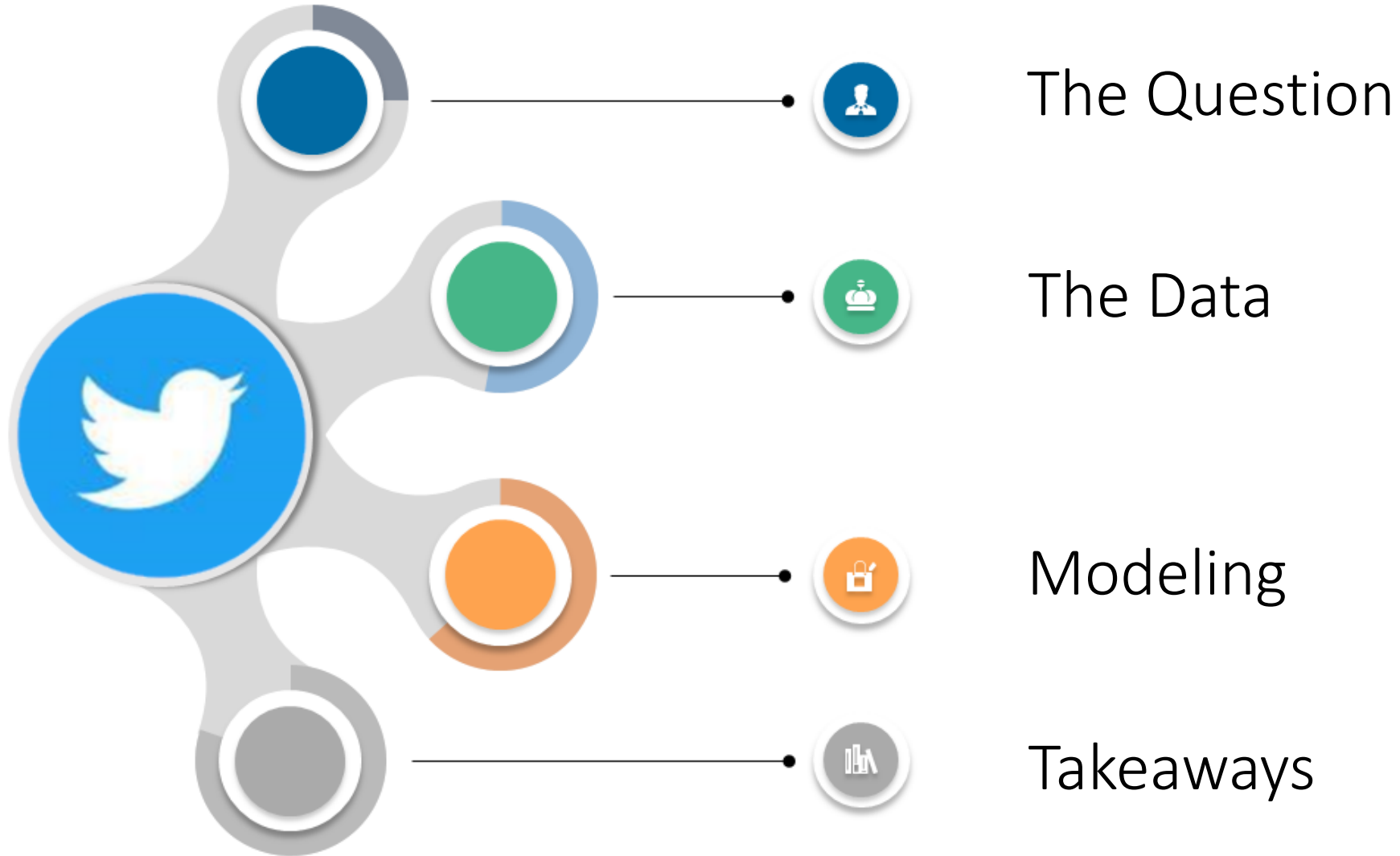




Final Capstone: Sentiment Analysis – Airline Tweets

Will Becerra

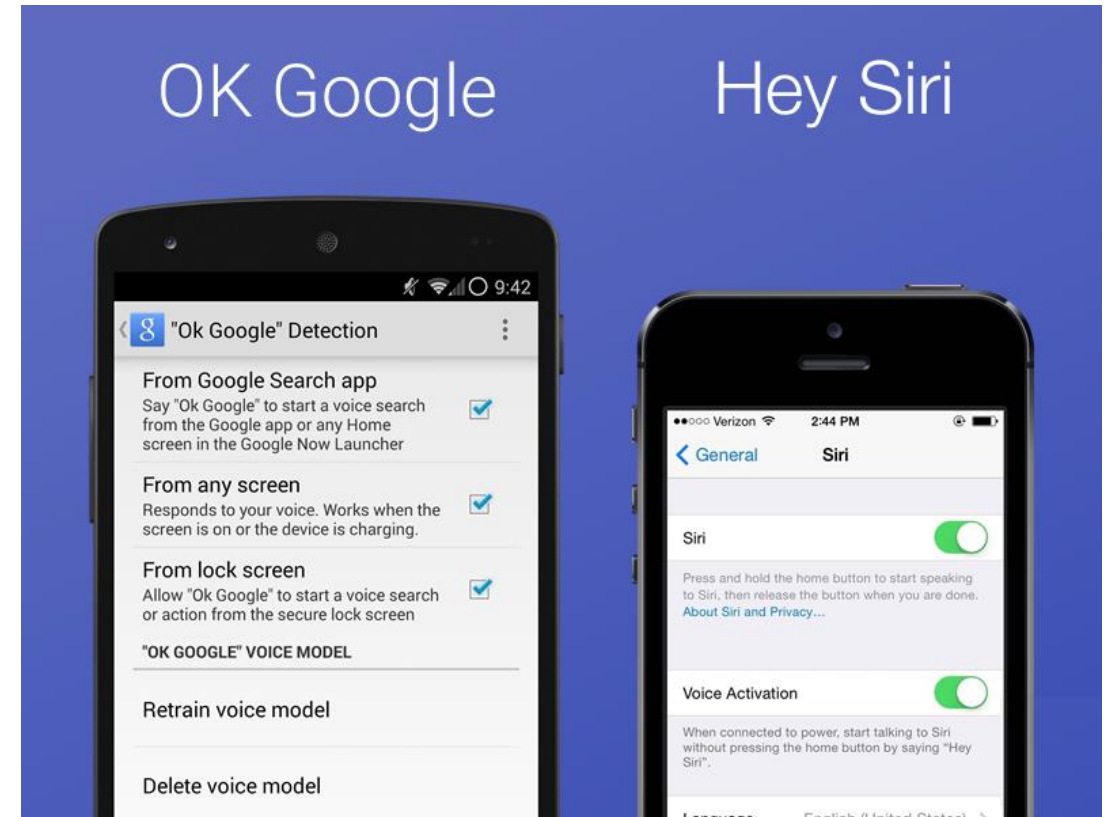
Agenda



Natural Language Processing (NLP)

NLP is the ability of a computer program to understand human language as it is spoken.

NLP is a component of artificial intelligence (AI).



What is Sentiment Analysis?

Definition: Type of data mining that measures the inclination of people's opinions through natural language processing (NLP), computational linguistics and text analysis

Companies can track:

- Brand reception and popularity
- New product perception and anticipation
- Company reputation
- Flame/rant detection

Data Set Characteristics

Source

- Airline Sentiment Analysis
- Figure Eight, Inc (Google Dataset Search)
- Twitter scraped February, 2015

Structure

- 14,640 tweets
- 16 features

Goal: Predict Classification

- Sentiment (target variable) : negative, neutral, positive

Features

- tweet id
- **airline sentiment**
- airline sentiment confidence
- negative reason
- negative reason confidence
- airline
- airline sentiment gold
- name
- negative reason gold
- retweet count
- text
- tweet coordinate
- tweet created
- tweet location
- user time zone

Preprocessing

Cleaning

- Use regular expressions to remove:
http, &, @, most punctuation
- Remove stop words (NLTK)
- Parse documents into words (tokens)

Feature Engineering

- Change categorical values to numeric
- Lowercase tweets
- Create new features based on date
- Demojization


```
print(emoji.demojize('Python is 👍'))
```

```
Python is :thumbs_up:
```

Exploratory Data Analysis



Top 5 Retweeted Tweets

@USAirways 5 hr flight delay and a delay when we land . Is that even real life ? Get me off this plane , I wanna go home  (3 heel clicks) – 44

@USAirways of course never again tho . Thanks for tweetin ur concern but not Doin anythin to fix what happened. I'll choose wiser next time – 32

STOP. USING.THIS.WORD. IF. YOU'RE. A. COMPANY. RT @JetBlue: Our fleet's on fleek. <http://t.co/Fd2TNYcTrB> - 31

