Learn Programming

Learn Machine Learning

Classification Modelling

Emily Kong

What exactly is Reddit and SubReddit?

In short, Reddit is the entire site of Reddit.com.

However, people were allowed to create their own reddit, which over time, was commonly called the SubReddits.

SubReddits is a specific online community, dedicated to a particular topic, denoted by /r/, followed by the subreddit's name.

For this presentation, posts from /r/learnmachinelearning and /r/learnprogramming are scraped.

Problem Statement

Select a classification model which best identifies posts from the subreddit Learn Programming and Learn Machine Learning from the various models selected for testing.

Business Objective

The application of Natural Language Processing Models ('NLP') to correctly classify the post contents to the named subreddits based on the words used most and related to respective subreddit. This will allow for more accurate search results of the related posts based on the keywords entered by users.

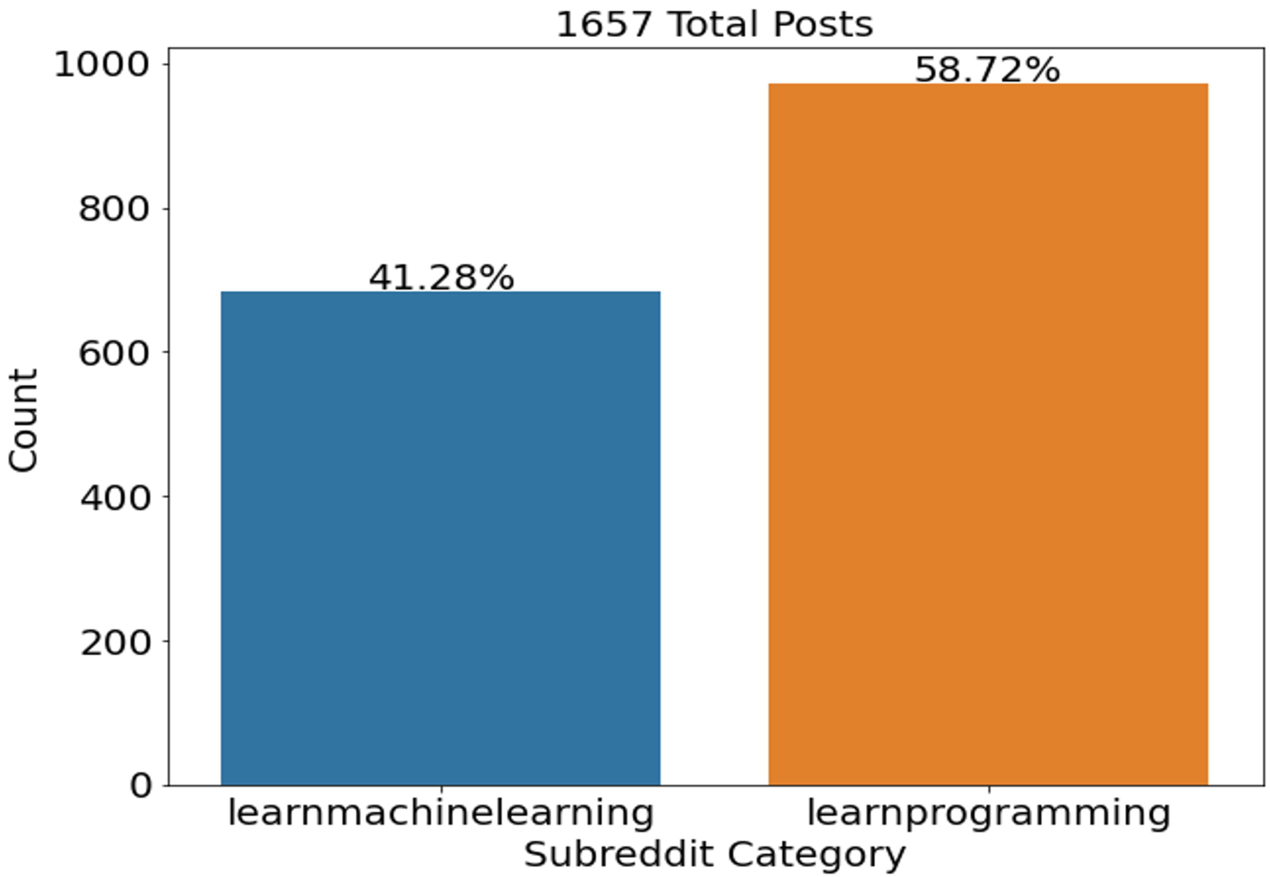
Data Gathering and Cleaning of Data

The process includes the following steps:

- **Web Scraping using Requests Library**
- **Data Cleaning using the following libaries:-**
    - Beautiful Soup
    - NLTK (Stopwords)
    - Regex
    - Python’s string manipulation

