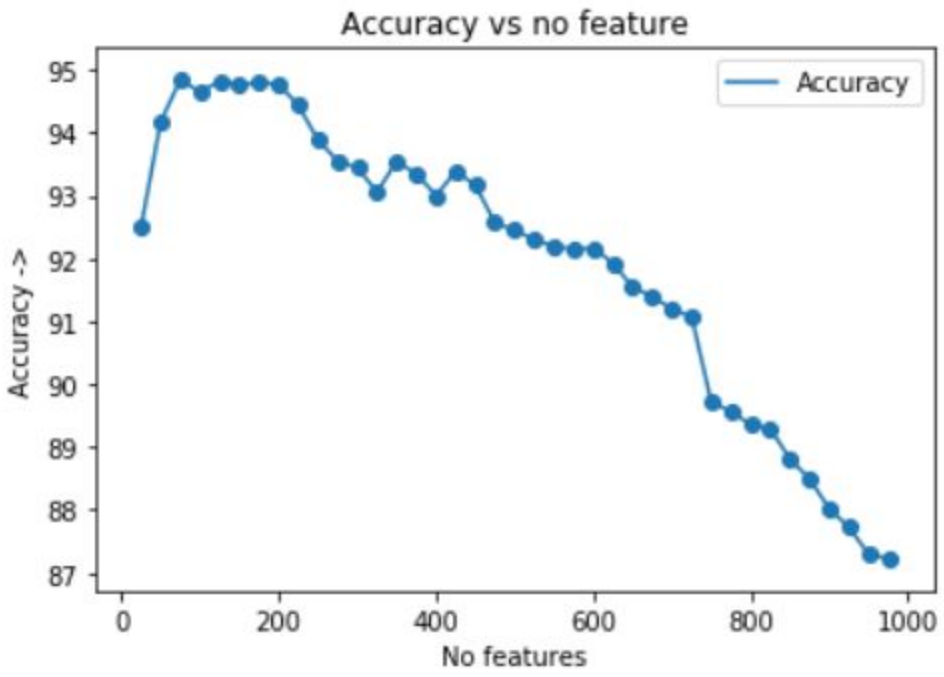
Assignment 5

Dibyendu Roy Chaudhuri
MT19034

Total number of unique word: 48004

Number of feature Selection



Selecting the best k value in top k features from each class is needed for model efficiency. Here (train: test ratio 70: 30) I have executed my code to find the best value of k for which overall model accuracy is high. We can see around k=100, we are getting our best accuracy.

Formula used for overall accuracy = (NB tf-idf + NB mi + k=1 knn tf-idf + k=1 knn mi + k=3 knn tf-idf + k=3 knn tf-idf + k=5 knn tf-idf + k=5 knn mi)/8

So the value of K is 100 here

Training: Test ratio- 50:50

Tf-Idf feature selection

1) Naive Bayes

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	456	6	7	12	10
Acc: 92.87169042769857					
rec.sport.hockey	0	506	0	0	0
Acc: 100.0					
sci.med	5	0	435	57	11
Acc: 85.62992125984252					
sci.space	3	0	14	476	3
Acc: 95.96774193548387					
talk.politics.misc	0	0	1	1	497
Acc: 99.59919839679358					

Tf-Idf feature based Naive Bayes accuracy- 94.8

2) KNN (k=1)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	482	0	0	0	0
Acc: 100.0					
rec.sport.hockey	0	516	0	0	0
Acc: 100.0					
sci.med	6	1	482	7	0
Acc: 97.17741935483872					
sci.space	2	0	76	419	2
Acc: 83.96793587174348					
talk.politics.misc	0	0	3	1	503
Acc: 99.21104536489152					

Tf-Idf feature based KNN (k= 1) accuracy- 96.08

3) KNN (k=3)

```
Confusion Matrix-
                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          477                  0                   5                   0                   0
Acc:  98.96265560165975

rec.sport.hockey       0                    516                 0                   0                   0
Acc:  100.0

sci.med                2                    1                   490                 3                   0
Acc:  98.79032258064517

sci.space              2                    0                   107                 389                 1
Acc:  77.9559118236473

talk.politics.misc     0                    0                   5                   1                   501
Acc:  98.81656804733728


Tf-Idf feature based KNN (k=  3 ) accuracy-  94.88
```

4) KNN (K=5)

```
Confusion Matrix-
                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          482                  0                   0                   0                   0
Acc:  100.0

rec.sport.hockey       0                    516                 0                   0                   0
Acc:  100.0

sci.med                6                    0                   487                 3                   0
Acc:  98.18548387096774

sci.space              3                    0                   133                 363                 0
Acc:  72.74549098196393

talk.politics.misc     0                    0                   3                   1                   503
Acc:  99.21104536489152


Tf-Idf feature based KNN (k=  5 ) accuracy-  93.84
```

We can observe the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes.