@USAirways with this livery back in the day. <http://t.co/EEqWVAMmiy> - 28

@SouthwestAir beautiful day in Seattle! <http://t.co/iqu0PPVq2S> - 22

Top 10 #Hashtags and @Mentions

Top 10 hashtags:

destinationdragons - 79

fail - 69

jetblue - 47

unitedairlines - 45

customerservice - 36

usairways - 30

neveragain - 27

americanairlines - 26

usairwaysfail - 26

united - 25

Top 10 mentions:

united - 3827

usairways - 2972

americanair - 2913

southwestair - 2426

jetblue - 2092

virginamerica - 506

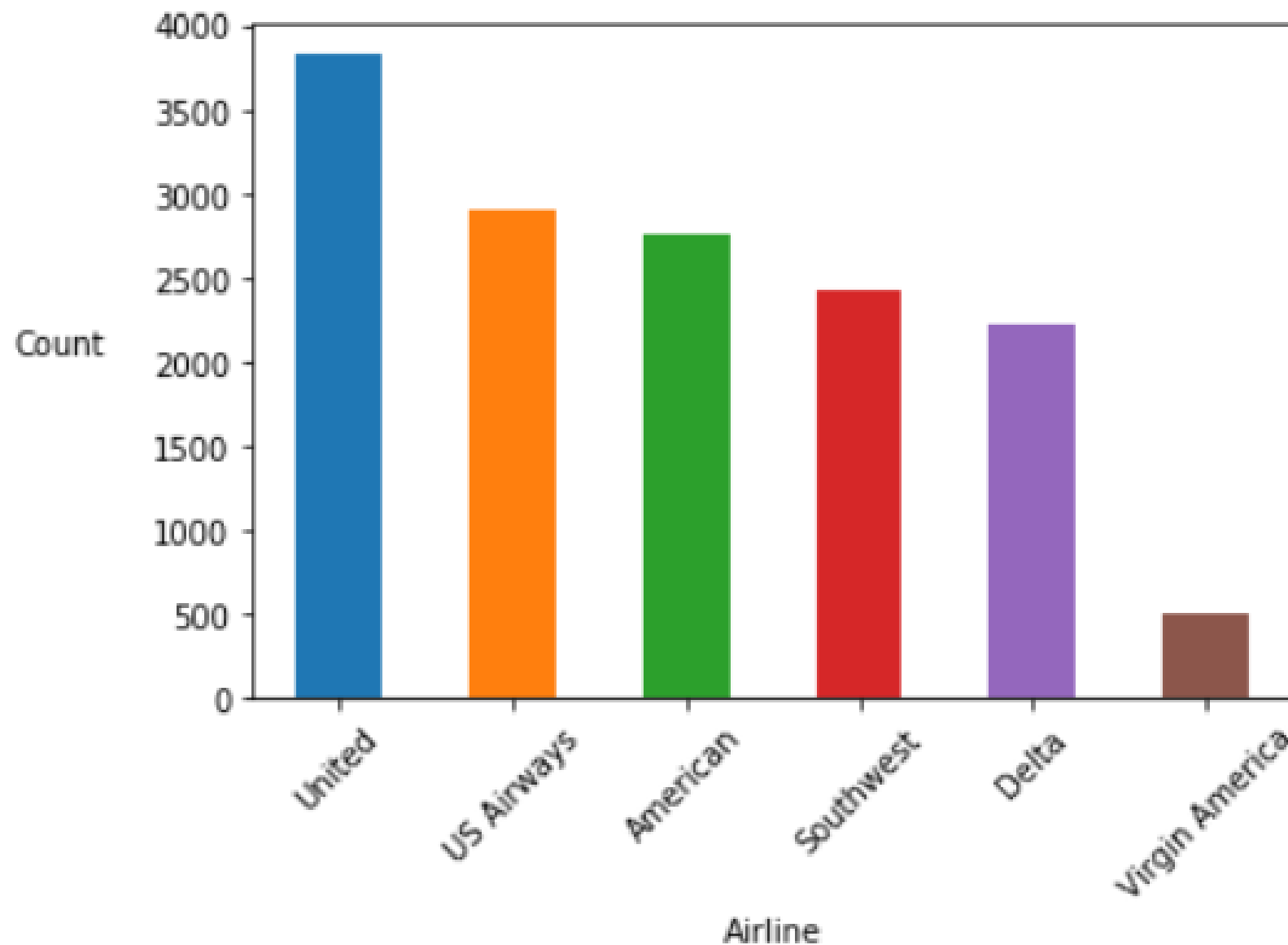
delta - 68

imaginedragons - 45

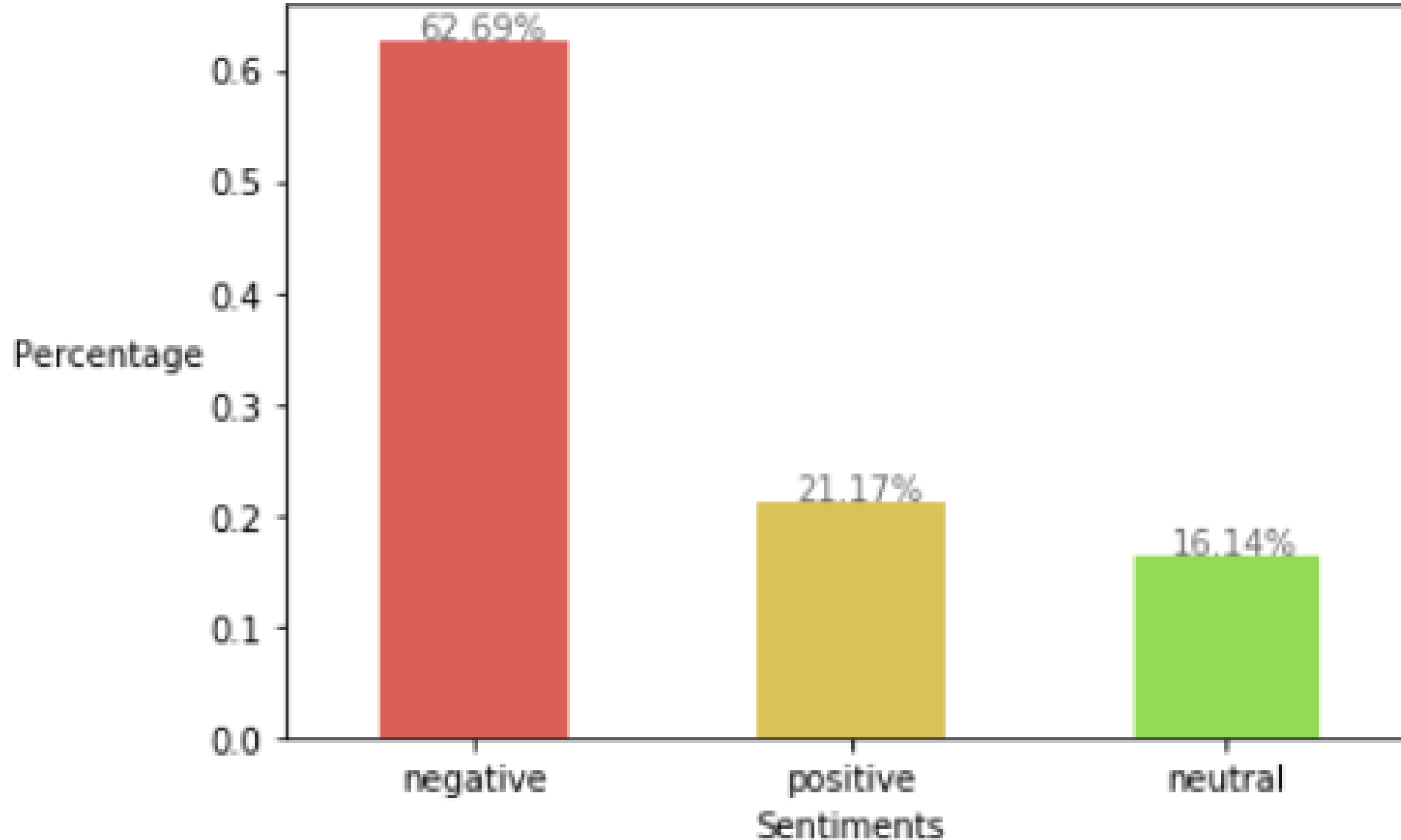
jetblues - 27

phlairport - 20

Tweets Per Airline

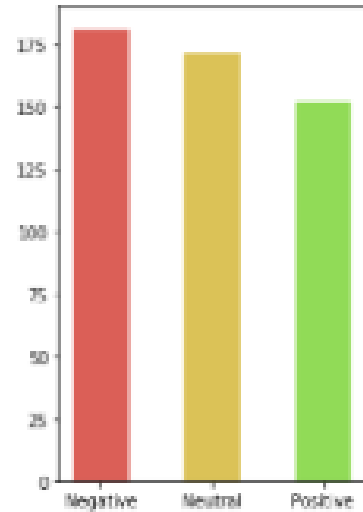


Total Sentiment Distribution

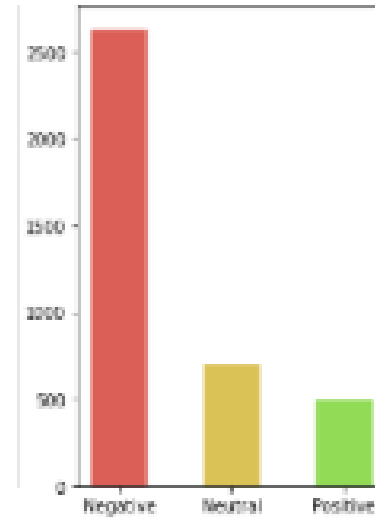


Sentiment by Airline

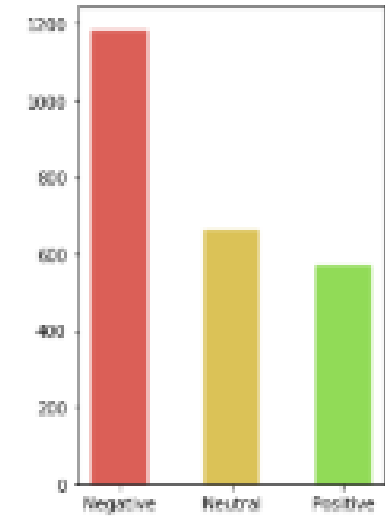
