NLP Twitter Sentiment Classification

Edward Cheng

Twitter

Our Mission

Twitter is a popular social media platform used by hundreds of millions of people around the globe. In fact, the current estimate of twitter users amount to approximately 330 million monthly active users and 145 million daily active users on Twitter. 63 percent of all Twitter users worldwide are between the ages of 35 and 65. You can find some more statistics about the usage of Twitter around the world below. The CEO of twitter has tasked me with building a model that has the capabilities of analyzing Twitter sentiments about Apple and Google products. The human raters rated the sentiment in over 9,000 Tweets as positive, negative, or neither.

01 **11.7 million**

App store downloads in the first quarter in 2019. Twitter enjoyed a year-over-year increase of 3.6% since 2015.

02 **500 million**

Tweets are sent each day, which equates to 5.787 tweets per second.

03 **40%**

Of Twitter users reported purchasing something after seeing it on the platform.

Process

01

Obtain the Data

- Import the csv file containing over 9000 tweets.

02

Text Pre-processing

- Check for NA values
- Stop word Removal:
  - Removal of punctuation and numbers
  - Remove capitalization
  - Regex modifications
- Lemmatization
- Noise removal
- Tokenization

03

Explore the Data

- Explore the distribution of the length of tweets
- Explore the top 20 most popular words
- Explore the top 20 most popular hashtags
- Explore the distribution of words based on the tweet's sentiment

04

Model the Data

- Run 6 different models:
  - Multinomial NB Classifier
  - Logistic Regression
  - Random Forest
  - Support Vector Classifier
  - Deep Neural Network (base model)
  - Deep Neural Network (regularized, with dropout)

Pre-processed text data

Non-tokenized Version

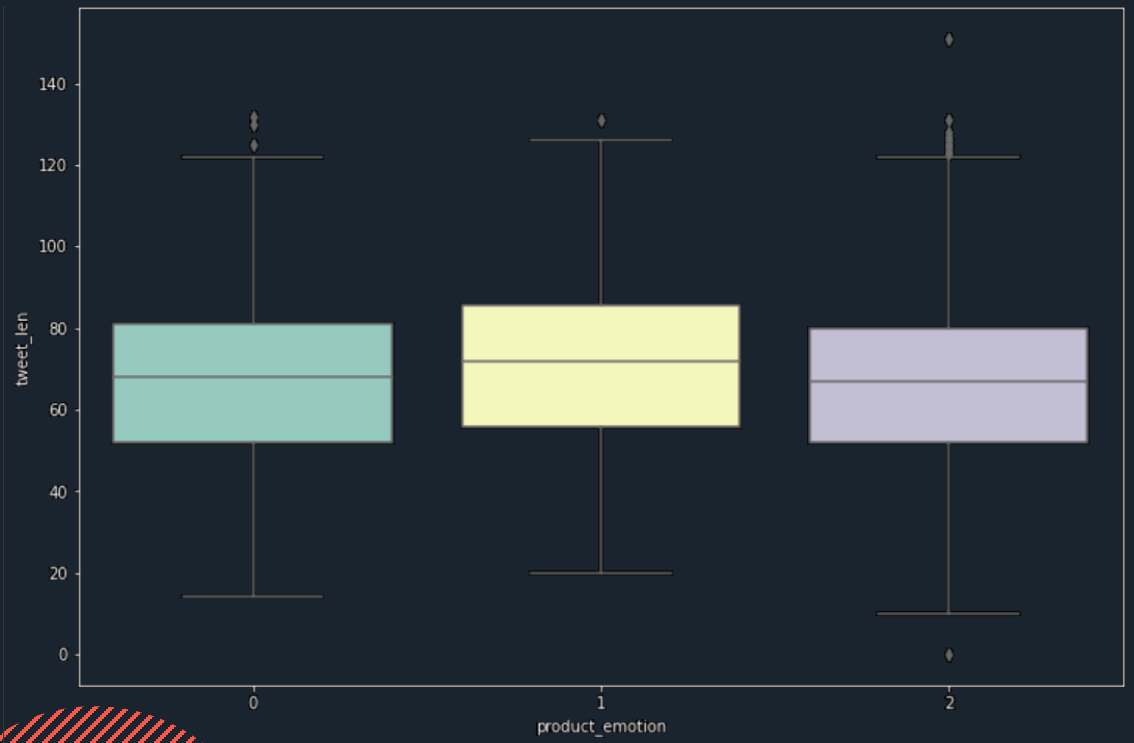
```
0     wesley83 3g iphone hr tweeting rise_austin dea...
1     jessedee know fludapp awesome ipad iphone app ...
2                  swonderlin wait ipad also sale sxsw
3     sxsw hope year festival crashy year iphone app...
4     sxtxstate great stuff fri sxsw marissa mayer g...
5     teachntech00 new ipad apps speechtherapy commu...
7     sxsw starting ctia around corner googleio hop ...
8     beautifully smart simple idea madebymany thene...
9     counting day sxsw plus strong canadian dollar ...
10    excited meet samsungmobileus sxsw show sprint ...
11    find amp start impromptu party sxsw hurricanep...
12    foursquare ups game time sxsw http j mp grn7pk...
13    got ta love sxsw google calendar featuring top...
14    great sxsw ipad app madebymany http tinyurl co...
15    haha awesomely rad ipad app madebymany http bi...
16    holler gram ipad itunes app store http co kfn3...
17    noticed dst coming weekend many iphone hour la...
18    added sxsw flight planely matching people plan...
19    must sxsw app malbonster lovely review forbes ...
20    need buy ipad2 austin sxsw sure need q austin ...
21    oh god sxsw app ipad pure unadulterated awesom...
22    okay really yay new foursquare android app 11 ...
23    photo installed sxsw iphone app really nice ht...
24    really enjoying change gowalla android looking...
25    laurieshook looking forward smcdallas pre sxsw...
26    haha awesomely rad ipad app madebymany http bi...
27    someone started austin partnerhub group google...
28    new 4sq3 look like going rock update iphone an...
```

