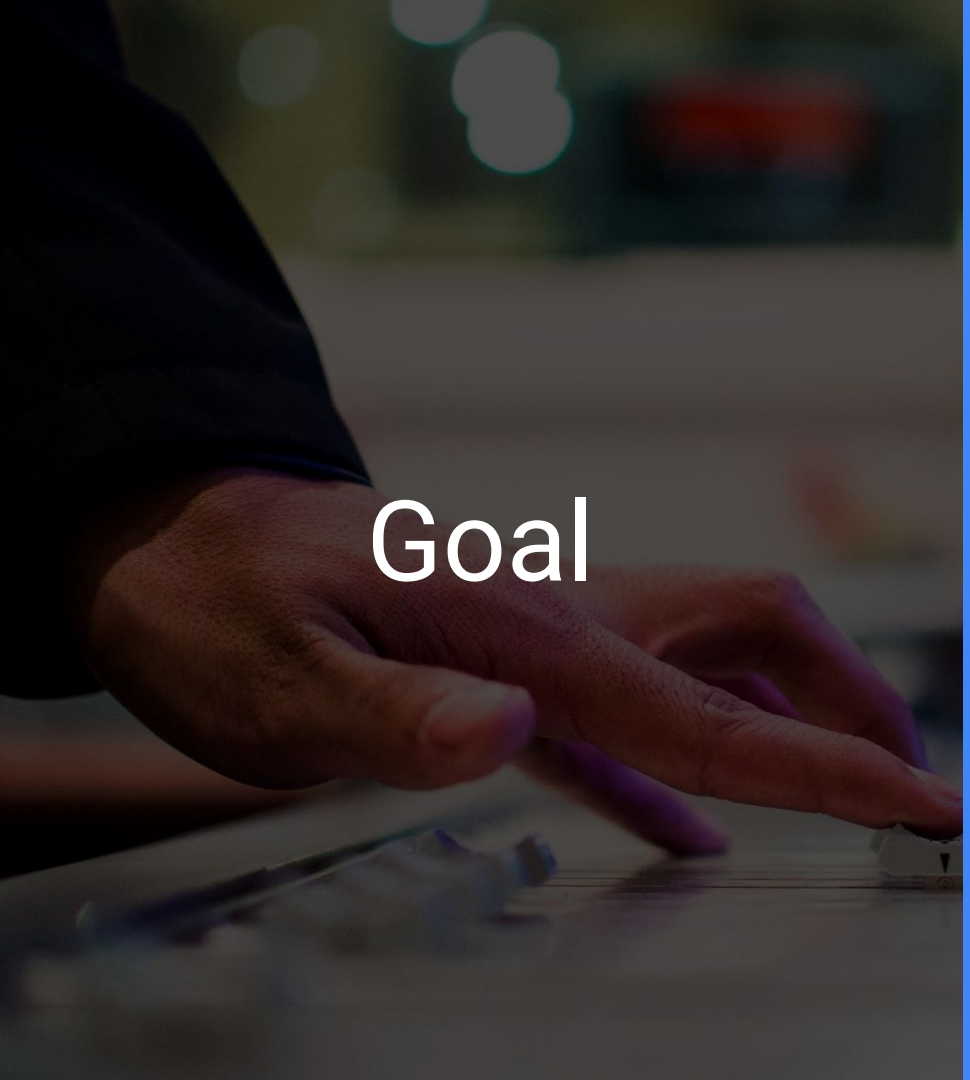
Sentiment Analysis of US Airlines

Rishab Ghose


Utilizing the data
from Social Media is
KEY!

