# ASSIGNMENT 5
# NAME:-Anand Sharma (MT19059)
# Analysis Report

**QUESTION 1:-**

**AIM:-**

1. **Build Naive Bayes Classifiers and KNN classifiers for Text Classification from scratch**
2. **Use TF-TDF and Mutual Inforamtion for Feature Selection**

**Dataset used are:-**

- 20_newsgroups/sci.med/59199
- 20_newsgroups/comp.graphics/38703
- 20_newsgroups/rec.sport.hockey/53700
- 20_newsgroups/talk.politics.misc/178450
- 20_newsgroups/sci.space/60196
- 5000

**Labels:-** {'rec.sport.hockey': 511, 'sci.space': 508, 'sci.med': 504, 'talk.politics.misc': 493, 'comp.graphics': 484}
Total Documents:-5000

**Tools Used:-**

- nltk
- matplotlib
- numpy ,pandas

**TRAIN-TEST SPLIT:-50:50,70:30,80:20**

**Graphs and Observations:-**

## NAIVE BAYES CLASSIFIER

| Classifiers Name | Feature Selection Technique | % of features selected | Train-Test Split | Accuracy | Confusion Matrix |
|---|---|---|---|---|---|
| Naive bayes | TF-IDF | 5% | 50-50 | 97.96 | [[483  11   0   2   4]<br> [  2 486   0   1   2]<br> [  2   1 491   2   1]<br> [  1   1   1 487   4]<br> [  3  11   0   2 502]] |
|  |  | 10% | 50-50 | 98.04 | [[483  11   0   2   4]<br> [  2 487   0   1   1]<br> [  1   1 492   2   1]<br> [  1   1   1 487   4]<br> [  2  12   0   2 502]] |
|  |  | 30% | 50-50 | 98 | [[483  11   0   2   4]<br> [  2 486   0   1   2]<br> [  1   1 492   2   1]<br> [  1   1   1 487   4]<br> [  2  12   0   2 502]] |
|  |  | 50% | 50-50 | 97.92 | [[483  11   0   2   4]<br> [  2 484   0   1   4]<br> [  1   1 492   2   1]<br> [  1   1   1 487   4]<br> [  2  12   0   2 502]] |
|  |  | 100% | 50-50 | 97.92 | [[483  11   0   2   4]<br> [  2 484   0   1   4]<br> [  1   1 492   2   1]<br> [  1   1   1 487   4]<br> [  2  12   0   2 502]] |
|  | No Feature | 100% | 50-50 | 95.28 | [[483  11   0   2   4]<br> [  2 484   0   1   4] |