Tokenized Version

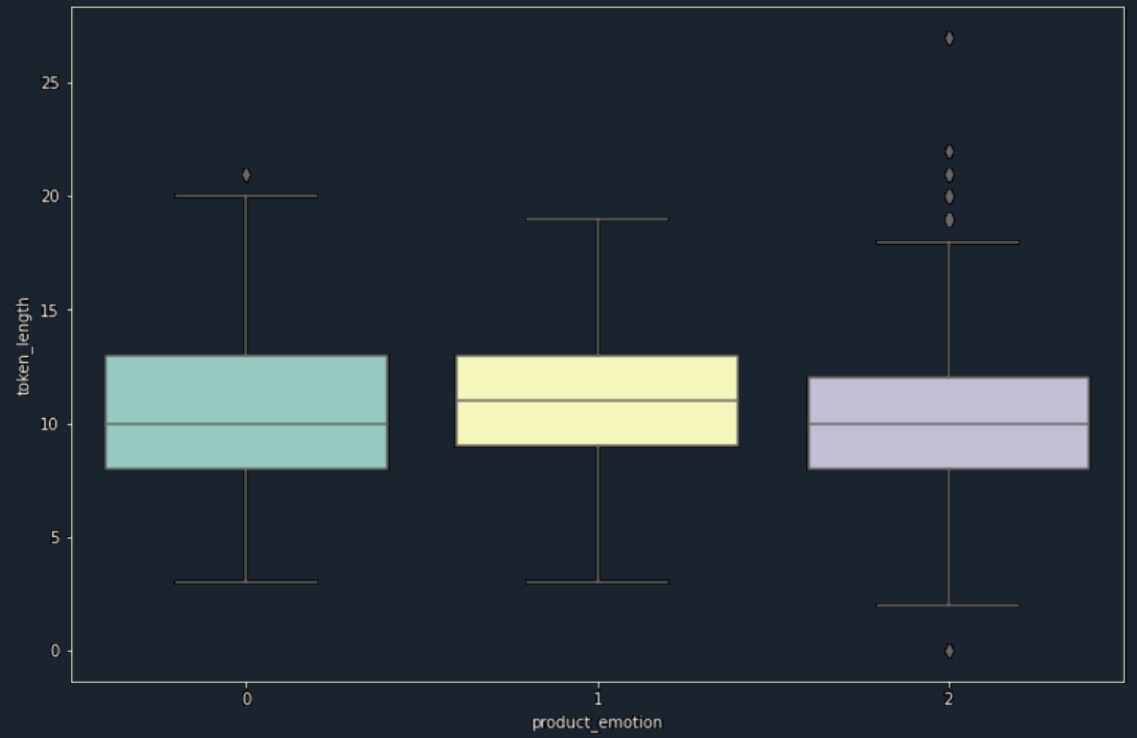
```
[('sxsw', 9372),
 ('google', 2594),
 ('ipad', 2457),
 ('apple', 2281),
 ('quot', 1657),
 ('iphone', 1541),
 ('store', 1498),
 ('new', 1078),
 ('austin', 949),
 ('amp', 834),
 ('app', 815),
 ('launch', 685),
 ('circle', 666),
 ('social', 637),
 ('android', 588),
 ('today', 576),
 ('network', 467),
 ('get', 451),
 ('line', 432),
 ('via', 428),
 ('party', 397),
 ('free', 387),
 ('called', 354),
 ('mobile', 345),
 ('sxswi', 337),
 ('one', 311),
 ('major', 303),
 ('time', 299),
 ('like', 291),
```