Introduction

- Many reasons for complaint in Airline Industry
- Twitter is nothing but Data!
- Sentiment Analysis is key in having returning passengers



Goal

Build model to classify tweets as positive or negative sentiment

Use model to determine reasons for negative sentiment

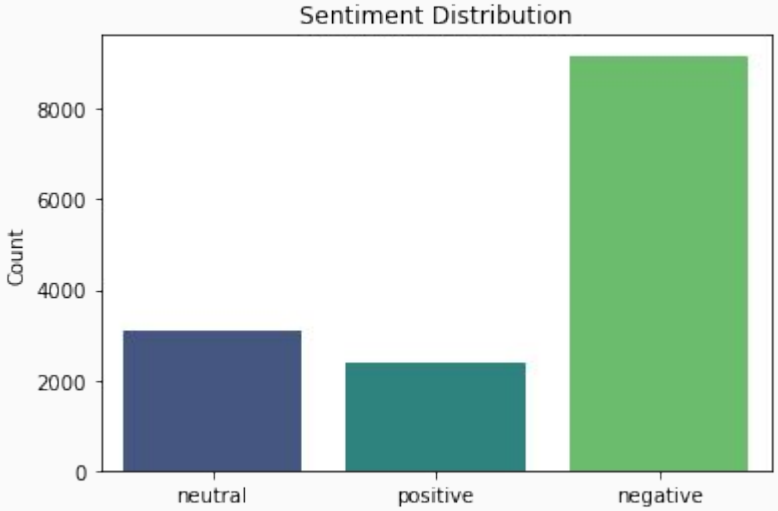
The Data

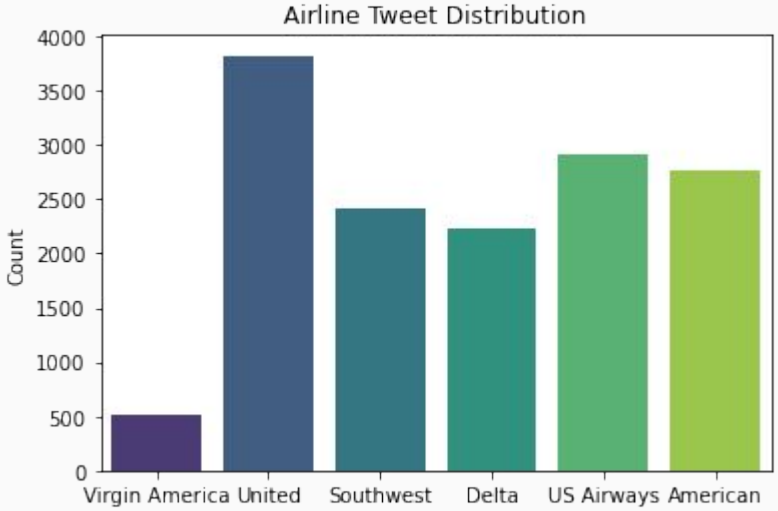
Kaggle Dataset

Twitter Data of Major US Airlines