Virgin America



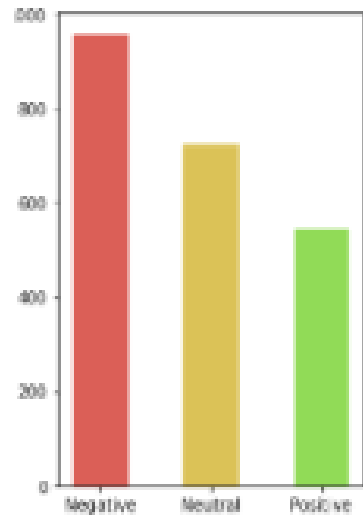
United



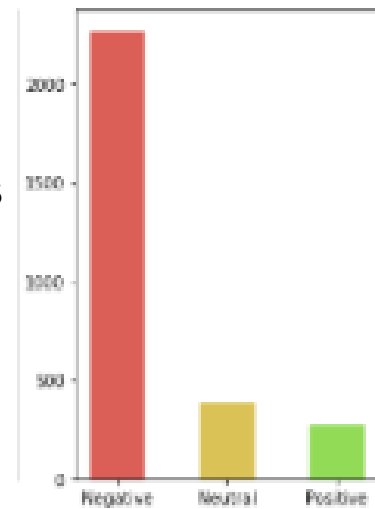
Southwest



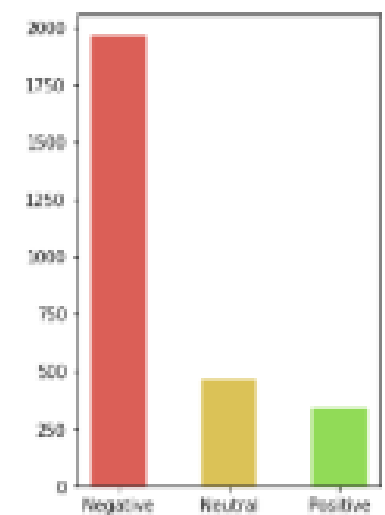
Delta



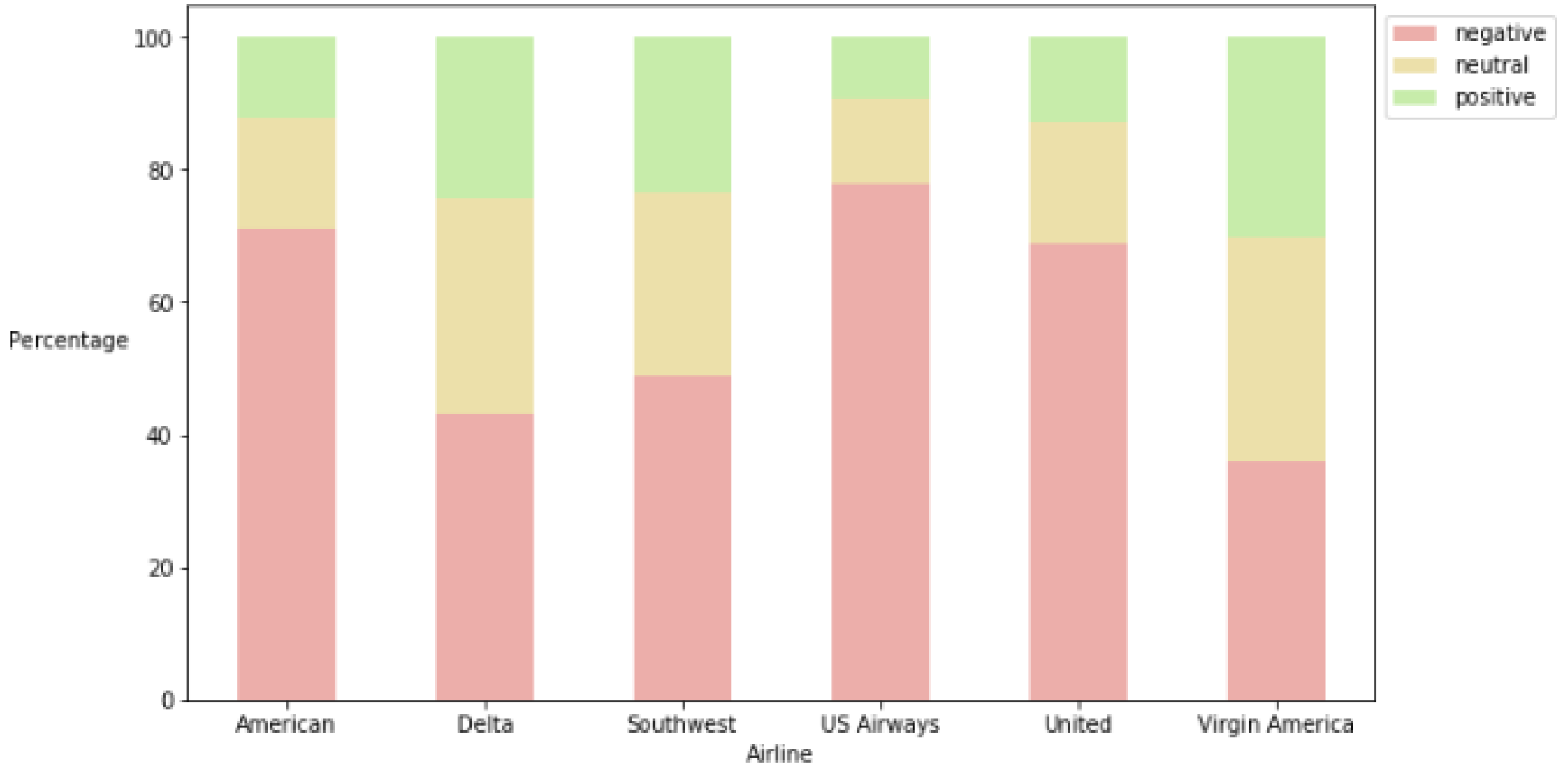
US Airways



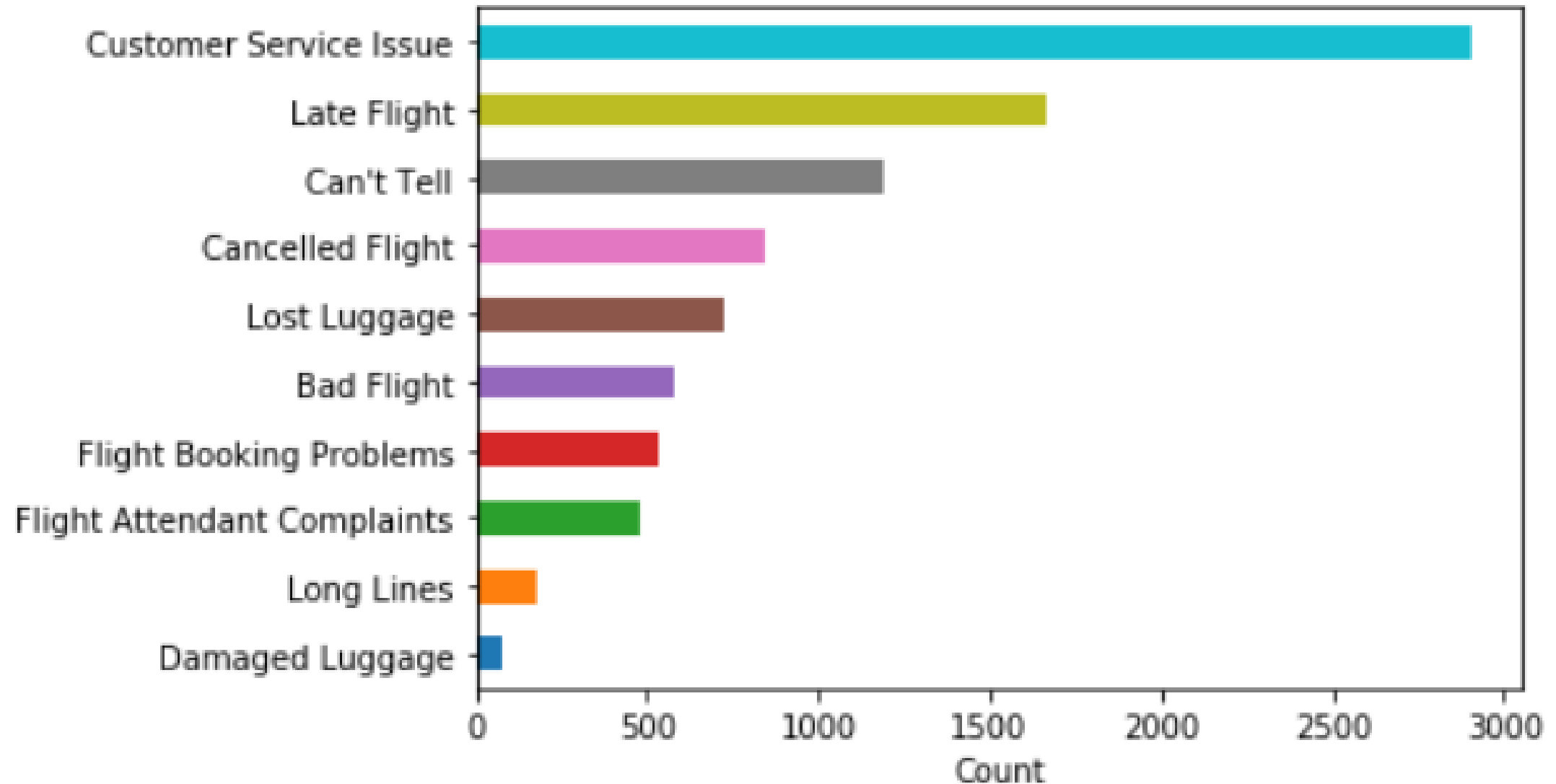
American



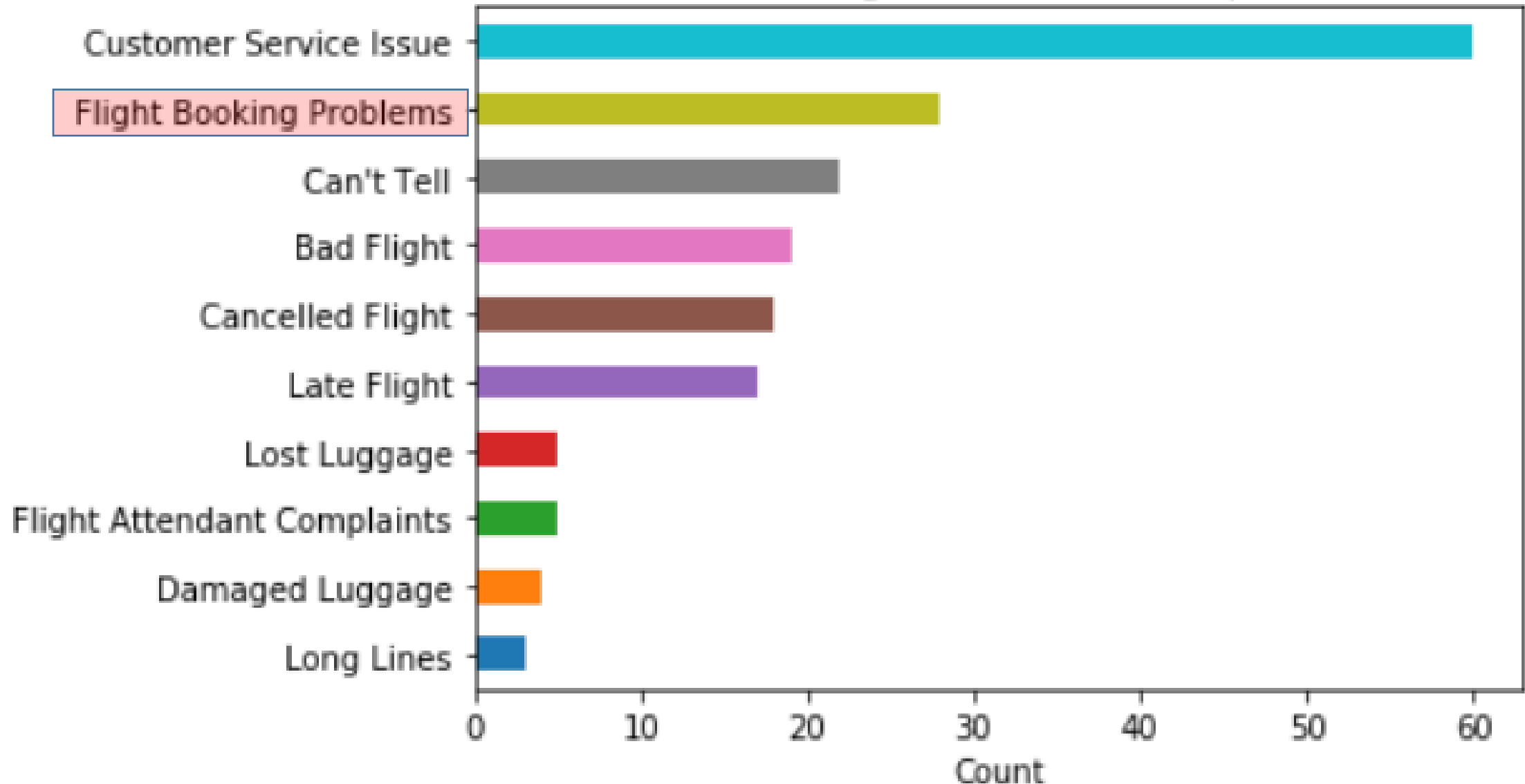
Sentiment by Airline - Ratio



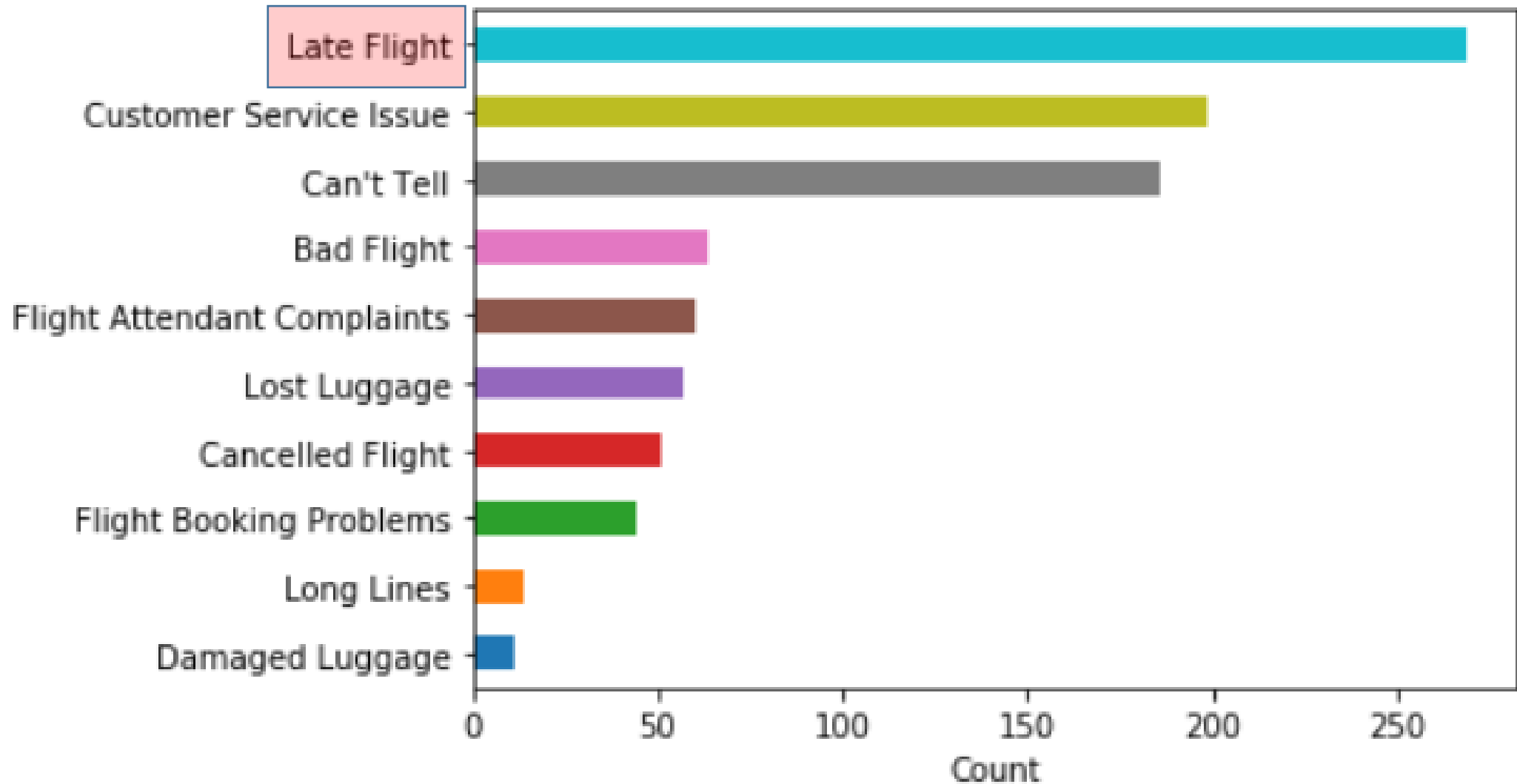
Categories of Negative Reasons - Overall



