Sentiment Analysis w/ Airline Tweets

Thinkful Final Capstone
By Matthew Huh

On Twitter and the spread of language

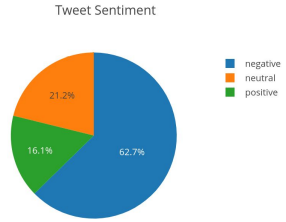
About the Dataset

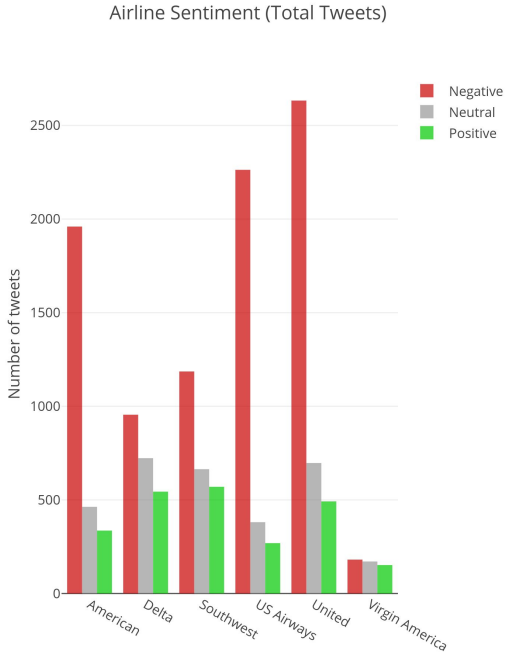
- Dataset 1 - obtained from Kaggle, all tweets have been validated with a positive / neutral / negative rating (and a negative reason if applicable) -
  - contains 6 airlines (American, Delta, Southwest, US Airways, United, Virgin)
  - From February 2015
  - 14,640 tweets
- Dataset 2 - dataset accessed using tweepy
  - 10 largest airlines in US, collected (Alaska, Allegiant, American, Delta, Frontier, Hawaiian, Jetblue, Southwest, Spirit, United)
  - From November 2018
  - 18,000 tweets

Notes: red text = present in both training and testing data sets

US Airways merged with American in late 2015, and Virgin merged with Alaska earlier this year, 2018

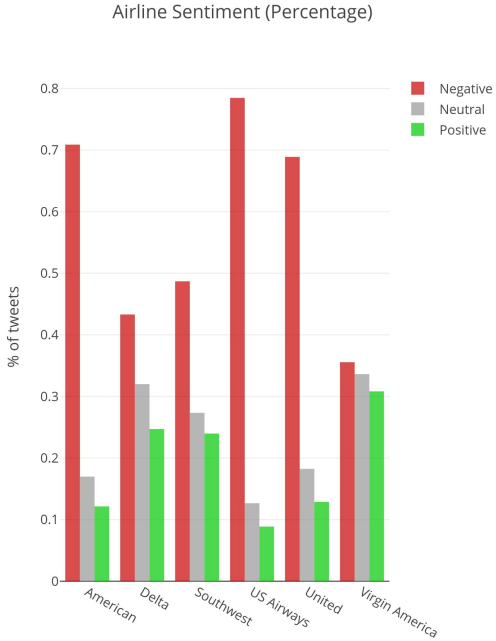
Tweet Distribution



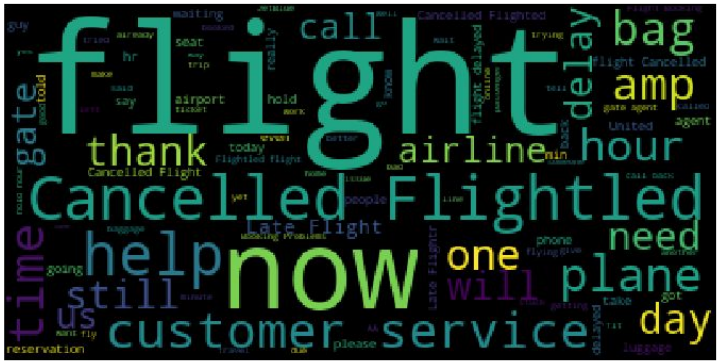


Tweet counts:

