Classifying Android & Iphone Reddit Posts

Problem

Android Inc

We are looking at how to better design and develop android phones. There are many variants of Android and we want to look at reddit forums to see what users prefer in a mobile phone. At the same time, Android has a close competitor, Iphone.

We would also want to look at their users comments on Iphone as well. However as there are still differences in the phones' users and usage, we would want to classify the two types of reddit posts to enhance our research.

Reddit Posts

786

https://www.reddit.com/r/Android

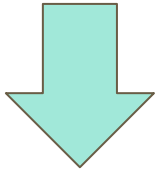
715

https://www.reddit.com/r/Iphone/new

Data Analysis & Cleaning

	approved_at_utc	subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title
0	NaN	Android	Note 1. Join us at /r/MoronicMondayAndroid, a ...	t2_6l4z3	False	NaN	0	False	Moronic Monday (Jan 20 2020) - Your weekly que...

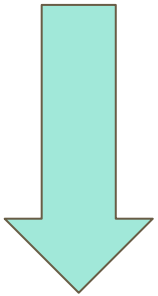
Remove columns all columns except these 3.



Selftext	Title	Subreddit
I like Android!	Android or Iphone	Android

Data Analysis & Cleaning

Selftext	Title	Subreddit
I like Android!	Android or Iphone	Android



- **Subreddit** will be our target as it indicates the post category.
- Combine Android and Iphone dataframes.
- Null values of are filled with blank.
- Concat selftext and title into one column **content**.
- **Subreddit** is map to 1 and 0.

content	Subreddit
I like Android! Android or Iphone	1
Iphone is great! Use Iphone	0

Data Analysis & Cleaning

Content column is cleaned:

- Removing all text characters and space.
- Removing stop words. (Words that will confuse our model)
- Lemmatizing (Combine words to their root word)

Comparing 2 Classification Models

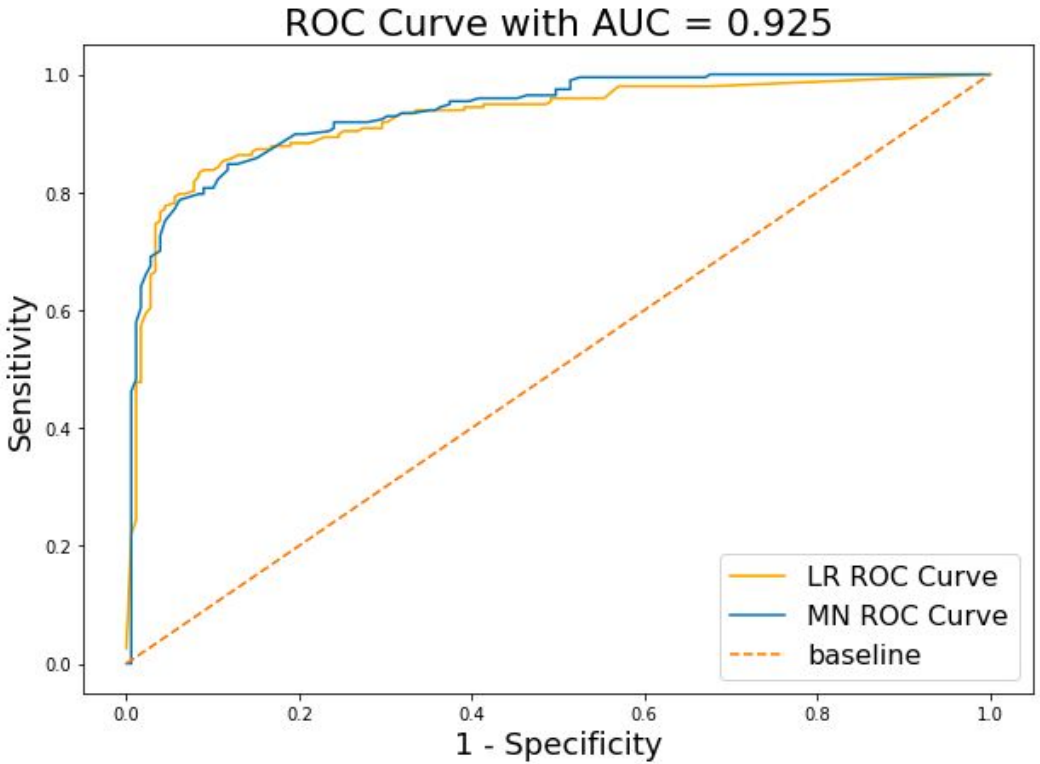
Logistic Regression

	Pred Iphone	Pred Android
Actual Iphone	134	45
Actual Android	16	181

Naive Bayes Multinomial

	Pred Iphone	Pred Android
Actual Iphone	136	43
Actual Android	19	178

Comparing 2 Classification Models



The more area under a curve means better better separated our distributions our model give.

When our ROC AUC is closer to 1, then our positive and negative populations are better separated which means the model is better.

From this graph, we can see that Logistic Regression gives a much better curve.

Conclusion

- From the model stats and ROC AUC curve, both models perform similarly and with acceptable accuracy.
- We can look at other models to see if they can do better than our current models.
- To further build on this project, we can look at sentiment analysis on the 2 topics. We can also look at specific mobile phone features that has more positive sentiment.