Length of Tweets and Tokenized Words

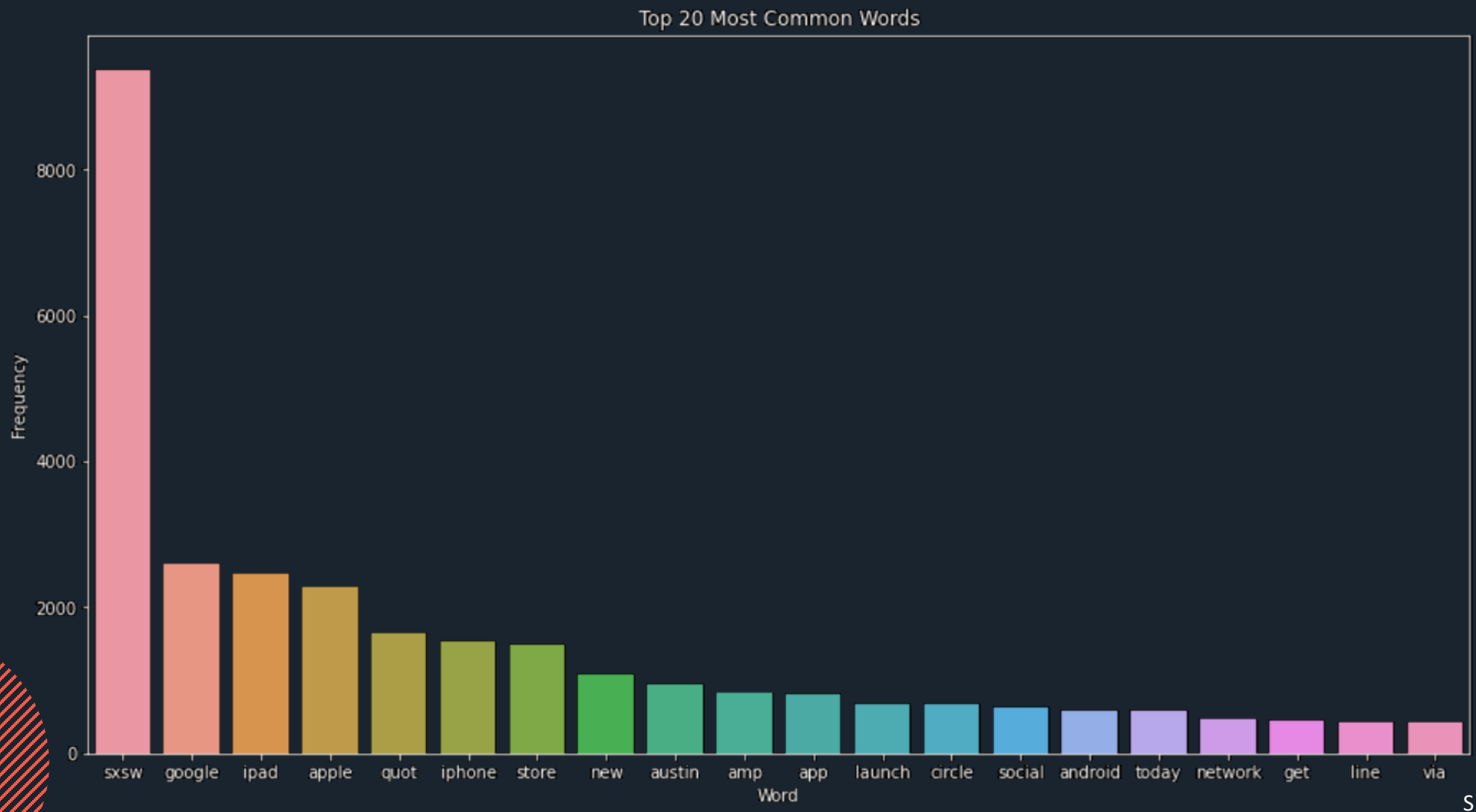
Tweet Length



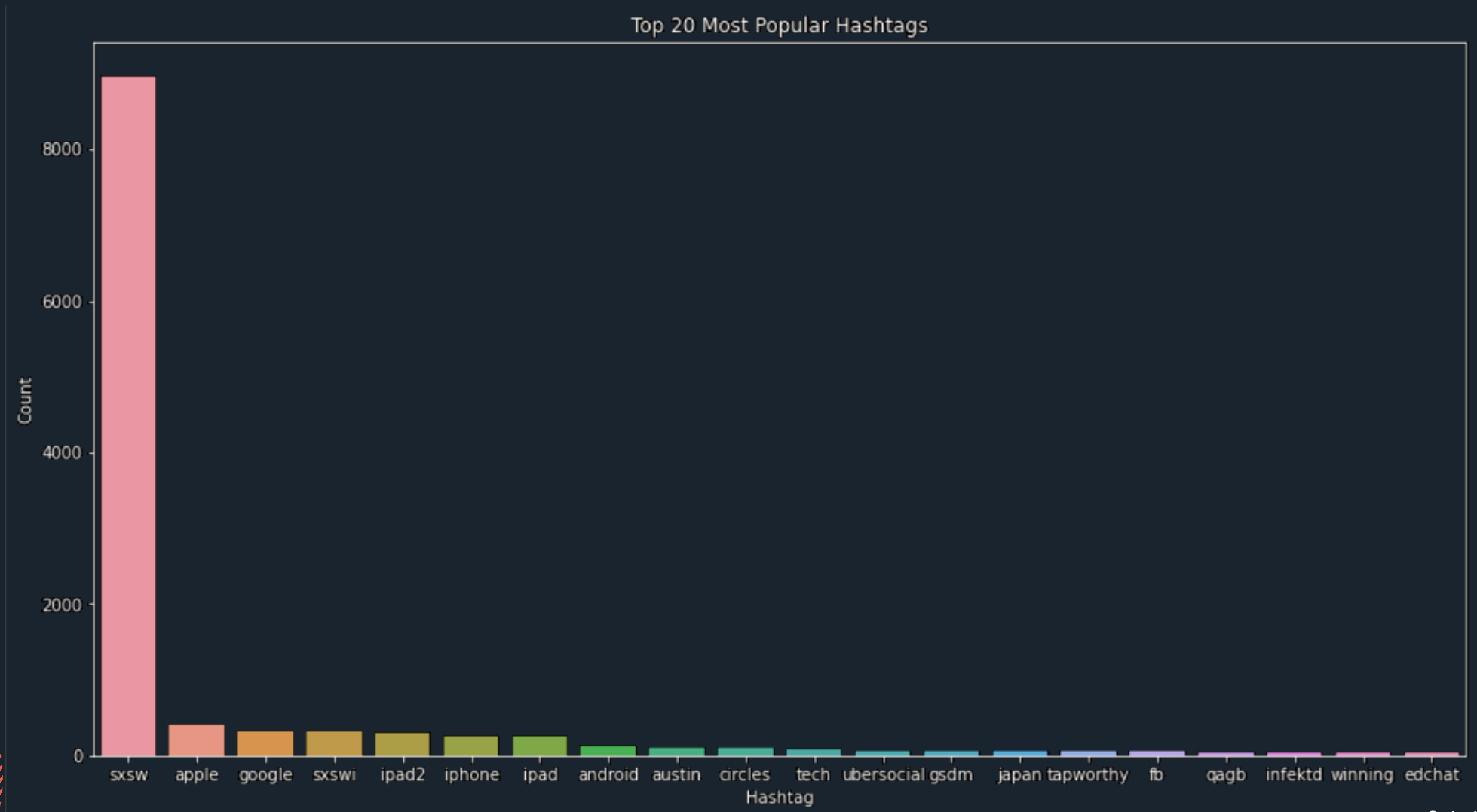
Token Length



Top 20 Most Popular Words



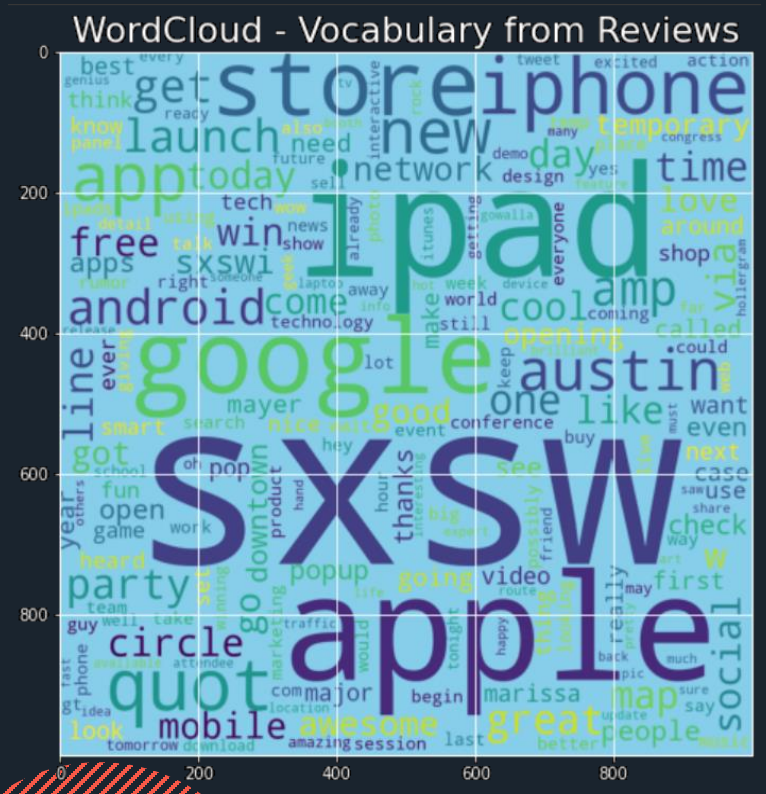
Top 20 Most Popular Hashtags



Most Popular Words

By Product Emotion

Positive



Negative

Neutral

Model Selection

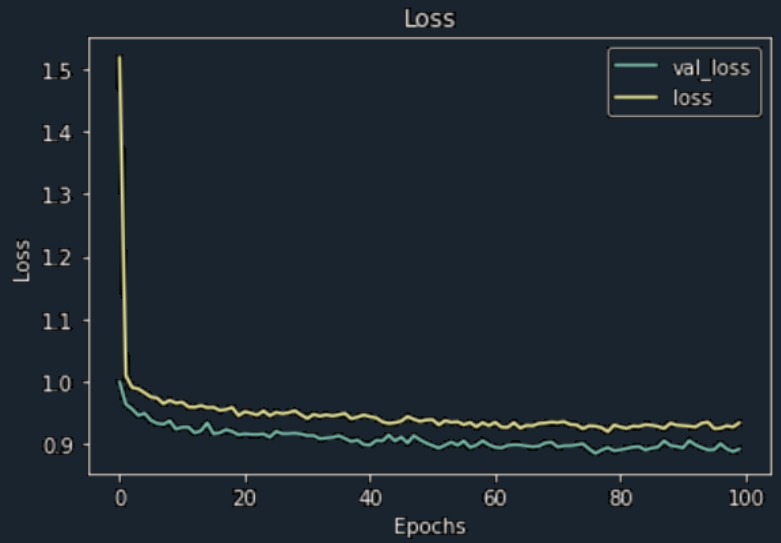
Multinomial NB, Logistic Regression, Random Forest, SVC, Deep Neural Network (base model), Deep Neural Network (regularized)

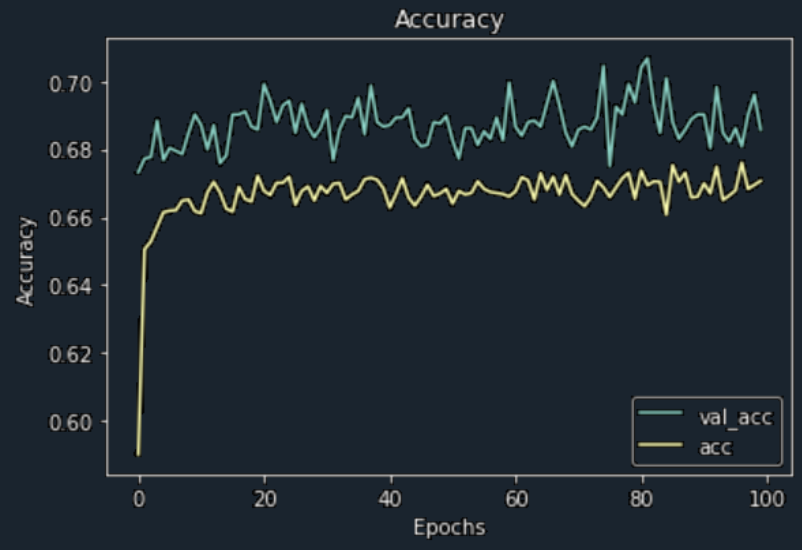
Best Performing Model: Deep Neural Network (regularized)

69.0% Accuracy Score

```
