Project Title

Spam Filtering using Support Vector Machine

STATEMENT OF THE PROBLEM

Spam is very annoying problem which is being faced by almost everyone having an email account.40are spam which around 15.4 billion email every day and that cost internet clients about $355 million every year. It is imperative to filtering of spam email before sending it to the inbox of users, indeed this has been very important and challenging task. Various Machine learning methods are being used to classify spammer's emails from legitimate emails. Now we are using machine learning algorithm Support vector machine (SVM) for solving this problem using different kernel-functions and also using different parameter, compare the performance of SVM for all different kernels and eventually we will optimize to get best result.

Analysing Using Library

Observations :

Kernels	Train Accuracy
Linear	1.0
Sigmoid	0.998205312275664
RBF	1.0
Polynomial	1.0

Kernels	Test Accuracy
Linear	0.9752333094041636
Sigmoid	0.9540559942569993
RBF	0.9727207465900933
Polynomial	0.9242641780330223

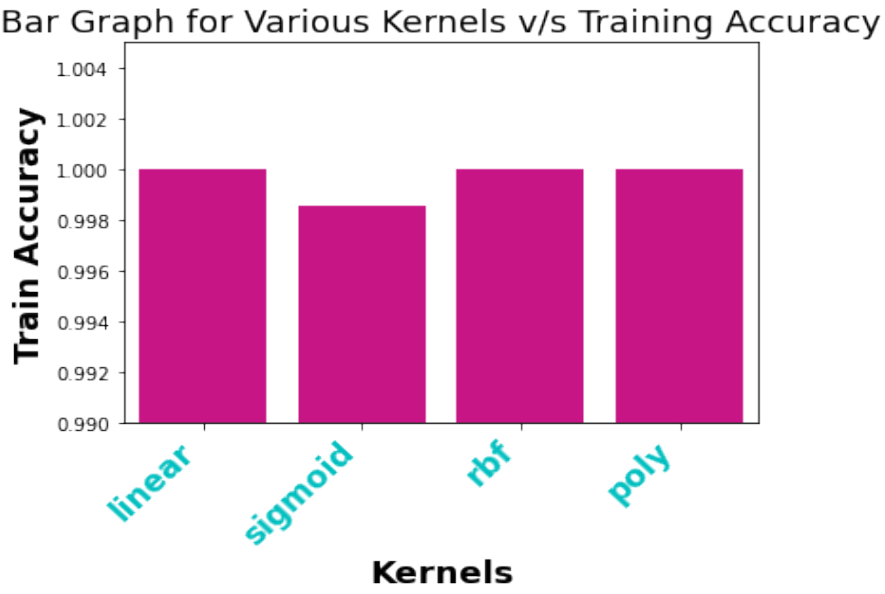

Figure 1: Kernels v/s Train Accuracy

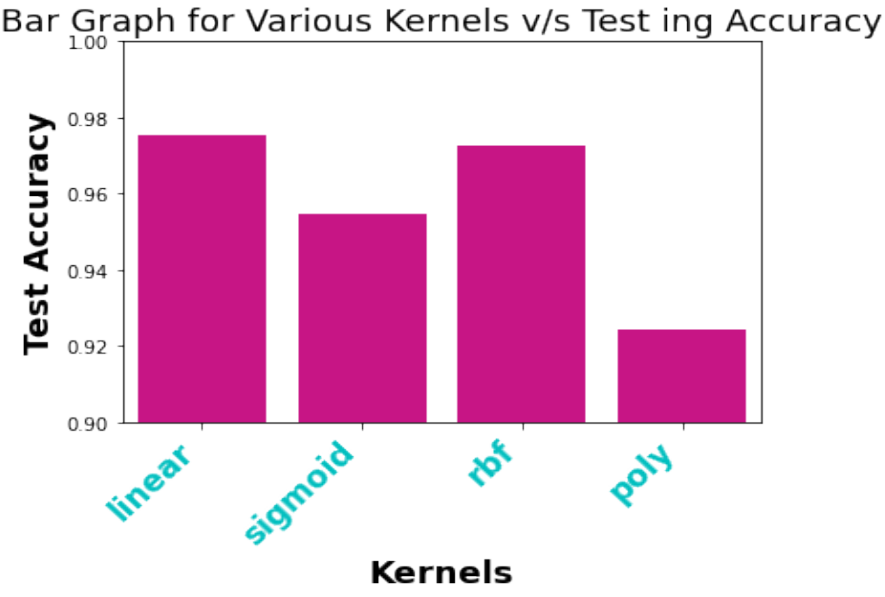

Figure 2: Kernels v/s Test Accuracy

Analysis:

Train Accuracy:
[1] All the kernels are able to train the model with 100% accuracy except Sigmoid kernel.(It all depends upon the dataset used)
If we talk with respect to our dataset taken sigmoid kernel is performing in-efficient in comparison to other kernels in terms of training.
All the kernels except Sigmoid (Linear, RBF, Polynomial) are able to classify the dataset with 100% accuracy.

Test Accuracy:
[2] For out Dataset Linear kernel is performing best with accuracy 97.52% which is slight equal to RBF kernel (97.27% Accuracy).
Polynomial Kernel is performing least with accuracy 92.42%.

Analysing against Train Test Split v/s Accuracy

Observations :

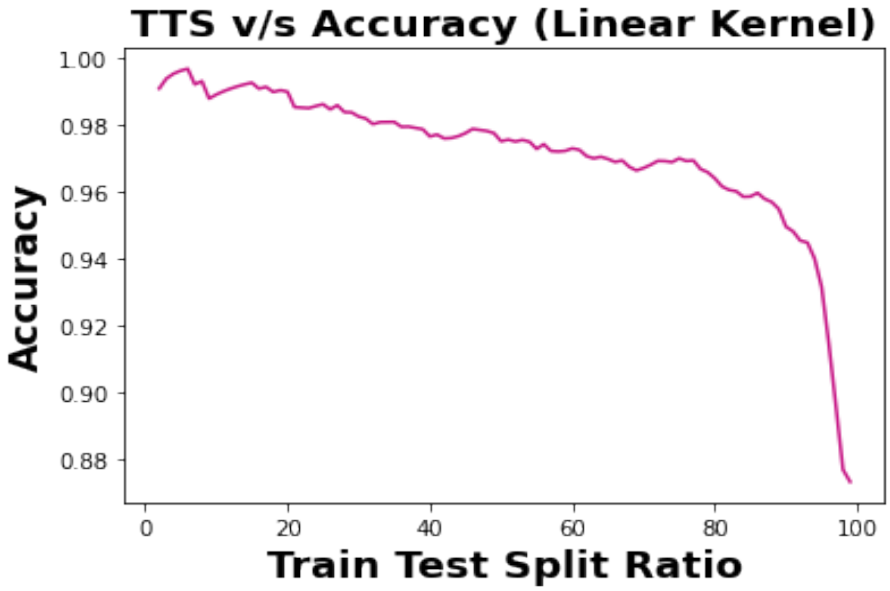

Figure 3: Train Test Split v/s Accuracy of the model for linear kernel

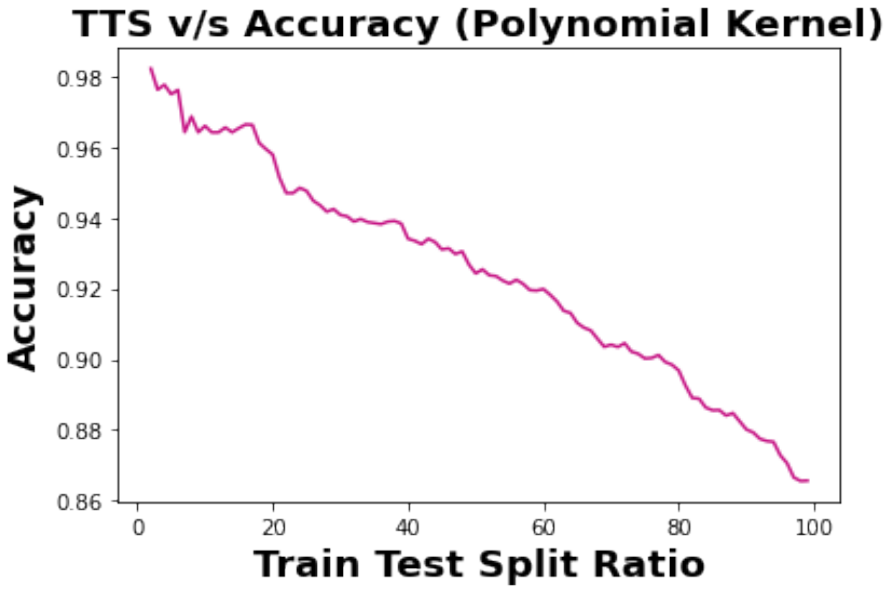

Figure 4: Train Test Split v/s Accuracy of the model for polynomial kernel

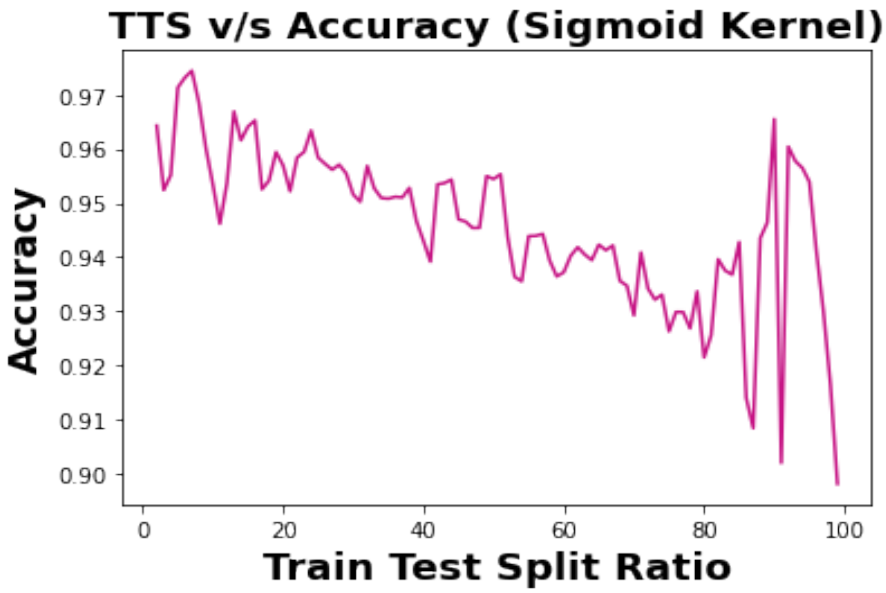

Figure 5: Train Test Split v/s Accuracy of the model for sigmoid kernel

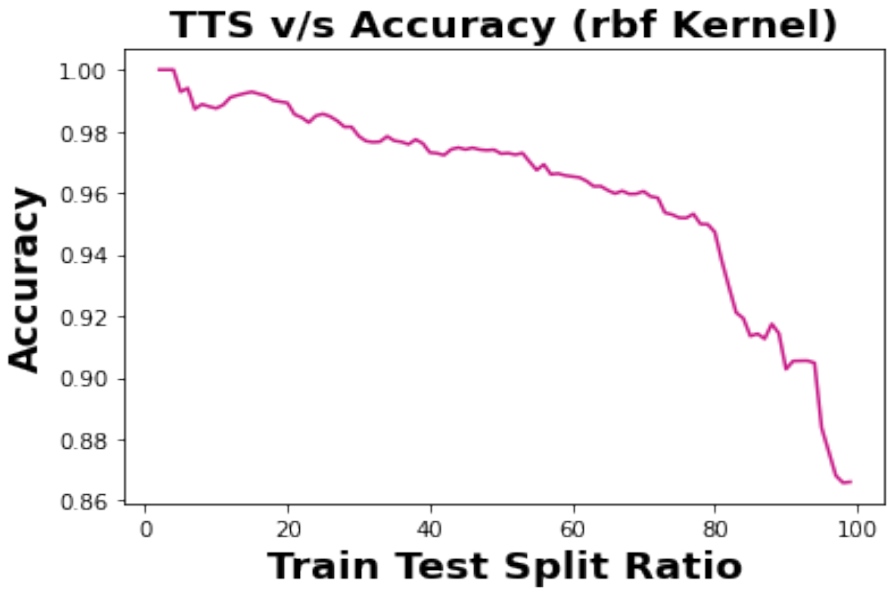

Figure 6: Train Test Split v/s Accuracy of the model for rbf kernel

	C	Train Accuracy	Test Accuracy	Test Recall	Test Precision
0	500.0	1.0	0.975233	0.828877	0.984127
1	600.0	1.0	0.975233	0.828877	0.984127
2	700.0	1.0	0.975233	0.828877	0.984127
3	800.0	1.0	0.975233	0.828877	0.984127
4	900.0	1.0	0.975233	0.828877	0.984127
5	1000.0	1.0	0.975233	0.828877	0.984127
6	1100.0	1.0	0.975233	0.828877	0.984127
7	1200.0	1.0	0.975233	0.828877	0.984127
8	1300.0	1.0	0.975233	0.828877	0.984127