Exploratory Data Analysis

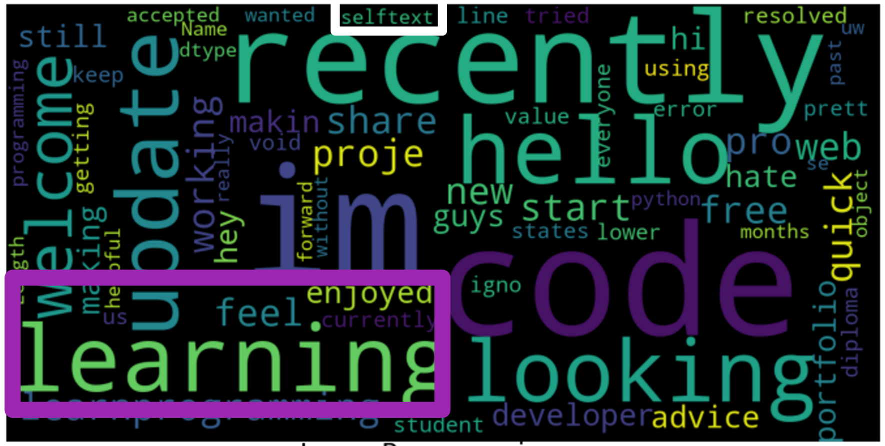
Exploratory Data Analysis



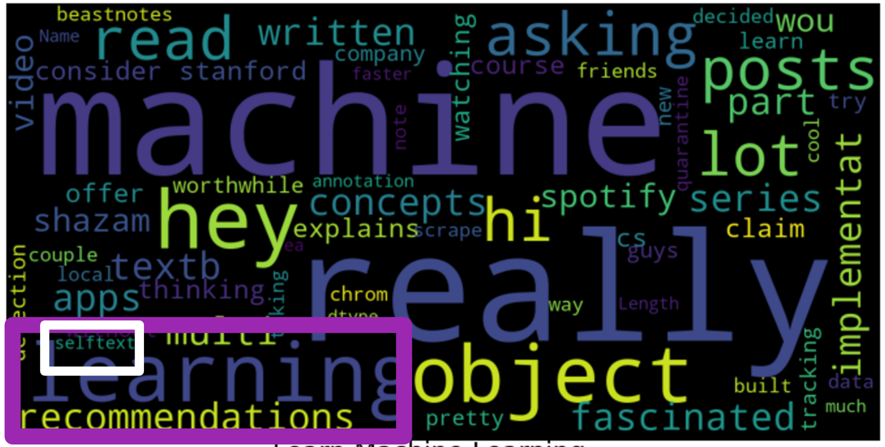
Baseline Accuracy:

