Assignment 1

Spam Classification Experiments

Valgard Gudni
Hans P. Reiser
5 September 2023

Experiment 1.1

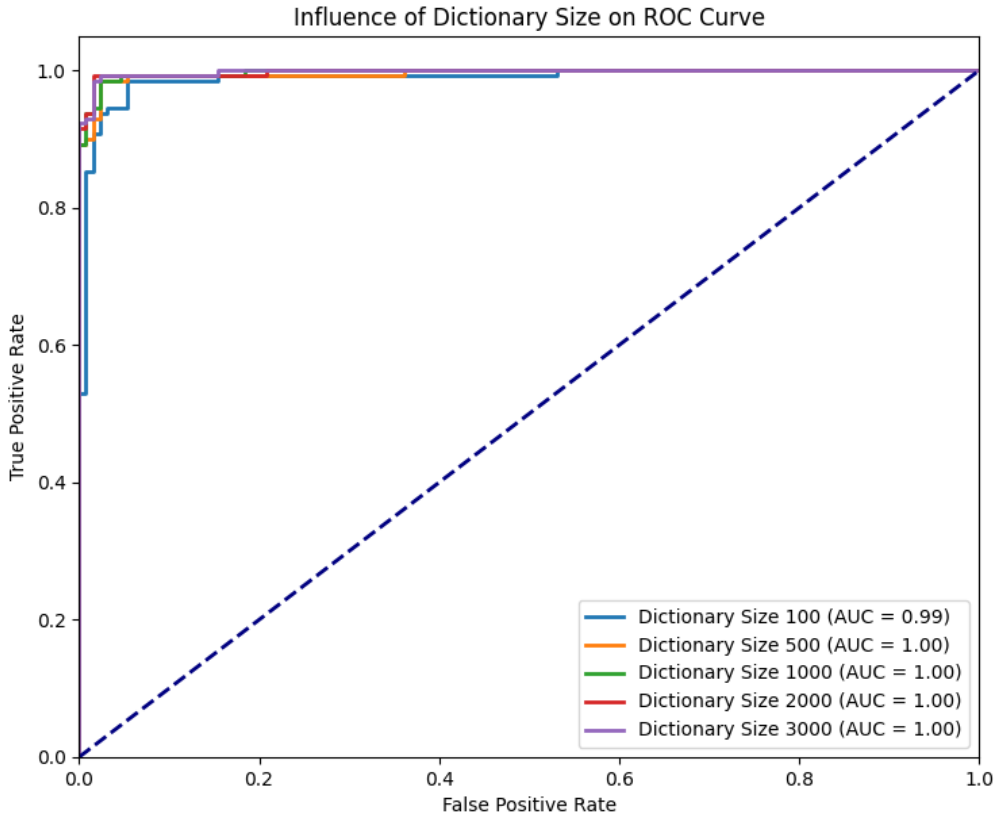

Figure 1.

Results:

Dictionary Size	Accuracy	Precision	Recall	F1-Score	AUC-ROC
100	0.9461538461538461	0.9833333333333333	0.9076923076923077	0.9444	0.9884023668639053
500	0.9538461538461539	0.9836065573770492	0.9230769230769231	0.9523809523809524	0.9950295857988166
1000	0.95	0.9915966386554622	0.9076923076923077	0.9477911646586346	0.9968639053254438
2000	0.9576923076923077	0.9917355371900827	0.9230769230769231	0.9561752988047809	0.9973964497041420
3000	0.9576923076923077	0.9917355371900827	0.9230769230769231	0.9561752988047809	0.9977514792899408

In the experiment on the influence of dictionary size on Naïve Bayes classifier performance, five different sizes were tested: 100, 500, 1000, 2000, and 3000. Across all metrics—accuracy, precision, recall, F1-score, and AUC-ROC—the classifier showed a generally increasing trend in performance with larger dictionary sizes. However, the performance gains diminished beyond a dictionary size of 2000. Specifically, the highest F1-score and AUC-ROC were achieved at dictionary sizes of 2000 and 3000, both scoring approximately 95.6% and 99.7% respectively. Given these results and considering computational efficiency, a dictionary size of 2000 was selected for subsequent experiments.

