Phase 4 Project
Leah Pope

NLP for Classifying Consumer Tweets

81720-ONL-DS-FT
12/11/2020

Hello!

I’m Leah Pope

I’m a Flatiron Data Science Student and Software Engineer

https://github.com/lspope/

“How are consumers feeling (Tweeting) about our brands and events?

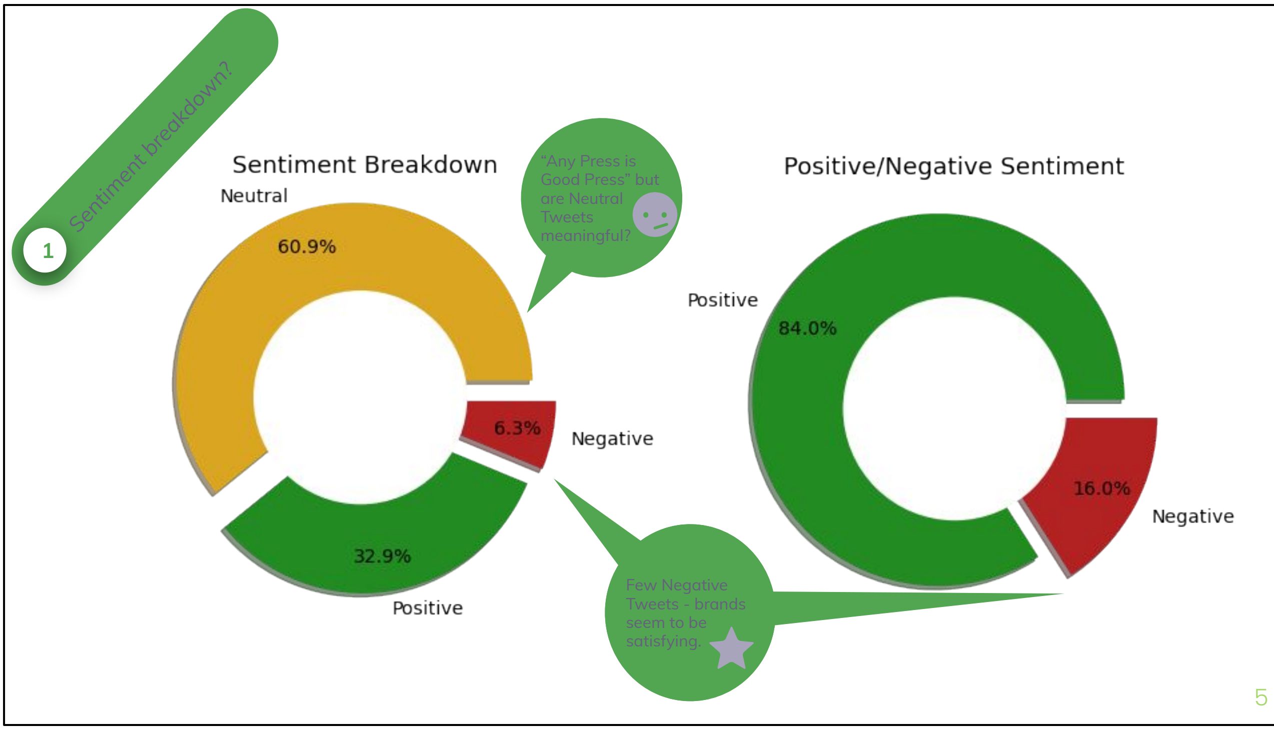
– Tech Companies & Conferences

Let’s Find Out

The Stakeholders for my project are marketing professionals in either company who are interested in learning consumer sentiment. It appears that these tweets were gathered during a session of the South by Southwest film, culture, music, and technology conference. This consumer sentiment would be of interest to the conference marketing professionals and vendor organizers.