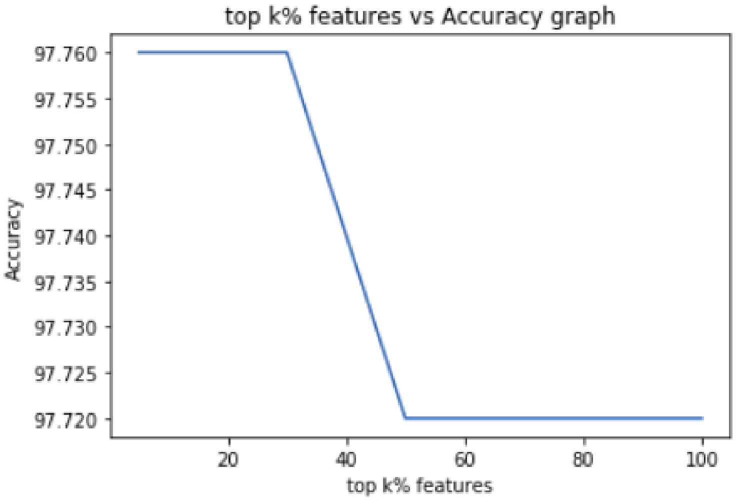
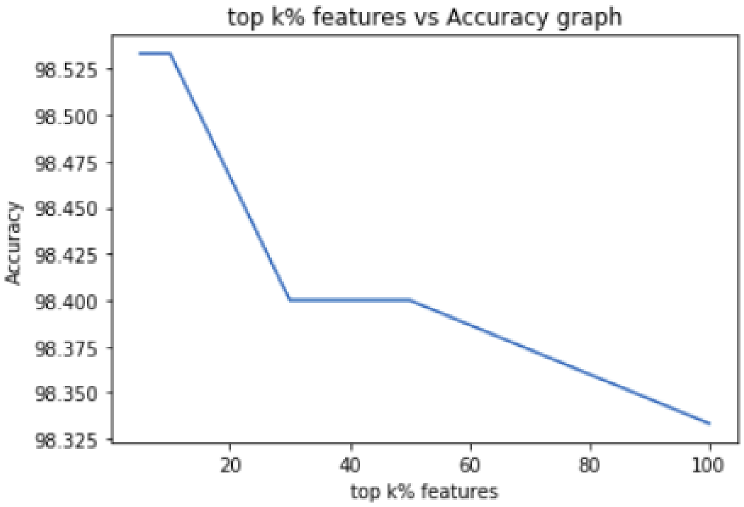
| | | | | | |
|---|---|---|---|---|---|
| | Selection | | | | [ 1 1 492 2 1]<br>[ 1 1 1 487 4]<br>[ 2 12 0 2 502]] |
| Naive Bayes | TF-IDF | 5% | 70-30 | 98.066 | [[312 4 0 1 1]<br>[ 2 283 2 0 3]<br>[ 1 3 304 1 0]<br>[ 2 0 0 282 1]<br>[ 1 4 0 3 290]] |
| | | 10% | 70-30 | 98.133 | [[312 4 0 1 1]<br>[ 2 283 2 0 3]<br>[ 1 2 305 1 0]<br>[ 2 0 0 282 1]<br>[ 1 4 0 3 290]] |
| | | 30% | 70-30 | 98.066 | [[311 4 0 1 2]<br>[ 2 283 2 0 3]<br>[ 1 2 305 1 0]<br>[ 2 0 0 282 1]<br>[ 1 4 0 3 290]] |
| | | 50% | 70-30 | 98.066 | [[311 4 0 1 2]<br>[ 2 283 2 0 3]<br>[ 1 2 305 1 0]<br>[ 2 0 0 282 1]<br>[ 1 4 0 3 290]] |
| | | 100% | 70-30 | 98.0 | [[310 4 0 1 3]<br>[ 2 283 2 0 3]<br>[ 1 2 305 1 0]<br>[ 2 0 0 282 1]<br>[ 1 4 0 3 290]] |
| Naive Bayes | TF-IDF | 5% | 80-20 | 98.3 | [[187 3 0 1 2]<br>[ 2 219 0 0 0]<br>[ 2 1 193 1 0]<br>[ 0 0 0 197 0]<br>[ 2 2 0 1 187]] |
| | | 10% | 80-20 | 98.3 | [[187 3 0 1 2]<br>[ 2 219 0 0 0]<br>[ 2 1 193 1 0]<br>[ 0 0 0 197 0]<br>[ 2 2 0 1 187]] |
| | | 30% | 80-20 | 98.3 | [[187 3 0 1 2]<br>[ 2 219 0 0 0]<br>[ 2 1 193 1 0]<br>[ 0 0 0 197 0]<br>[ 2 2 0 1 187]] |
| | | 50% | 80-20 | 98.2 | [[187 3 0 1 2]<br>[ 2 219 0 0 0] |