Experiment 1.2

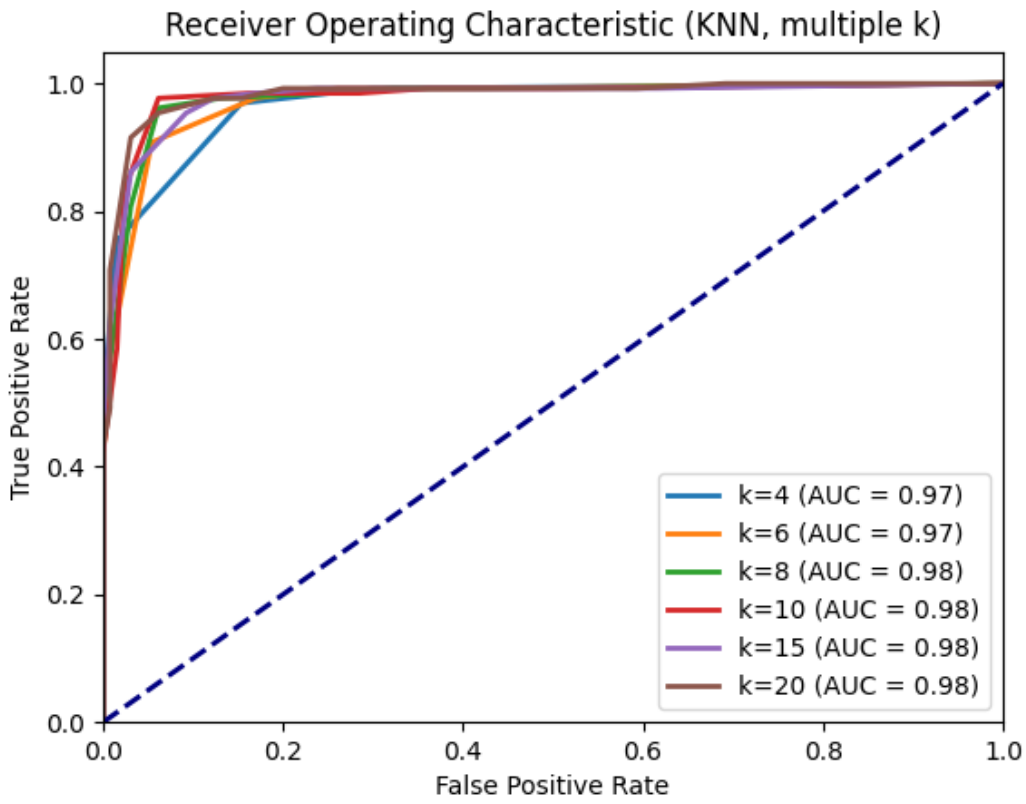

Figure 2.

Results:

K	Accuracy	Precision	Recall	F1-Score	AUC-ROC
4	0.8692307692307693	0.98	0.7538461538461538	0.8521739130434782	0.9695857988816568
6	0.9269230769230770	944	0.9076923076923077	0.9254901960784314	0.9728994082840237
8	0.95	0.9398496240601504	0.9615384615384616	0.9505703422053231	0.9773372781065088
10	0.9576923076923077	0.9407407407407408	0.9769230769230769	0.9584905660377359	0.9787278106508877
15	0.9269230769230770	0.8881118881118881	0.9769230769230769	0.9304029304029305	0.9775147928994090
20	0.9269230769230770	0.8881118881118881	0.9769230769230769	0.9304029304029305	0.9825443786982249

The experiment aimed to understand the impact of different K-values on the K-NN classifier's performance, using the following metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. The table above shows the results of the experiment.

- Increasing 'K' initially led to improvements in model performance across all metrics but started to plateau after 'K=10'.
- The best overall performance was observed at 'K=10', where the model achieved the highest accuracy, precision, and F1-Score, as well as a balanced recall value.
- The AUC-ROC value increased in value as k increased, topping at 0.983 for 'K=20'.