0.5867

Exploratory Data Analysis



Learn Programming

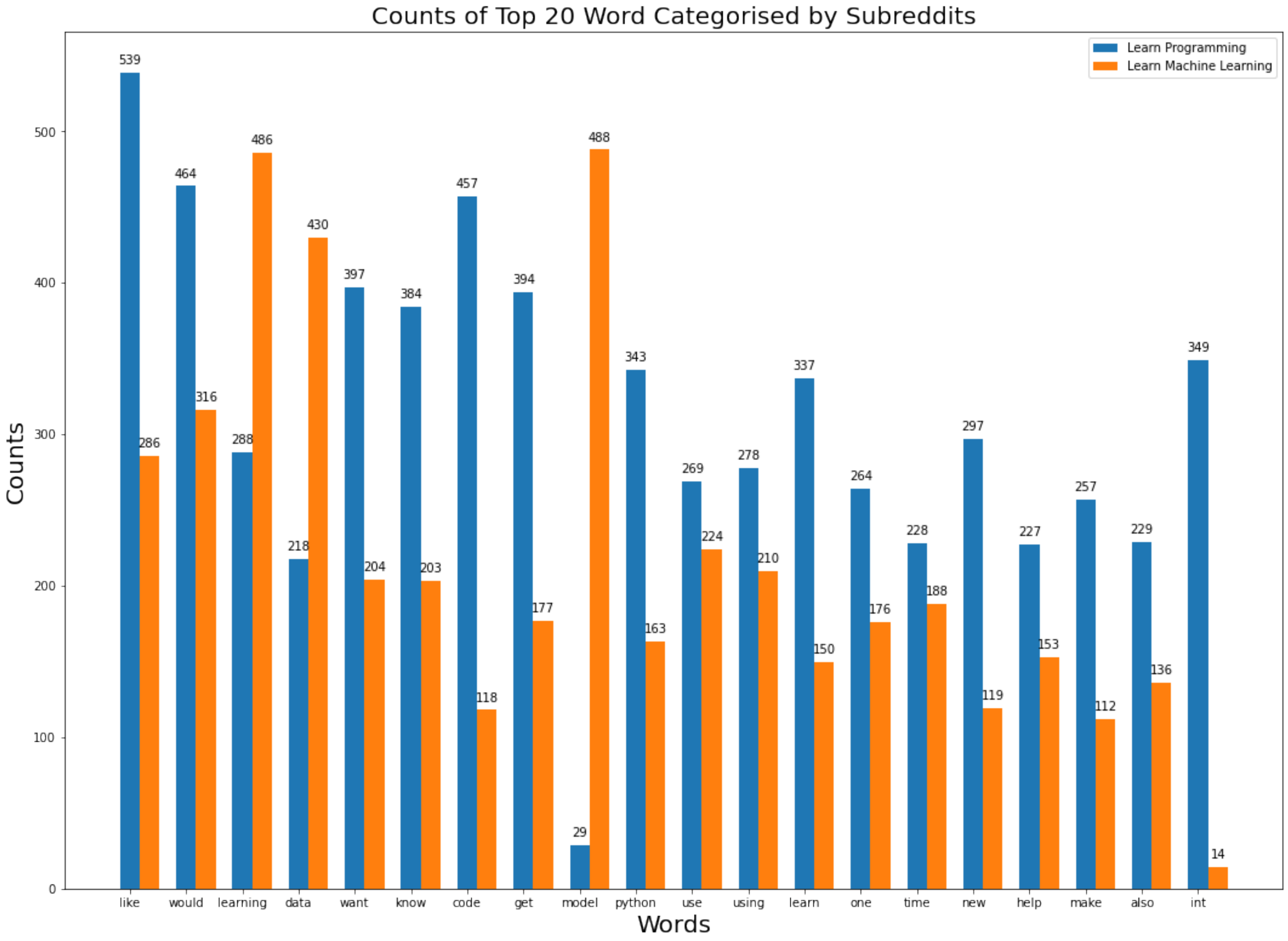


Learn Machine Learning

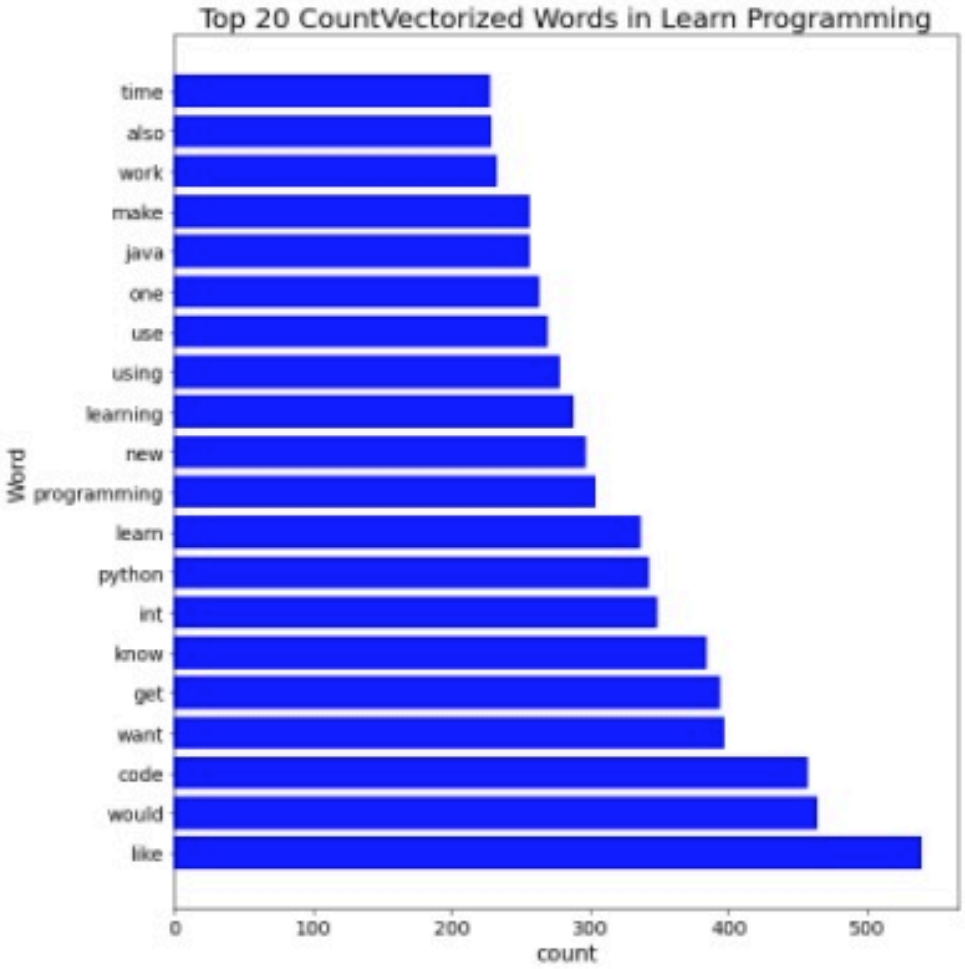
- "Learning" and "Selftext" are some of the common words which appears frequently in both subreddits.

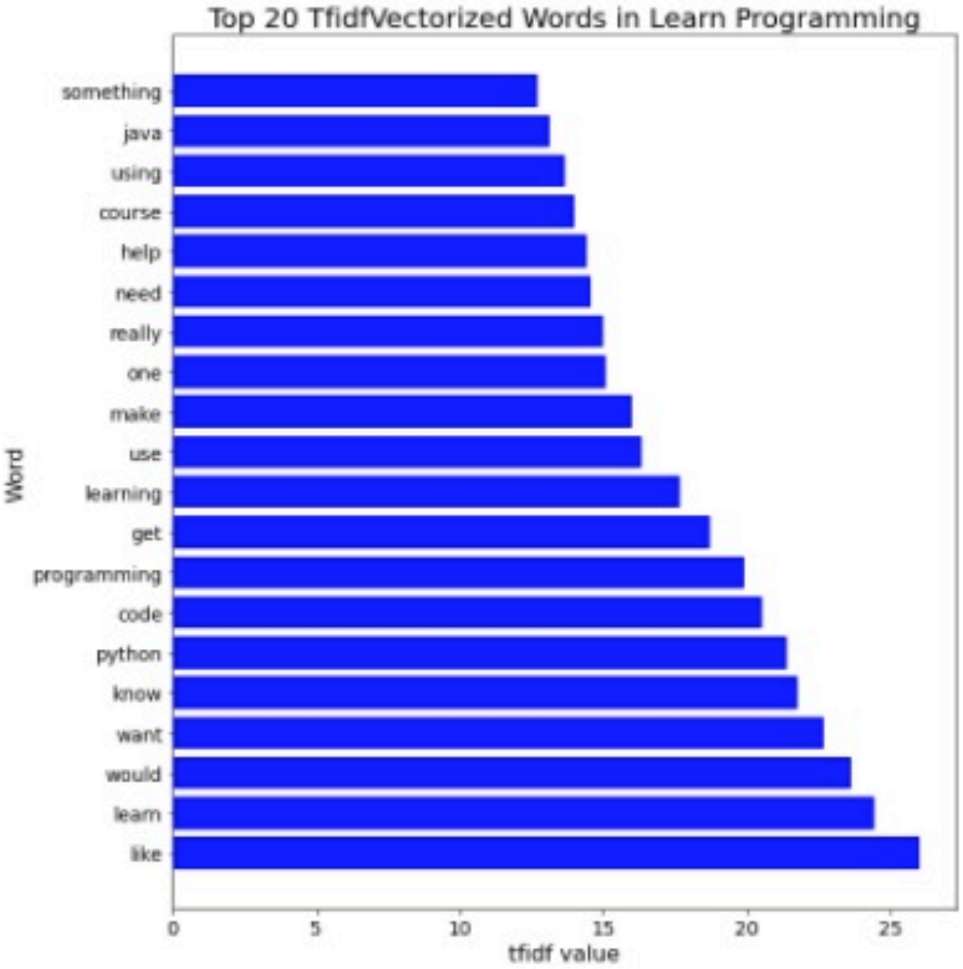
- As the common words may affect the classification of the posts, we will explore these further.