MI feature Selection

1) Naive Bayes

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	466	0	2	1	0
Acc: 99.36034115138592					
rec.sport.hockey	1	512	0	0	0
Acc: 99.80506822612085					
sci.med	5	0	482	3	3
Acc: 97.76876267748479					
sci.space	9	0	1	485	2
Acc: 97.58551307847083					
talk.politics.misc	0	0	0	1	527
Acc: 99.81060606060606					

MI feature based Naive Bayes accuracy- 98.88

2) KNN (k=1)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	508	0	2	1	0
Acc: 99.412915851272					
rec.sport.hockey	0	487	0	0	0
Acc: 100.0					
sci.med	4	0	474	1	0
Acc: 98.95615866388309					
sci.space	0	0	16	486	1
Acc: 96.62027833001989					
talk.politics.misc	0	0	0	0	520
Acc: 100.0					

MI feature based KNN (k= 1) accuracy- 99.0

3) KNN (k=3)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	508	0	2	1	0
Acc: 99.412915851272					
rec.sport.hockey	0	487	0	0	0
Acc: 100.0					
sci.med	3	0	476	0	0
Acc: 99.37369519832986					
sci.space	0	0	12	490	1
Acc: 97.4155069582505					
talk.politics.misc	0	0	0	0	520
Acc: 100.0					

MI feature based KNN (k= 3) accuracy- 99.16

4) KNN (K=5)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	509	0	2	0	0
Acc: 99.60861056751467					
rec.sport.hockey	0	487	0	0	0
Acc: 100.0					
sci.med	4	0	475	0	0
Acc: 99.16492693110646					
sci.space	1	0	8	494	0
Acc: 98.2107355864811					
talk.politics.misc	0	0	0	0	520
Acc: 100.0					

MI feature based KNN (k= 5) accuracy- 99.4

We can observe the accuracy of the model using MI feature selection are higher than the model using Tf-Idf feature selection. We can inference from above observation that MI is more suitable for selecting features. Also, the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes.

Training: Test ratio- 70:30

Tf-Idf feature selection

1) Naive Bayes

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	265	2	4	9	0
Acc: 94.64285714285714					
rec.sport.hockey	0	298	0	0	0
Acc: 100.0					
sci.med	2	0	244	40	4
Acc: 84.13793103448276					
sci.space	1	0	4	303	0
Acc: 98.37662337662337					
talk.politics.misc	6	2	2	2	312
Acc: 96.29629629629629					

Tf-Idf feature based Naive Bayes accuracy- 94.8

2) KNN (k=1)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	200	0	0	0	0
Acc: 100.0					
rec.sport.hockey	0	196	0	0	0
Acc: 100.0					
sci.med	2	0	204	3	0
Acc: 97.60765550239235					
sci.space	2	0	34	172	0
Acc: 82.6923076923077					
talk.politics.misc	0	0	3	1	183
Acc: 97.86096256684492					

Tf-Idf feature based KNN (k= 1) accuracy- 95.5

3) KNN (K=3)

```
Confusion Matrix-

                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          299                  0                   1                   0                   0
Acc:  99.66666666666667

rec.sport.hockey       0                    305                 0                   0                   0
Acc:  100.0

sci.med                5                    0                   294                 5                   0
Acc:  96.71052631578947

sci.space              3                    0                   44                  247                 2
Acc:  83.44594594594594

talk.politics.misc     0                    0                   7                   2                   286
Acc:  96.94915254237289


Tf-Idf feature based KNN (k=  3 ) accuracy-  95.13333333333334
```

4) KNN (K=5)

```
Confusion Matrix-

                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          300                  0                   0                   0                   0
Acc:  100.0

rec.sport.hockey       0                    305                 0                   0                   0
Acc:  100.0

sci.med                4                    0                   298                 2                   0
Acc:  98.02631578947368

sci.space              3                    0                   52                  240                 1
Acc:  81.08108108108108

talk.politics.misc     1                    0                   6                   2                   286
Acc:  96.94915254237289


Tf-Idf feature based KNN (k=  5 ) accuracy-  95.19999999999999
```

We can observe the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes. We can inference from the above two occasions that KNN can work better in text classifications using the lower number of features than Naive Bayes.

