

REPORT

Parameter Values:

Total number of training samples: 463	Length of englishstopwordslst: 174	λ -values = [0.001, 0.1, 1]
Total number of test samples: 478	Learning rate (eta): 0.1	Number of iterations: 100

Naïve Bayes Results:

Using stopwordslist	HAM Accuracy	SPAM Accuracy	Combined Accuracy
NO	95.11 %	97.69 %	95.82 %
YES	94.54 %	96.92 %	95.19 %

Logistic Regression Results:

Using stopwordslist	HAM Accuracy	SPAM Accuracy	Combined Accuracy	λ -value
NO	97.41 %	90.00 %	95.19 %	0.001
NO	97.12 %	89.23 %	94.77 %	0.1
NO	97.12 %	88.46 %	94.56 %	1
YES	95.68 %	96.15 %	95.61 %	0.001
YES	96.25 %	94.62 %	95.61 %	0.1
YES	97.41 %	90.77 %	95.40 %	1

Conclusion:

In Naïve Bayes algorithm, we see that the accuracy of HAM/SPAM is slightly decreased in case of using stopwordslist which might be due to fact that some useful words are removed which were classified to a label as HAM or SPAM.

Generally, in case of Logistic Regression, when we decrease the learning rate from 0.1, the weight computation will take a longer time to converge to a solution, so we will have to take higher number of iterations for less learning rate. In this case, keeping the learning rate fixed at "0.1" and the number of iterations to be "100". We can see that when we increase the λ -value from 0.001 to 1, the accuracies of HAM or SPAM decreases in case of not using any stopwordslist with overall accuracy to reduce with increase in value of regularization parameter, but when we use stopwordslist, the accuracy of HAM increases but the accuracy of SPAM decreases, while keeping the overall accuracy to reduce with increase in the value of regularization parameter, this might be due to the removal of words which happen to occur in the stopwordslist.