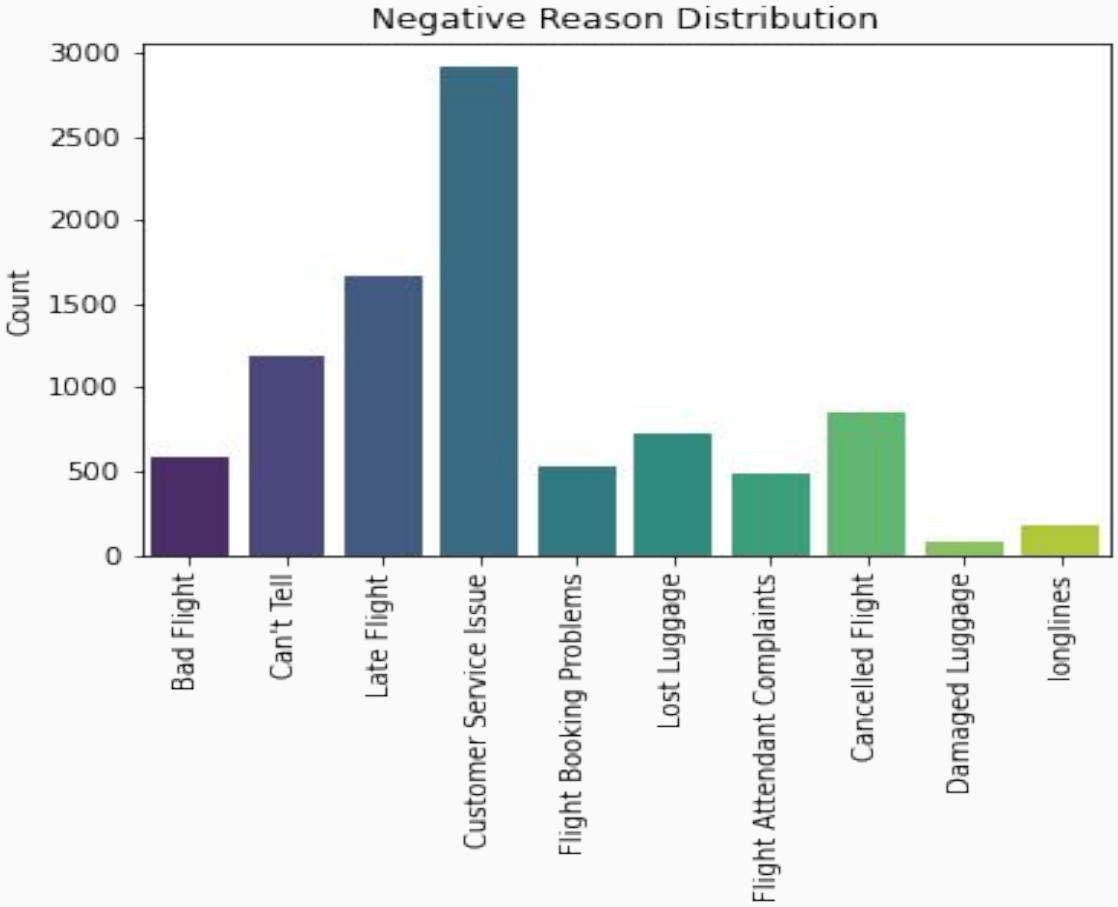
Almost 15K labeled tweets

Data Exploration



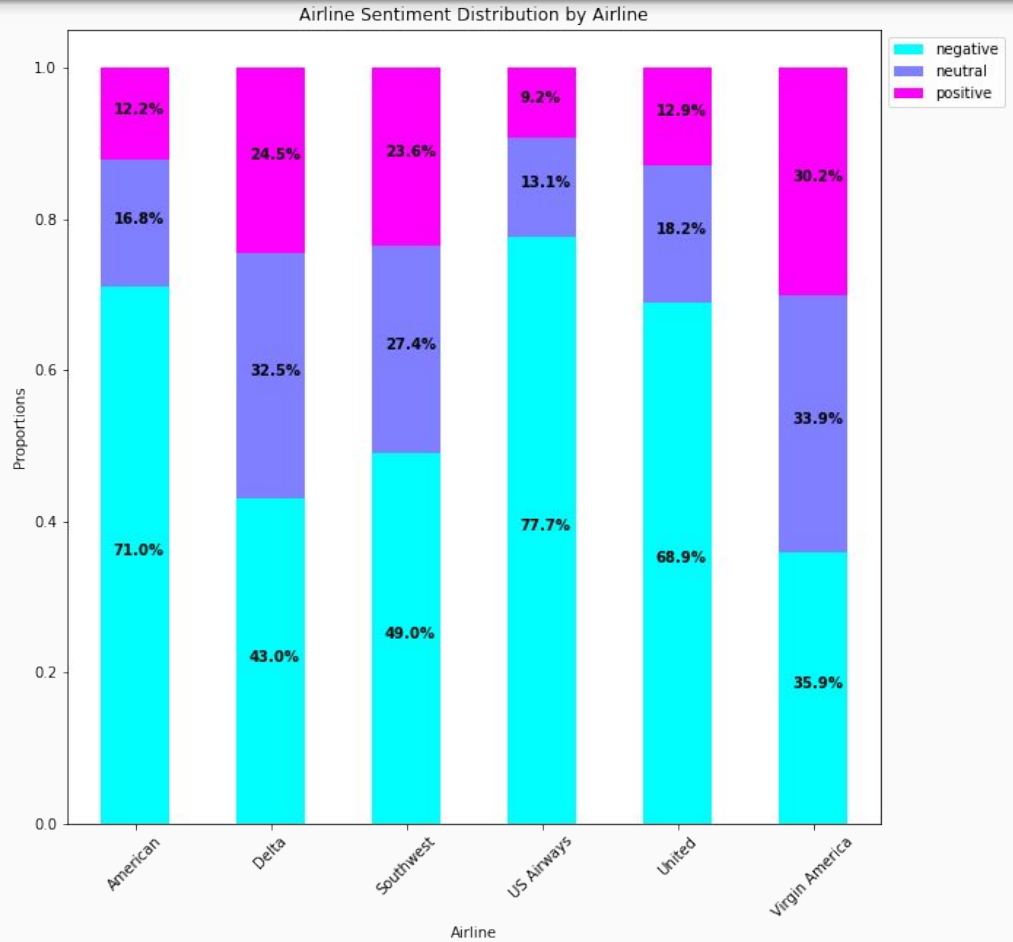


Data Exploration

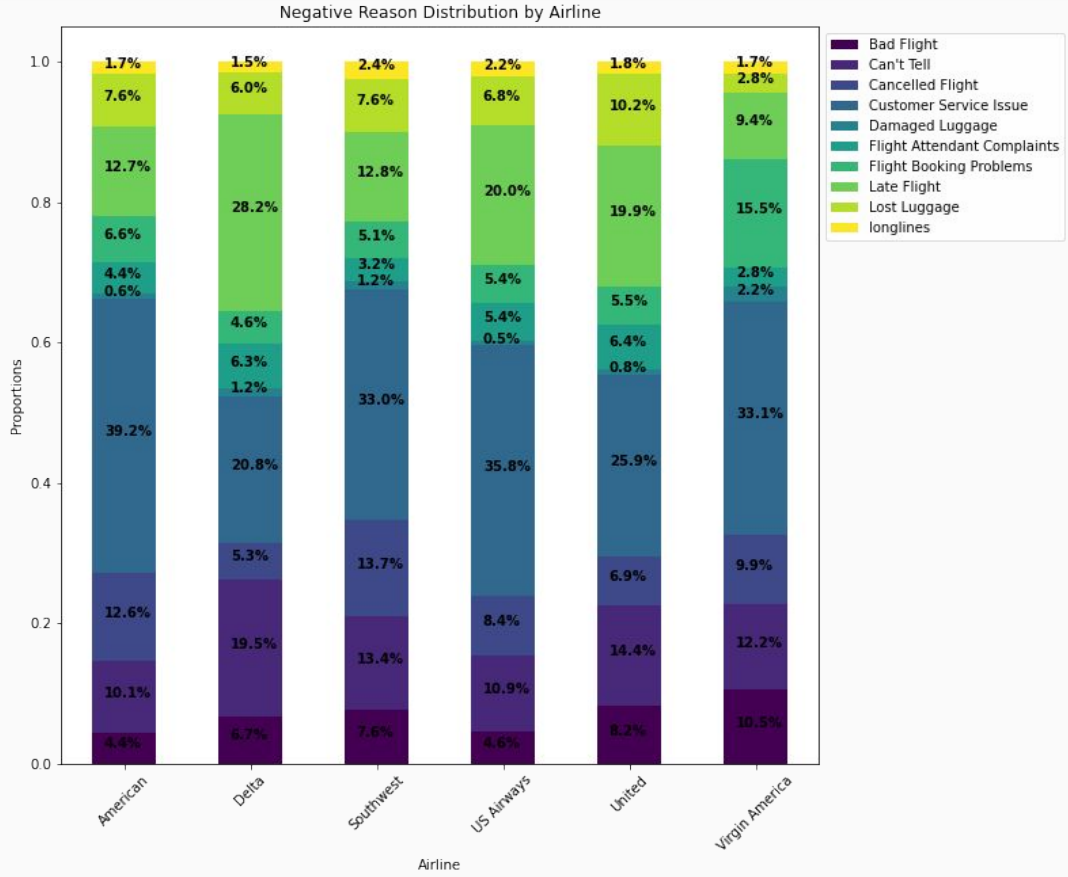


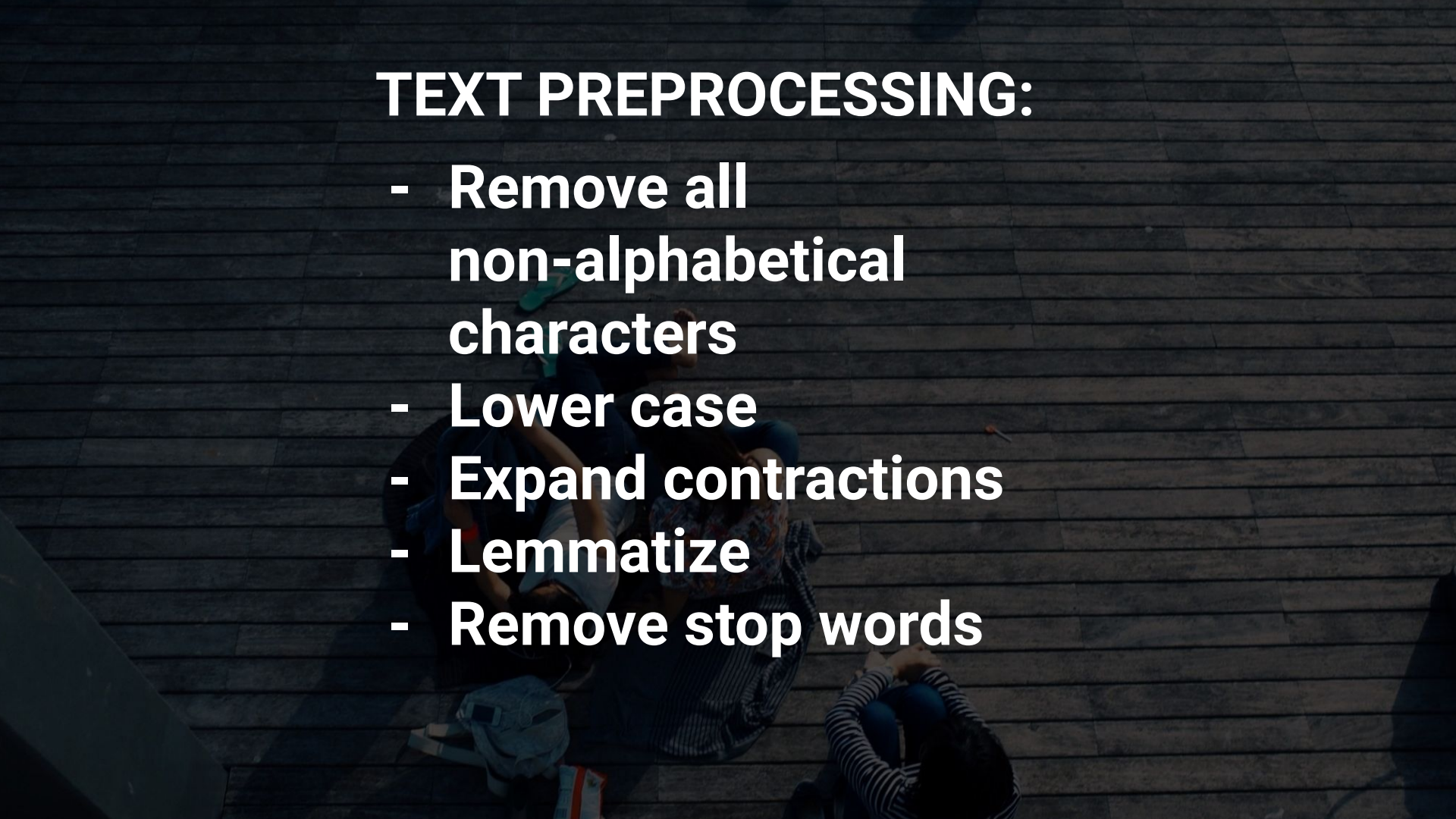

