826
           1       0.56      0.68      0.62       611
           2       0.72      0.70      0.71       491

    accuracy                           0.78      2928
   macro avg       0.72      0.74      0.73      2928
weighted avg       0.79      0.78      0.78      2928


Confusion Matrix:
[[1512  242   72]
 [ 129  418   64]
 [  65   82  344]]
```

Tweet Time

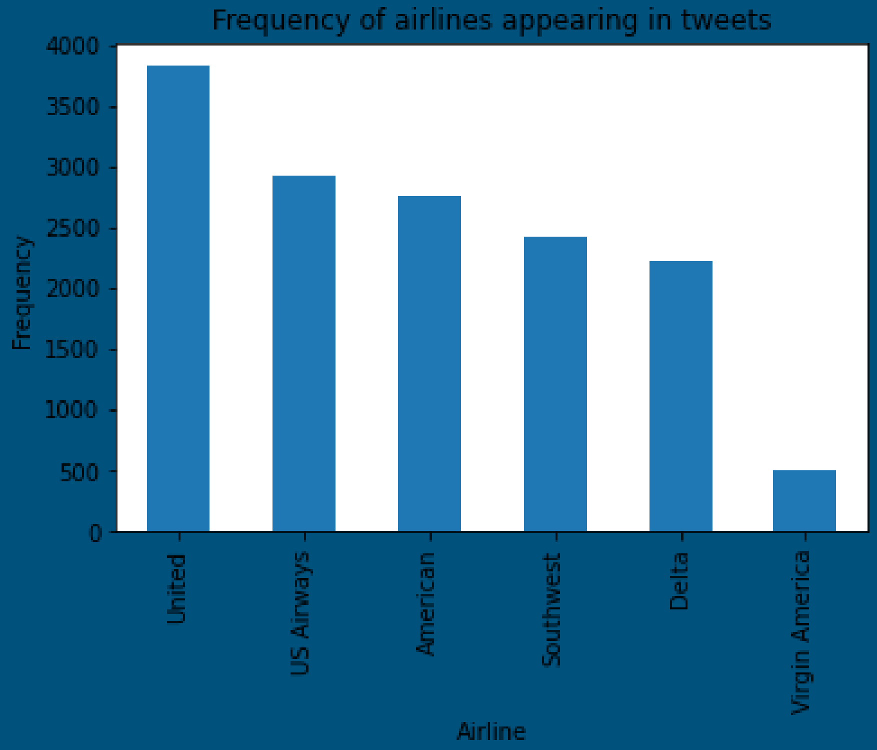
```
Accuracy:  0.7759562841530054

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1826
           1       0.56      0.68      0.62       611
           2       0.72      0.70      0.71       491

    accuracy                           0.78      2928
   macro avg       0.72      0.74      0.73      2928
weighted avg       0.79      0.78      0.78      2928