Exploratory Data Analysis

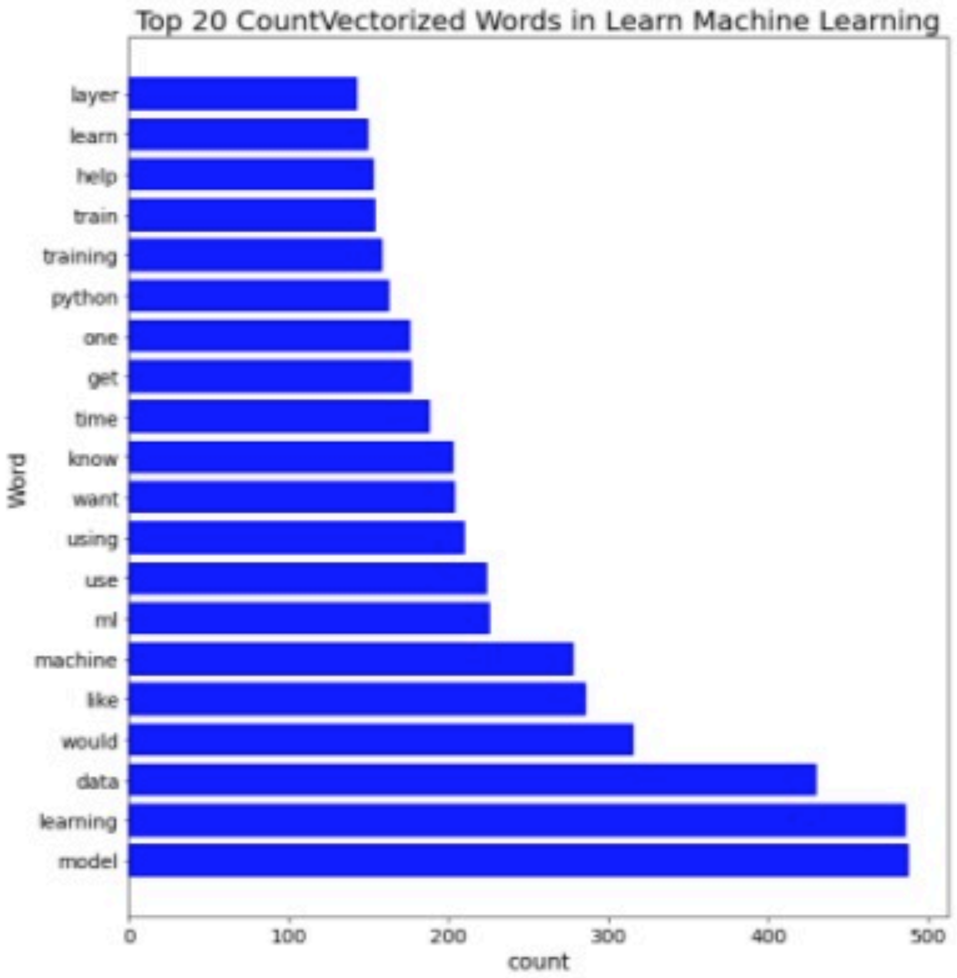


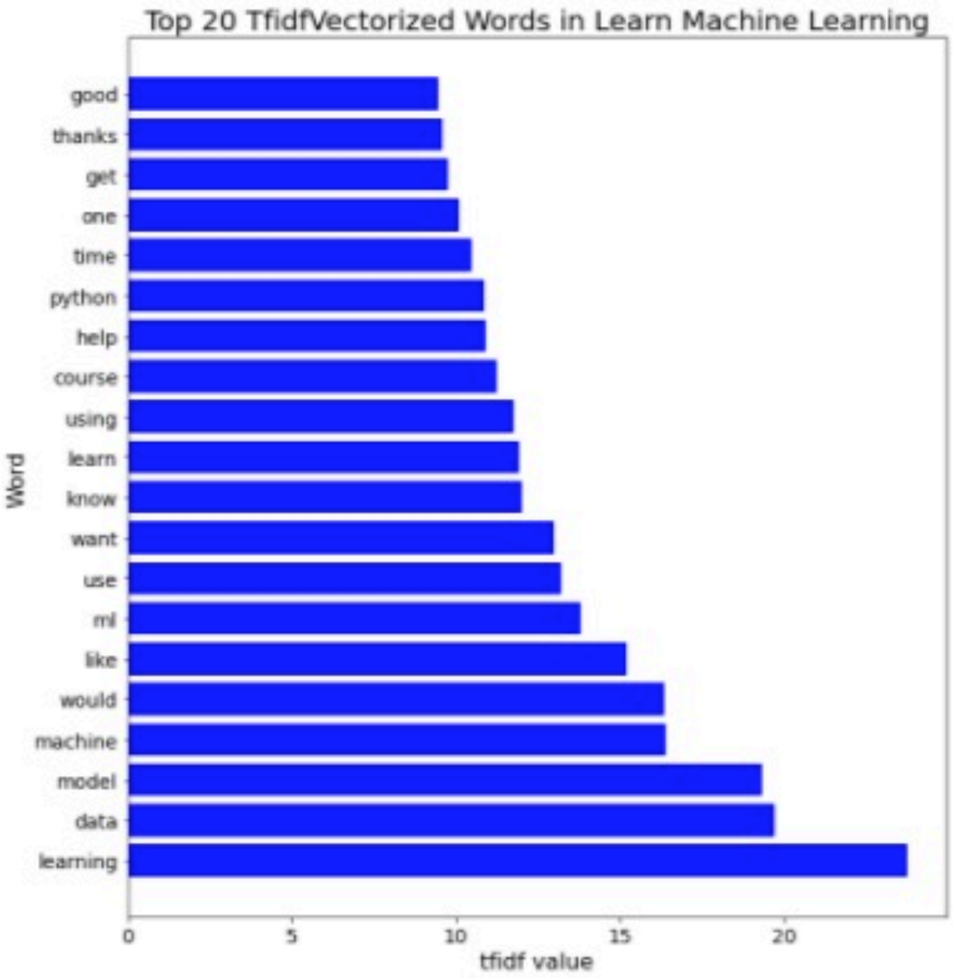
Exploratory Data Analysis





Exploratory Data Analysis





Features Engineering

Features Engineering

- Stemming

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language

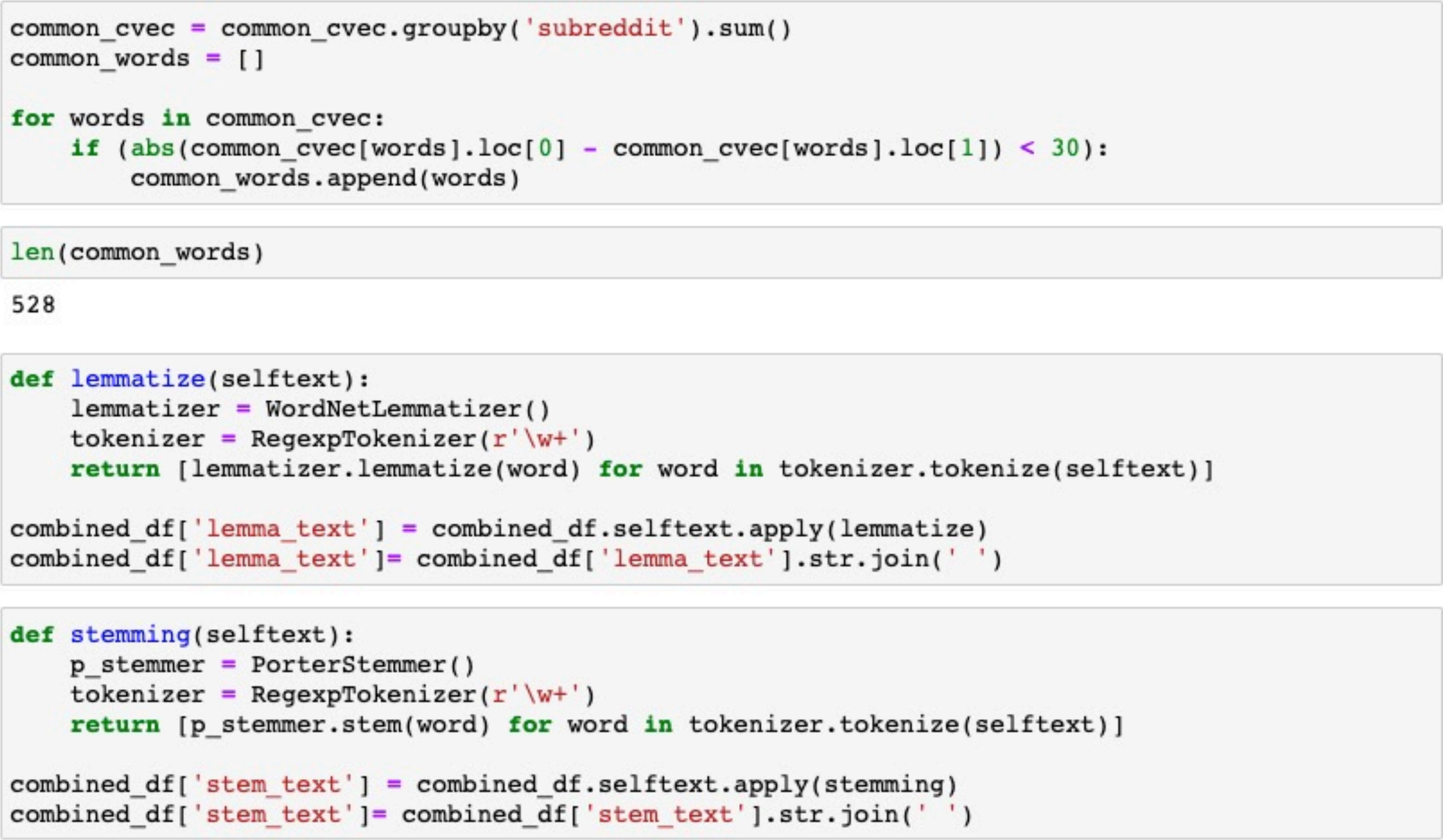
- Lemmatization

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language

- Stop Words

Stop Words are words which do not contain important significance to be used in Search Queries. Usually, these words are filtered out from search queries because they return a vast amount of unnecessary information. Mostly they are words that are commonly used in the English language such as 'as, the, be, are' etc.