Categories of Negative Reasons – Virgin Airways



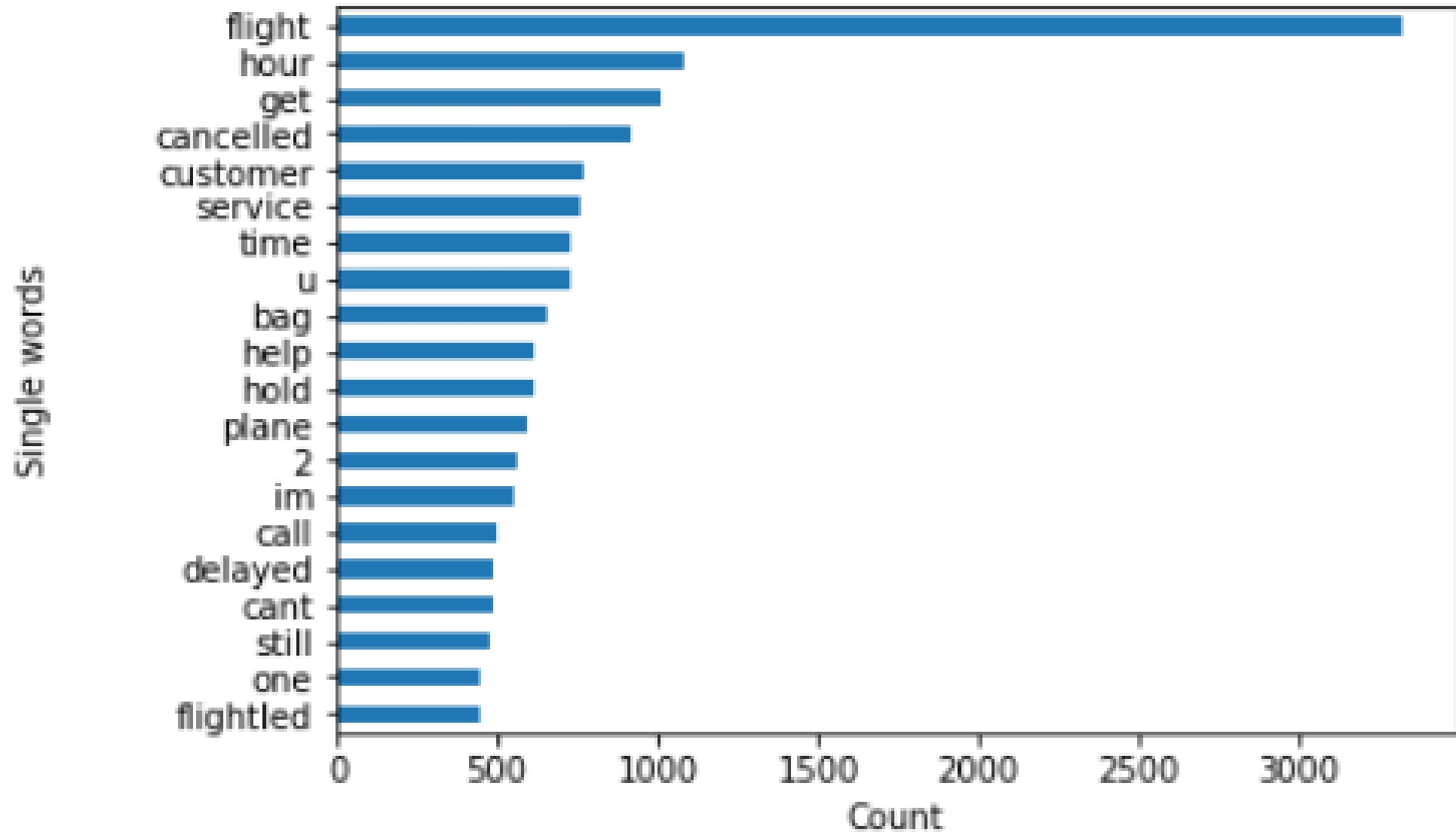
Categories of Negative Reasons - Delta



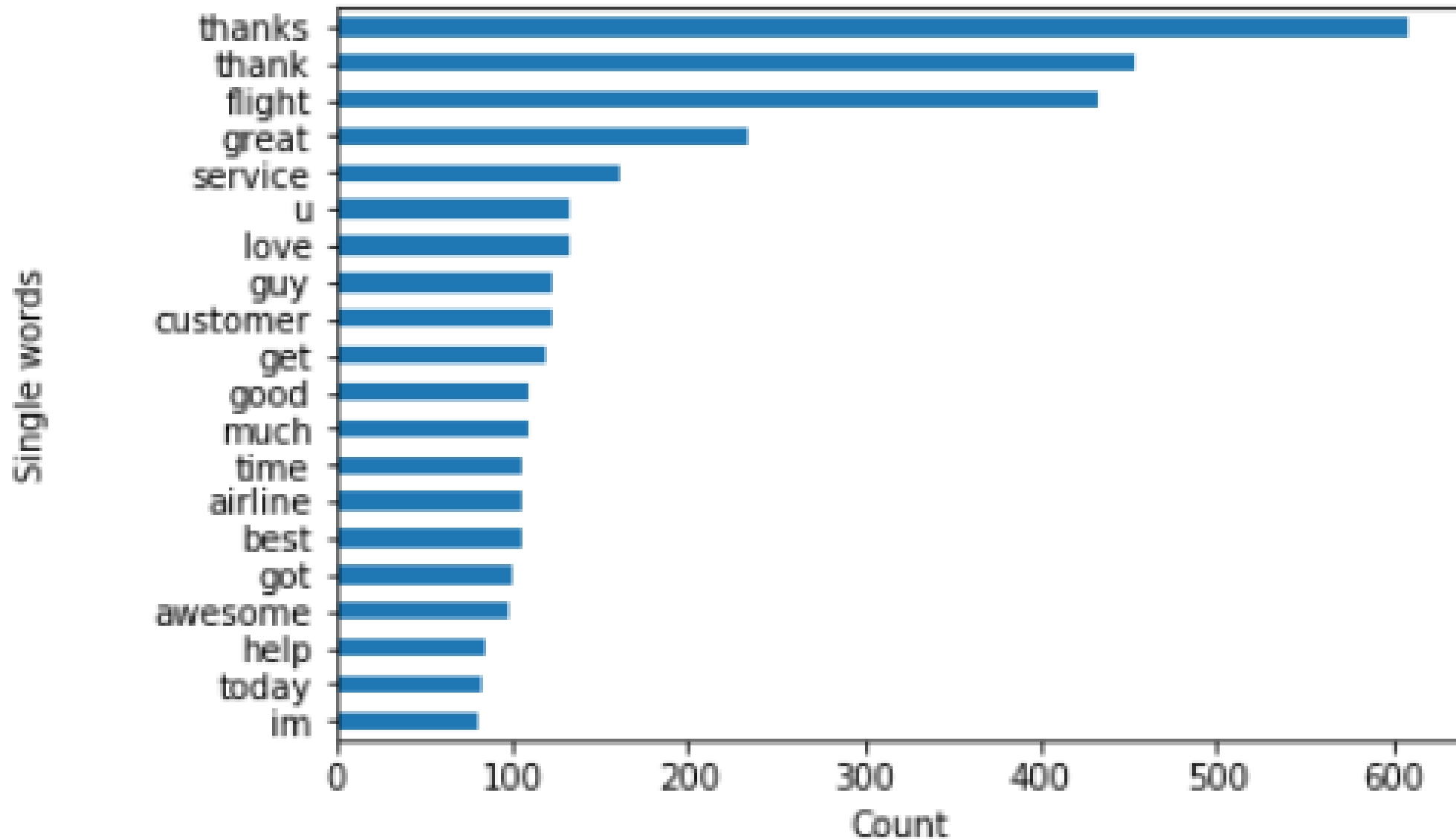
Top Negative Reasons – Word Cloud



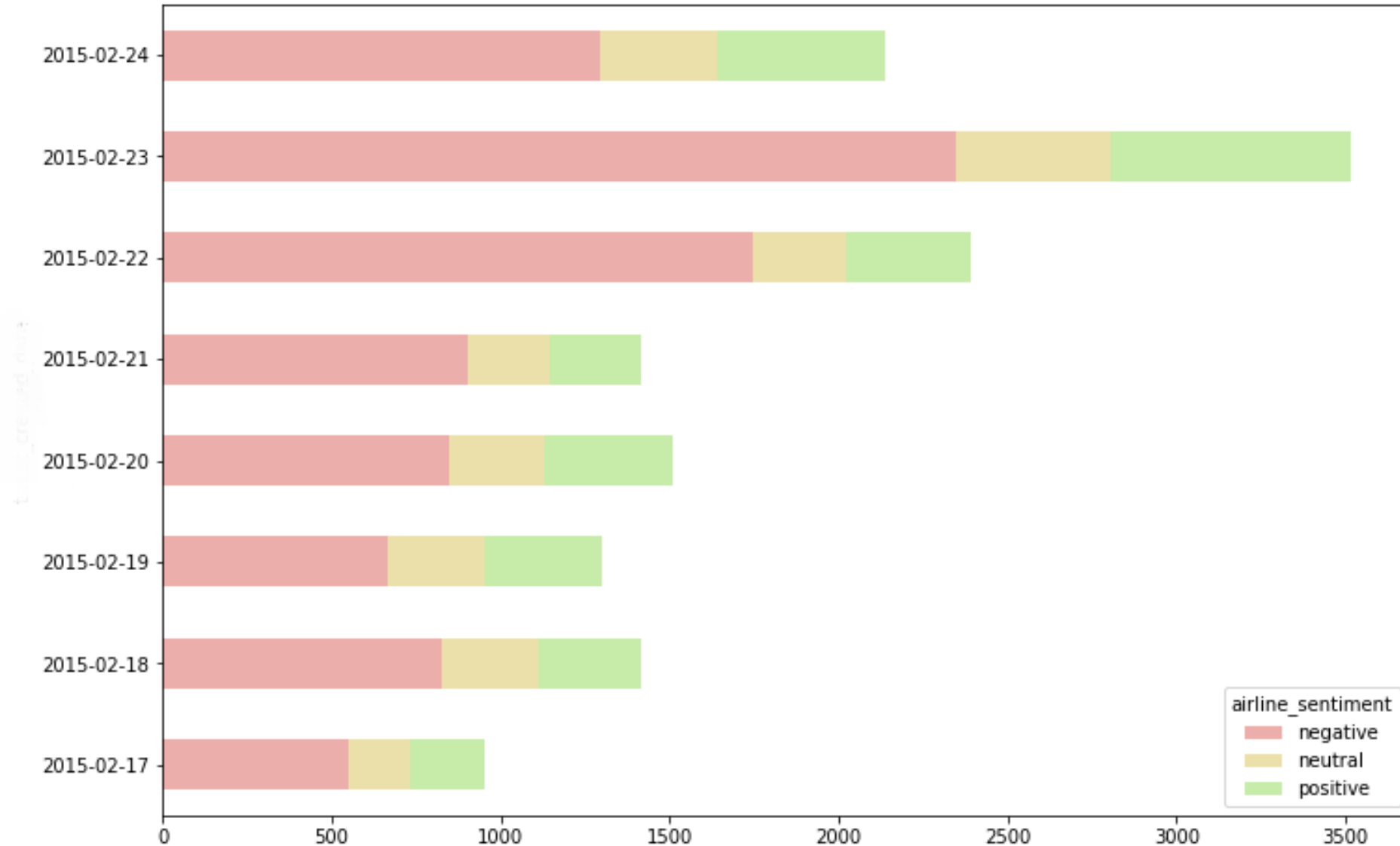
Top Negative Reasons – Single Words