Based on these metrics, a 'K' value of 10 appears to be the optimal choice for this K-NN classifier, as it provides the highest accuracy, precision, f1-score, as well as a balanced recall value, and will be used as the value for K for the comparison in part 1.5

Experiment 1.3

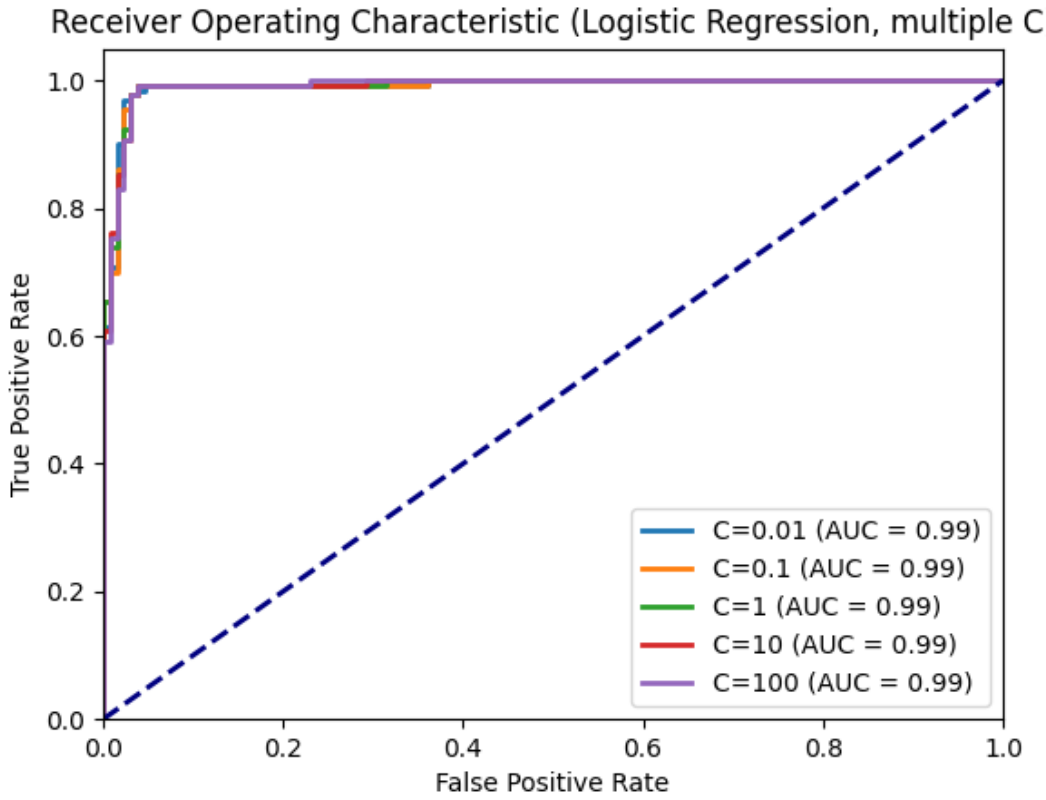

Figure 3.

The objective in experiment 1.3 was to investigate the influence of the regularization parameter C on the performance of the Logistic Regression classifier for spam classification. Varying values of C were used from the set {0.01, 0.1, 1, 10, 100}.

Results:

C	Accuracy	Precision	Recall	F1-Score	AUC-ROC
0.01	0.9692307692307692	0.9692307692307692	0.9692307692307692	0.9692307692307692	0.9910650887573964
0.1	0.9730769230769231	0.9624060150375939	0.9846153846153847	0.9733840304182510	0.9909467455621301
1	0.9730769230769231	0.9624060150375939	0.9846153846153847	0.9733840304182510	0.9913017751479290
10	0.9692307692307692	0.9621212121212122	0.9769230769230769	0.9694656488549619	0.9911834319526627
100	0.9692307692307692	0.9621212121212122	0.9769230769230769	0.9694656488549619	0.9913017751479289

The results suggest that the classifier performance is fairly stable across different C values, but the best performance is observed for C values of 0.1 and 1 in terms of the F1-Score and AUC-ROC. Therefore, either of these C values could be considered the optimal for this case.

Experiment 1.4

Results:

C	Train Recall	Train Precision	Test Recall	Test Precision
0.01	0.994302	1.0	0.969231	0.969231
0.10	1.000000	1.0	0.984615	0.962406
1.00	1.000000	1.0	0.984615	0.962406
10.00	1.000000	1.0	0.976923	0.962121
100.00	1.000000	1.0	0.976923	0.962121

From the results we can see that the recall and precision values are perfect for all values of C, which suggests that the model is overfitting the training data. This is further supported by the difference in performance between the training and test sets.