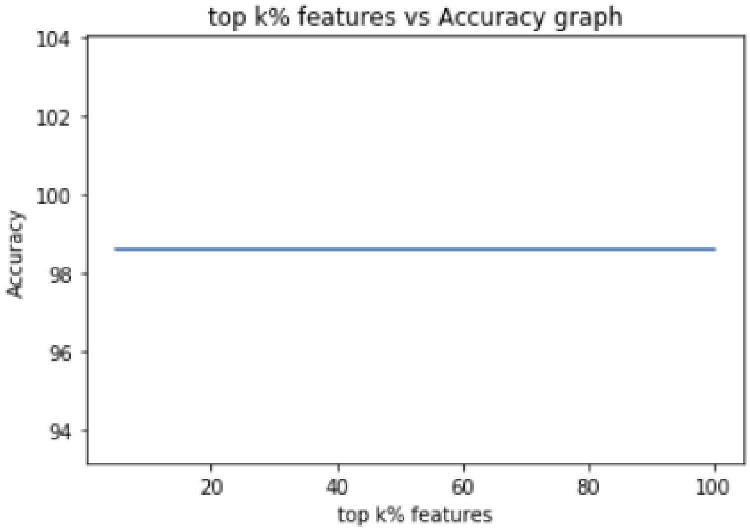
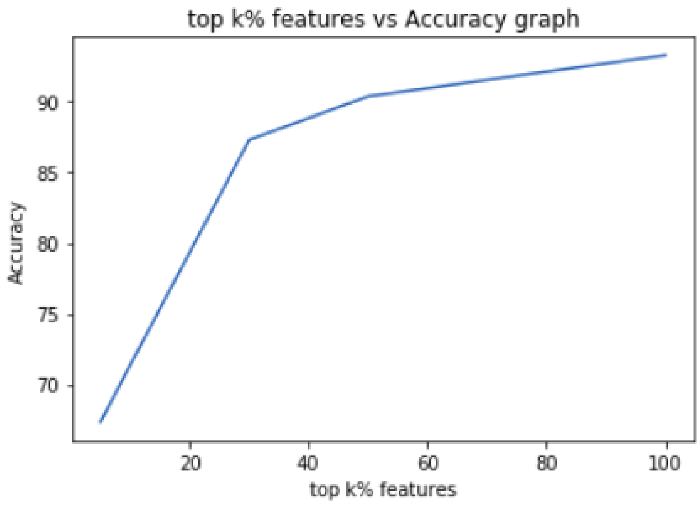
| | | | | | |
|---|---|---|---|---|---|
| | | | | | [ 2 1 193 1 0]<br>[ 0 0 0 197 0]<br>[ 2 2 0 1 187]] |
| | | 100% | 80-20 | 98.2 | [[187 3 0 1 2]<br>[ 2 219 0 0 0]<br>[ 2 1 193 1 0]<br>[ 0 0 0 197 0]<br>[ 2 2 0 1 187]] |
| Naive Bayes | Mutual-Information | 5% | 50-50 | 68.6 | [[284 143 28 33 16]<br>[ 12 425 14 3 28]<br>[ 8 51 418 11 6]<br>[ 92 47 18 365 9]<br>[ 74 126 27 39 223]] |
| | | 10% | 50-50 | 79.6 | [[402 44 15 8 35]<br>[ 18 425 13 3 23]<br>[ 4 9 474 5 2]<br>[139 13 9 308 62]<br>[ 40 44 7 8 390]] |
| | | 30% | 50-50 | 88.92 | [[415 26 9 25 29]<br>[ 21 417 5 10 29]<br>[ 4 8 478 2 2]<br>[ 11 12 9 490 9]<br>[ 20 26 10 10 423]] |
| | | 50% | 50-50 | 91.12 | [[422 32 7 17 26]<br>[ 15 427 12 3 25]<br>[ 5 3 481 5 0]<br>[ 2 5 5 505 14]<br>[ 13 18 4 11 443]] |
| | | 100% | 50-50 | 93.479 | [[464 17 11 4 8]<br>[ 13 446 16 1 6]<br>[ 1 3 489 0 1]<br>[ 1 6 16 504 4]<br>[ 7 36 7 5 434]] |
| Naive Bayes | Mutual-Information | 5% | 70-30 | 80.4 | [[228 45 7 13 27]<br>[ 10 254 6 4 5]<br>[ 6 18 258 14 13]<br>[ 28 11 6 232 21]<br>[ 27 30 0 3 234]] |
| | | 10% | 70-30 | 86.66 | [[266 20 3 14 17]<br>[ 15 243 3 11 7]<br>[ 4 7 284 10 4]<br>[ 16 4 1 266 11]<br>[ 18 22 3 10 241]] |
| | | 30% | 70-30 | 89.33 | [[263 14 3 21 19]<br>[ 9 245 3 8 14]<br>[ 0 4 291 8 6]<br>[ 5 1 1 280 11] |