MI feature Selection

1) Naive Bayes

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	306	0	0	0	0
Acc: 100.0					
rec.sport.hockey	0	290	0	0	0
Acc: 100.0					
sci.med	4	0	309	2	2
Acc: 97.47634069400631					
sci.space	4	0	2	286	0
Acc: 97.94520547945206					
talk.politics.misc	0	0	0	1	294
Acc: 99.66101694915255					

MI feature based Naive Bayes accuracy- 99.0

2) KNN (K=1)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	278	0	3	0	0
Acc: 98.93238434163702					
rec.sport.hockey	0	301	0	0	0
Acc: 100.0					
sci.med	2	0	303	3	0
Acc: 98.37662337662337					
sci.space	1	0	3	314	0
Acc: 98.74213836477988					
talk.politics.misc	0	0	0	0	292
Acc: 100.0					

MI feature based KNN (k= 1) accuracy- 99.2

3) KNN (K=3)

```
Confusion Matrix-
                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          278                     0                   3                   0                   0
Acc:  98.93238434163702

rec.sport.hockey       0                       301                 0                   0                   0
Acc:  100.0

sci.med                2                       0                   305                 1                   0
Acc:  99.02597402597402

sci.space              0                       0                   5                   313                 0
Acc:  98.42767295597484

talk.politics.misc     0                       0                   0                   0                   292
Acc:  100.0


MI feature based KNN (k=  3 ) accuracy-  99.26666666666667
```

4) KNN (K=5)

```
Confusion Matrix-
                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          277                     0                   4                   0                   0
Acc:  98.57651245551602

rec.sport.hockey       0                       301                 0                   0                   0
Acc:  100.0

sci.med                0                       0                   306                 2                   0
Acc:  99.35064935064936

sci.space              0                       0                   4                   314                 0
Acc:  98.74213836477988

talk.politics.misc     0                       0                   0                   0                   292
Acc:  100.0


MI feature based KNN (k=  5 ) accuracy-  99.33333333333333
```

We can observe the accuracy of the model using MI feature selection is higher than the model using Tf-Idf feature selection. We can inference from the above observation that MI is more suitable for selecting features. Also, the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes.

Training: Test ratio- 80:20

Tf-Idf feature selection

1) Naive Bayes

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	171	1	1	7	2
Acc: 93.95604395604396					
rec.sport.hockey	0	203	0	0	0
Acc: 100.0					
sci.med	3	0	166	27	6
Acc: 82.17821782178217					
sci.space	2	0	7	210	1
Acc: 95.45454545454545					
talk.politics.misc	0	0	0	0	193
Acc: 100.0					

Tf-Idf feature based Naive Bayes accuracy- 94.3

2) KNN (k=1)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	200	0	0	0	0
Acc: 100.0					
rec.sport.hockey	0	196	0	0	0
Acc: 100.0					
sci.med	2	0	204	3	0
Acc: 97.60765550239235					
sci.space	2	0	34	172	0
Acc: 82.6923076923077					
talk.politics.misc	0	0	3	1	183
Acc: 97.86096256684492					

Tf-Idf feature based KNN (k= 1) accuracy- 95.5

3) KNN (K=3)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	199	0	1	0	0
Acc: 99.5					
rec.sport.hockey	0	196	0	0	0
Acc: 100.0					
sci.med	2	0	204	3	0
Acc: 97.60765550239235					
sci.space	2	0	43	163	0
Acc: 78.36538461538461					
talk.politics.misc	0	0	6	1	180
Acc: 96.2566844919786					

Tf-Idf feature based KNN (k= 3) accuracy- 94.19999999999999

4) KNN (K=5)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	200	0	0	0	0
Acc: 100.0					
rec.sport.hockey	0	196	0	0	0
Acc: 100.0					
sci.med	2	0	204	3	0
Acc: 97.60765550239235					
sci.space	2	0	47	159	0
Acc: 76.4423076923077					
talk.politics.misc	0	0	4	1	182
Acc: 97.32620320855615					

Tf-Idf feature based KNN (k= 5) accuracy- 94.1

We can observe the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes. We can inference from the above occasions that KNN can work better in text classifications using the lower number of features than Naive Bayes.