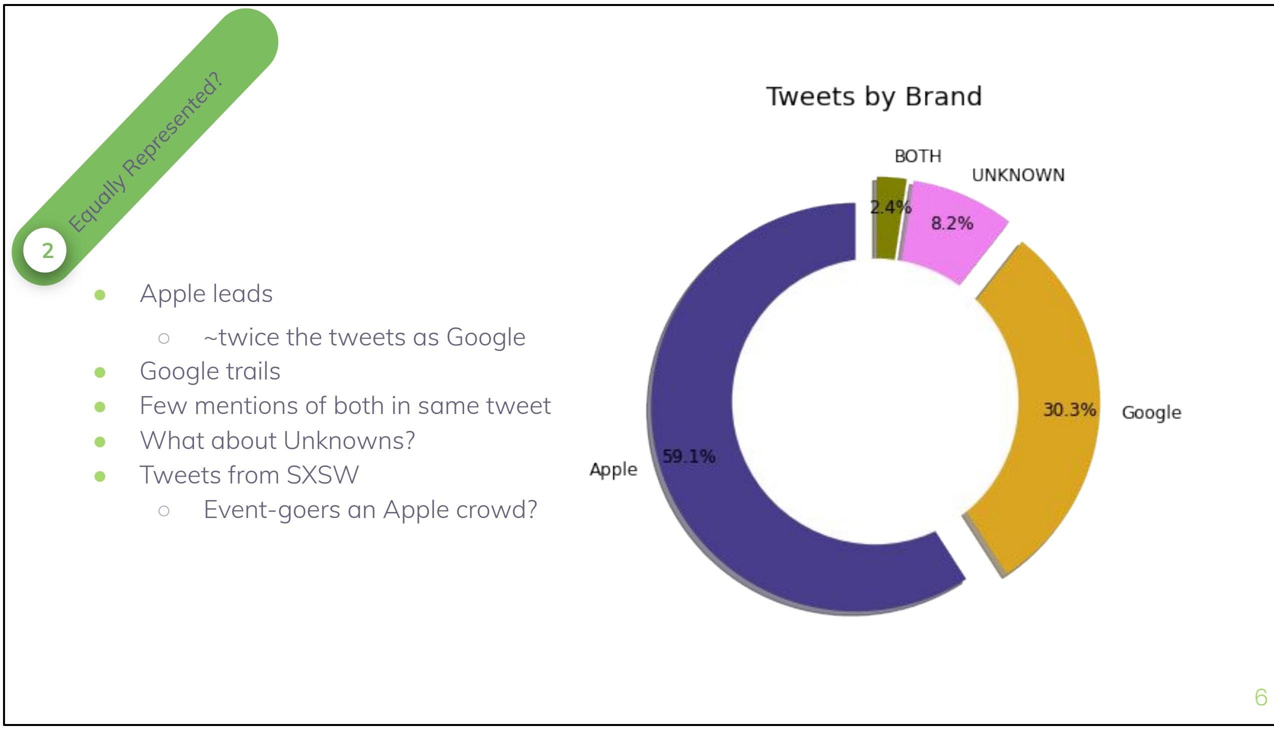
Business Questions

1. Sentiment breakdown?
2. Equally represented?
3. Brand insights?



Overall, the majority of the tweets are neutral. For Pos/Neg, the large majority are Positive.

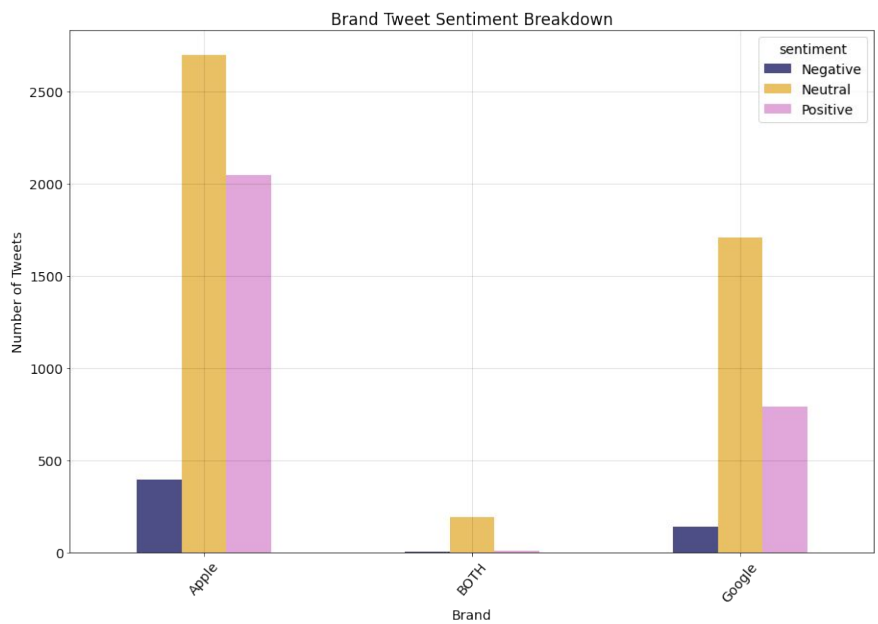
This also impacts my text classification: I am looking to create a binary classifier using Positive/Negative sentiment and a Multiclass classifier for all 3.
There is a class imbalance issue here with more Positive than Negative labeled tweets AND more Neutral than Pos or Neg. Be aware of this when training the classifier.