United	3822
US Air	2913
American	2759
Southwest	2420
Delta	2222
Virgin	504



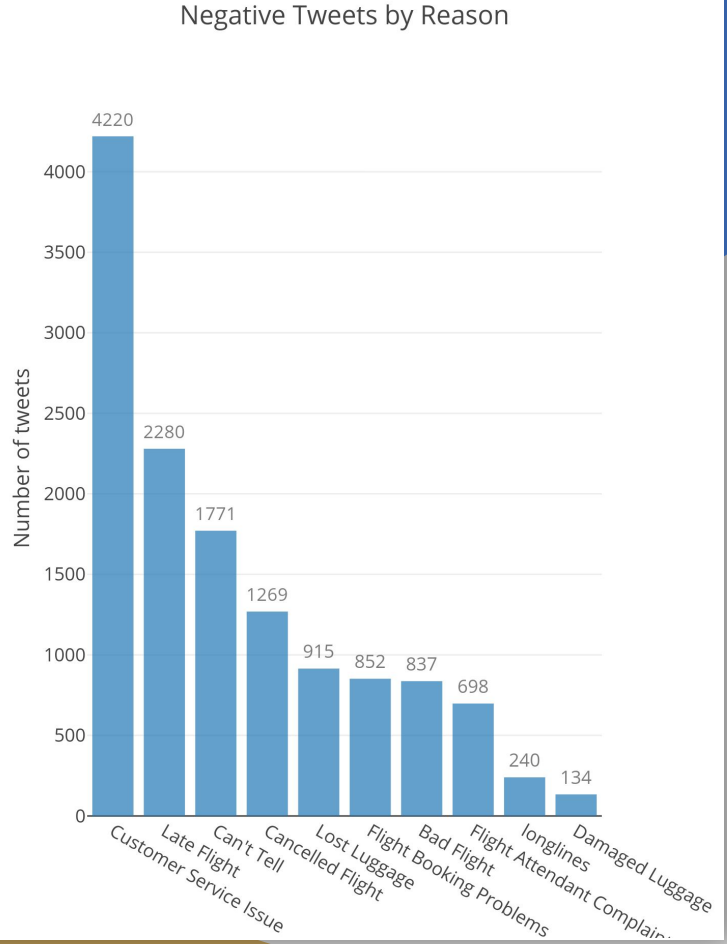
What People are Saying



@united I did start a claim but 8-10 weeks is unrealistic, am I really supposed to go that long with out a car seat for my child.Ridiculous!

@VirginAmerica you guys messed up my seating.. I reserved seating with my friends and you guys gave my seat away ... 😎 I want free internet

@AmericanAir I understand weather is not your fault but ur cs reps are atrocious. I am NOT happy nor will I EVR fly w/ u again.



Feature Selection / Natural Language Processing

Text cleaning -

- Regex (stop words, characters)
- spacy

Feature selection process

- Spacy (parts of speech - adverb, verb, noun, adjective)
- Tweet characteristics (word count, character count, stop count, special char count)
- Tf-idf vectorization (1200/ 6000+ features selected)

Modeling Process

Target outcome (Positive, neutral, negative) -> 1, 0, -1

Models Selected

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting
- Neural Networks

Logistic Regression

+ Easy to build and interpret
+ Fast runtimes / low complexity
- Difficult to predict multiple outcomes

Train Score:	81.92
Test Score:	77.26
Runtime:	9.90 s

Optimal Parameters

- Penalty: ‘l1’, ‘l2’
- Solver = ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’
- Multi_class = ‘ovr’, ‘multinomial’

Random Forest Classifier

+ Feature importances
+ Reduced variance
+ Parallelizable
- “Black box”
- Computationally expensive

Train Score:	88.28
Test Score:	79.64
Runtime:	19.52 s

Optimal Parameters:

- N_estimators: 10, 20, 40, 60
- Class_weight: none, ‘balanced’, ‘balanced_subsample
- Oob_socre: True, False

Gradient Boosting Classifier

\+ Feature importance
\+ Strong learner / performer
\- Sequential operation
\- Prone to overfitting

Train Score:	75.44
Test Score:	72.93
Runtime:	27.98 s

Tested parameters:

- Number of estimators: 100
- Max depth: 3
- Min number of samples in leaves: 1

Neural Networks

+ Very strong performer
- “Black box”
- Very computationally expensive

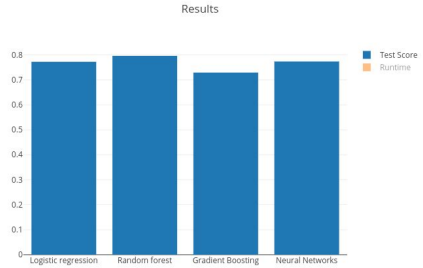
Train Score: 88.16
Test Score: 77.38
Runtime: 59.18 s