Features Engineering



```
common_cvec = common_cvec.groupby('subreddit').sum()
common_words = []

for words in common_cvec:
    if (abs(common_cvec[words].loc[0] - common_cvec[words].loc[1]) < 30):
        common_words.append(words)
```

```
len(common_words)
```

```
528
```

```
def lemmatize(selftext):
    lemmatizer = WordNetLemmatizer()
    tokenizer = RegexpTokenizer(r'\w+')
    return [lemmatizer.lemmatize(word) for word in tokenizer.tokenize(selftext)]

combined_df['lemma_text'] = combined_df.selftext.apply(lemmatize)
combined_df['lemma_text']= combined_df['lemma_text'].str.join(' ')
```

```
def stemming(selftext):
    p_stemmer = PorterStemmer()
    tokenizer = RegexpTokenizer(r'\w+')
    return [p_stemmer.stem(word) for word in tokenizer.tokenize(selftext)]

combined_df['stem_text'] = combined_df.selftext.apply(stemming)
combined_df['stem_text']= combined_df['stem_text'].str.join(' ')
```

Modelling

Modelling

- Count Vectorizer with Naïve Bayesian's Multinomial NB
- Tfidf Vectorizer with Naïve Bayesian's Multinomial NB
- Tfidf Vectorizer with Logistic Regression
- Count Vectorizer with Logistic Regression
- Count Vectorizer with Random Forest
- Tfidf Vectorizer with Random Forest

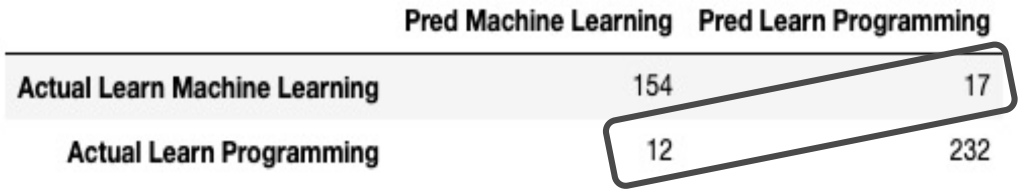
Scores

- Precision can be seen as a measure of exactness or quality. High precision means that an algorithm returned substantially more relevant results than irrelevant ones.

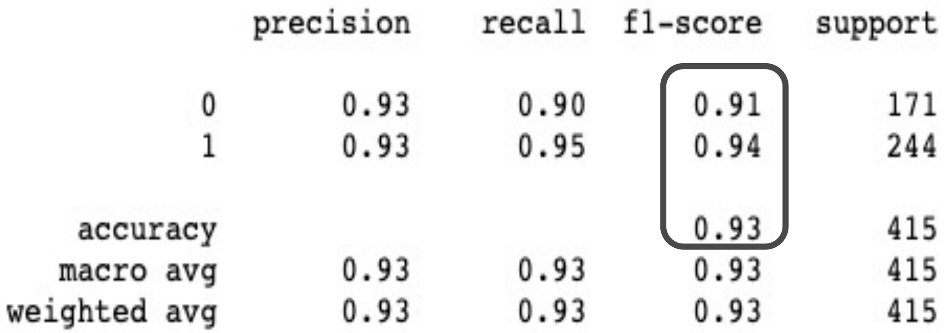
- Recall is a measure of completeness or quantity. High recall means that an algorithm returned most of the relevant results.

- F-Measure provides a single score that balances both the concerns of precision and recall in one number

Model 1

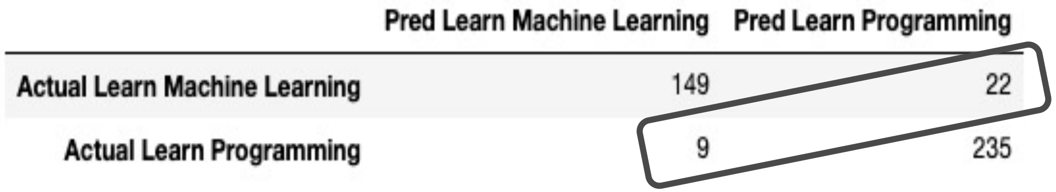


	Pred Machine Learning	Pred Learn Programming
Actual Learn Machine Learning	154	17
Actual Learn Programming	12	232

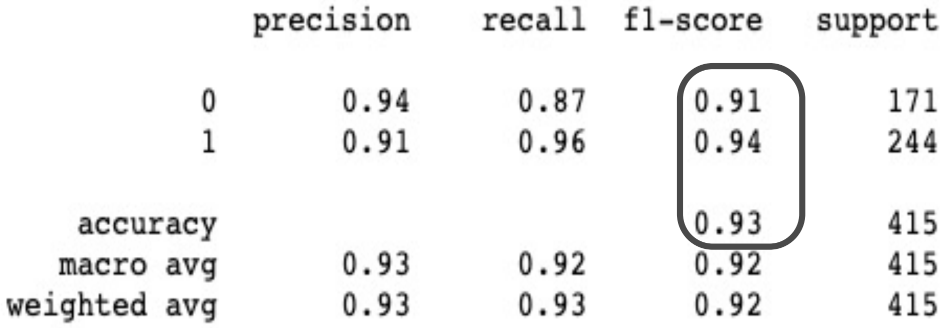


```
              precision    recall  f1-score   support

           0       0.93      0.90      0.91       171
           1       0.93      0.95      0.94       244

    accuracy                           0.93       415
   macro avg       0.93      0.93      0.93       415
weighted avg       0.93      0.93      0.93       415
```

