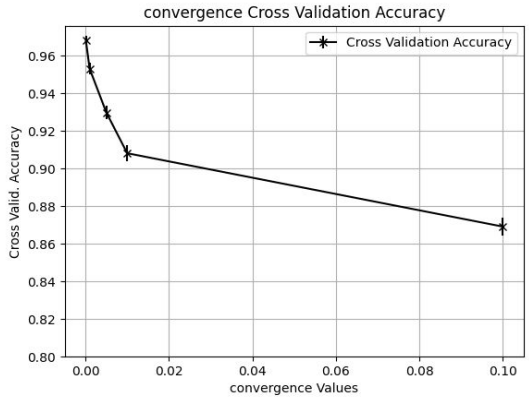
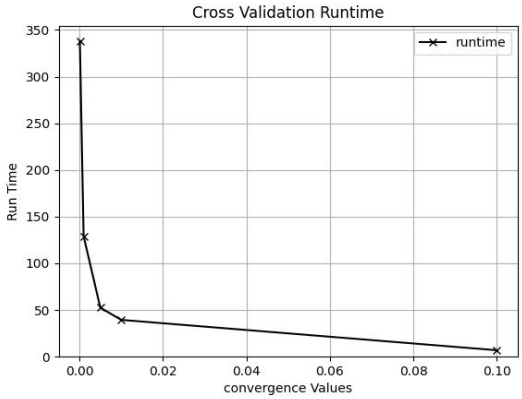
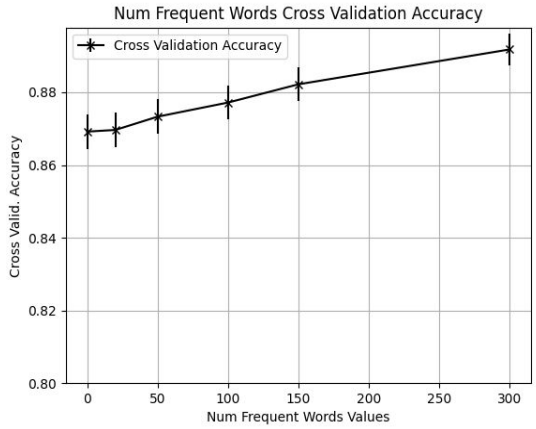
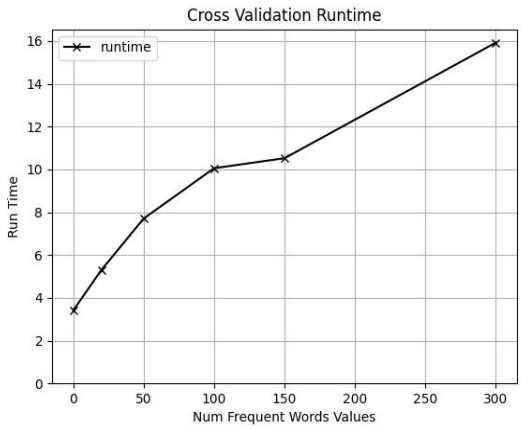
The plots you produce for the first 4 steps of optimization (one sweep on each of the hyperparameters).



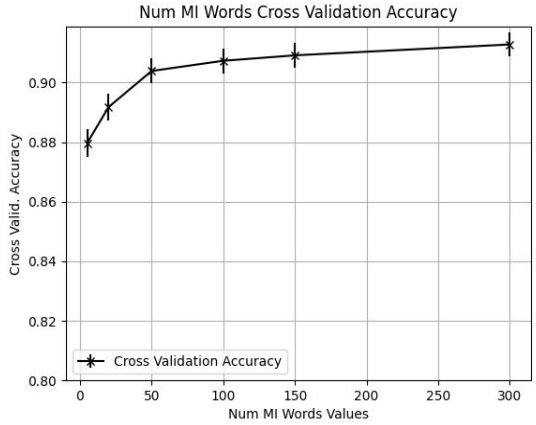


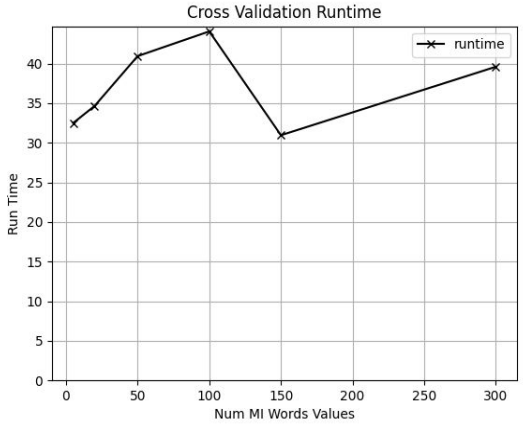
Convergence is very costly in terms of runtime but can improve accuracy quite significantly. The problem was if I had the compute, I would have tried to push the convergence limit to lower than 0.0001.