Airline Sentiment Distribution by Airline
negative
neutral
positive
12.2%
16.8%
71.0%
24.5%
32.5%
43.0%
23.6%
27.4%
49.0%
9.2%
13.1%
77.7%
12.9%
18.2%
68.9%
30.2%
33.9%
35.9%
1.0
Airline

Sentiment Distribution Per Airline



Negative Reason Distribution per Airline





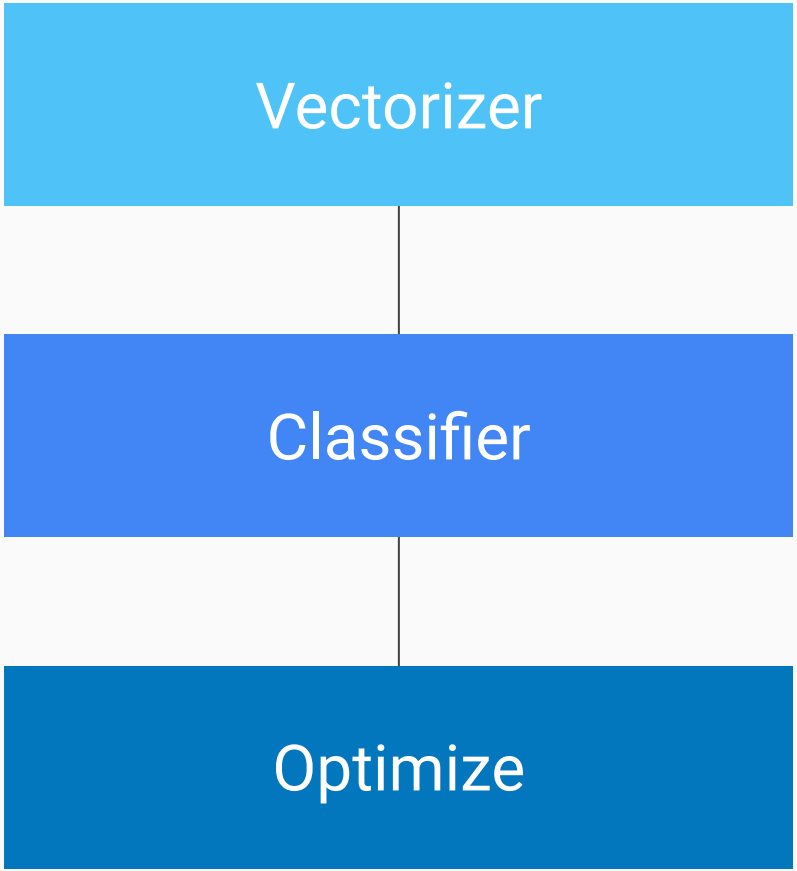
TEXT PREPROCESSING:

- Remove all non-alphabetical characters
- Lower case
- Expand contractions
- Lemmatize
- Remove stop words