9	1400.0	1.0	0.975233	0.828877	0.984127

Figure 7: Train Accuracy, Test Accuracy, Test Recall, Test Precision v/s Regularization parameter(C)

Analysis:

[1] Train Test Split ratio effects the accuracy of model.
[2] Linear Kernel is giving best accuracy when train and testing part is split into 80:20 ratio.
[3] As the testing part is increased accuracy decreases approximate linearly till 80-85% but decreases drastically after 80-85%.
[4] In Polynomial Kernel, accuracy decreases approximately linearly from 100-1% part of training data, but gives some peak at approx 80% of training data.
[5] Most shocking result came in case of sigmoid Kernel, In case of Sigmoid Kernel the accuracy is very fluctuating with the training part.
In this kernel sometimes it gives very less accuracy but besides that split ratio, it gives very high accuracy with respect to that split value.
[6] In case of RBF Kernel accuracy variation with respect to split ratio is similar to Linear Kernel, but the difference b/w these two is that in linear kernel it drastically after 80-85% training data but in case of RBF Kernel rate of declination is less for that 80-85% split part also.

It was expected that accuracy will decrease with decrease of training data, but the observations were interesting when we trained the model using different kernels.

Analysing without using Library:

Till now we have used Libraries for analysing the model but now we are going to implement the Support Vector Machine from scratch.

