IR ASSIGNMENT 5 ANALYSIS

MT19021

1. NAÏVE BAYES ALGORITHM

- The plot having the number of selected features and the accuracy obtained using Naïve Bayes algo for 80:20 train:test split is shown below for the 2 feature selection techniques.
  It can be seen that when the no. of selected features is around 100-150 then the accuracy obtained by mutual independence feature selection is maximum. On further selecting more no. of features the accuracy drops.

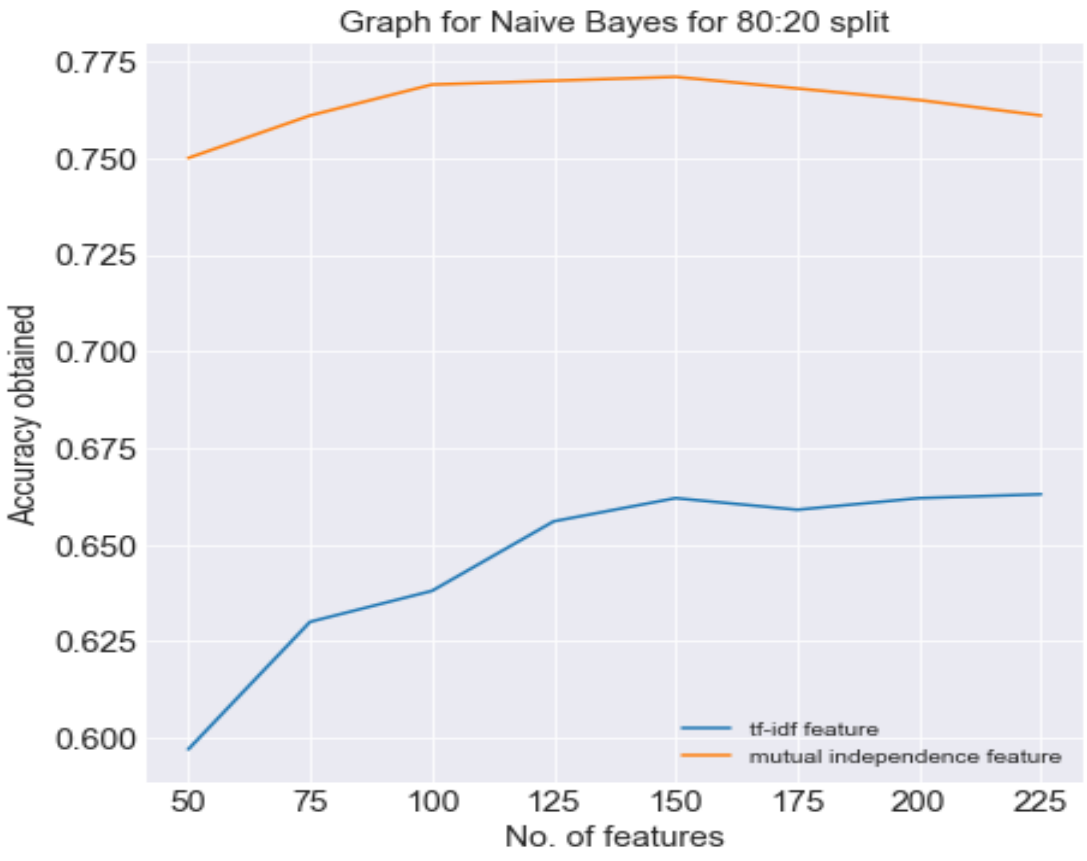

Fig.: Accuracy obtained vs No. of features selected

Therefore, I have selected the no. of selected feature as 100 for both the Naïve Bayes and KNN implementation using both tf-idf and mutual exclusion based feature extraction techniques.

- Now plotting a graph for Naïve Bayes classification with different train:test split ratios on the X-axis and the accuracies obtained on Y-axis.

