Assignment 2 Text Classification Report

KNN: value of n	1	3	5
Hamming	36.5%	31.7272727272%	31.04545454545%
Euclidean	55.2727272727%	50.5909090909%	50.090909090909%
Cosine Distance with TF-IDF	81.18181818%	81.7727272727%	83%

Smoothing factor	Naïve Bayes Accuracy
0.1	91.954545454545%
0.2	92%
0.25	92.181818181818%
0.3	92.1363636363636%
0.35	92.0454545454545%
0.4	91.95454545454545%
0.45	91.72727272727272%
0.5	91.63636363636363%
0.55	91.59090909090909%
0.6	91.3181818181818%

Here, the best performance of validation set in KNN algorithm is Cosine distance with TF-IDF with accuracy of 83%

For Naïve Bayes Algorithm, the best performance is by the smoothing factor 0.25 with accuracy 92.18181818%.

Iteration	Naïve Bayes, smoothing factor = 0.25 (%)	KNN, n = 5, cosine distance with TF-IDF(%)
1	92.72727272727272	87.27272727272727
2	95.45454545454545	86.36363636363636
3	92.72727272727272	81.81818181818183
4	89.0909090909091	82.72727272727273
5	96.36363636363636	85.45454545454545
6	90.9090909090909	88.18181818181819
7	94.54545454545455	85.45454545454545
8	95.45454545454545	88.18181818181819
9	91.81818181818183	80.0
10	90.9090909090909	83.63636363636363
11	93.63636363636364	83.63636363636363
12	96.36363636363636	84.54545454545455
13	89.0909090909091	78.18181818181819
14	89.0909090909091	80.9090909090909
15	93.63636363636364	78.18181818181819
16	88.18181818181819	79.0909090909091
17	90.0	79.0909090909091
18	91.81818181818183	81.81818181818183
19	90.0	75.45454545454545
20	87.27272727272727	78.18181818181819
21	84.54545454545455	79.0909090909091
22	96.36363636363636	83.63636363636363
23	93.63636363636364	85.45454545454545
24	93.63636363636364	79.0909090909091
25	90.0	82.72727272727273
26	88.18181818181819	83.63636363636363
27	92.72727272727272	82.72727272727273
28	88.18181818181819	76.36363636363637
29	88.18181818181819	84.54545454545455
30	93.63636363636364	78.18181818181819
31	91.81818181818183	82.72727272727273
32	91.81818181818183	81.81818181818183
33	93.63636363636364	84.54545454545455
34	93.63636363636364	80.9090909090909

35	90.9090909090909	80.9090909090909
36	90.0	79.0909090909091
37	90.0	80.0
38	90.9090909090909	75.45454545454545
39	92.72727272727272	86.36363636363636
40	95.45454545454545	89.0909090909091
41	90.9090909090909	78.18181818181819
42	96.36363636363636	89.0909090909091
43	87.27272727272727	75.45454545454545
44	92.72727272727272	80.9090909090909
45	93.63636363636364	80.0
46	93.63636363636364	87.27272727272727
47	91.81818181818183	84.54545454545455
48	90.0	78.18181818181819
49	89.0909090909091	83.63636363636363
50	92.72727272727272	81.81818181818183
Average	91.74545454545456	82.07272727272726

The t-statistic value I get is (using ttest_rel, first parameter was knn, second was NB)

statistic=-22.02971880524802, pvalue=4.503554545628819e-27

for significance value of 0.005, 0.01, 0.05, pvalue is significantly lower. So, we can say that two algorithms do not perform same on the same dataset, one performs much better than other. As statistic value is negative, so algorithm in 2nd parameter performs much better than the first one. So Naïve Bayes works better than KNN algorithm.