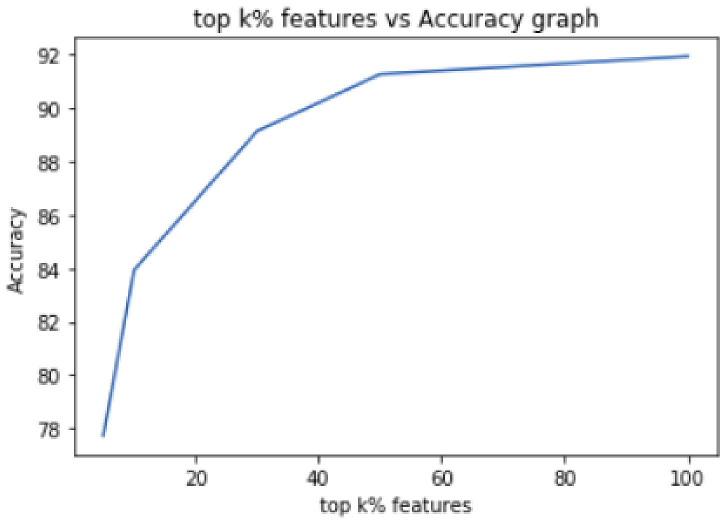
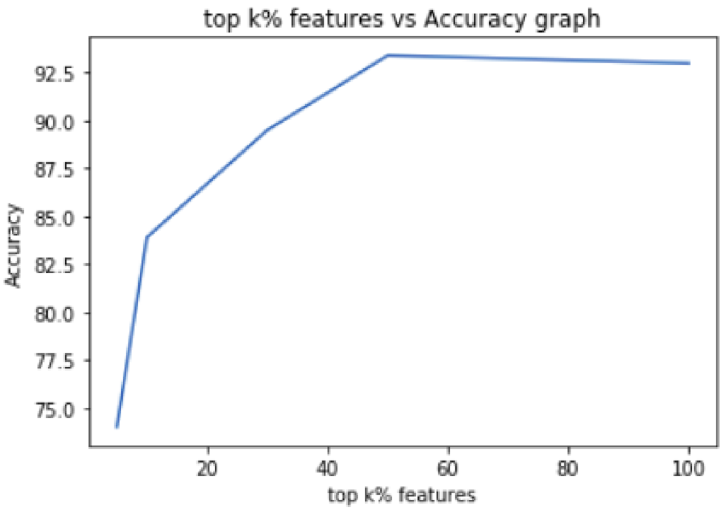
|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  |  |  |  | [   5  15   3  10 261]] |
|  |  | 50% | 70-30 | 89.8 | [[269  16  11  11  13]<br>[ 11 237  10   3  18]<br>[   1   4 304   0   0]<br>[   4   3   6 277   8]<br>[   5  23   3   3 260]] |
|  |  | 100% | 70-30 | 90.933 | [[266  10  21   2  21]<br>[   2 244  15   1  17]<br>[   0   2 307   0   0]<br>[   0   3  24 262   9]<br>[   1   4   4   0 285]] |
| Naive Bayes | Mutual-Information | 5% | 80-20 | 74.0 | [[112  21   4  28  40]<br>[   5 135   3   9  34]<br>[   1   3 169  22  13]<br>[ 11   3   1 187  10]<br>[   2  31   1  18 137]] |
|  |  | 10% | 80-20 | 83.899 | [[156  11   3  18  17]<br>[ 15 146   0   4  21]<br>[   7   1 192   6   2]<br>[ 13   5   4 185   5]<br>[   7  12   0  10 160]] |
|  |  | 30% | 80-20 | 89.5 | [[178   4   3  12   8]<br>[ 18 151   2   1  14]<br>[   0   0 206   1   1]<br>[   6   1   3 201   1]<br>[   5   7   4  14 159]] |
|  |  | 50% | 80-20 | 93.4 | [[181   5   4   6   9]<br>[ 11 165   1   1   8]<br>[   0   0 205   2   1]<br>[   2   1   4 203   2]<br>[   2   2   1   4 180]] |
|  |  | 100% | 80-20 | 93.0 | [[182   6  10   1   6]<br>[   3 173   7   0   3]<br>[   0   1 206   1   0]<br>[   0   0  19 190   3]<br>[   1   4   5   0 179]] |