Model 4

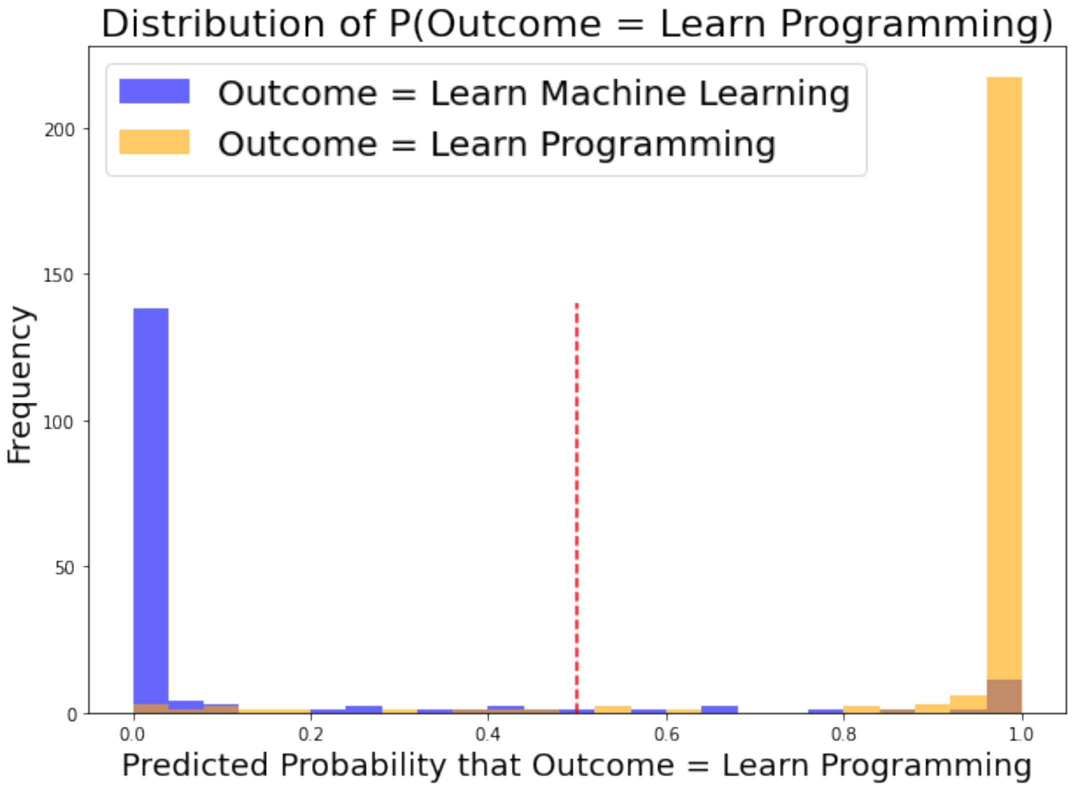


	Pred Learn Machine Learning	Pred Learn Programming
Actual Learn Machine Learning	149	22
Actual Learn Programming	9	235

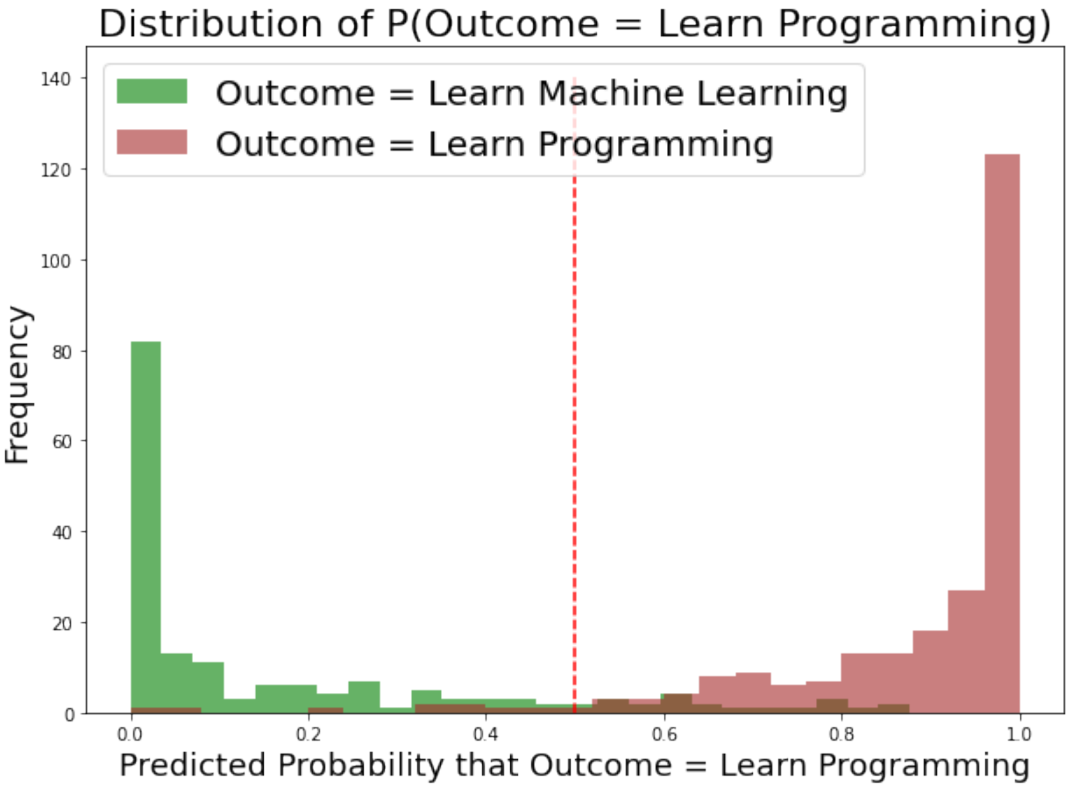


```
              precision    recall  f1-score   support

           0       0.94      0.87      0.91       171
           1       0.91      0.96      0.94       244

    accuracy                           0.93       415
   macro avg       0.93      0.92      0.92       415
weighted avg       0.93      0.93      0.92       415
```