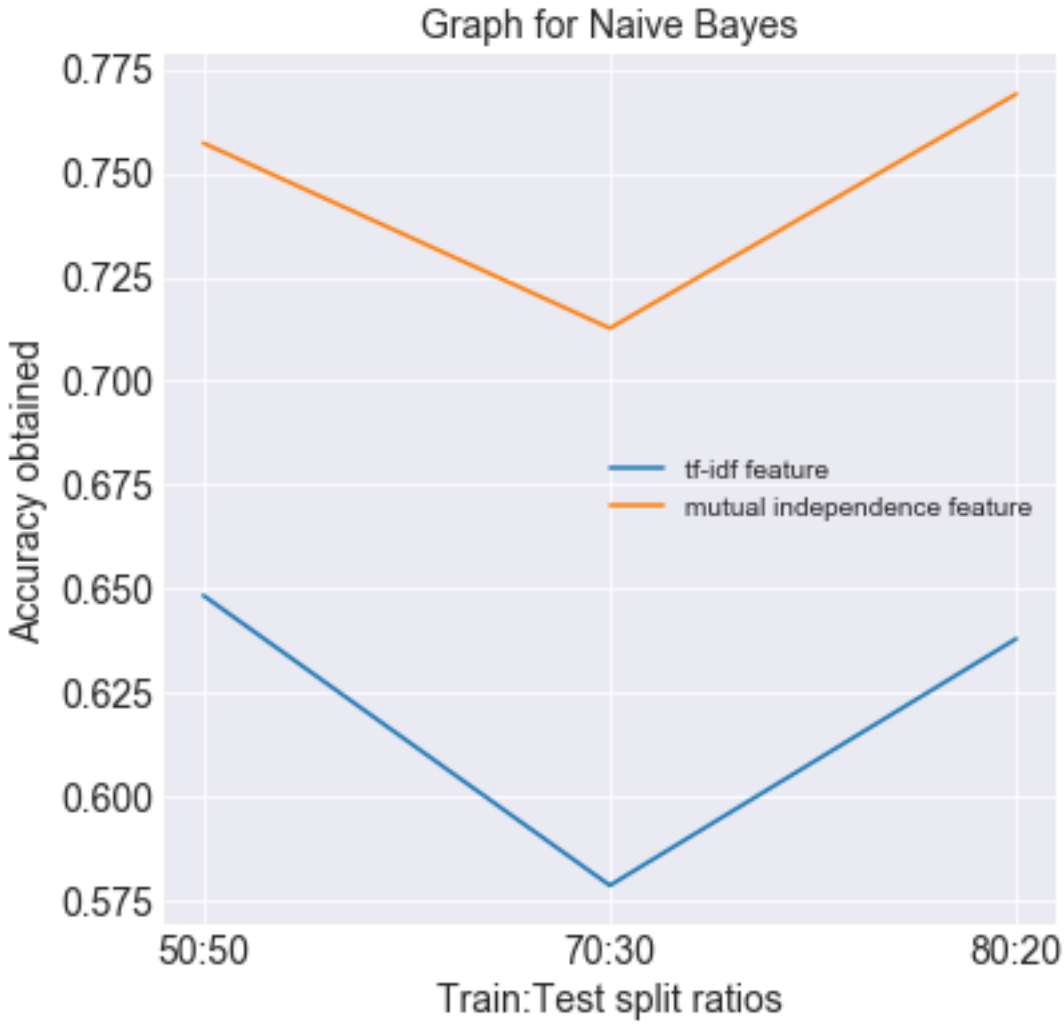

Fig.: Accuracy vs train:test split ratio

Inferences using Naïve Bayes algo implementation:

1. It can be seen from the plot that using tf-idf feature maximum accuracy is obtained for 50:50 split and the minimum accuracy for 70:30 split.
2. Also, it can be seen that on extracting features based on mutual independence maximum accuracy is obtained for 80:20 split and minimum for 70:30 split.
3. The accuracy obtained on extracting features using mutual independence value is quite high as compared to that obtained using tf-idf value.

2. KNN ALGORITHM:

- The graph having the different values of k in KNN algo and the accuracy obtained on the Y-axis is shown below:

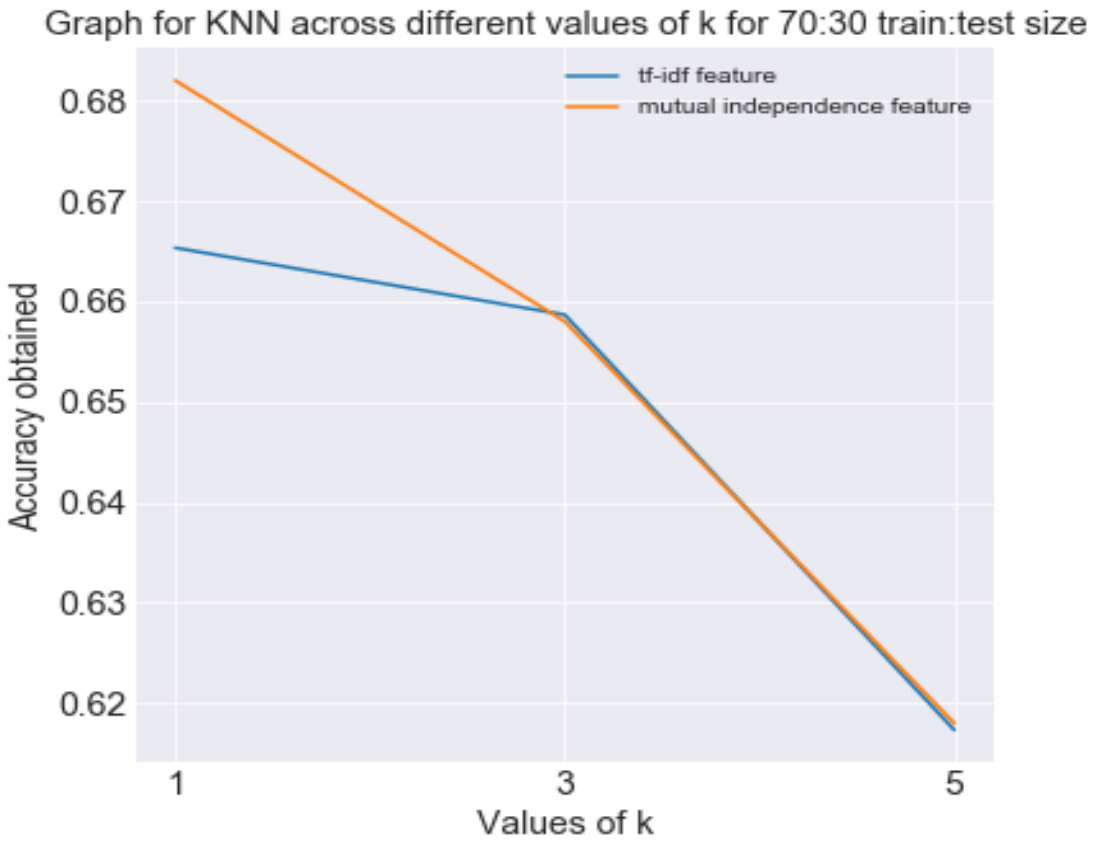

Fig. : Accuracy obtained vs value of k in KNN

The inferences that can be drawn are:

1. It can be seen that using both feature extraction techniques the accuracy obtained on taking k = 1 is maximum and accuracy using k = 5 is minimum.
2. Also, the accuracy obtained on extracting features using mutual information is more than that obtained using tf-idf based extraction for k = 1

- The graph having the different train:test ratio on X-axis and the accuracy obtained on Y-axis for KNN implementation for k = 3 is shown below:

Fig. : Accuracy obtained vs train:test ratio

The inferences that can be drawn from above plot are:

1. For mutual independence based feature extraction maximum accuracy is obtained for 50:50 split and minimum for 80:20 split.
2. For tf-idf based feature extraction maximum accuracy is obtained for 50:50 split and minimum for 80:20 split.

- **<u>PERFORMANCE COMPARISON OF NAÏVE BAYES AND KNN:</u>**

1. **Using tf-idf based feature extraction:**

- The graph having the different split ratio on X-axis and the accuracy obtained on Y-axis for the 2 classifiers is shown below:
  k = 3 in KNN classifier