Apple is the most mentioned brand, followed by Google, then Unknown. A small percentage of tweets mention both brands.

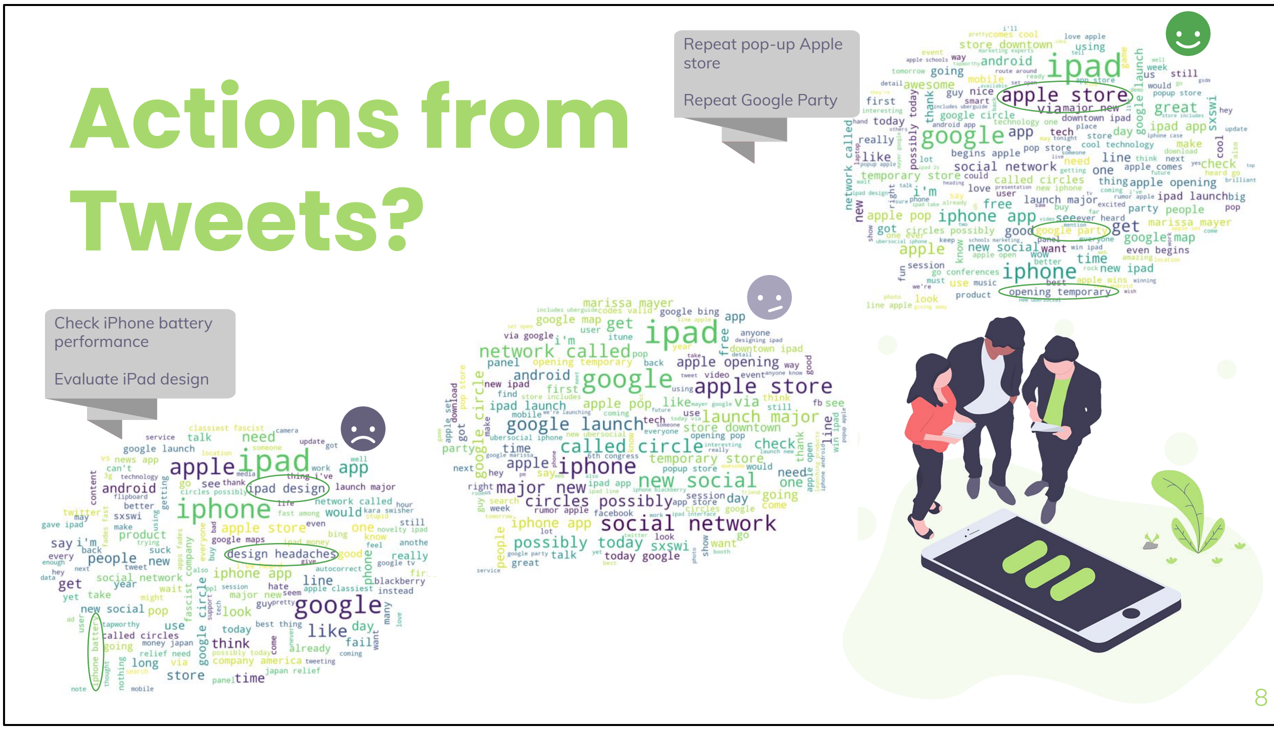
3 Brand insights?

