Predicting Subreddits

Wesley Osborne

Introduction

Problem Statement

Maker Faire Conference wants to do a fun interactive app for their attendees at the upcoming conference. For the attendees who are fans of Arduinos and Raspberry Pi they want to build a classification model that will identify which person is a fan of either device based on the text they enter into the app. The goal of the model is to be as accurate as possible. The hope for this project is to delight the attendees and retain attendance for future conferences.

image source

Data Source

r/raspberry_pi

image source

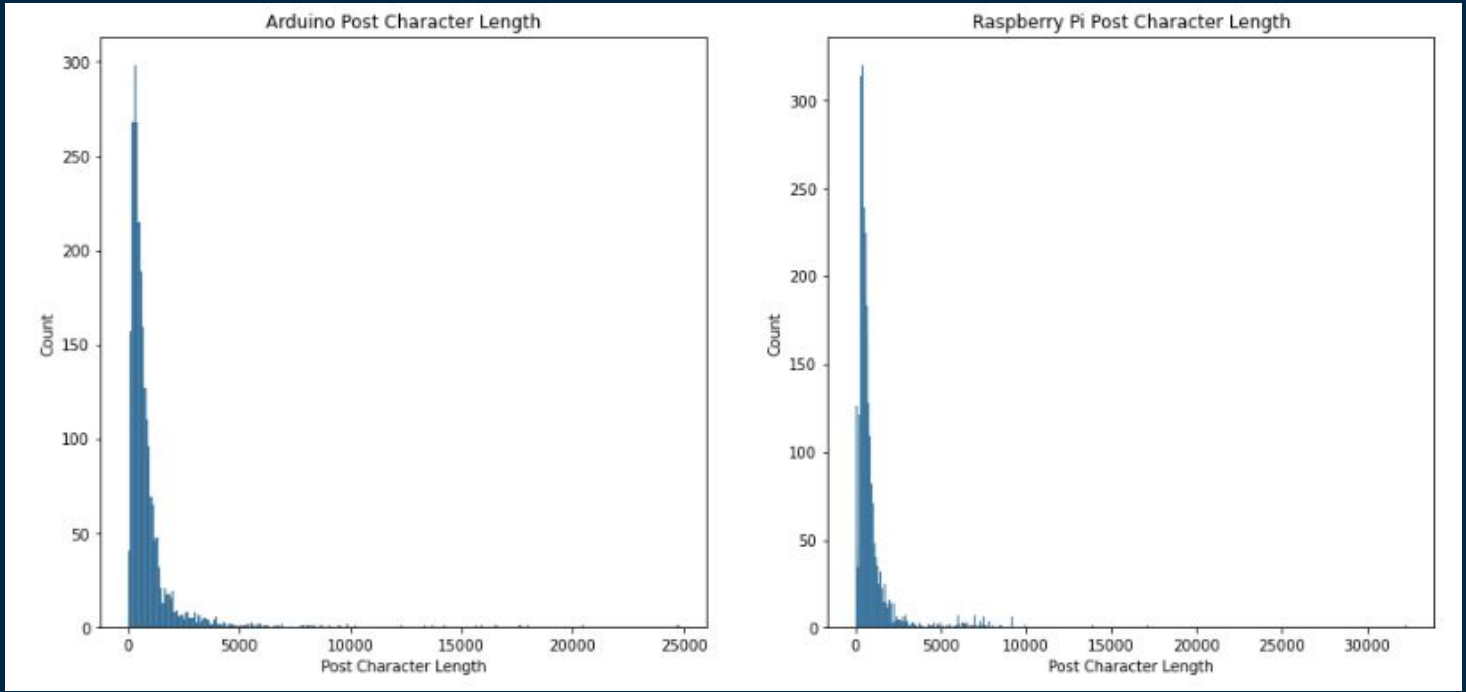
image source

r/arduino

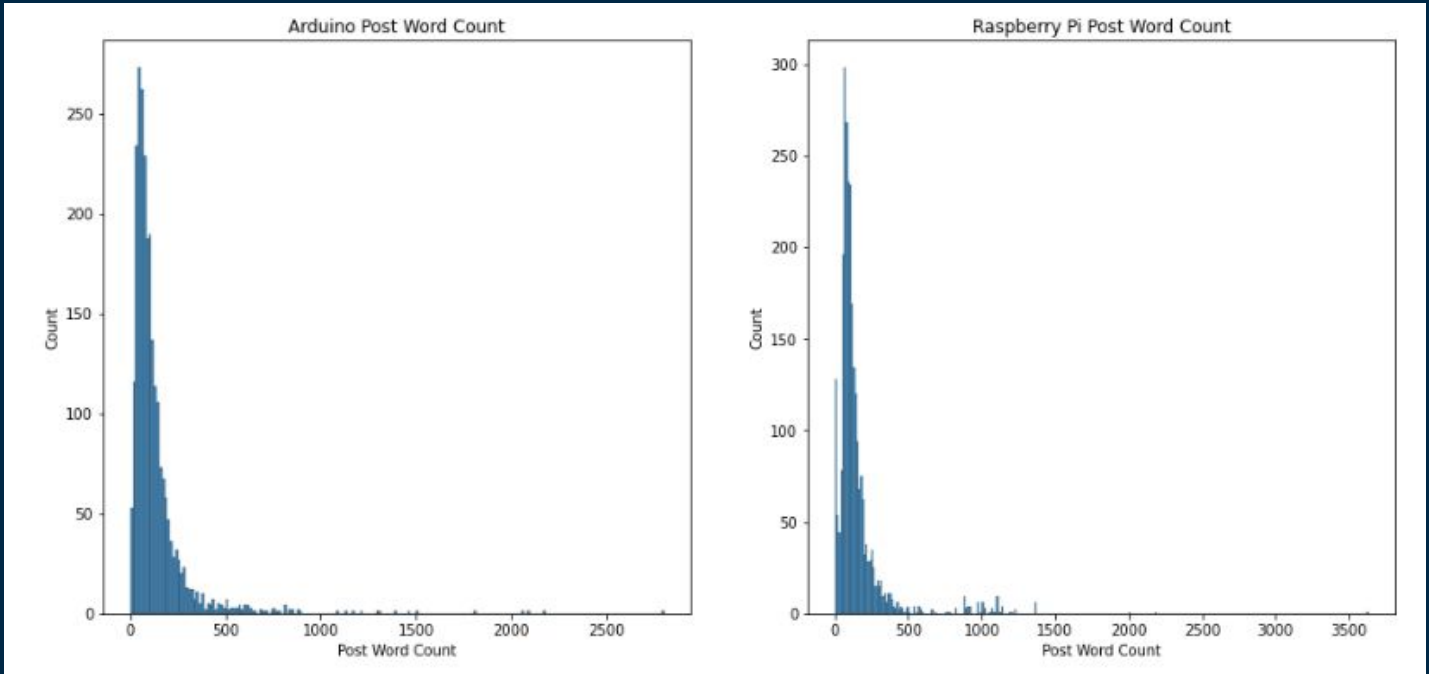
image source

Exploratory Data Analysis

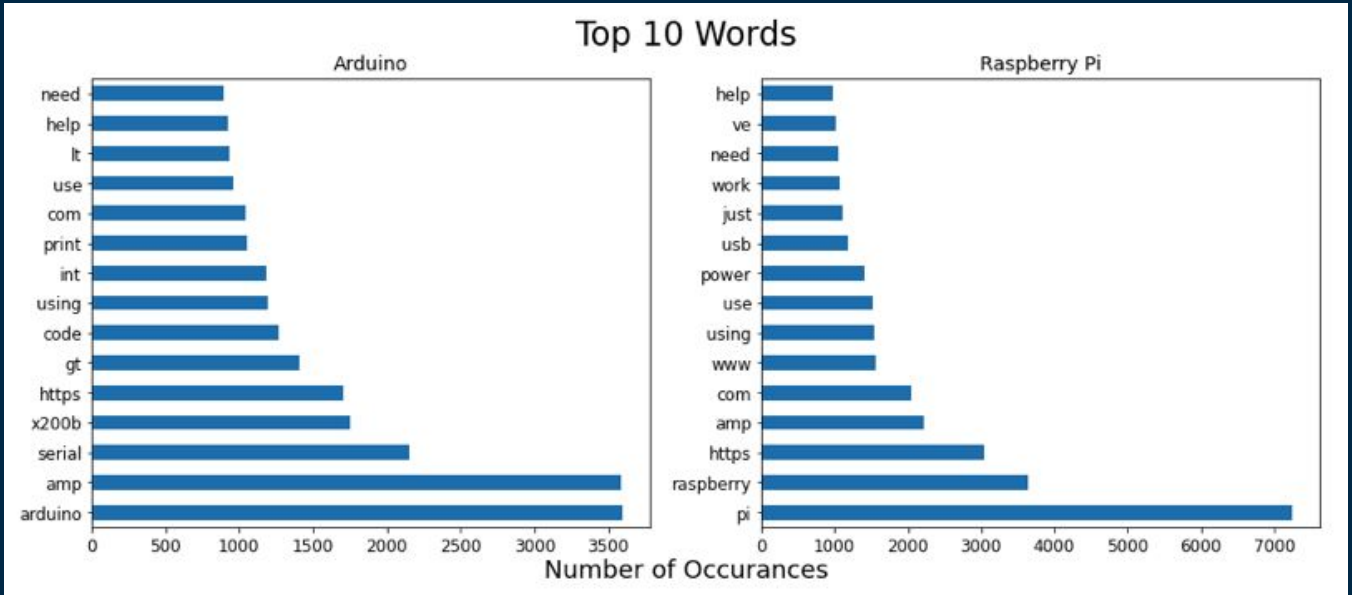
Insights


Arduino Post Character Length
Count
Post Character Length
0
50
100
150
200
250
300
0
5000
10000
15000
20000
25000
Raspberry Pi Post Character Length
Count
Post Character Length
0
50
100
150
200
250
300
0
5000
10000
15000
20000
25000
30000

Insights


Arduino Post Word Count
Count
Post Word Count
0
50
100
150
200
250
0
500
1000
1500
2000
2500
Raspberry Pi Post Word Count
Count
Post Word Count
0
50
100
150
200
250
300
0
500
1000
1500
2000
2500
3000
3500

Insights


Top 10 Words
Arduino
need
help
lt
use
com
print
int
using
code
gt
https
x200b
serial
amp
arduino
0
500
1000
1500
2000
2500
3000
3500
Raspberry Pi
help
ve
need
work
just
usb
power
use
using
www
com
amp
https
raspberry
pi
0
1000
2000
3000
4000
5000
6000
7000
Number of Occurances