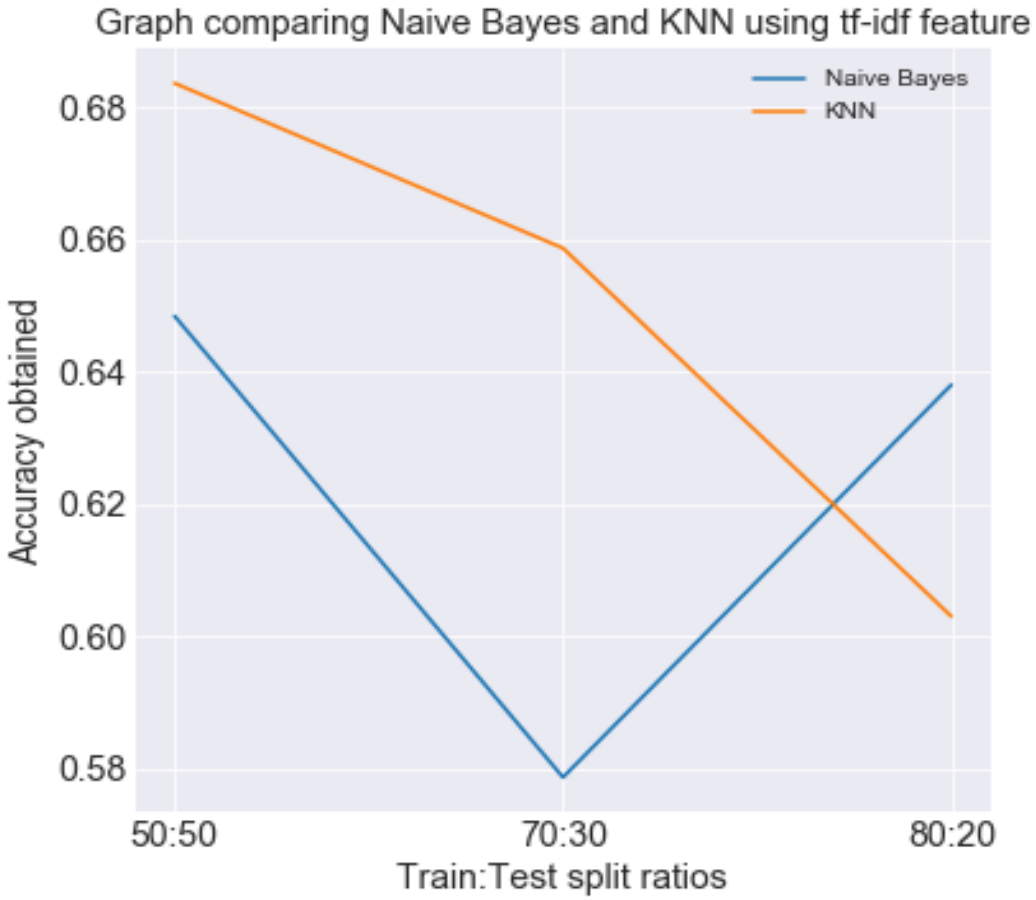

Fig. : Accuracy obtained vs split ratio

Inference: The accuracy obtained using KNN classifier is quite more than that obtained using Naïve Bayes classifier except in case of 80:20 split where is slightly less than that of Naïve Bayes.

So, if features are extracted using the tf-idf values then KNN is a better classifier than Naïve Bayes classifier for the given dataset.

2. **Using mutual independence based feature extraction:**

- The graph having the different split ratio on X-axis and the accuracy obtained on Y-axis is shown below:
  K = 3 in KNN classifier

