Web APIs & Classification

Classifying reddit posts into subreddits
“Developing Android Apps” and “iOS Programming”

Overview

1. Problem Statement
2. Exploratory Data Analysis
3. Modeling
4. Further Modeling
5. Conclusions and Recommendations

Problem Statement

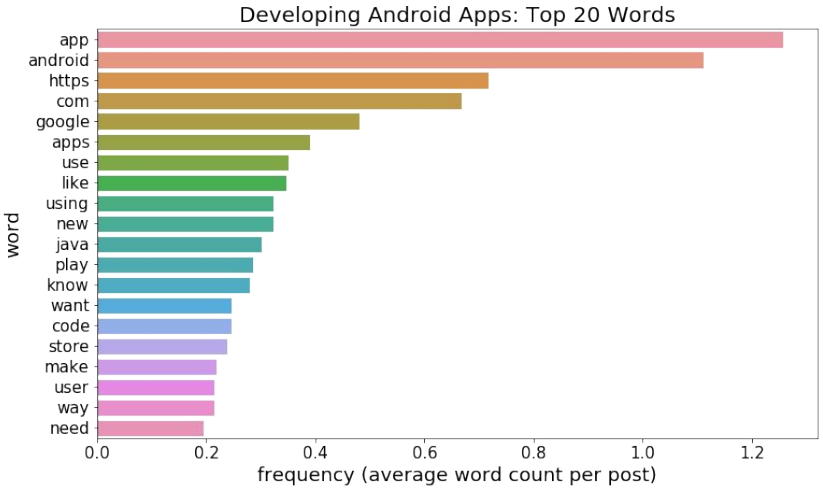
To identify the better model between Logistic Regression and Naive Bayes, based on their accuracies in classifying reddit posts into two categories, namely "iOS Programming" subreddit and "Developing Android Apps" subreddit.

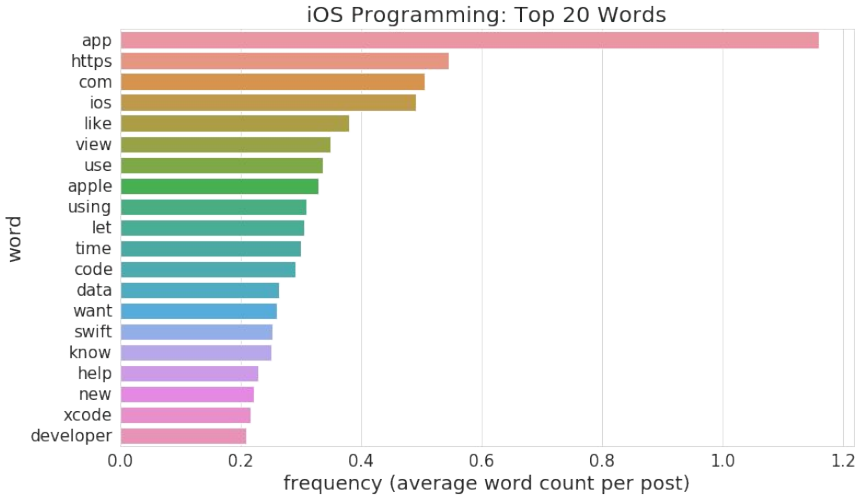
Exploratory Data Analysis

Pre-processing of selftext feature:

1. HTML, non-letters, urls were removed
2. Lowercased
3. Tokenized
4. Stopwords removed
5. Lemmatized

Exploratory Data Analysis





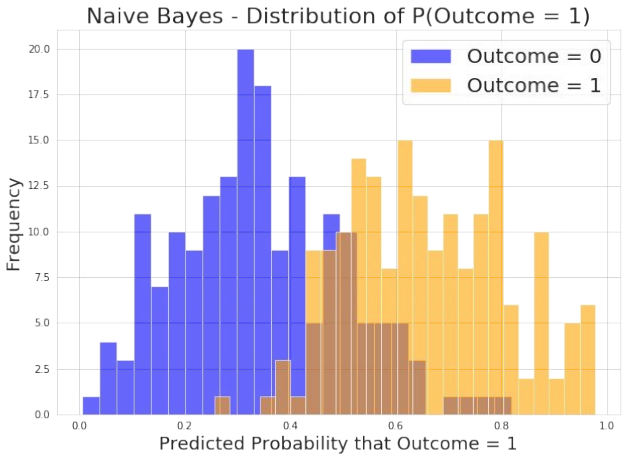
Modeling

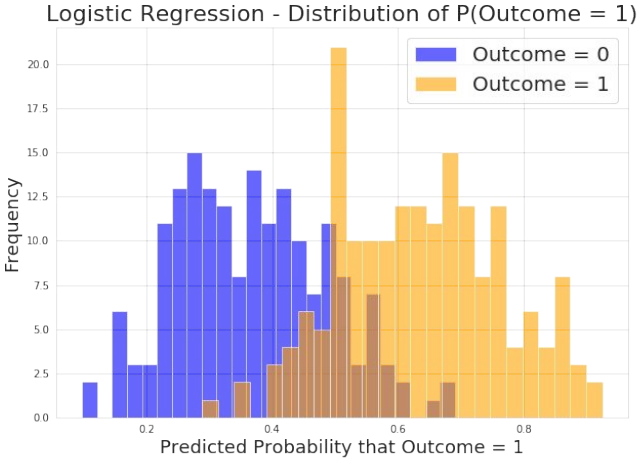
1. selftext
2. CountVectorized
3. Tf-idf transformed
4. Naive Bayes and Logistic Regression

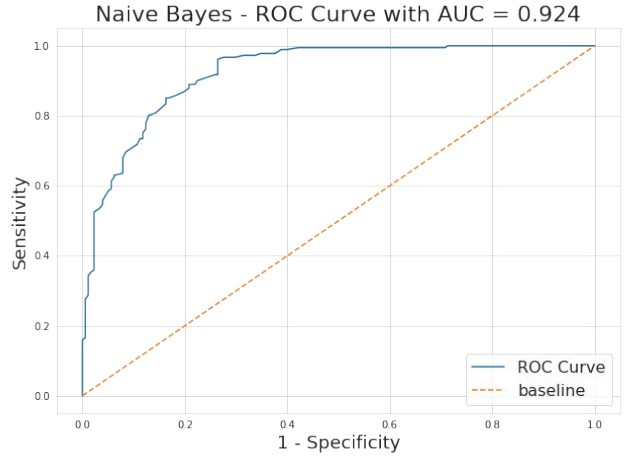
Modeling

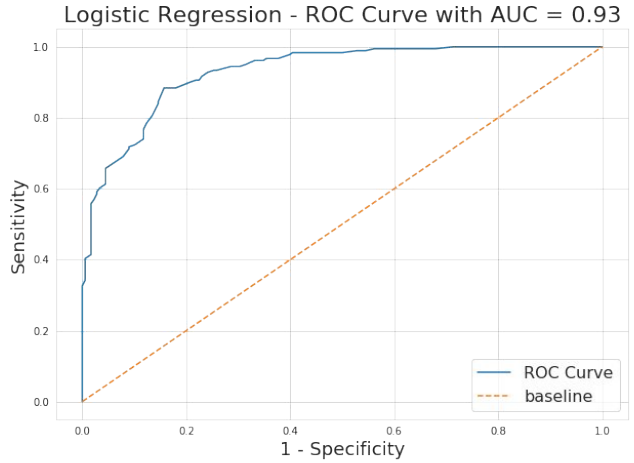
Metrics	Naive Bayes	Logistic Regression
Accuracy	0.841	0.850
Misclassification	0.159	0.150
Sensitivity	0.851	0.845
Specificity	0.831	0.854
Precision	0.837	0.855

Modeling


Naive Bayes - Distribution of P(Outcome = 1)
Outcome = 0
Outcome = 1
Frequency
Predicted Probability that Outcome = 1


Logistic Regression - Distribution of P(Outcome = 1)
Outcome = 0
Outcome = 1
Frequency
Predicted Probability that Outcome = 1


Naive Bayes - ROC Curve with AUC = 0.924
Sensitivity
1 - Specificity
ROC Curve
baseline


Logistic Regression - ROC Curve with AUC = 0.93
Sensitivity
1 - Specificity
ROC Curve
baseline

Modeling

Top 20 Features	
Naive Bayes	**Logistic Regression**
zip	android
maintain subreddit	google
major	java
make effort	play
make top	kotlin
manifest	admob
claim	activity
may open	play store
mediation	fragment
members	android studio
message compose	ad
messaging	google play
messaging app	studio
checking sidebar	game
midi	recyclerview
clean architecture	phone
migrate	mobile
migration code	apk
mistake	etc
mod	images

