Assignment 2: Text Classification

Name : Sanjay Malakar
Roll : 1505057

KNN Validation

Metric k	1	3	5
Hamming	37.864	39.500	37.773
Euclidean	54.909	54.545	54.545
Cosine	79.409	82.955	83.591

NB Validation

Smoothing Factor	Accuracy
0.05	92.045
0.10	91.955
0.20	92.136
0.40	92.227
0.60	92.045
0.80	91.773
1.00	91.273
1.20	90.955
1.40	90.818
1.60	90.773

Best Classifiers

For **KNN**,

K = **5**

Method = **Cosine**

For **NB**,

Smoothing Factor = **0.40**

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
1	89.091	92.727
2	86.364	97.273
3	83.636	95.455
4	78.182	90.909
5	85.455	96.364
6	82.727	91.818
7	90.000	93.636

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
8	86.364	94.545
9	80.000	90.000
10	84.545	90.000
11	83.636	92.727
12	83.636	93.636
13	76.364	91.818
14	80.000	88.182

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
15	80.909	91.818
16	78.182	90.000
17	81.818	91.818
18	76.364	93.636
19	81.818	90.909
20	77.273	88.182
21	80.000	83.636

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
22	89.091	97.273
23	87.273	95.455
24	81.818	91.818
25	83.636	86.364
26	85.455	92.727
27	82.727	87.273
28	77.273	89.091

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
29	80.000	88.182
30	81.818	92.727
31	83.636	90.909
32	80.909	93.636
33	80.909	93.636
34	79.091	94.545
35	81.818	90.000

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
36	80.000	89.091
37	76.364	90.909
38	80.000	89.091
39	85.455	94.545
40	88.182	95.455
41	79.091	90.000
42	82.727	92.727

Test Accuracy

Test Number	KNN Accuracy	NB Accuracy
43	78.182	89.091
44	81.818	89.091
45	80.909	92.727
46	80.909	93.636
47	82.727	88.182
48	82.727	90.000
49	82.727	89.091
50	86.364	92.727

t-statistic

t-statistic score (nb, knn): 14.846482532520774

p-value (nb, knn): 8.060758741458655e-27

p-value is smaller than the threshold (0.005, 0.01, 0.05), so we can reject the null hypothesis of equal averages.

Justification of result

For this dataset NB performs better than KNN.

As the total size of data is quite large it is expected to perform better on NB.

KNN is most likely to overfit, and hence adjusting ‘k’ to maximise test set performance is the way to go. As the complexity of the space grows, the accuracy of KNN comes down and you would need more data, but the order of this classifier is n^2 and it becomes too slow.

NB is an eager learning classifier and it is much faster than KNN, runs in O(1). Thus, it could be used for prediction in real time.