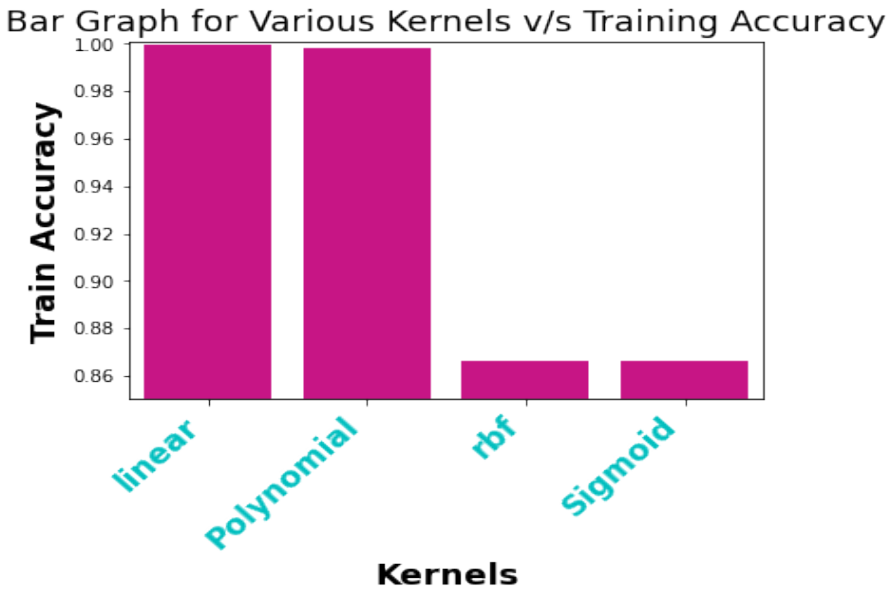

Figure 8: Train Accuracy v/s various Kernels Without using libraries

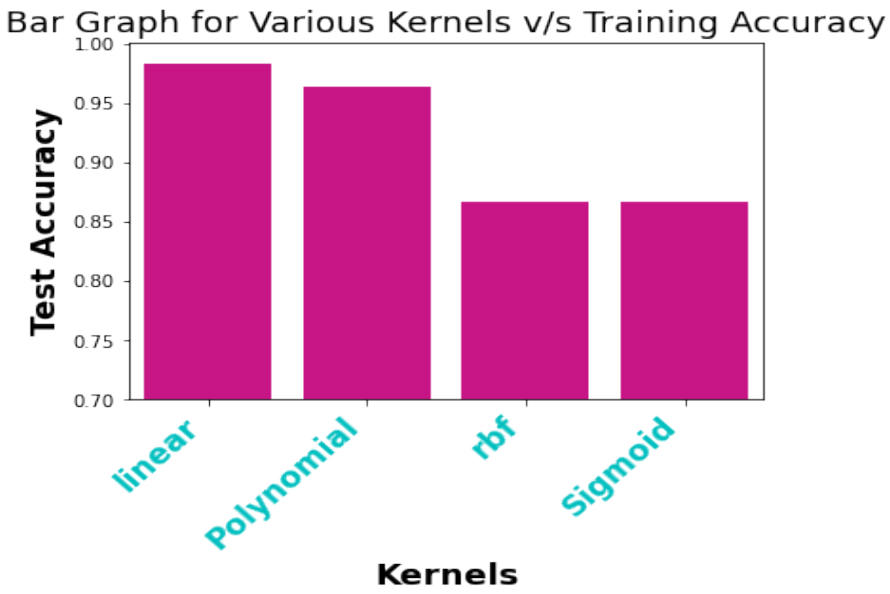

Figure 9: Test Accuracy v/s various Kernels Without using libraries

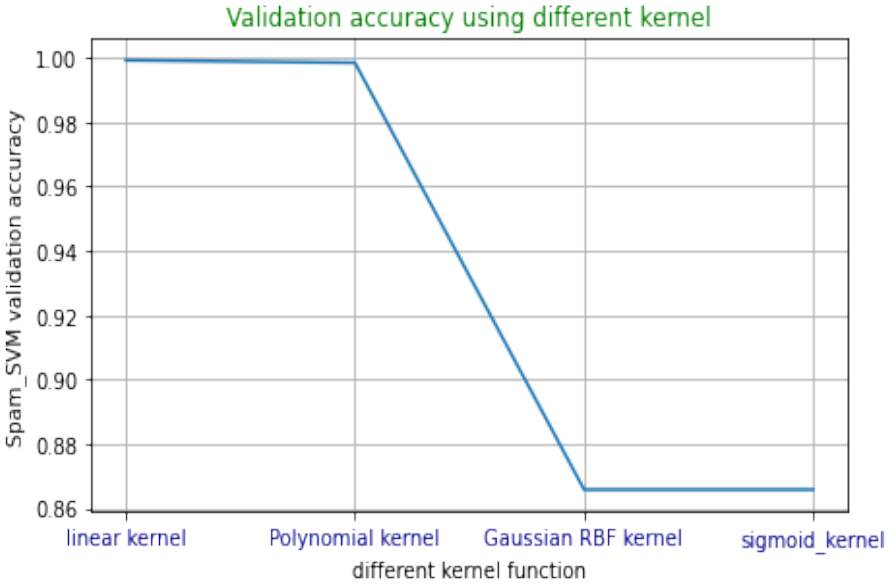

Figure 10: Validation Accuracy v/s different Kernels

Analysis:

[1] Implementing the SVM from scratch is taking much more time then using library.
[2] In case of training the model, use of library is much efficient then implementing from scratch.
Linear and polynomial Kernels are training the model very well with approximate 100% accuracy while Sigmoid and RBF Kernel in not performing that much better.
[3] In case of testing the model, Linear Kernel is performing best with approximate 98% accuracy while polynomial Kernel is also performing very good with approximately 96% accuracy.
Polynomial Kernel implementation from scratch is performing better then the using of Library, in case of testing the model.
[4] Sigmoid and RBF kernels are performing similar in case of testing the model.
[5] Validation accuracy of Linear and polynomial Kernel is approximate 100% while in case of other Kernels it is less, around 86.5%.

Conclusion:

In this report we analysed the Support Vector Machine for spam filtering, we analysed the model on various aspects like what is training and testing accuracy pattern of the model using libraries and without using libraries.
How SVM performs when we very the train test split ratio on different Kernels and we get to know the interesting patterns.