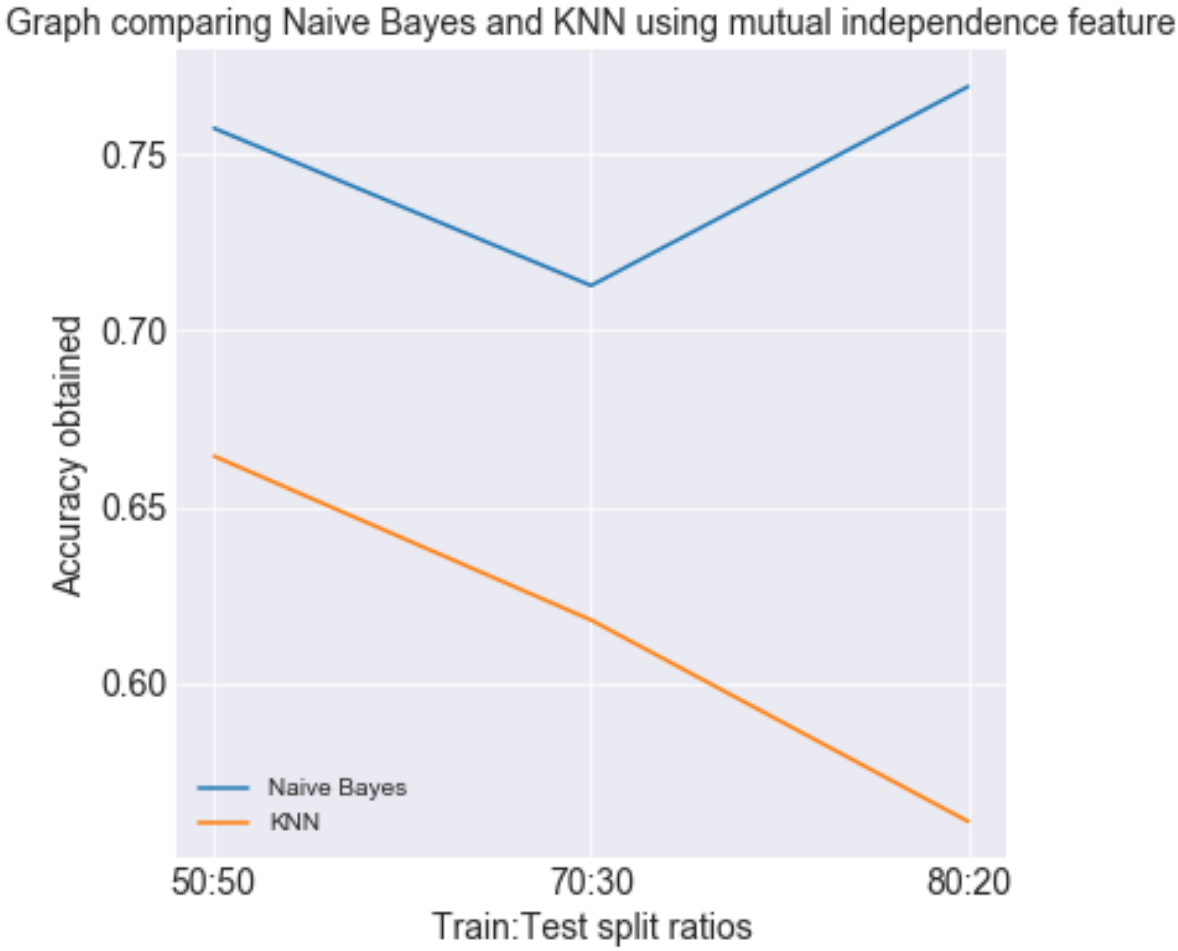

Fig. : Accuracy obtained vs split ratio

Inference: The accuracy obtained using Naïve Bayes classifier is quite more than that obtained using KNN classifier for all split ratios.

So, if features are extracted using the mutual independence value then Naïve Bayes is a better classifier than KNN for the given dataset.

- **CONFUSION MATRIX AND ACCURACY OBTAINED:**

1. **Using Naïve Bayes algorithm for different splits:**

For 50:50 using tf-idf features :
The confusion matrix is:
[[187 0 2 313 0]
[0 398 2 111 1]
[1 0 258 232 5]
[1 0 1 485 2]
[0 2 6 200 293]]
The accuracy score is:
0.6484

For 50:50 using mutual independence features :
The confusion matrix is:
[[288 11 3 200 0]
[0 479 0 31 2]
[3 7 325 155 6]
[0 10 1 477 1]
[5 12 9 151 324]]
The accuracy score is:
0.7572

For 70:30 using tf-idf features :
The confusion matrix is:
[[128 175 0 0 0]
[0 282 0 0 0]
[0 165 137 0 3]
[3 156 1 145 4]
[0 122 2 1 176]]
The accuracy score is:
0.5786666666666667

For 70:30 using mutual independence features :
The confusion matrix is:
[[171 121 0 11 0]
[0 282 0 0 0]
[1 117 176 7 4]

[0 57 2 250 0]
[0 100 3 8 190]]
The accuracy score is:
0.7126666666666667

For 80:20 using tf-idf features :
The confusion matrix is:
[[64 0 149 1 0]
[0 158 47 0 0]
[0 0 178 0 3]
[0 0 93 113 1]
[0 3 65 0 125]]
The accuracy score is:
0.638

For 80:20 using mutual independence features :
The confusion matrix is:
[[106 5 94 8 1]
[0 199 5 1 0]
[1 2 175 2 1]
[0 2 44 160 1]
[0 2 58 4 129]]
The accuracy score is:
0.769

2. Using KNN algorithm for different splits and different k values:

For 50:50 split and k = 1 using tf-idf features :
The confusion matrix is:
[[252 1 233 13 3]
[0 301 208 0 3]
[2 1 476 4 13]
[1 0 147 325 16]
[2 4 128 6 361]]
The accuracy score is:
0.686

For 50:50 split and k = 1 using mutual independence features :
The confusion matrix is:
[[300 4 174 8 16]
[0 137 90 0 285]
[2 3 427 13 51]
[0 3 79 265 142]
[7 5 105 10 374]]
The accuracy score is:
0.6012

