



Twitter Sentiment Analysis

William Gast

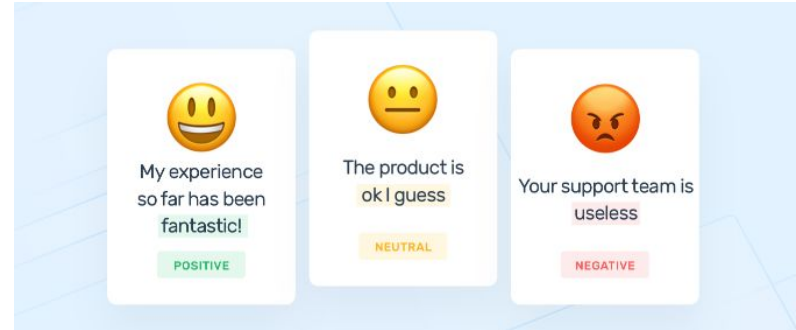
About Me

- William Gast
- Carthage College Class of 2020
 - Marketing
- Big Star Wars fan
- Hobbies include
 - Any type of sport
 - Travel



Motivation

- Sentiment analysis or opinion mining is an NLP technique that lets you determine the attitude (positive, negative, or neutral) of text.
- Given the model I am creating, an airline company could take result and understand how their customers are reacting to them and their services from their twitter data.
- This can be used to improve a company's decision making, customer satisfaction and more



Data

- Source: Kaggle
- Name : Twitter US Airline Sentiment
- Columns in use:
 - Text
 - Airline Sentiment
 - Airline

```
Data columns (total 15 columns):
tweet_id          14640 non-null int64
airline_sentiment 14640 non-null object
airline_sentiment_confidence 14640 non-null float64
negativereason    9178 non-null object
negativereason_confidence 10522 non-null float64
airline           14640 non-null object
airline_sentiment_gold 40 non-null object
name              14640 non-null object
negativereason_gold 32 non-null object
retweet_count     14640 non-null int64
text              14640 non-null object
tweet_coord       1019 non-null object
tweet_created     14640 non-null object
tweet_location    9907 non-null object
user_timezone     9820 non-null object
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

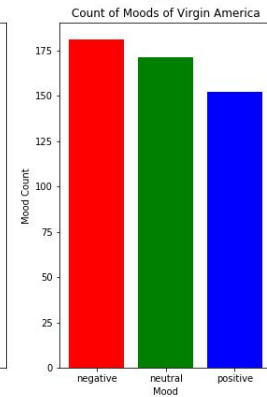
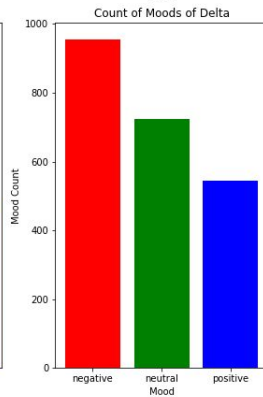
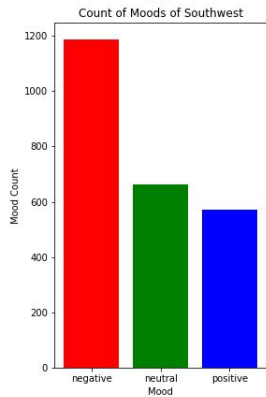
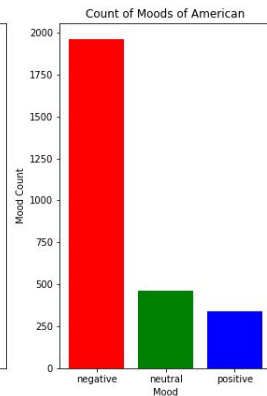
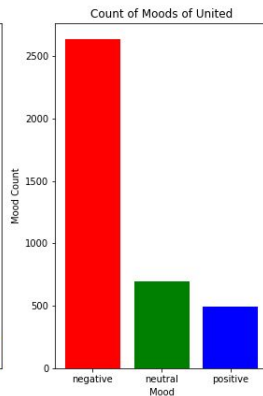
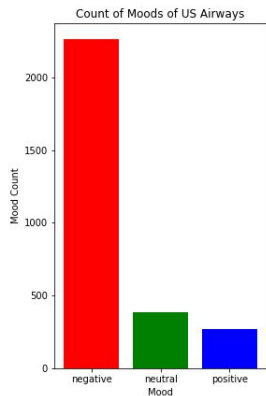
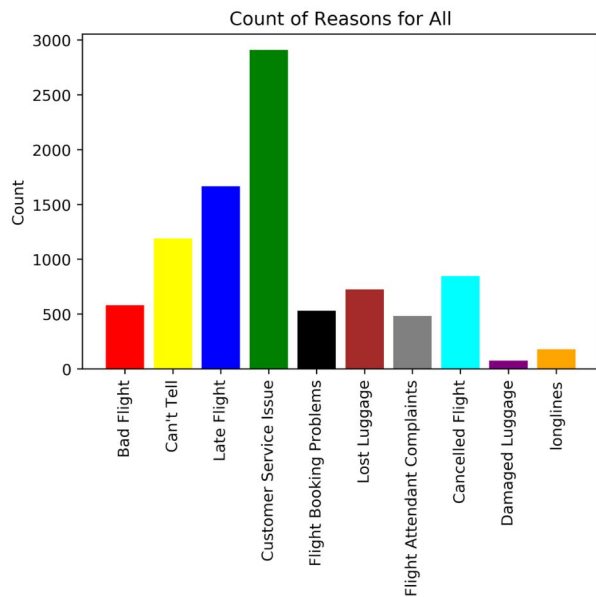
column_name	percent_missing
tweet_id	0.000000
airline_sentiment	0.000000
airline_sentiment_confidence	0.000000
negativereason	37.308743
negativereason_confidence	28.128415
airline	0.000000
airline_sentiment_gold	99.726776
name	0.000000
negativereason_gold	99.781421
retweet_count	0.000000
text	0.000000
tweet_coord	93.039617
tweet_created	0.000000
tweet_location	32.329235
user_timezone	32.923497



Cleaning and Preprocessing

1. Cleaning
 - a. Taking out the @ and # signs while also removing the RT tags
2. Tokenization
3. Lower casing
4. Stop words removal
5. Stemming
6. Lemmatization
7. Vectorizing
 - a. Convert a collection of text documents to a matrix of token counts

EDA





Modeling

- VADER sentiment
 - lexicon and rule-based sentiment analysis tool specifically tuned to be used for social media.
- Naive Bayes
 - The intuition behind Naive Bayes is to find the probability of classes assigned to given text by using the joint probabilities of words and classes.
- Random Forest
 - Ability to handle large data sets with higher dimensionality.
 - with more trees, it won't allow-overfitting trees in a model.



Choosing an evaluation metric

- I used the F1-score for my evaluation metric
- Gives a good measure of the incorrectly classified cases.
- F-1 score is used when the False Negatives and False Positives are the most important.
- Does a good job of combating class imbalance



Results

	Precision	Recall	f1-score
VADER	.70	.55	.58
Textblob	.67	.46	.48
Naive Bayes	.84	.77	.79
Random Forest	.77	.75	.76



Questions?