Confusion Matrix:
[[1511  246   69]
 [ 130  418   63]
 [  66   82  343]]
```

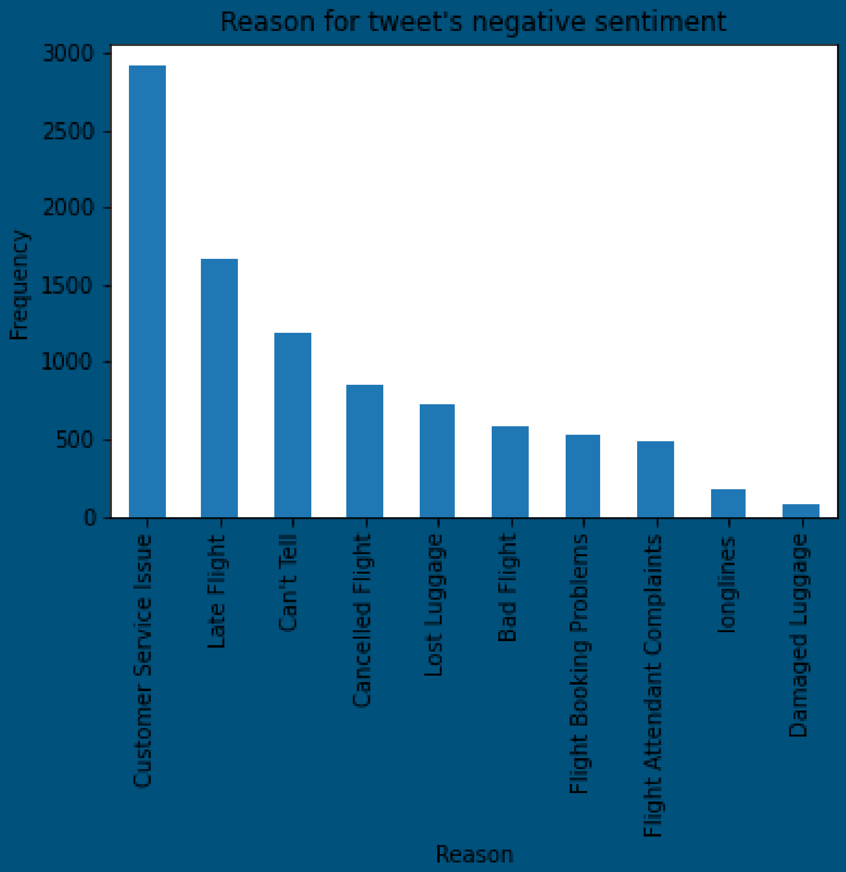
With retweet

Challenges

- Choosing which columns of the dataset to feed into the model and assessing how much of an impact they had (feature importance)
- Knowing how much to clean and prepare the data for processing
- Pre-Trained models for sentiment analysis are not easily accessible for free experimentation and keep their parameters hidden
- A problem beyond our control left us without a neural model, which we initially predicted to be our best performing model
- Our ‘baseline’ model (Naive Bayes) performed similarly to our other models but in a shorter time, meaning that these more complicated models may not actually be better

Sentiment Conclusion



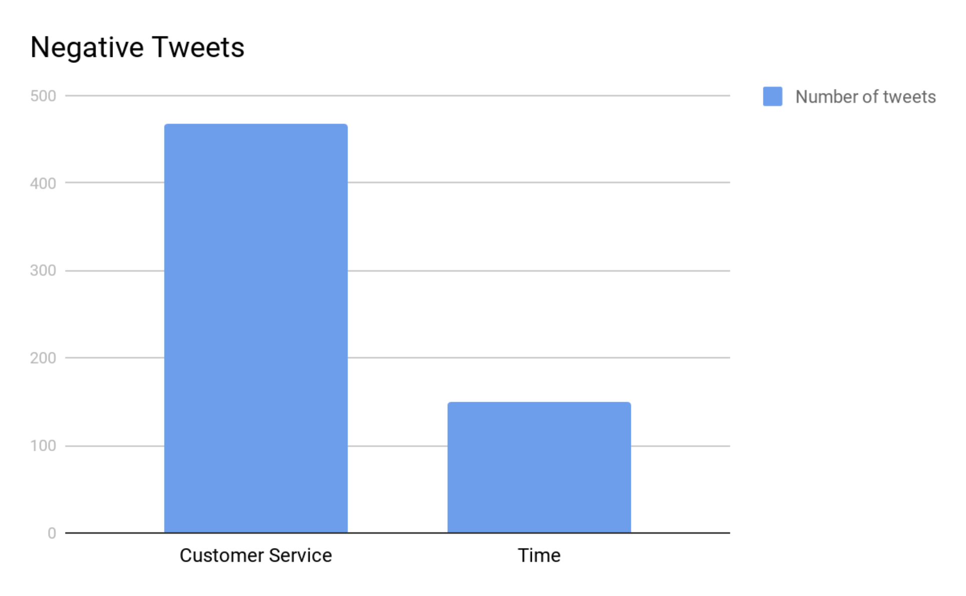


Technical Next Steps

- Building a cloud model that classifies using our findings
- Balance and expand dataset

Business Next Steps

- 1035 of the tweet were related to Customer service
- 22% of those tweets mentioned time
- Quite a simple problem to fix



Teamwork

- Video calls 1-2 times a week and messaging in between helped delegate work and check up on progress.
- We split up to explore different areas of the project, sharing our results in Google Drive using Google Colab for our code.
- A setback towards the end of the project meant we did not have a completed Neural Network solution, but we were able to regroup quickly and reevaluate our approach to the project.
- Splitting up the project into stages and using an agile methodology allowed for smooth re-planning and adaptation.

Learning Outcomes

Data exploration is key to good results

Introduction to NLP

Learning Python

There is no one size fits all model

Impact of cleaning a dataset

ML Evaluation Metrics

Power of visualisation

Questions?