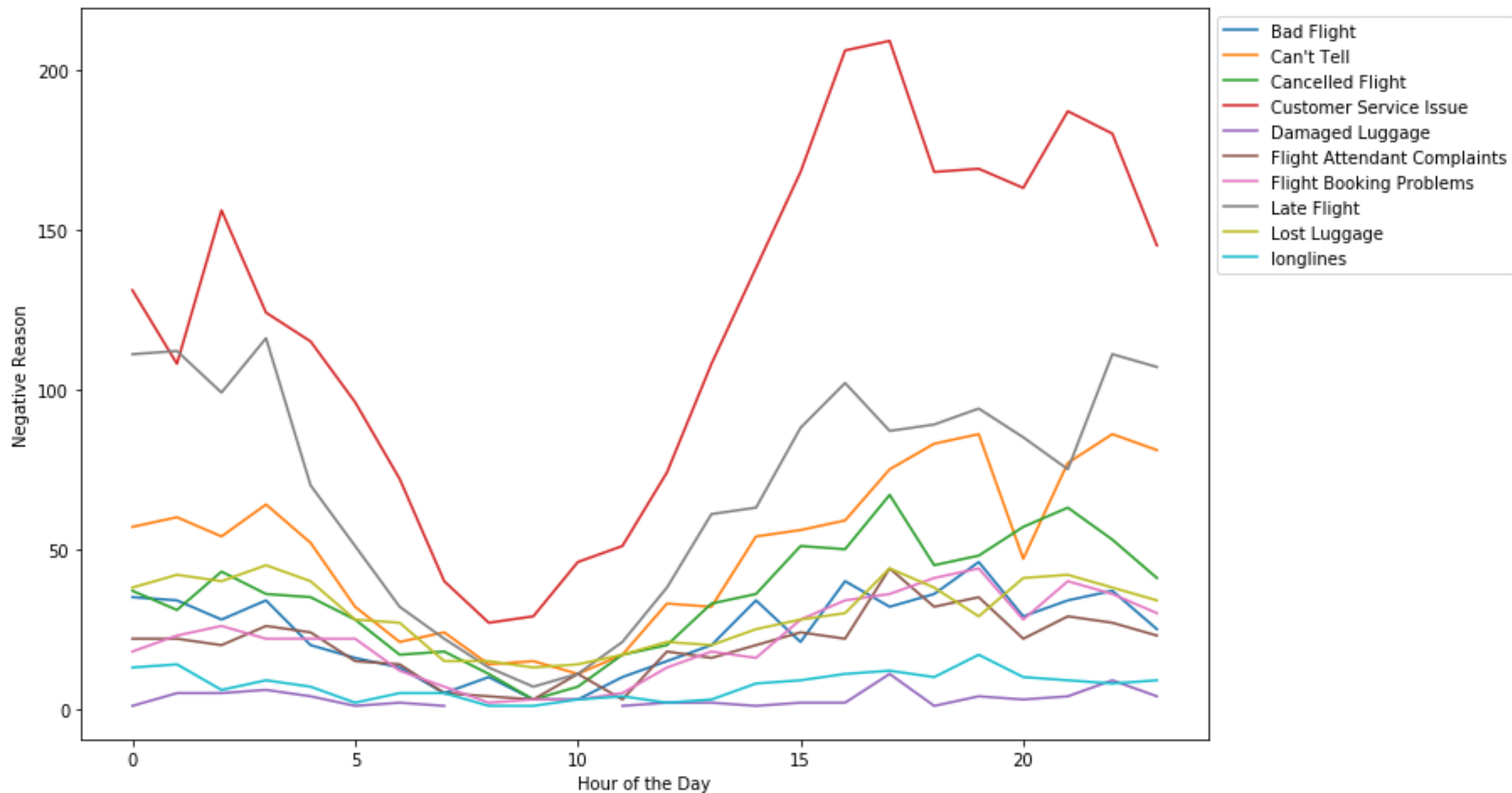
Top Positive Reasons – Single Words



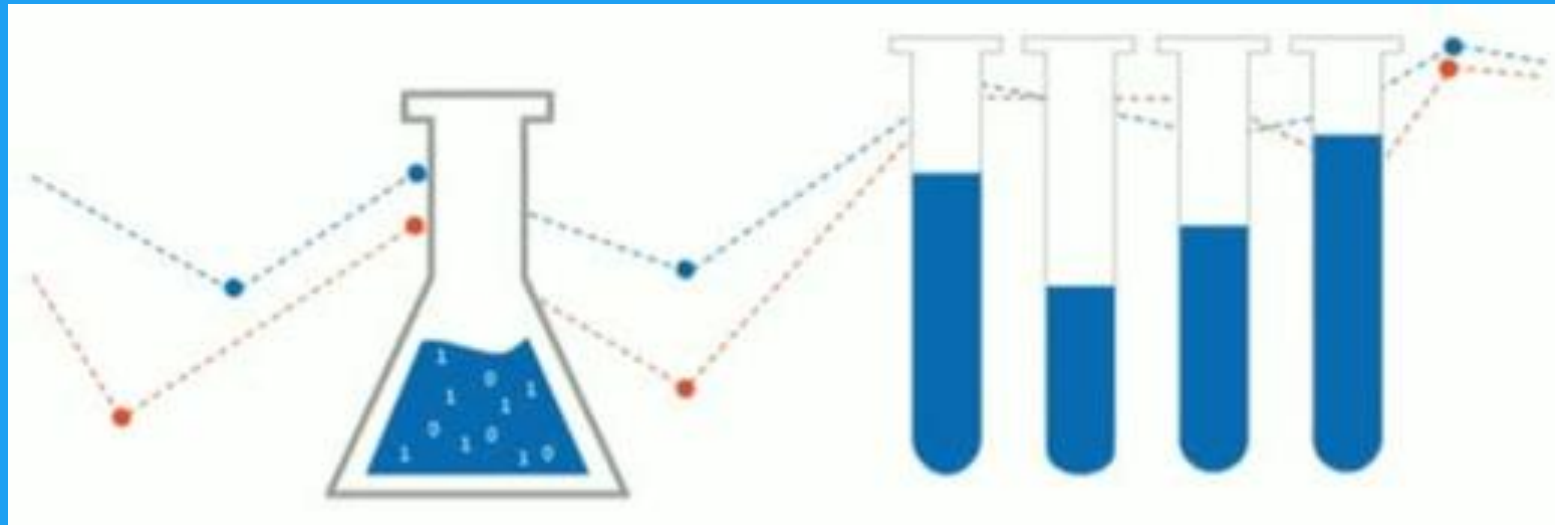
Sentiment by Date



Negative Reasons by Time of Day



Data Preparation



Tokenization

Definition: the process of splitting the text into smaller parts called tokens.