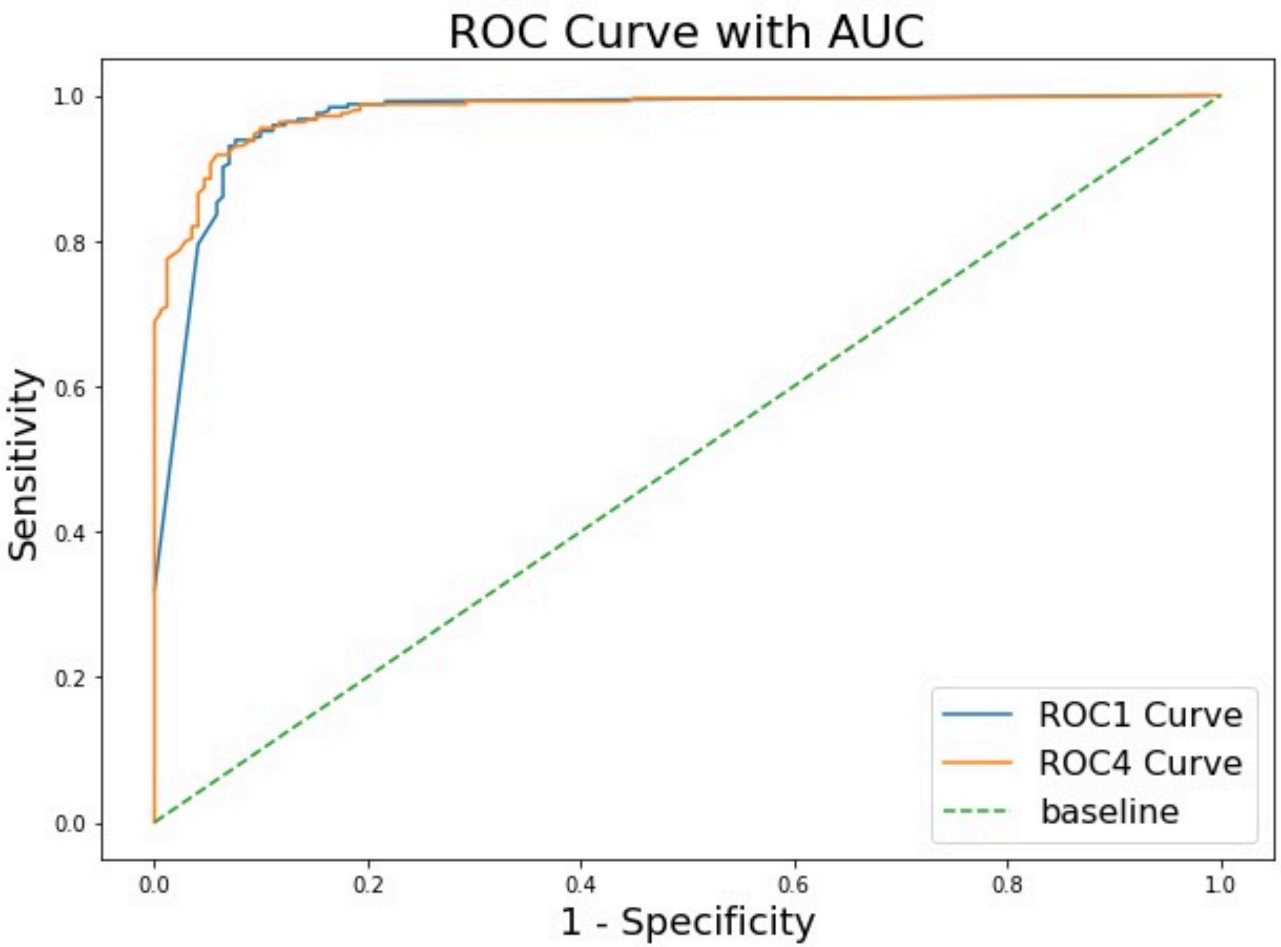
Model 1



Model 4



ROC Curve with AUC



AUC Score for Model 1 = 0.973

AUC Score for Model 4 = 0.978

Conclusions

Observations

- All 6 models have higher cross-validated mean train scores when compared to their respective test scores which could indicate overfitting and the test scores were also in the approximately 0.9 range. This could have been a result of the 2 subreddits being vastly unrelated.

- For the next iteration, the classifier model should be applied on unseen data to validate the scoring and further validated using other subreddits.

- Apply lemmatization and use other stop words to lower the overfitting results