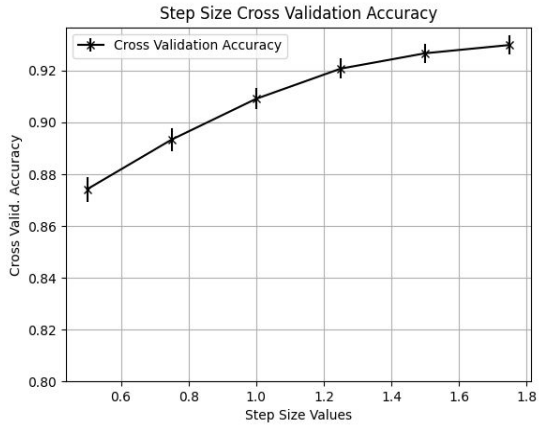


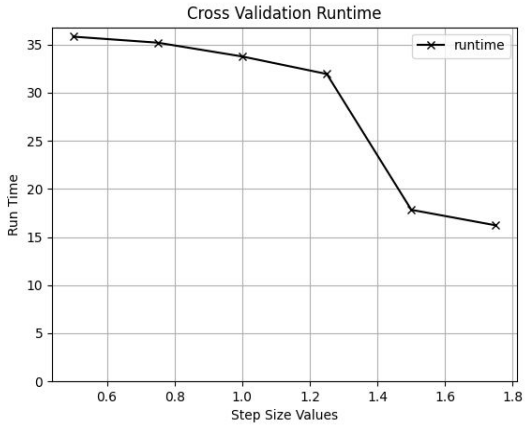
Performance continues to increase with more frequent words at a very slow pace. It seems that the run-time increases linearly with the number of features in this area, so it might not be worth the effort to keep adding such features as the return on investment is not very high.





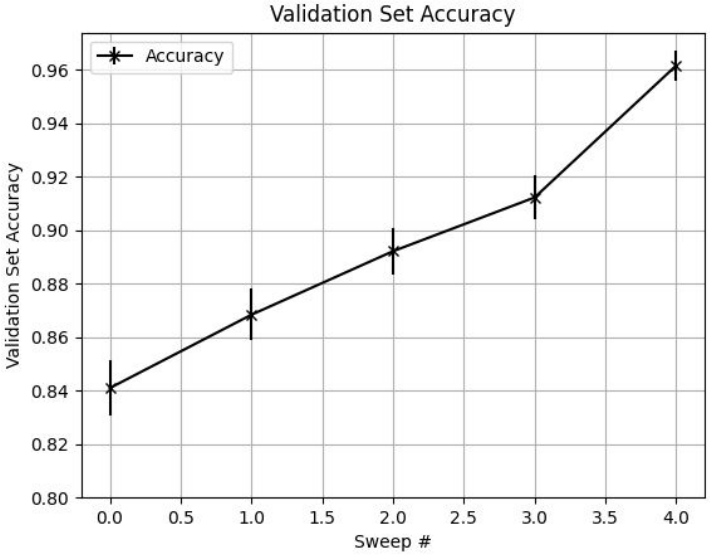
Mutual information features first increase accuracy quite a bit for a small increase in the number of features and then the relationship becomes more linear. Again for the runtime costs and the diminishing returns, I would not try to increase the number of words further.





This is an interesting trade-off, it seems there is significant improvement in runtime by increasing the step size without sacrificing accuracy during cross-validation. Again if I had more compute, I would try different learning rate schedules to further optimize runtime against classification accuracy.

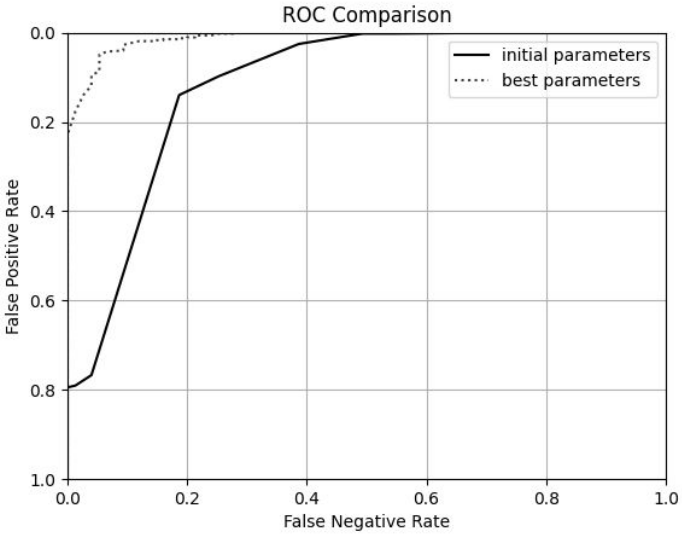
The final plot you produce showing how validation accuracy improves throughout the entire optimization run with x-axis being the sweep number.



The final hyperparameters you learned.

{'stepSize': 1.75, 'convergence': 0.0001, 'numFrequentWords': 300, 'numMutualInformationWords': 150, 'accuracy': 0.9616087751371115, 'accuracyErrorBound': 0.005504226953045976}

The ROC plot you produced on the test set (comparing your best model to the initial model).



A short analysis of how the initial model relates to the optimized model, including confidence intervals and performance at low false positive rates.

The optimized model is significantly better across all FNR and FPR. Further the more optimized model with higher accuracy has significantly narrower confidence intervals on the validation set. I only did 4 sweeps because once the convergence was set to 0.0001, it became too hard to do more sweeps.