- Few co-mentions may show “Two Camps”
- Similar patterns
  - Pos > Neg
  - ++ Neutral
  - Neg:Pos Ratios Apple-19% Google-18%



Both Apple and Google have many more Positive tweets than Negative
Both brands have many more Neutral tweets than Positive or Negative
How do the ratios of Negative to Positive tweets compare for Apple and Google?

  * 398:2018 Apple, 149:790 Google
  * 0.19 Apple, 0.18 Google   __Similar Ratios__



Word clouds can give a quick look at what consumers are saying.
Word clouds are not suitable for doing any deep analysis of word frequencies. :)
They can reveal some actionable recommendations to the two companies and the conference. From the Negative Tweets, we see “iPad design, design headaches, and iphone battery” as common terms. Recommend to check iPhone battery performance and evaluate iPad design in light of this Negative sentiment. From the Positive Tweets, we see “Apple store, opening temporary, and Google party” as common terms. Recommend to repeat those well received events.

Some common words are in all three sentiment word clouds..included in all 3 classes of Tweets. I used stopwords and TD-IDF in my text processing to make sure that only the words that better indicators of a unique class are used to train the classifier. I also removed stopwords (super common words w/ little information like a, the an) and performed Lemmatization (getting ‘word’ bases) on the text. I also used a specialized TweetTokenizer to process the Tweet text.

Tweet Classifier

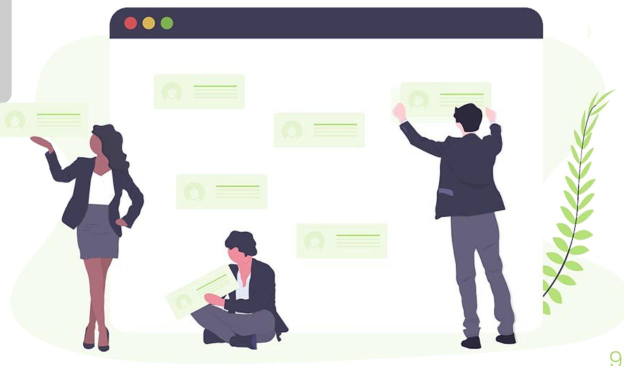
All 3 Sentiments?

Many Neutral tweets. Consider them.

Positive or Negative?

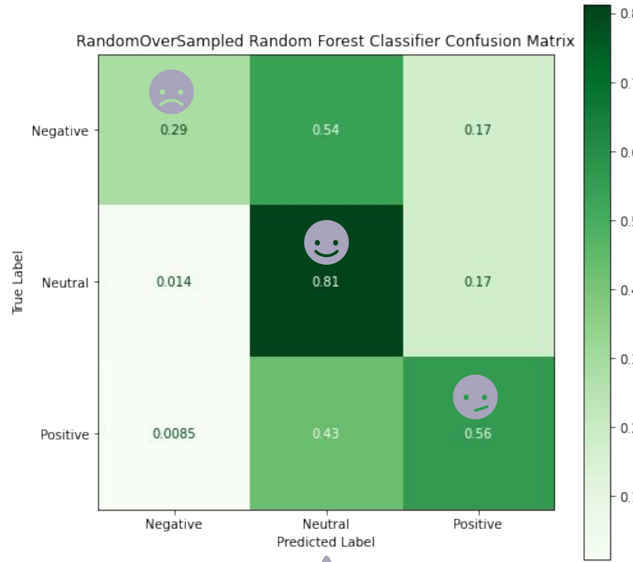
Ignore Neutral, we want strong opinions.

Create Models for Both!