MI feature Selection

1) Naive Bayes

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	181	0	0	1	0
Acc: 99.45054945054946					
rec.sport.hockey	0	203	0	0	0
Acc: 100.0					
sci.med	3	0	193	3	3
Acc: 95.54455445544554					
sci.space	3	0	0	217	0
Acc: 98.63636363636363					
talk.politics.misc	0	0	0	0	193
Acc: 100.0					

MI feature based Naive Bayes accuracy- 98.7

2) KNN (K=1)

Confusion Matrix-

	comp.graphics	rec.sport.hockey	sci.med	sci.space	talk.politics.misc
comp.graphics	179	0	3	0	0
Acc: 98.35164835164835					
rec.sport.hockey	0	203	0	0	0
Acc: 100.0					
sci.med	0	0	201	1	0
Acc: 99.5049504950495					
sci.space	1	0	4	215	0
Acc: 97.72727272727273					
talk.politics.misc	0	0	0	0	193
Acc: 100.0					

MI feature based KNN (k= 1) accuracy- 99.1

3) KNN (K=3)

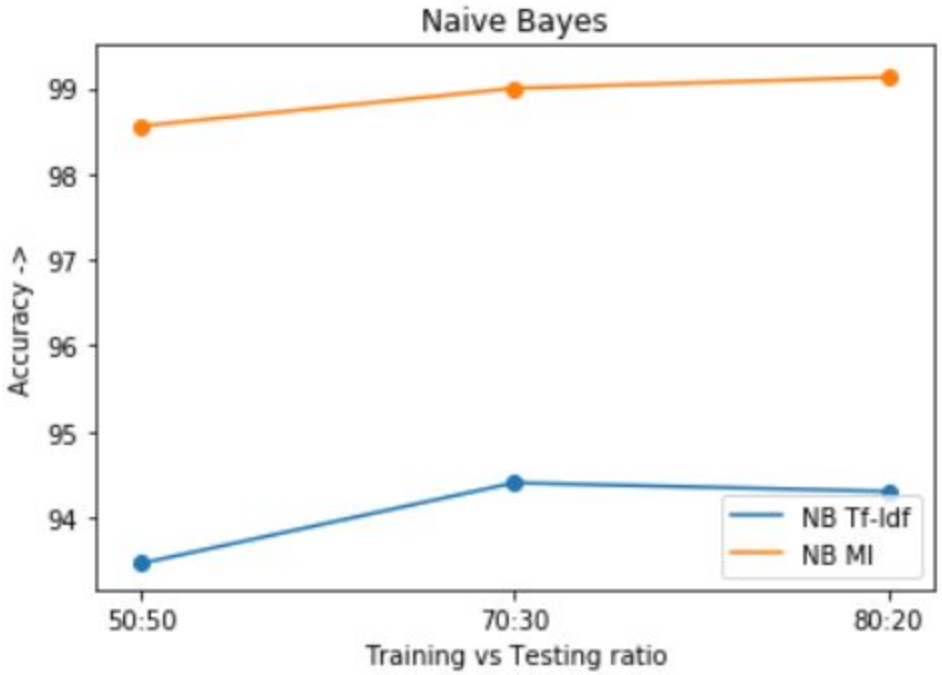
```
Confusion Matrix-
                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          179                     0                   3                   0                   0
Acc:  98.35164835164835

rec.sport.hockey       0                       203                 0                   0                   0
Acc:  100.0

sci.med                0                       0                   200                 2                   0
Acc:  99.00990099009901

sci.space              1                       0                   3                   216                 0
Acc:  98.18181818181819

talk.politics.misc     0                       0                   0                   0                   193
Acc:  100.0


MI feature based KNN (k=  3 ) accuracy-  99.0
```

4) KNN (K=5)

