Modeling Tweet Sentiment

Charlotte Basch

Introduction

- Twitter has 166 million users
- 22% of American adults use twitter, with those users more likely to be more affluent and younger
- This makes Twitter an excellent resource for gauging consumer sentiment

- https://s22.q4cdn.com/826641620/files/doc_financials/2020/q1/Q1-2020-Earnings-Press-Release.pdf
- https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/
- https://www.stickpng.com/img/icons-logos-emojis/tech-companies/twitter-logo

Data

- 9,000 tweets from CrowdFlower about Apple and Google products
- Rated as positive, negative, neutral, and unknown
- Used 3,500 positive and negative tweets

Example Tweets

Negative Tweet:

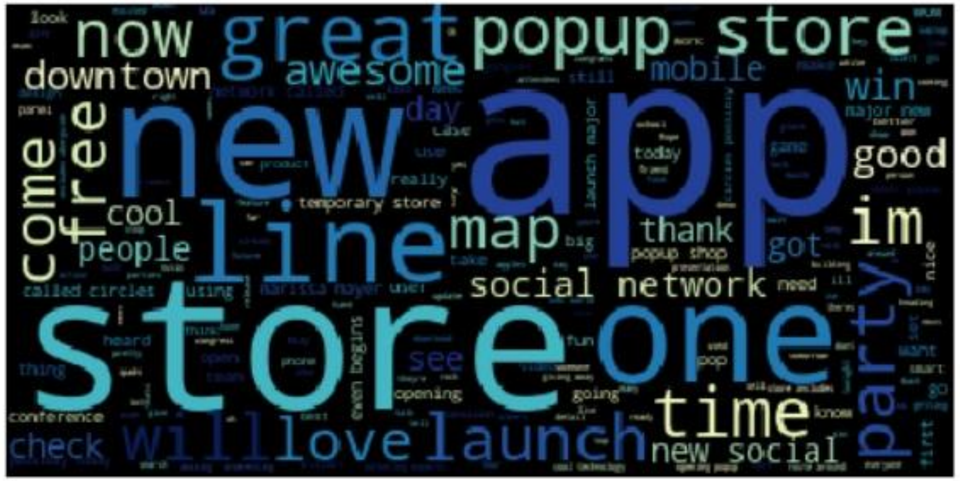
.@[username] I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plugin stations at #SXSW.

Positive Tweet:

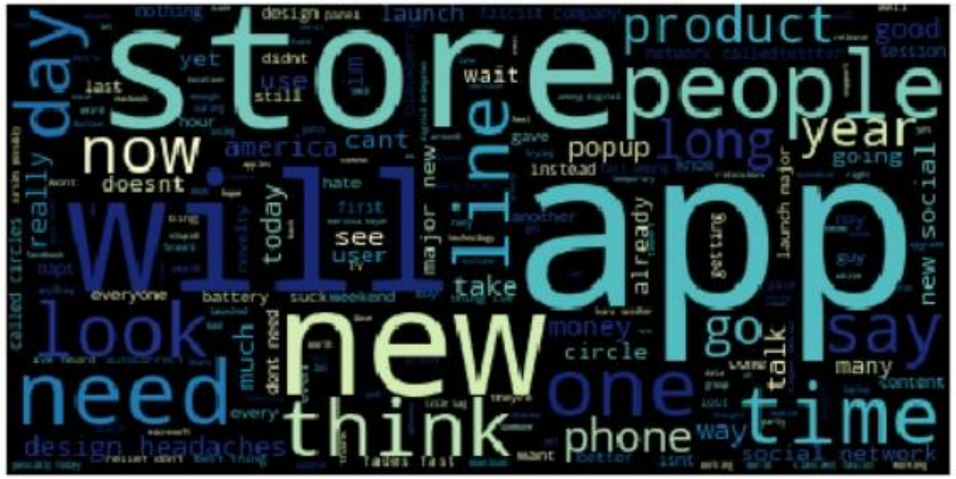
@[username] Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW

Wordclouds

Wordcloud for Positive Tweets

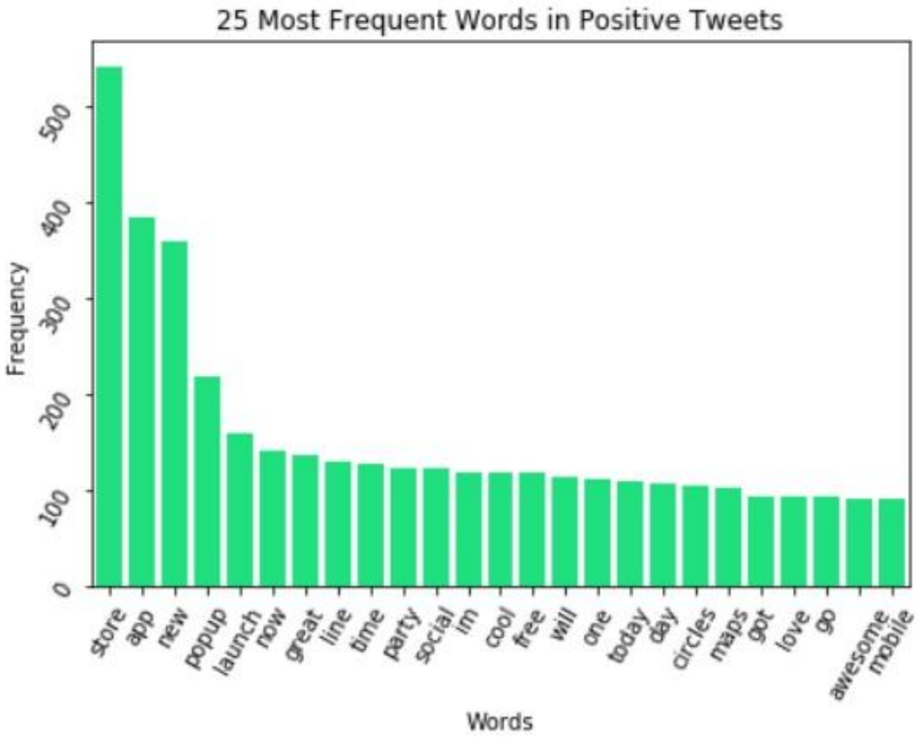


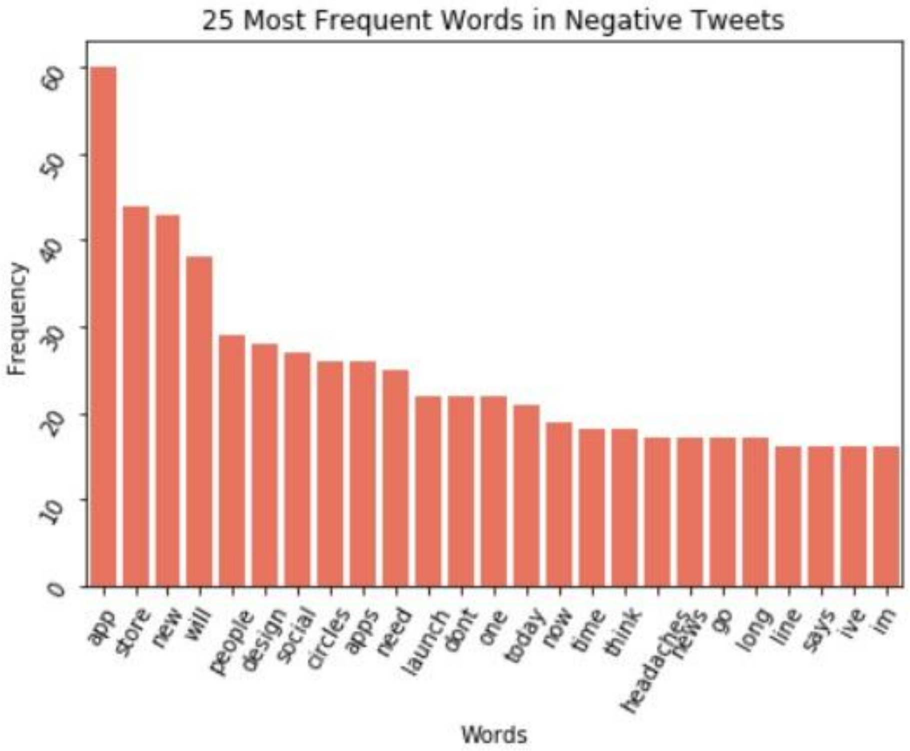
Wordcloud for Negative Tweets



- There is some overlap in some words
- There are more positive words (i.e. great) in the positive tweets
- The positive tweets seem to reference new products

Most Common Words

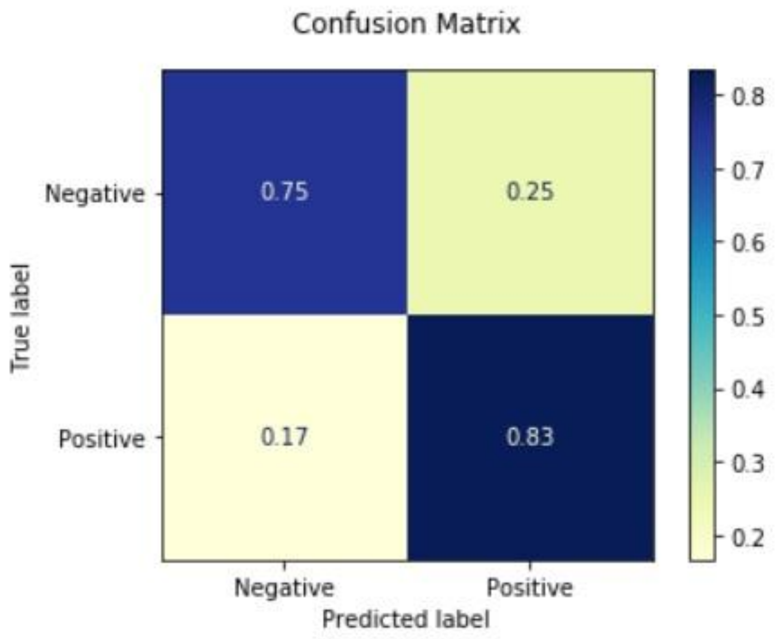

25 Most Frequent Words in Positive Tweets


25 Most Frequent Words in Negative Tweets

- Verbs appear more prominently in the negative tweets
- There are multiple references to new products (i.e. popup, party) in the positive tweets
- Overall the words are fairly similar, possibly related to where the tweets were collected