## GRAPHS OF ACCURACY  FOR NAIVE BAYES

| **Train-Test Split** | **Feature Selection Technique** | **Accuracy vs top k% features** |
|---|---|---|

| | | |
|---|---|---|
| 50-50 | TF-IDF |  |
| 70-50 | TF-IDF |  |

| | | |
|---|---|---|
| 80-50 | TF-IDF |  |
| 50-50 | MI |  |

|  |  |  |
|---|---|---|
| 70-50 | MI |  |
| 80-50 | MI |  |

## KNN Classifiers

| Classifiers Name | Feature Selection Technique | % of features selected | Train-Test Split | Accuracy | Confusion Matrix |
|---|---|---|---|---|---|
| K=1 | TF-IDF | 1 | 50-50 | 88.24 | [[429  41   8  17  16]<br>[ 23 454   5   1   6]<br>[ 10   9 482   4   2]<br>[ 12   6   2 474   4]<br>[ 81  18  16  13 367]] |
| | TF-IDF | 1 | 70-30 | 89.266 | [[251  19  10   9   7]<br>[ 17 282   1   2   6]<br>[  2   3 297   0   1]<br>[ 10   6   7 284   1]<br>[ 37  16   5   2 225]] |
| | TF-IDF | 1 | 80-20 | 86.4 | [[177  13   8   4   7]<br>[ 10 185   3   1   2]<br>[  1   5 198   1   0]<br>[  9   9   1 157   3]<br>[ 32  17   7   3 147]] |
| | TF-IDF | 5 | 70-30 | 84.533 | [[260  26  13  10   1]<br>[ 20 270   8   2   4]<br>[  7  21 283   0   3]<br>[ 12  11  17 253   3]<br>[ 29  31   6   8 202]] |
| | MI | 1 | 50-50 | 89.2 | [[427  33   8   5  41]<br>[ 22 452   3   3  14]<br>[ 11   4 484   1   4]<br>[ 10  15   4 466   3]<br>[ 43  17   6   5 419]] |
| | MI | 1 | 70-30 | 88.533 | [[271  17   4   2  15]<br>[ 18 274   3   3   7]<br>[  6   5 283   1   2]<br>[ 14  11   2 271   8]<br>[ 34  12   7   1 229]] |
| | MI | 1 | 80-20 | 90.8 | [[175   7   0   3   8]<br>[ 10 188   4   3   2]<br>[  4   2 210   0   0]<br>[  3   4   2 186   2]<br>[ 27   8   3   0 149]] |
| | MI | 5 | 70-30 | 85.133 | [[265  26   9   2   7]<br>[ 16 276   7   1   5] |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | [ 11 6 278 1 1]<br>[ 20 18 9 253 6]<br>[ 44 26 6 2 205]] |
| K=3 | TF-IDF | 1 | 50-50 | 82.0 | [[266 93 29 37 95]<br>[ 0 424 30 15 36]<br>[ 1 2 470 13 16]<br>[ 0 4 4 438 28]<br>[ 10 15 11 11 452]] |
| | TF-IDF | 1 | 70-30 | 85.33 | [169 54 23 20 30]<br>[ 1 262 18 7 20]<br>[ 0 0 293 6 4]<br>[ 1 0 2 296 9]<br>[ 5 10 8 2 260]] |
| | TF-IDF | 1 | 80-20 | 84.2 | [[117 38 20 10 24]<br>[ 0 173 20 2 6]<br>[ 0 0 202 3 0]<br>[ 0 0 1 172 6]<br>[ 4 12 8 4 178]] |
| | TF-IDF | 5 | 70-30 | 79.533 | [[147 77 42 28 16]<br>[ 0 243 31 22 8]<br>[ 0 0 295 14 5]<br>[ 0 1 7 274 14]<br>[ 1 16 15 10 234]] |
| | MI | 1 | 50-50 | 84.92 | [[281 73 20 17 123]<br>[ 0 425 13 13 43]<br>[ 0 0 484 3 17]<br>[ 1 1 1 462 33]<br>[ 7 6 4 2 471]] |
| | MI | 1 | 70-30 | 83.866 | [[172 41 16 11 69]<br>[ 0 250 15 12 28]<br>[ 0 0 280 3 14]<br>[ 0 1 1 286 18]<br>[ 3 6 3 1 270]] |
| | MI | 1 | 80-20 | 87.12 | [[118 25 4 8 38]<br>[ 0 179 10 9 9]<br>[ 0 0 211 2 3]<br>[ 0 0 0 187 10]<br>[ 4 3 1 2 177]] |
| | MI | 5 | 70-30 | 82.933 | [[182 57 26 7 37]<br>[ 0 255 19 6 25]<br>[ 0 0 283 3 11]<br>[ 0 3 7 267 29]<br>[ 4 11 7 4 257]] |
| K=5 | TF-IDF | 1 | 50-50 | 74.44 | [[141 105 56 67 142]<br>[ 1 340 57 35 56]<br>[ 0 0 456 27 24]<br>[ 0 0 0 449 49] |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | [   1   5   6   8 475]] |
| | TF-IDF | 1 | 70-30 | 76.2 | [[ 89  63  52  30  62]<br>[   0 222  35  12  39]<br>[   0   0 286   9   8]<br>[   0   1   0 273  34]<br>[   0   3   5   4 273]] |
| | TF-IDF | 1 | 80-20 | 76.6 | [[ 69  46  38  12  44]<br>[   0 145  31  10  15]<br>[   0   0 194   5   6]<br>[   0   0   0 163  16]<br>[   0   1   8   2 195]] |
| | TF-IDF | 5 | 70-30 | 71.733 | [ 94  81  52  49  34]<br>[   0 196  47  47  14]<br>[   0   0 275  33   6]<br>[   0   0   3 264  29]<br>[   0   6  14   9 247]] |
| | MI | 1 | 50-50 | 77.16 | [[195  67  22  24 206]<br>[   0 368  28  17  81]<br>[   0   0 462   6  36]<br>[   1   0   0 419  78]<br>[   0   0   4   1 485]] |
| | MI | 1 | 70-30 | 74.333 | [[108  40  27  16 118]<br>[   0 203  31  19  52]<br>[   0   0 272   3  22]<br>[   0   0   0 255  51]<br>[   0   2   3   1 277]] |
| | MI | 1 | 80-20 | 80.5 | [[ 72  35   8  14  64]<br>[   0 160  14  13  20]<br>[   0   0 207   2   7]<br>[   0   0   0 181  16]<br>[   0   0   0   2 185]] |
| | MI | 5 | 70-30 | 73.666 | [[101  78  45  15  70]<br>[   0 224  24  12  45]<br>[   0   0 262  10  25]<br>[   0   1   2 250  53]<br>[   0   8   6   1 268]] |

