CSE 532 Final Programming Portion
Sima Shafaei
Spring 2023

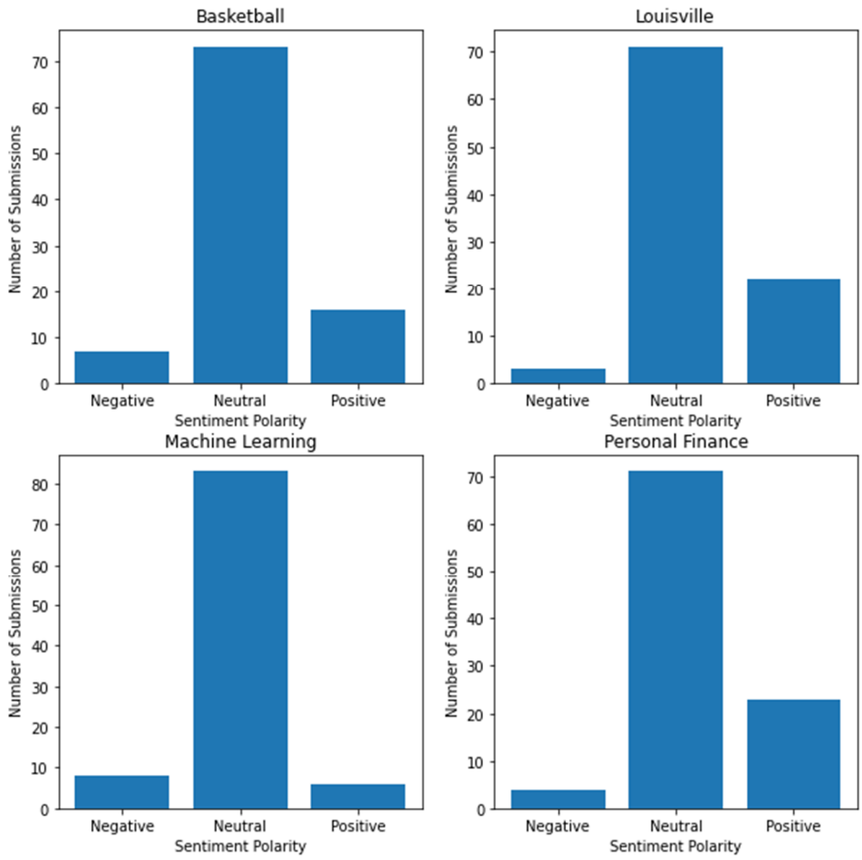
Part 1

After removing the submission records with one or more missing information number of remaining submission in each class is as follows:

- Basketball: 96
- Louisville: 96
- Machine Learning: 97
- Personal Finance: 98

Part 2

The results show that most titles are neutral. However, Louisville and Personal Finance have more positive titles than Basketball and Machine Learning.



Part 3

Created Class Labels are as follows:

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3]
```

Part 4

Sample of obtained features are depicted below:

```
[46.00, -0.20, 1356.00, 249.00, 2020.00]
[21.00, 0.00, 717.00, 103.00, 2022.00]
[10.00, 0.35, 584.00, 13.00, 2017.00]
[25.00, 0.00, 504.00, 17.00, 2019.00]
[68.00, 0.00, 505.00, 26.00, 2019.00]
[24.00, 0.00, 474.00, 15.00, 2019.00]
```

Part 5

I used three distinct classifiers in this part including (KNN, Linear SVM and Gaussian Naïve Bayes).

The accuracy of each classifier are represented below:

```
K Nearest Neighbors  Classifier Accuracy: 0.9328165374677002
Support Vector Machine  Classifier Accuracy: 0.9405684754521964
Gaussian Naive Bayes  Classifier Accuracy: 0.8733850129198967
```

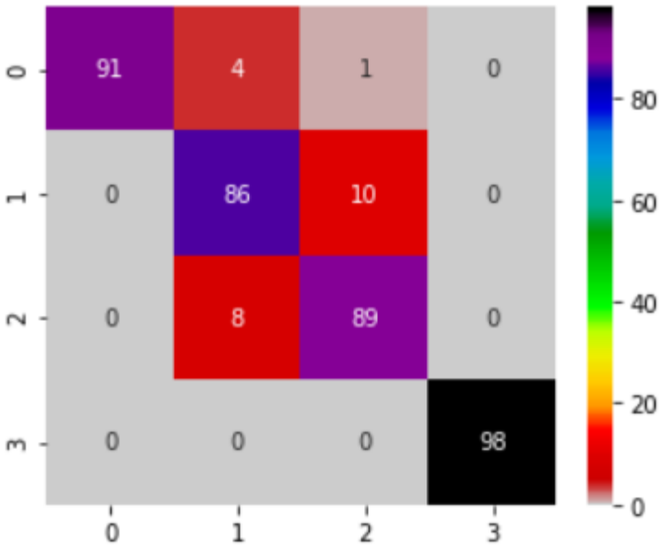
Accuracy Analysis:

Based on the given classifier accuracies, it appears that the Support Vector Machine (SVM) classifier provides the highest accuracy of 0.940, followed by the K Nearest Neighbors (KNN) classifier with an accuracy of 0.933, and the Gaussian Naive Bayes (GNB) classifier with an accuracy of 0.873.

Whether or not these accuracies are acceptable for the task depends on the context of the task at hand and its requirements. I believe that an accuracy of 0.9 or higher is considered acceptable for the reddit submission classification, particularly based on the few number of features we used for classifications.

Confusion Matrix Analysis:

The Confusion Matrix of this Classifier is illustrated in the following picture：



According to the confusion matrix, the Personal Finance class has been classified with 100% accuracy, which indicates high performance of the classification model for this class. However, the Louisville and Machine Learning classes appear to be causing some confusion. Specifically, eight Machine Learning submissions have been incorrectly classified as Louisville, while ten Louisville submissions have been classified as Machine Learning. Additionally, in the Basketball class, four submissions have been erroneously grouped with Louisville, while one submission has been classified as Machine Learning, despite belonging to the Basketball class. No misclassification has been observed for any of the other classes.