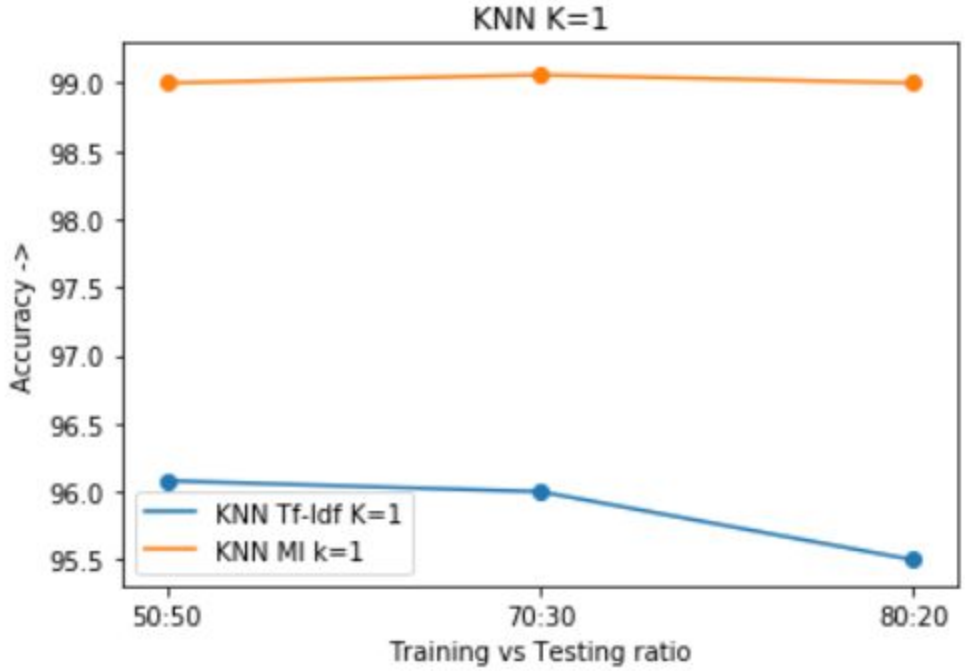
```
Confusion Matrix-
                      comp.graphics       rec.sport.hockey    sci.med             sci.space           talk.politics.misc
comp.graphics          179                     0                   3                   0                   0
Acc:  98.35164835164835

rec.sport.hockey       0                       203                 0                   0                   0
Acc:  100.0

sci.med                0                       0                   201                 1                   0
Acc:  99.5049504950495

sci.space              0                       0                   2                   218                 0
Acc:  99.0909090909091

talk.politics.misc     0                       0                   0                   0                   193
Acc:  100.0


MI feature based KNN (k=  5 ) accuracy-  99.3
```

We can observe the accuracy of the model using MI feature selection is higher than the model using Tf-Idf feature selection. We can inference from the above observation that MI is more suitable for selecting features. Also, the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes.