210/210 [==============================] - 0s 1ms/step - loss: 0.8691 - accuracy: 0.6941
Training Loss: 0.869
Training Accuracy: 0.694
--------------------------------------------------------------------------------
70/70 [==============================] - 0s 1ms/step - loss: 0.8926 - accuracy: 0.6858
Test Loss: 0.893
Test Accuracy: 0.686
```

```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dropout_4 (Dropout)          (None, 2229)              0
_________________________________________________________________
dense_18 (Dense)             (None, 25)                55750
_________________________________________________________________
dropout_5 (Dropout)          (None, 25)                0
_________________________________________________________________
dense_19 (Dense)             (None, 3)                 78
=================================================================
Total params: 55,828
Trainable params: 55,828
Non-trainable params: 0
```





Analysis

- Final model details:
  - Two layers
  - Early stopping
  - L1 regularization
  - Dropout layers
  - Data augmentation
  - RMS prop optimizer algorithm

- Working with a multi-categorical classification problem, our model performed significantly better than random guessing
  - With random guessing, one would predict the tweet's sentiment 48% (benchmark accuracy score) of the time correctly
  - The 48% was calculated by squaring the percentage of values belonging to each of the tweet's sentiment class: 0.60% ** 2 (neutral sentiment) + 0.33% ** 2 (positive sentiment) + 0.06% ** 2 (negative sentiment) = 48%

- ***With our model, it would predict the tweet's correct sentiment 69.0% of the time***

Future Work

- Collect more data to train the models
- Spend more time preprocessing the text data
- Trying out more modelling algorithms such as XG boost
- Applying weight initializations to the neural networks
- Trying out more deep neural networks such as GRU

Thank You

Flatiron School Online Data Science Bootcamp