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

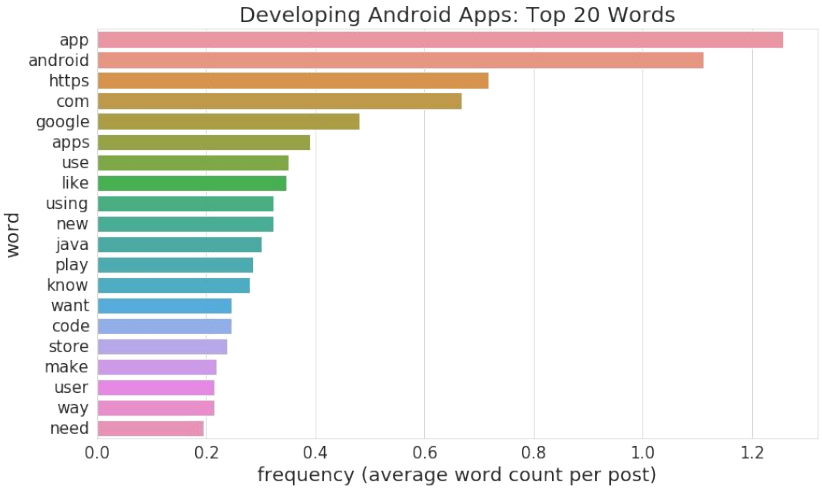
To identify the better model between Logistic Regression and Naive Bayes, based on their accuracies in classifying reddit posts into two categories, namely "iOS Programming" subreddit and "Developing Android Apps" subreddit.

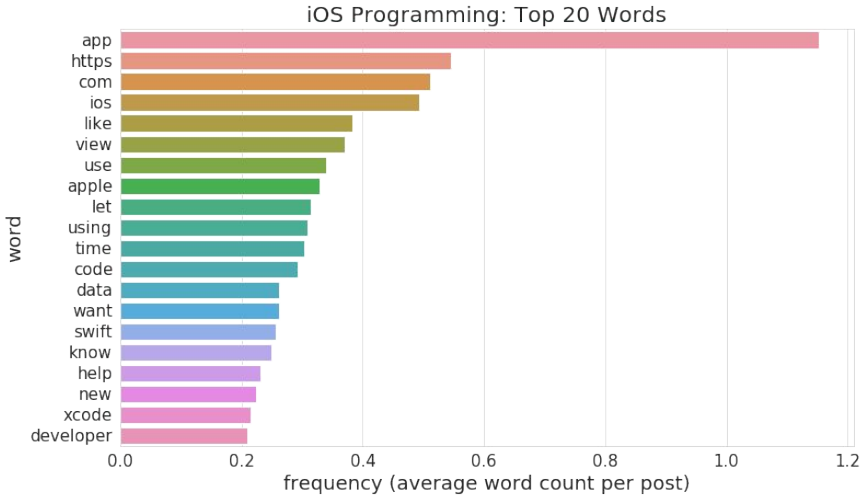
Exploratory Data Analysis

Pre-processing of selftext feature:

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

Exploratory Data Analysis





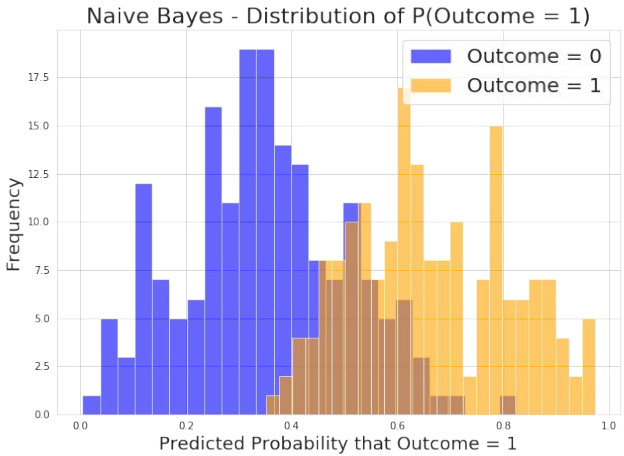
Modeling

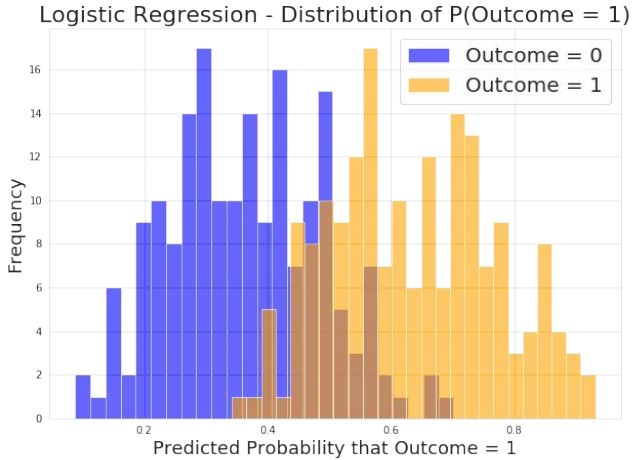
1. selftext
2. CountVectorized
3. Tf-idf transformed
4. Naive Bayes and Logistic Regression

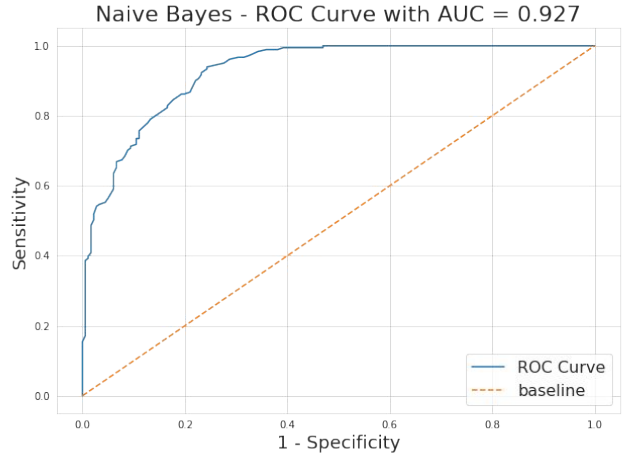
Modeling

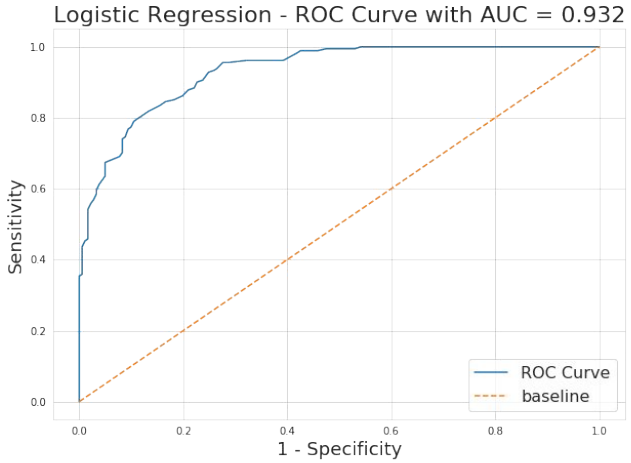
Metrics	Naive Bayes	Logistic Regression
Accuracy	0.837	0.848
Misclassification	0.163	0.152
Sensitivity	0.856	0.829
Specificity	0.818	0.867
Precision	0.824	0.862

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 = 0.927
Sensitivity
1 - Specificity
ROC Curve
baseline


Logistic Regression - ROC Curve with AUC = 0.932
Sensitivity
1 - Specificity
ROC Curve
baseline