## INFERENCES FROM RESULTS:-

### NAIVE BAYES Classifiers

1. In Naive Bayes Classifiers top 5% features according to TF-IDF are giving very good results in all train test splits ie approx 98%. If we increase our no of features further there is no improvements in result ie they are same . It means that there is only 5% top features that are important based on TF-IDF
2. In Naive Bayes Classifiers when we are taking features according to Mutual-Information . We can see from graphs for all train-test splits there is improvement in accuracy results when we increase no of features . After 40-50% top features results are nearly constant or overfitted .
3. We are getting 98% accuracy in TF-IDF and 93% accuracy in Mutaul Information because in MI we have considered top k% for each class where as in TF-IDF top k % same for each class . And TF-IDF are considered according to classwise and MI according to document wise
4. In TF_IDF we can clearly see overfitting as accuracy is nearly same or less

### KNN Classifiers

1. With increase in value of K in KNN accuracy start decreasing . With increase in K It is including noise also . So testing accuracy start decreasing
2. IN TF-IDF and MI top 1% features are giving good results . If we increase it to top 5% accuracy start decreasing slightly . With more features accuracy start decreasing as it is overfitting and including noise also .
3. With 80-20 split we have achieved highest accuracy ie 90% . As increase in data KNN start increasing its accuracy . With mode data KNN performs better

### ADDITIONAL INFERENCES:-

1. Naive Bayes performs better than KNN because Naive works better on small datatset and its assumption of conditional independence
2. Naive Bayes work on principle of conditional independence so it performs better
3. Naive Bayes is very fast when compared to KNN as it dosenot calculate cosine similarity and distance . Execution time is very fast when compared to KNN