Model Workflow

51%

Baseline Accuracy

Model 1: Count Vectorizer

Random Forest Classifier

Best Parameters

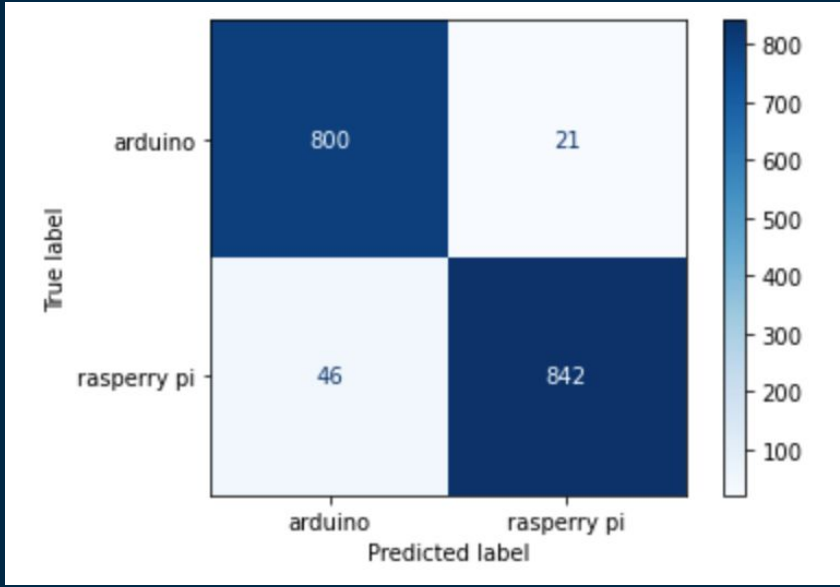
- Cvec max df: 0.9
- Cvec max features: 2000
- Cvec min df: 3
- Cvec ngram range: (1, 1)
- Rf max depth: None
- Rf max features: 'sqrt'
- Rf n estimators: 100
- Best Score: 95.39%

Metrics

Accuracy Score

96.08%

Confusion Matrix



Total Errors

67

Model 2: Count Vectorizer

K-Nearest Neighbors Classifier

Best Parameters

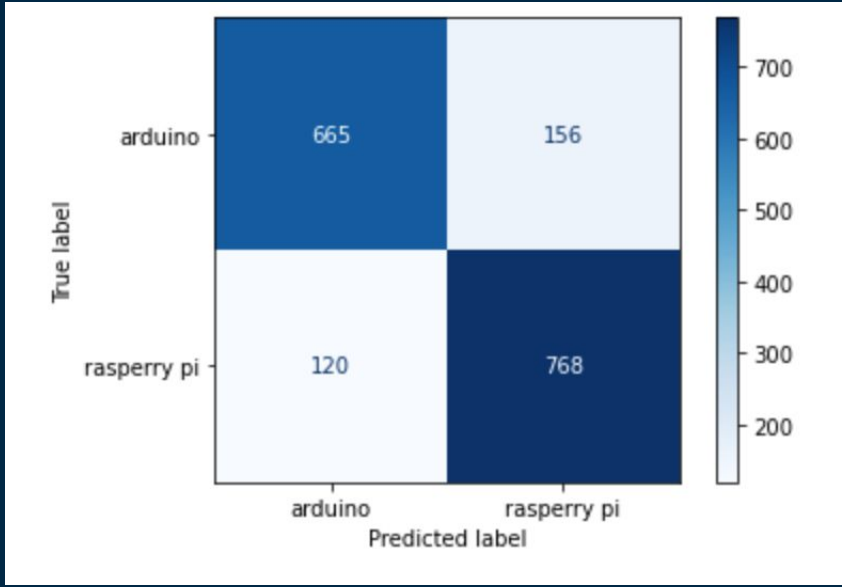
- Cvec max df: 0.9
- Cvec max features: 2000
- Cvec min df: 2
- Cvec ngram range: (1, 2)
- Knn n neighbors: 5
- Knn p: 2
- Knn weights: distance
- Best Score: 81.43%