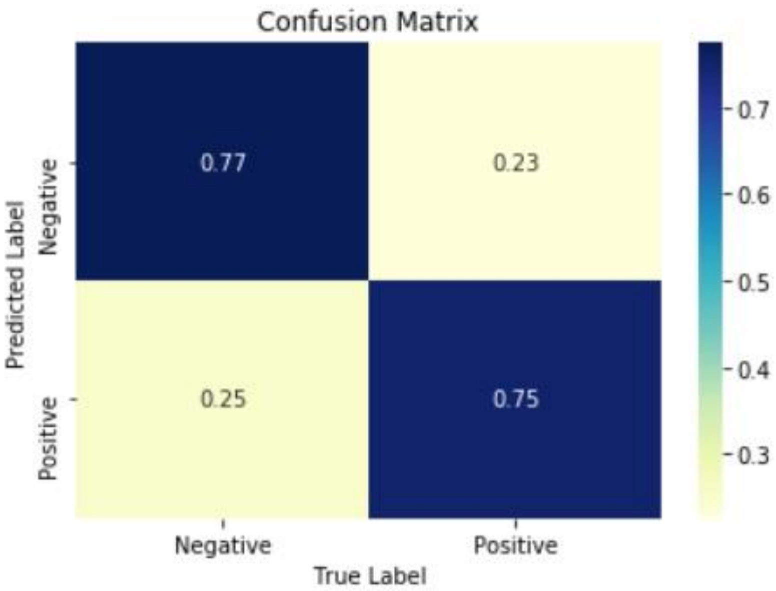
Modeling – Logistic Regression

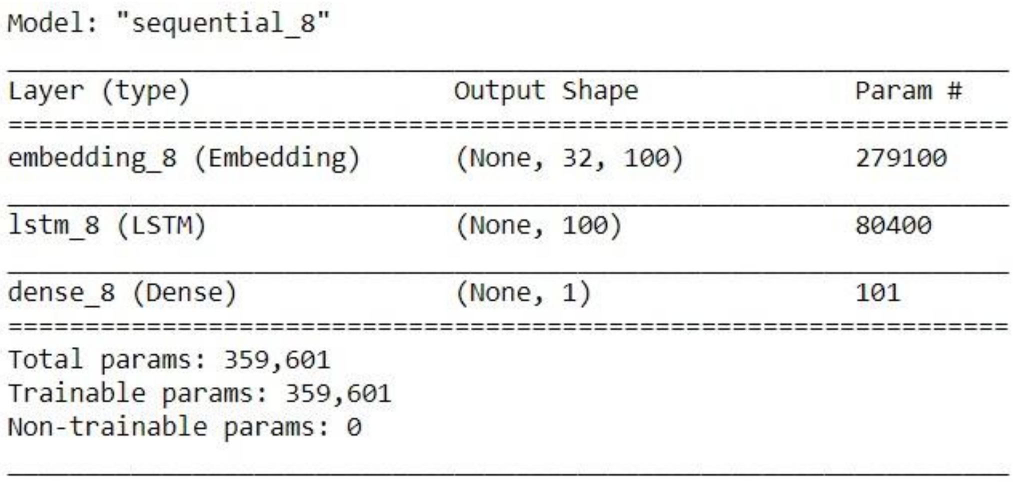
- Predicting the probability that an observation belongs to the negative class
- Logistic regression is a fairly simple model
- The best model was 82% accurate



Modeling – Neural Networks

- Long Short Term Memory neural networks
  - This is a network that is able to remember but also able to throw away information it does not need
- The best model was approximately 75% accurate





```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_8 (Embedding)      (None, 32, 100)           279100
_________________________________________________________________
lstm_8 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 101
=================================================================
Total params: 359,601
Trainable params: 359,601
Non-trainable params: 0
_________________________________________________________________
```

Recommendations

- More complex is not always better
- Use a logistic regression model to classify tweets as positive or negative
- Because the model only correctly classifies negative tweets 75% of the time and positive tweets 83% of the time, it is important to check the model against human verification
- Tweets using this model should have urls, hashtags, and usernames removed

Future Work

- Collect more data
- Add in neutral category
- Get data from a wider time period

Summary

- People tended to tweet more positively about the excitement of new products
- Action words, i.e. will, are more common in negative tweets, perhaps indicating that people are tweeting about their intention to stop using a product
- While the current model has fairly good accuracy, steps can be taken to improve the classification
- The logistic regression model outperforms the neural network model in classifying positive and negative tweets
- This model is faster, more consistent, and requires fewer resources

Thank you!