For 50:50 split and k = 3 using tf-idf features :
The confusion matrix is:
[[251 1 233 14 3]
[0 298 209 0 5]
[1 0 477 6 12]
[1 0 151 323 14]
[2 4 130 5 360]]
The accuracy score is:
0.6836

For 50:50 split and k = 3 using mutual independence features :
The confusion matrix is:
[[297 4 169 16 16]
[0 345 96 1 70]
[2 3 450 12 29]
[0 8 70 378 33]
[7 8 120 9 357]]
The accuracy score is:

0.7308

For 50:50 split and k = 5 using tf-idf features :
The confusion matrix is:
[[188 0 295 16 3]
 [0 142 365 0 5]
 [0 0 481 5 10]
 [0 0 151 322 16]
 [0 2 132 6 361]]
The accuracy score is:
0.5976

For 50:50 split and k = 5 using mutual independence features :
The confusion matrix is:
[[218 3 247 14 20]
 [0 266 115 1 130]
 [1 2 449 14 30]
 [0 8 70 367 44]
 [6 5 114 15 361]]
The accuracy score is:
0.6644

For 70:30 split and k = 1 using tf-idf features :
The confusion matrix is:
[[179 0 0 0 124]
 [2 148 1 8 123]
 [1 2 184 3 115]
 [8 0 2 191 108]
 [0 1 2 2 296]]
The accuracy score is:
0.6653333333333333

For 70:30 split and k = 1 using mutual independence features :
The confusion matrix is:
[[175 8 0 16 104]
 [0 124 0 117 41]
 [2 4 207 16 76]
 [4 5 5 240 55]
 [0 4 6 14 277]]
The accuracy score is:

0.682

For 70:30 split and k = 3 using tf-idf features :
The confusion matrix is:
[[178 0 0 1 124]
 [2 141 0 17 122]
 [1 2 182 4 116]
 [8 0 1 191 109]
 [0 1 2 2 296]]
The accuracy score is:
0.6586666666666666

For 70:30 split and k = 3 using mutual independence features :
The confusion matrix is:
[[168 9 0 25 101]
 [0 95 0 157 30]
 [1 4 201 26 73]
 [3 5 3 256 42]
 [0 4 3 27 267]]
The accuracy score is:
0.658

For 70:30 split and k = 5 using tf-idf features :
The confusion matrix is:
[[178 0 0 0 125]
 [0 75 0 24 183]
 [1 2 182 4 116]
 [5 0 1 197 106]
 [0 1 2 4 294]]
The accuracy score is:
0.6173333333333333

For 70:30 split and k = 5 using mutual independence features :
The confusion matrix is:
[[172 6 0 21 104]
 [3 45 11 163 60]
 [5 3 204 15 78]
 [15 3 3 245 43]
 [8 2 4 26 261]]
The accuracy score is:

0.618

For 80:20 split and k = 1 using tf-idf features :
The confusion matrix is:
[[114 0 2 3 95]
 [0 110 1 0 94]
 [1 0 102 0 78]
 [1 0 1 124 81]
 [0 3 0 1 189]]
The accuracy score is:
0.639

For 80:20 split and k = 1 using mutual independence features :
The confusion matrix is:
[[124 1 2 7 80]
 [23 13 0 1 168]
 [2 3 69 0 107]
 [11 3 1 62 130]
 [3 3 3 1 183]]
The accuracy score is:
0.451

For 80:20 split and k = 3 using tf-idf features :
The confusion matrix is:
[[121 0 1 0 92]
 [0 138 1 3 63]
 [1 0 79 0 101]
 [1 0 1 76 129]
 [0 3 0 1 189]]
The accuracy score is:
0.603

For 80:20 split and k = 3 using mutual independence features :
The confusion matrix is:
[[129 2 1 3 79]
 [0 71 0 2 132]
 [1 4 63 2 111]
 [0 1 2 72 132]
 [0 1 1 1 190]]
The accuracy score is:

0.525

For 80:20 split and k = 5 using tf-idf features :
The confusion matrix is:
[[102 2 1 0 109]
 [0 110 1 0 94]
 [1 0 38 0 142]
 [0 0 1 56 150]
 [0 3 0 0 190]]
The accuracy score is:
0.496

For 80:20 split and k = 5 using mutual independence features :
The confusion matrix is:
[[102 3 1 9 99]
 [0 107 0 1 97]
 [1 0 55 2 123]
 [0 1 1 109 96]
 [0 1 1 3 188]]
The accuracy score is:
0.561