Modeling

The next step was to model the data

- Binary Classification
- Bag of Words feature set
- 70/30 Train Test Split



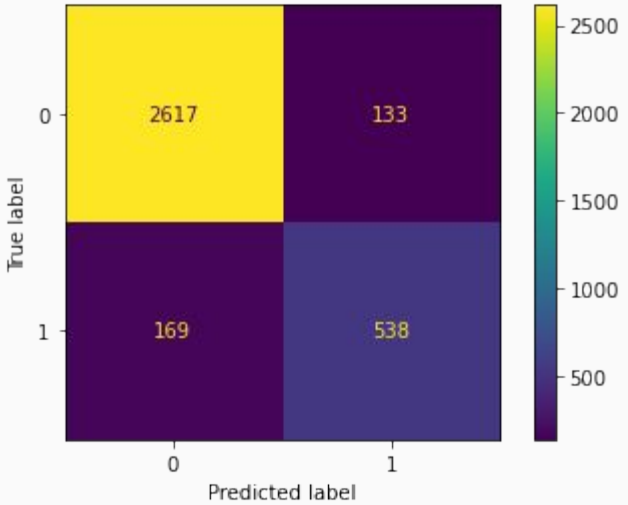
Vectorizers and Classifiers

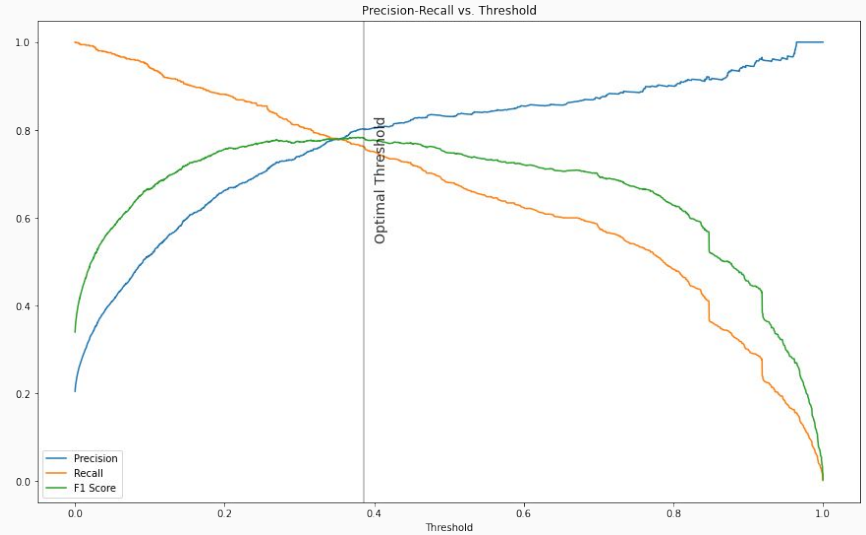
- CountVectorizer and TfidfVectorizer
- Logistic Regression, Multinomial Naive Bayes, Support Vector Machine
- GridSearch Cross Validation
- ROC-AUC as metric for Evaluation

Results

Count Vectorizer & Logistic Regression performed best

Optimal Threshold found from F1 Score





	precision	recall	f1-score	support
0	0.94	0.95	0.95	2750
1	0.80	0.76	0.78	707
accuracy			0.91	3457
macro avg	0.87	0.86	0.86	3457
weighted avg	0.91	0.91	0.91	3457

Identifying Predictive Words

Used trained model to find strength of each predictive word

Found top 10 good and bad words across tweets for all airlines

Good words	P(fresh | word)
thank	0.92
thanks	0.85
amazing	0.84
awesome	0.81
great	0.81
kudos	0.79
excellent	0.78
love	0.76
wonderful	0.76
thankful	0.76
Bad words	**P(fresh | word)**
paid	0.07
online	0.07
disappointed	0.06
hold	0.06
rude	0.06
delayed	0.05
hour	0.05
website	0.05
luggage	0.04
worst	0.02

Predictive Words for each Airline

UNITED Airlines

US Airways

American Airlines

Southwest Airlines

Delta Airlines

Virgin America

Further Goals

- More features such as Emotion, length of tweets, Capitalization, etc.
- Sarcasm Detection
- Multi-class classification to include neutral