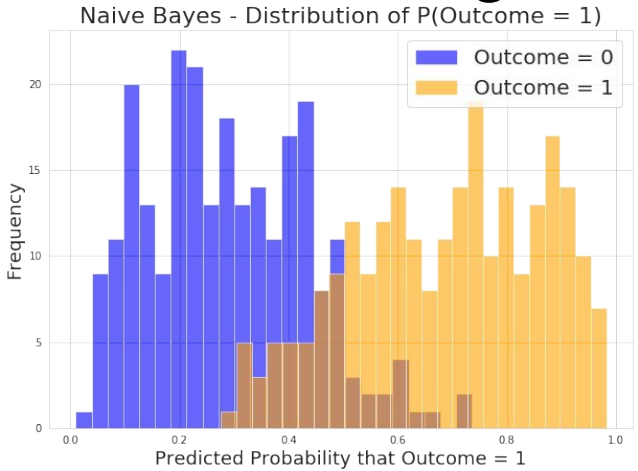
Further Modeling

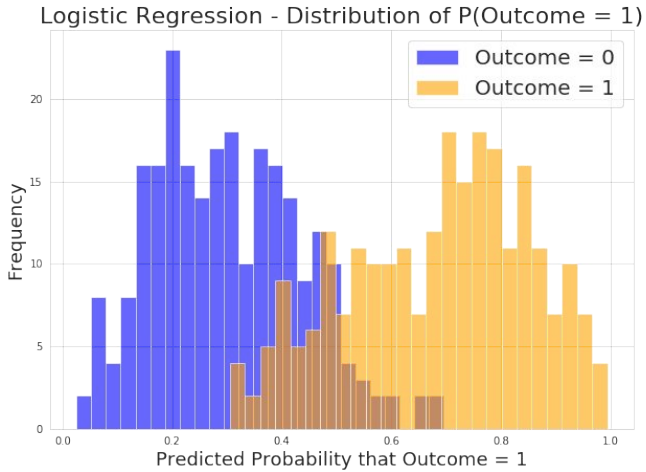
1. selftext + title
2. CountVectorized
3. Tf-idf transformed
4. Naive Bayes and Logistic Regression

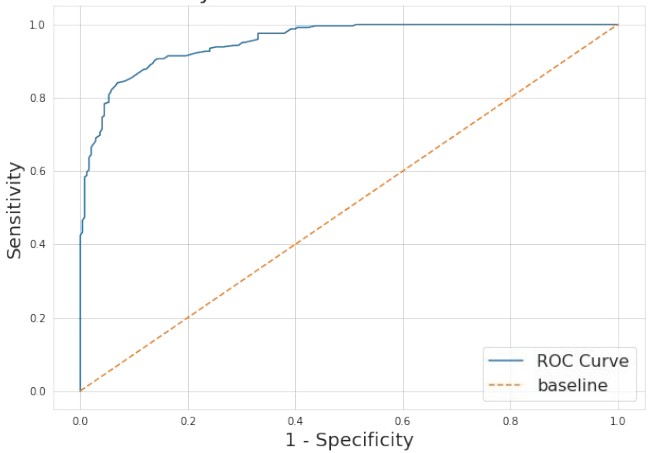
Further Modeling

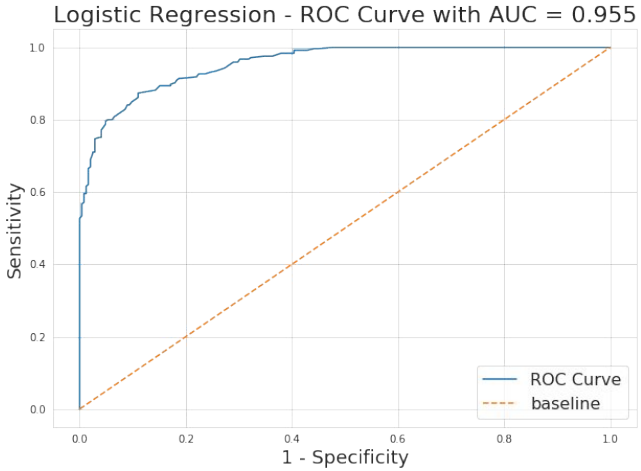
Metrics	Naive Bayes	Logistic Regression
Accuracy	0.888	0.869
Misclassification	0.112	0.131
Sensitivity	0.841	0.820
Specificity	0.935	0.918
Precision	0.928	0.910

Further Modeling









Further Modeling

Top 20 Features	
Naive Bayes	**Logistic Regression**
inject	android
com posting	google
name android	kotlin
please paste	java
initialize	admob
stores	play
commenters	emulator
comments consider	library
store pages	activity
store listing	play store
ndk	android studio
industry	studio
still allowed	google play
steal	dagger
indexing	room
compose	fragment
compose fr	fragments
stats	jetpack
strike	ad
income	game

Conclusions and Recommendations

- Before feature engineering, accuracy of Logistic Regression > accuracy of Naive Bayes
- After feature engineering, accuracy of Naive Bayes > accuracy of Logistic Regression
- Recommended that the content of title and of selftext be combined
- Train Naive Bayes classifier