1 Multiclass Classifier Results

- Weighted F1 Score as Evaluation Metric for the (4) built models
  - Useful for class imbalance
  - Balances Precision and Recall
- RandomForest w/ Oversampling generated the “best” model
  - Good for Neutral tweets
  - Mediocre for Positive tweets
  - Poor for Negative tweets



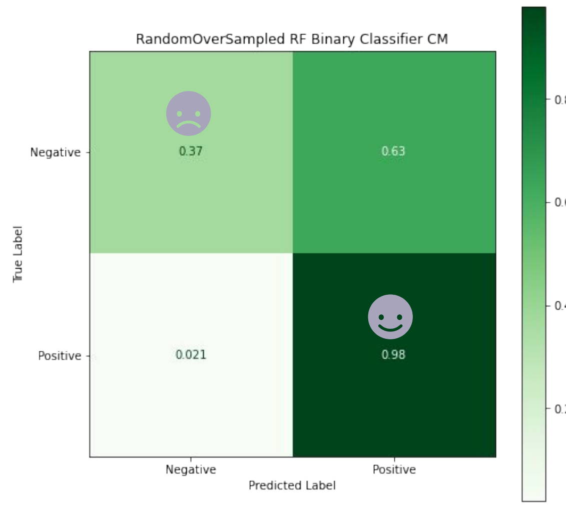
Score
0.68

Remember when I highlighted the SMALL number of Negative Tweets? That’s a problem.Class imbalance problem where we don’t have enough examples of the minority class in the data. But there are workarounds. I chose to use a technique called RandomOverSampling where you re-choose examples to train your model. Let’s see how that worked out.

RandomForest (and RandomForest with RandomOverSampling) had the highest weighted F1 scores of all models I trained. I'm calling RandomForest with RandomOverSampling the **winner** as it has a slightly better True Positive rate for correctly identifying Negative Tweets (0.29 vs 0.25).
I imagine Negative tweets are most interesting to stakeholders
It also has a better True Positive rate for correctly identifying Positive Tweets (0.56 vs 0.48).
These numbers are still pretty poor.