Metrics

Accuracy Score

83.85%

Confusion Matrix



Total Errors

276

Model 3: TF-IDF Vectorizer

Random Forest Classifier

Best Parameters

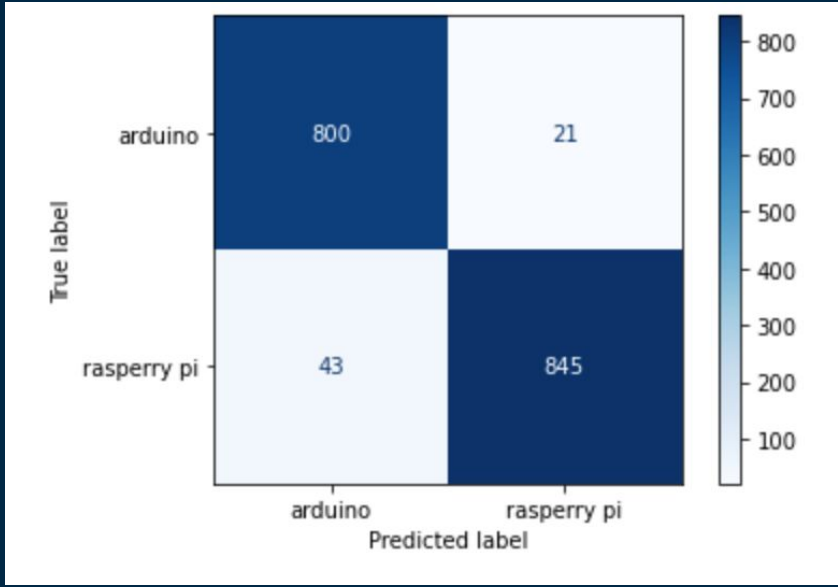
- Tvec max df: 0.9
- Tvec max features: 3000
- Tvec min df: 3
- Tvec ngram range: (1, 1)
- Rf max depth: None
- Rf max features: 'sqrt'
- Rf n estimators: 100
- Best Score: 95.27%

Metrics

Accuracy Score

96.26%

Confusion Matrix



Total Errors

64

Model 4: TF-IDF Vectorizer

K-Nearest Neighbors Classifier

Best Parameters

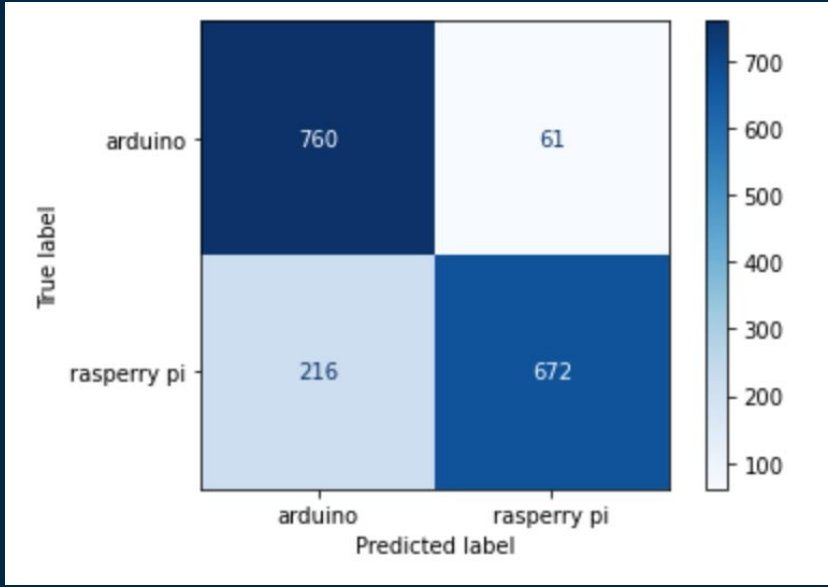
- Tvec max df: 0.9
- Tvec max features: 3000
- Tvec min df: 2
- Tvec ngram range: (1, 1)
- Knn n neighbors: 5
- Knn p: 2
- Knn weights: distance
- Best Score: 84.16%

Metrics

Accuracy Score

83.80%

Confusion Matrix



Total Errors

277

Recommendation

Random Forest Classifier: TF-IDF Vectorizer

Best Parameters

- Tvec max df: 0.9
- Tvec max features: 3000
- Tvec min df: 3
- Tvec ngram range: (1, 1)
- Rf max depth: None
- Rf max features: 'sqrt'
- Rf n estimators: 100
- Best Score: 95.27%

Metrics

Accuracy Score

96.26%

Further Investigations

- AdaBoost
- Gradient Boost
- XGBoost

Questions?