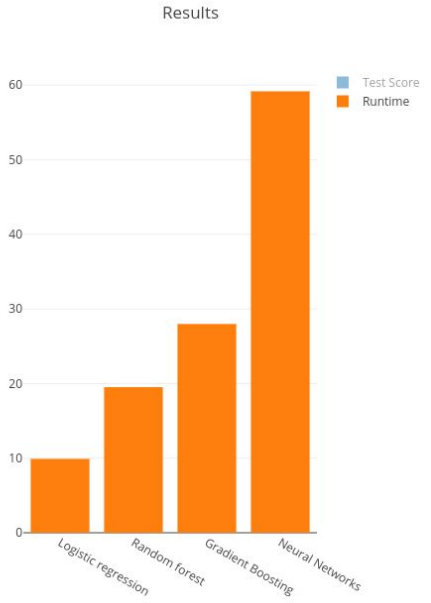
Tested parameters:

- Hidden layer size: (100, 10)
- Max iterations: 200
- Learning rate: ‘constant’

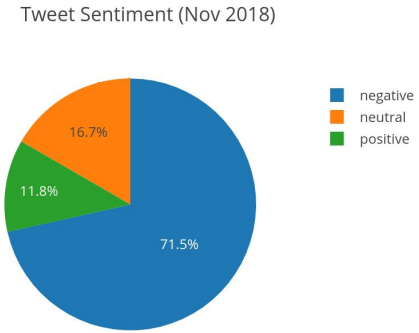
Model Comparison

	Logistic Regression	Random Forest	Gradient Boosting	Neural Networks
Train accuracy	81.92	88.28	75.44	88.16
Validation accuracy	77.26	79.63	72.93	77.38
Runtime	9.90 s	19.52 s	27.98 s	59.18 s



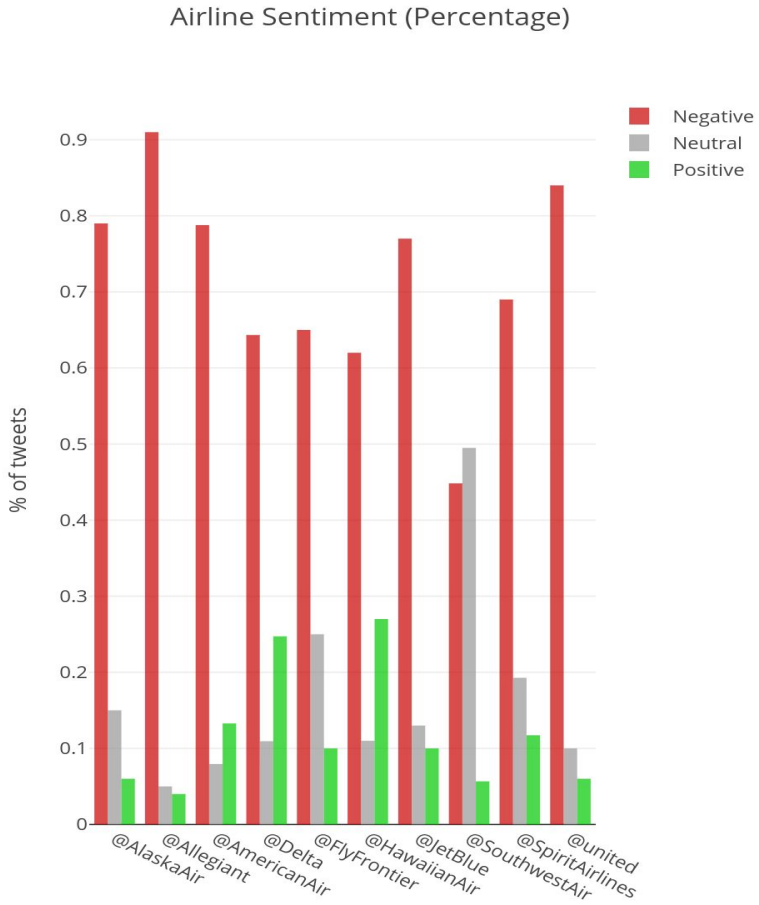


Testing Set Results



RT @AlaskaAir: Happy birthday to our Chief Football Officer, @DangeRussWilson! We hope your thirties are as fly as you! 🎂 https://t.co/XfTm…

@AlaskaAir Stupid, stupid decision!!



Conclusion & Next Steps

- Model behaves fairly well when training and testing within the dataset, but applying the results to another has mixed results
- NLP requires large data sets (14,000 tweets in training set can provide fair results, but having a larger training set may be necessary to perform this task)
- Compare the results to a word2vec model

Links

Link to capstone project folder:

https://github.com/mhuh22/Portfolio/tree/master/Sentiment%20Analysis%20with%20Airline%20Tweets

Link to kaggle dataset

https://www.kaggle.com/crowdflower/twitter-airline-sentiment