Tokens can be:

- Words
- Phrases
- Symbols
- other meaningful elements

Modeling Approaches

Feature Engineering

- Lemmatization
- Stemming
- Bigrams
- Lemmatization + SMOTE
- Lemmatization + Demojization
- Lemmatization + Feature Weight

Dimension Reduction - Lemmatization

Lemma definition: *morphological root of a word*

Consider the words "think", "thought", and "thinking". They're related.

They all share the same root word: the verb "think".

	text	normalized_tweet
0	@VirginAmerica What @dhepburn said.	[said]
1	@VirginAmerica plus you've added commercials to the experience... tacky.	[plus, youve, added, commercial, experience, tacky]
2	@VirginAmerica I didn't today... Must mean I need to take another trip!	[didnt, today, must, mean, need, take, another, trip]
3	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse	[really, aggressive, blast, obnoxious, entertainment, guest, face, little, recourse]
4	@VirginAmerica and it's a really big bad thing about it	[really, big, bad, thing]

Dimension Reduction - Stemming

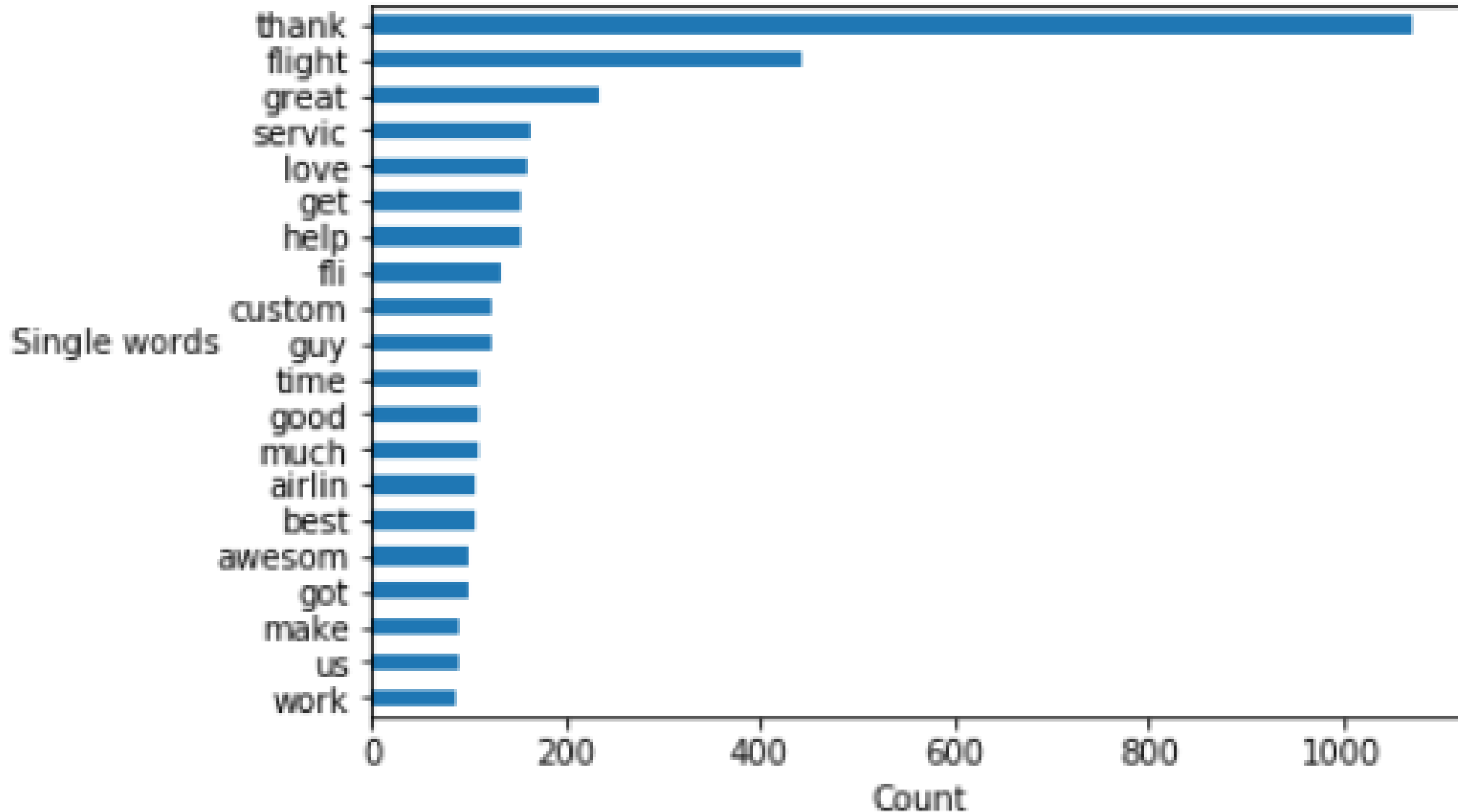
Stem definition: *reduction of variant form of word to common form*

The words “connection”, “connections”, and “connected” are related.

They all share the same stem: “connect”.

text	normalized_tweet_stem
@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but great things about Virgin America. #29DaysToGo	[excit, first, cross, countri, flight, lax, mco, ive, heard, noth, great, thing, virgin, america, 29daystogo]
@VirginAmerica I flew from NYC to SFO last week and couldn't fully sit in my seat due to two large gentleman on either side of me. HELP!	[flew, nyc, sfo, last, week, couldnt, fulli, sit, seat, due, two, larg, gentleman, either, side, help]
I ❤️ flying @VirginAmerica. 🍷👍	[fli]
@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!! I want to fly with only you.	[know, would, amaz, awesom, bosfill, pleas, want, fli]

Top Positive Reasons – Single Words (Stems)



Bigrams

Definition: A **bigram** is a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words.

Example - *'The quick brown fox jumps.'*

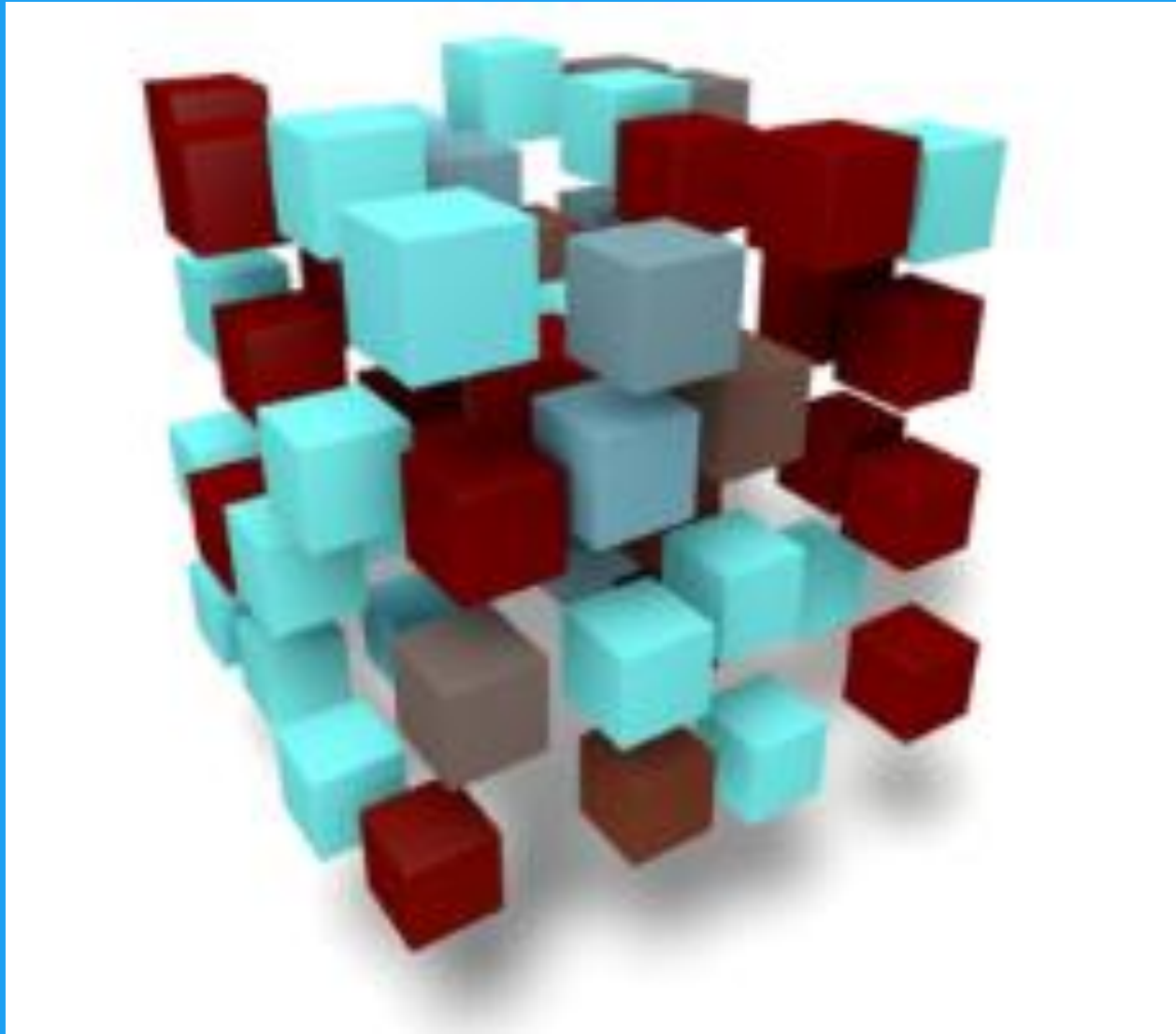
Bigrams:

- The quick
- quick brown
- brown fox
- fox jumps

Demojized Tweets

text	normalized_tweet_stem_emoji
I ❤️ flying @VirginAmerica. ☺️👍	[redheart, fli, smilingfacethumbsup]
@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!! I want to fly with only you.	[know, would, amaz, awesom, bosfll, pleas, want, fli]
@VirginAmerica why are your first fares in May over three times more than other carriers when all seats are available to select???	[first, fare, may, three, time, carrier, seat, avail, select]
@VirginAmerica I love this graphic. http://t.co/UT5GrRwAaA	[love, graphic]
@VirginAmerica I love the hipster innovation. You are a feel good brand.	[love, hipster, innov, feel, good, brand]
@VirginAmerica will you be making BOS>LAS non stop permanently anytime soon?	[make, bosgtla, non, stop, perman, anytim, soon]
@VirginAmerica you guys messed up my seating.. I reserved seating with my friends and you guys gave my seat away ... 😡 I want free internet	[guy, mess, seat, reserv, seat, friend, guy, gave, seat, away, poutingfac, want, free, internet]

Modeling



Feature Extraction: Vectorization of words (TF-IDF)

Term frequency - how many times a word appears in a tweet

Document frequency - counts how many tweets word appears in

Inverse document frequency $idf_t = \log \frac{N}{df_t}$

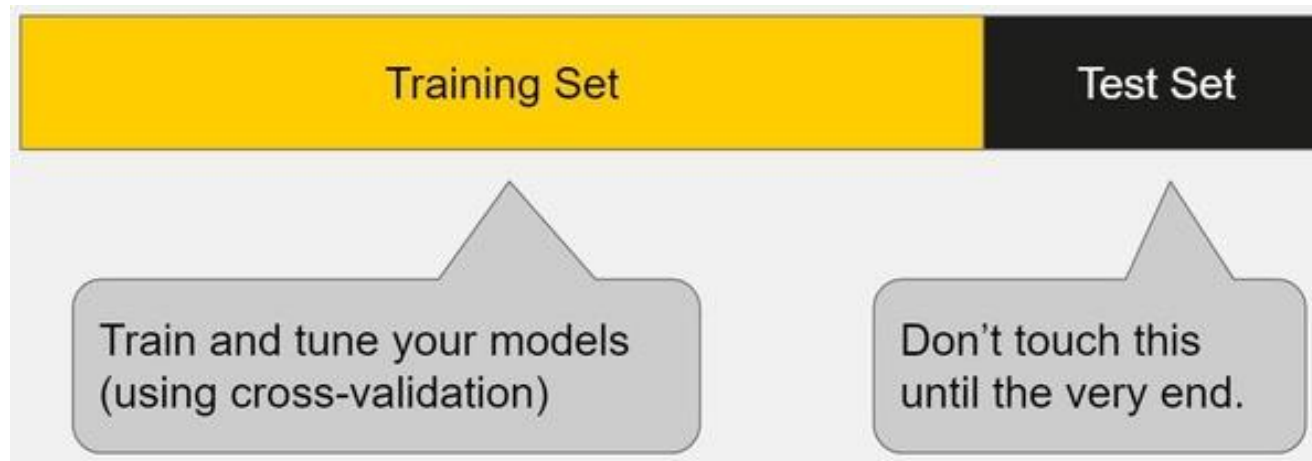
Term frequency weights

1. "The funniest Monty Python sketch is the one about the dead parrot, I laughed so hard."
2. "I laugh when I think about Python's Ministry of Silly Walks sketch, it is funny, funny, funny, the best!"

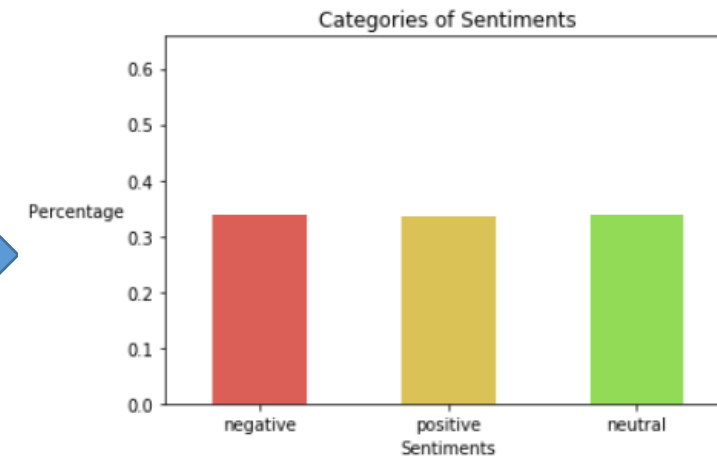
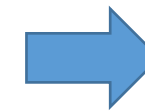
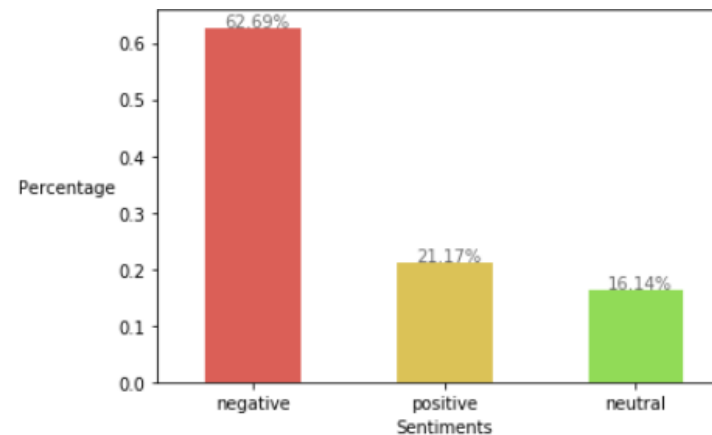
$$tf - idf_{t,d} = (tf_{t,d})(idf_t)$$

Split & SMOTE

Divided data set train / test - 80% / 20%



Synthetic Minority
Oversampling Technique
(SMOTE)



Models

- Naïve Bayes
- Logistic Regression
- Random Forest
- Support Vector Machine
- Gradient Boosting

Evaluation Criteria

- Grid search to select best hyperparameters
- Cross validation
- Accuracy, precision, recall, F1 score



Model Evaluation

		Training	Test	<i>difference</i>	Cross val	+ / -
	Model	Accuracy	Accuracy	<i>Training-Test</i>	mean	%
Lemmas	Naive Bayes	0.737	0.69	0.047	0.676	0.451
	Logistic Regression	0.947	0.775	0.172	0.775	0.455
	Random Forest	0.993	0.758	0.235	0.753	0.532
	SVM	0.901	0.78	0.121	0.773	0.593
	Gradient Boosting	0.762	0.732	0.03	0.723	0.46

Model	Sentiment	Precision	Recall	F1 score
SVM	negative	0.81	0.92	0.86
	neutral	0.78	0.62	0.69
	positive	0.64	0.47	0.54

Lemmas model scored highest.

Support Vector Machine
scored high and scored best on
precision and recall.



Takeaways

- Set threshold triggers for negative keywords in tweets
- Focus \$ in improving customer service could improve brand good will.
- NLP gives insights on competitors strengths and weaknesses.
- Changes over time can show where competitors are strategically investing.

Future Analysis

- Explore sentiment geographically
- Use Emoji package to separate emojis before modeling
- Use pre-trained vectors (trained on large data sets)
- Find bigger data set or scrape more tweets
 - word2vec
- Scrape more tweets
 - Twitter API – real time analysis (in process)