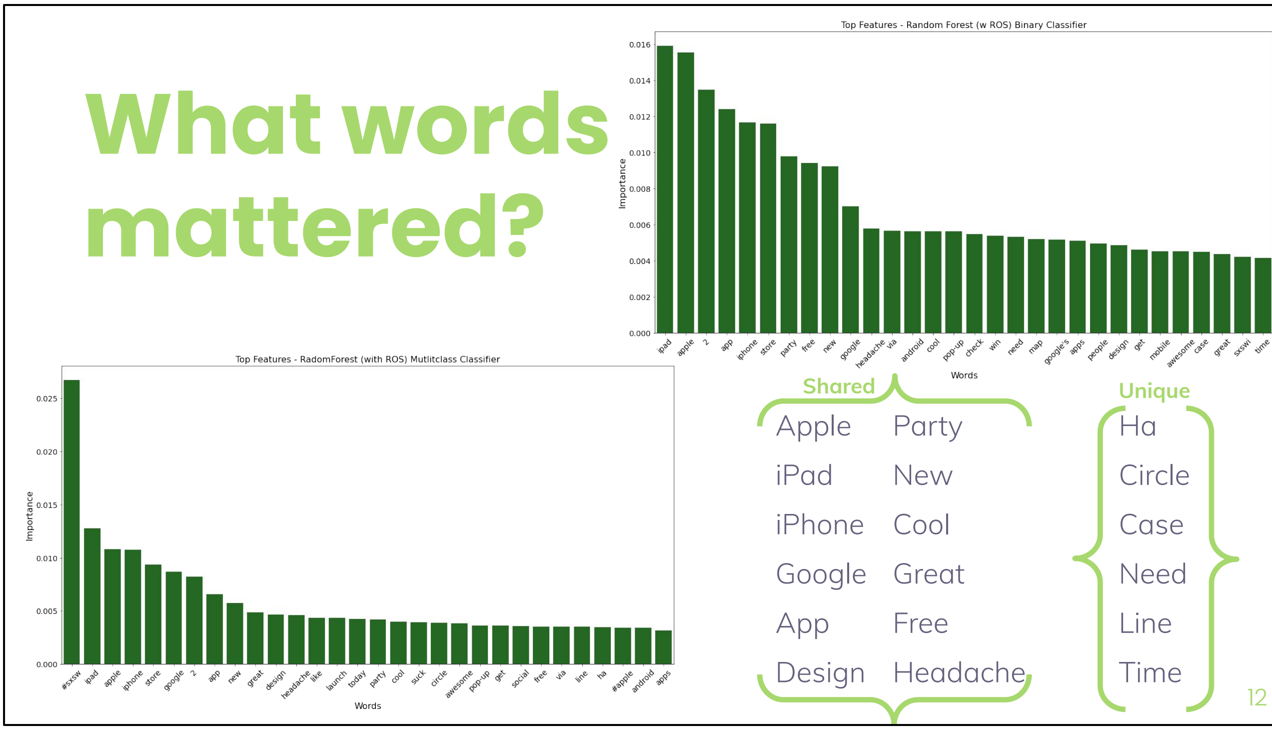
2 Binary Classifier Results

- Weighted F1 Score as Evaluation Metric for the (4) built models
- RandomForest w/ Oversampling generated the “best” model
  - Fantastic for Positive tweets
  - Poor for Negative tweets
- Better Score than Multiclass Model



Score
0.87

None of the binary classifiers did a good job with classifying Negative tweets, even with RandomOverSampling. RandomForest and RandomForest with RandomOverSampling had the highest weighted F1 scores of all models I trained. I'm calling RandomForest with RandomOverSampling the winner as it has a slightly better True Positive rate for correctly identifying Negative Tweets (0.37 vs 0.3). This is still crummy.



Showing top features: the words that were important to determine class.

Future Work

1. Punctuation as a Feature?
2. Hyper-parameter tuning
3. More Tweets!

See In-Depth Future Work in Appendix for more technical version of Future Work

Thanks!

Any questions?

You can find me at:

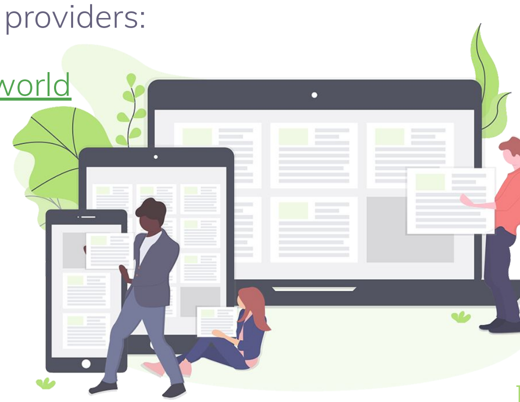
- https://github.com/lspope/
- https://leahspope7.medium.com/
- https://www.linkedin.com/in/leahspope/
- leah@metisconsultingllc.com

Appendix

Credits

Special thanks to (awesome & free) resources providers:

- Data provided by CrowdFlower via data.world
- Presentation template by SlidesCarnival
- Photographs by Unsplash
- Illustrations by Undraw.co

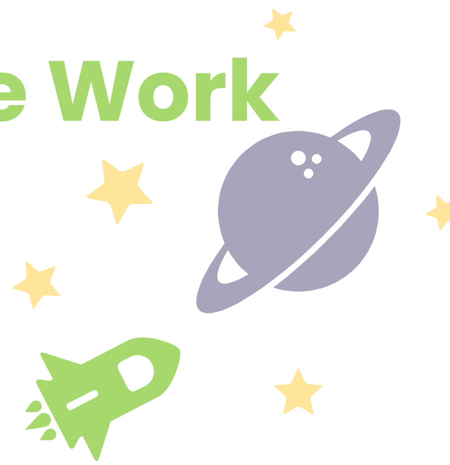


Special Thanks

To special people:

- Family & Friends
- Fellow Cohort members (We're almost there!)
- Flatiron Instructors

In Depth Future Work



1. Punctuation as a Feature?
2. Use spaCy in place of nltk
3. SMOTE for class imbalance
4. Hyper-parameter tuning
5. Better understanding of LIME
6. Transfer Learning with GloVe
7. More Tweets!
8. VADER and/or Text Blob to generate sentiment?
    a. *See my blog*