Experiment 1.5

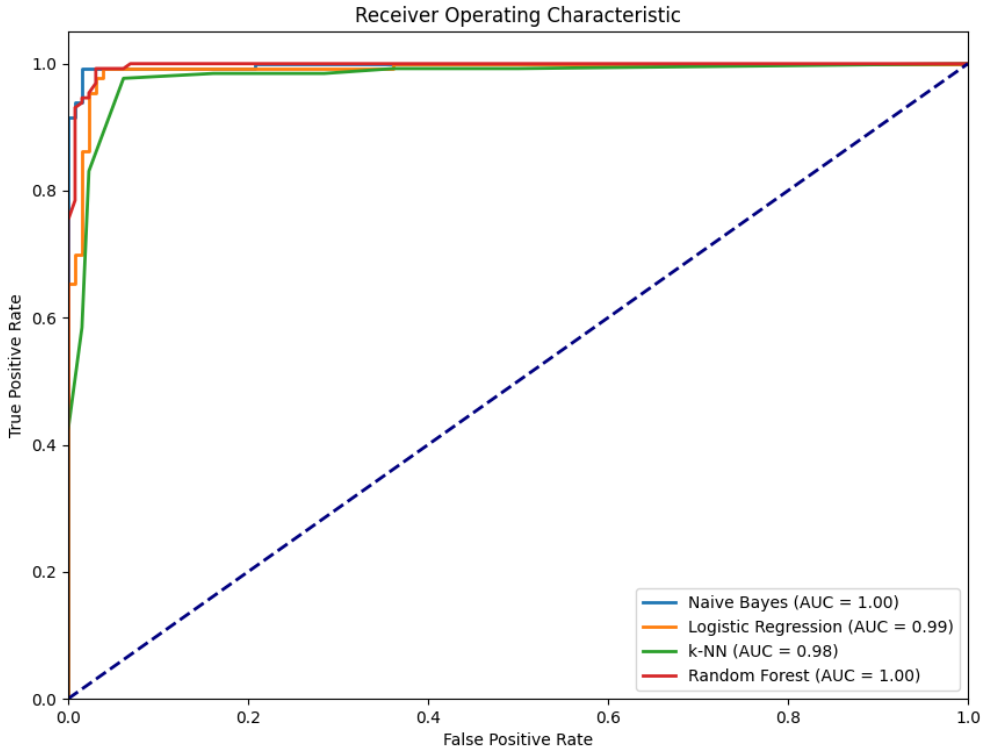

Figure 4.

Results:

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naive Baynes	0.9576923076923077	0.9917355371900827	0.9230769230769231	0.9561752988047809	0.9973964497041420
k-NN	0.9576923076923077	0.9407407407407408	0.9769230769230769	0.9584905660377359	0.9787278106508877
Logistic Regression	0.9730769230769231	0.9624060150375939	0.9846153846153847	0.9733840304182510	0.9909467455621301
Random Forest	0.9769230769230769	0.9626865671641791	0.9923076923076923	0.9772727272727273	0.9967455621301775

Model Performances:

- Naive Bayes: The Naive Bayes classifier has a good AUC-ROC score, suggesting that it distinguishes well between positive and negative classes. Its recall is the lowest among the three, suggesting that it might be missing some true positive cases.
- k-NN: The k-NN model has the lowest AUC-ROC score but a decent f1-score. Its recall is high, suggesting that it identifies most of the true positive cases. However, its precision is the lowest among the three, indicating that it might be classifying some negative samples as positive.
- Logistic Regression: The Logistic Regression model appears to have an impressive overall performance among the four, with the second highest accuracy, recall, and f1-score, only barely behind the Random Forest classifier. Its AUC-ROC score is also impressive, though slightly less than that of the Naive Bayes model.
- Random Forest: The Random Forest classifier model has the strongest accuracy value, only slightly ahead of Logistic Regression, as well as the highest recall and f1-score. It also has a very respectable AUC-ROC value, making the random forest classifier the most optimal out of the chosen classifiers.

Summary:

Highest Accuracy: Random Forest

Highest Precision: Naive Bayes

Highest Recall: Random Forest

Highest F1-Score: Random Forest

Highest AUC-ROC: Naive Bayes

Interpretation:

Naive Bayes: This model might be a good choice when high precision is wanted, in cases where i.e false positives are costly for the application.

Logistic Regression: Overall, this appears to be the second-best-performing model, with high scores across all metrics, only slightly behind the Random forest classifier.

k-NN: Could be useful in cases where high recall is essential, although it comes at the expense of precision.

Random Forest: This classifier performed the best overall of the four classifiers, scoring highest in accuracy, recall, and f1. Additionally, it posted high the second highest AUC-ROC value.