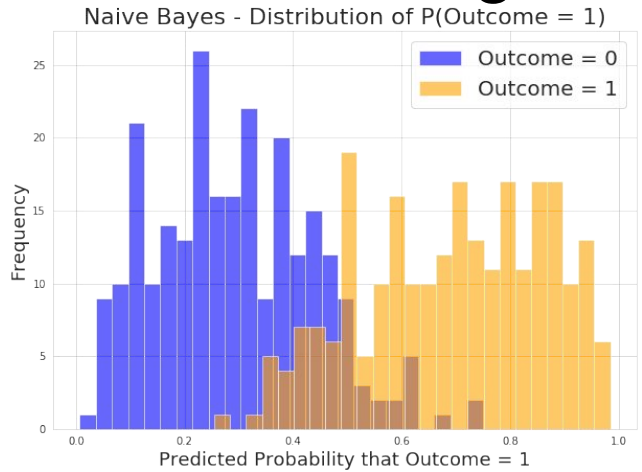
Further Modeling

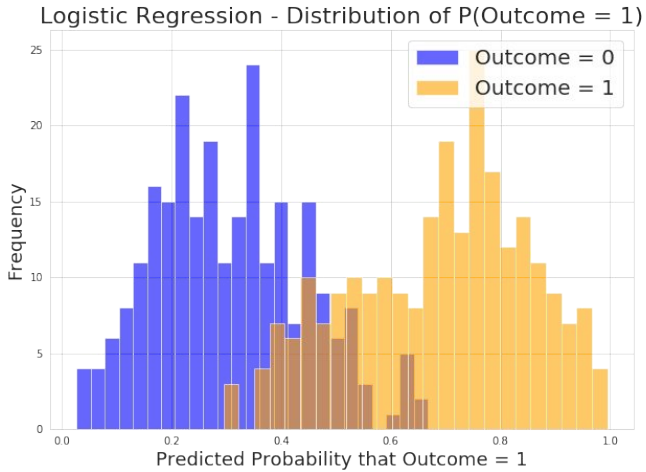
1. selftext + title
2. CountVectorized
3. Tf-idf transformed
4. Naive Bayes and Logistic Regression

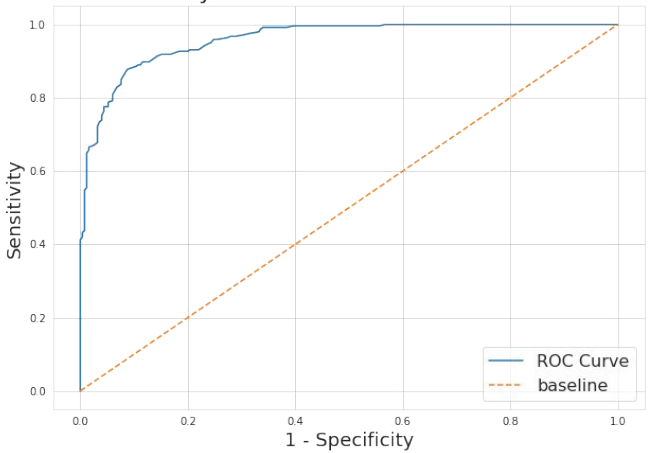
Further Modeling

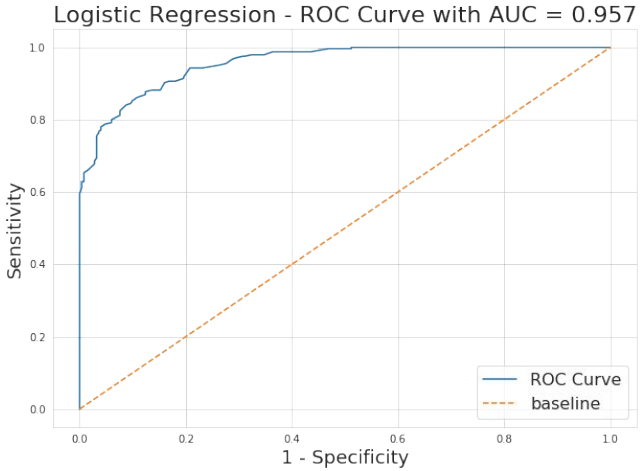
Metrics	Naive Bayes	Logistic Regression
Accuracy	0.885	0.877
Misclassification	0.115	0.123
Sensitivity	0.845	0.841
Specificity	0.924	0.912
Precision	0.916	0.904

Further Modeling









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