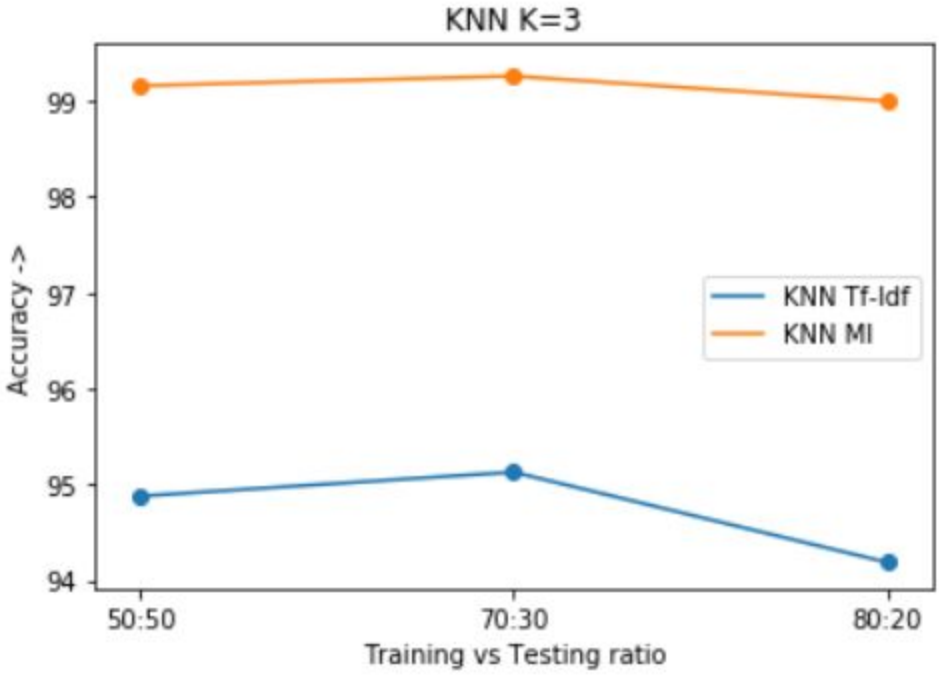
Naive Bayes comparison



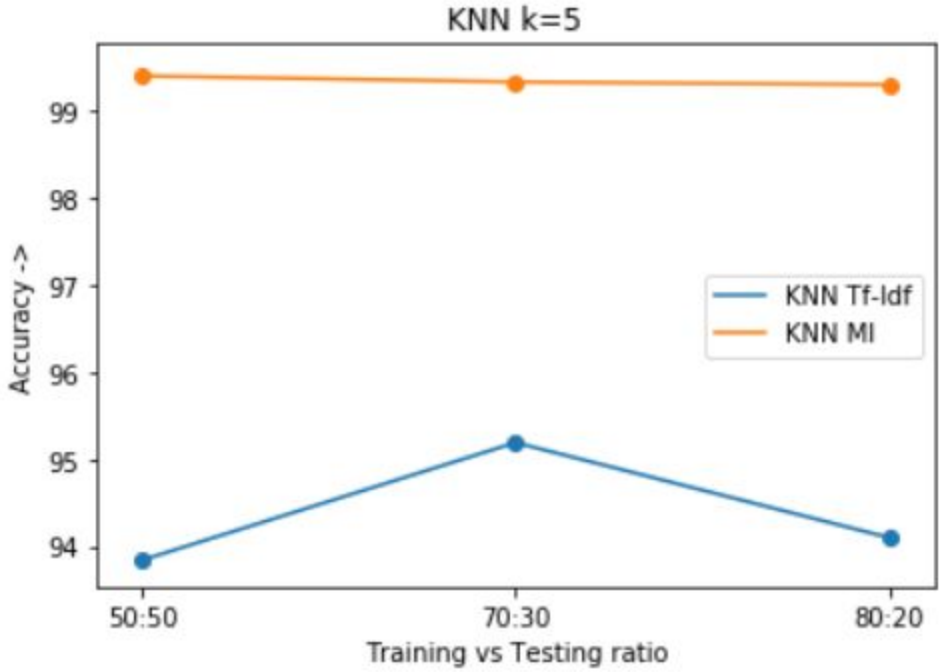
KNN (k=1) comparison



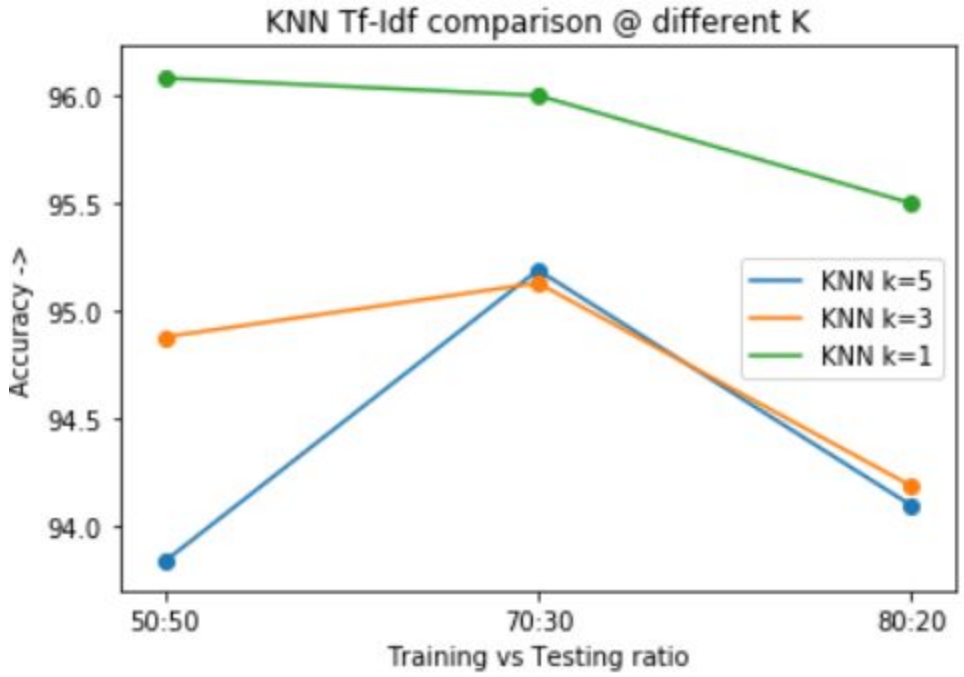
KNN (k=3) comparison



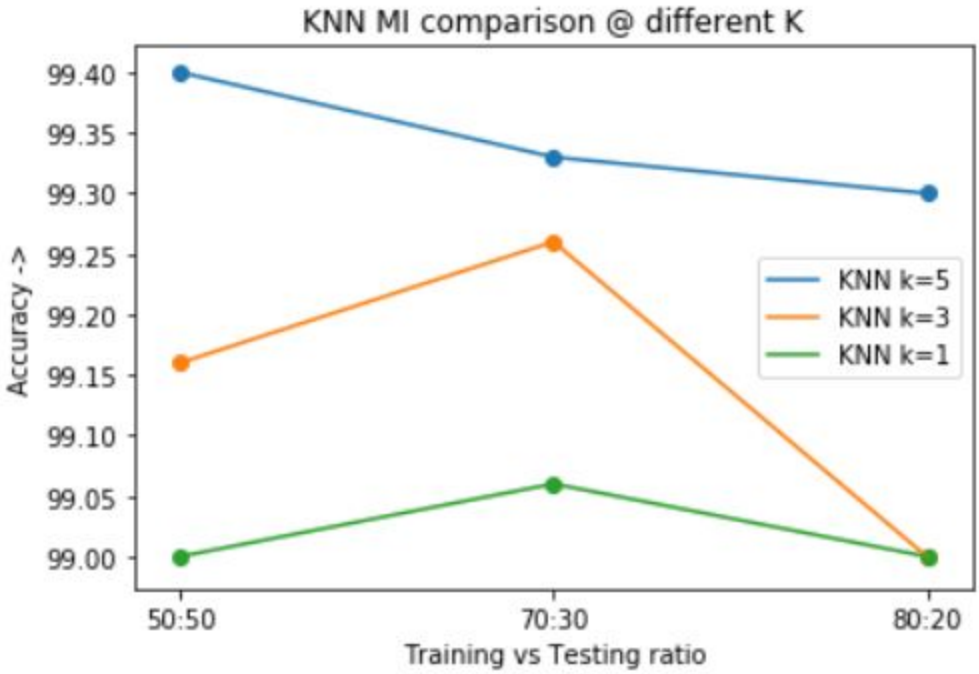
KNN (k=5) comparison



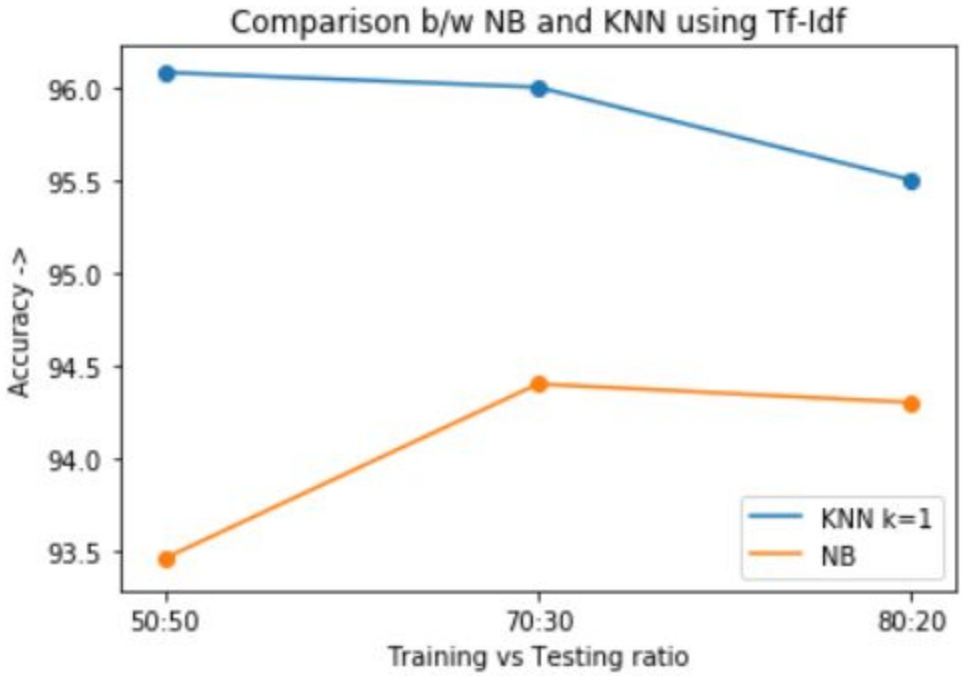
KNN Tf-Idf comparison



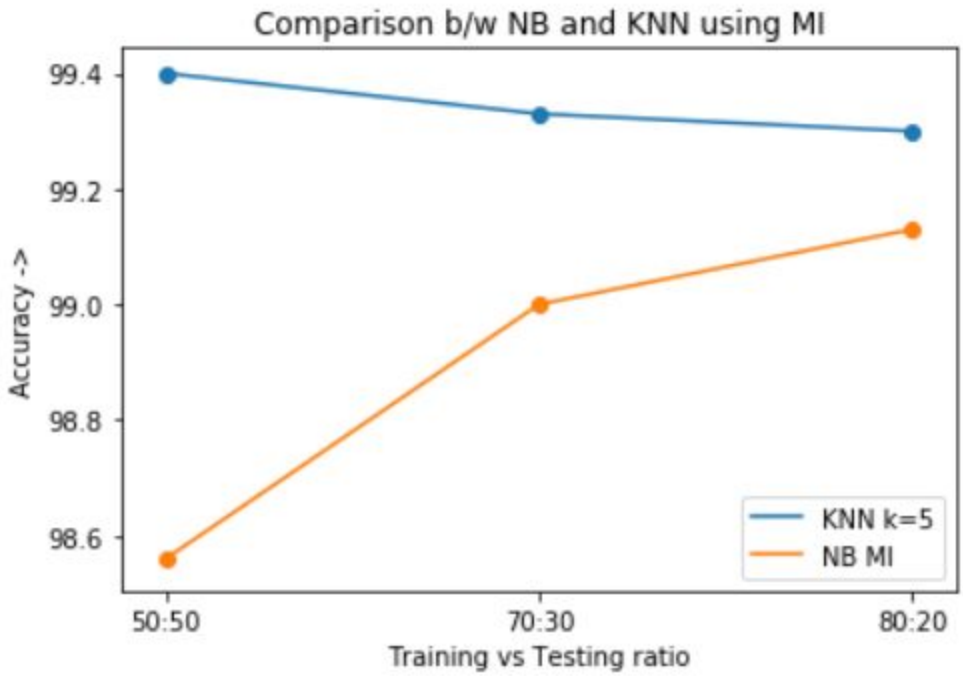
KNN MI comparison



Tf-Idf KNN best model vs Naive Bayes



MI KNN best model vs Naive Bayes



We can observe the accuracy of the model using MI feature selection is higher than the model using Tf-Idf feature selection. We can inference from the above observation that MI is more suitable for selecting features. Also, the accuracy of KNN at different k points are slightly higher or equal to Naive Bayes. We can inference from the above occasions that KNN can work better in text classifications using the lower number of features than Naive Bayes. Although when increase value of k in top k feature selection, we can the time when the naive Bayes accuracy actually overcome the accuracy of KNN. So we can inference Naive Bayes is more suitable when the number of features word are more. We can also see the accuracy of Naive Bayes is increased with more training documents where the accuracy of KNN first increased but later slightly decreased. We can from that point KNN started overfitted. In general with the increased value of k in KNN, the model becomes more efficient.