*Notes: While cross validating, I used 5 folds. I only used the training set. In logistic regression, I checked for convergence every 10 steps.*

3.0 points - The plots you produce in step 2 for the first 4 steps of optimization (one sweep on each of the hyperparameters). Label clearly so the TA can follow the steps (and potentially reproduce them)!
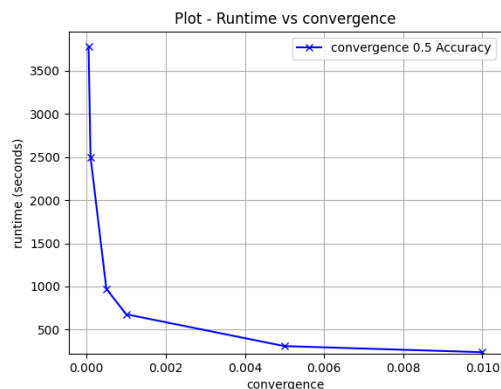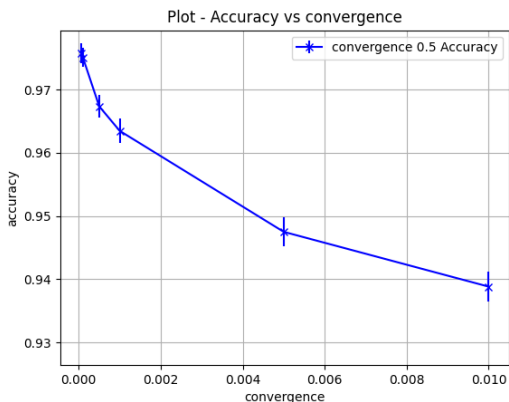
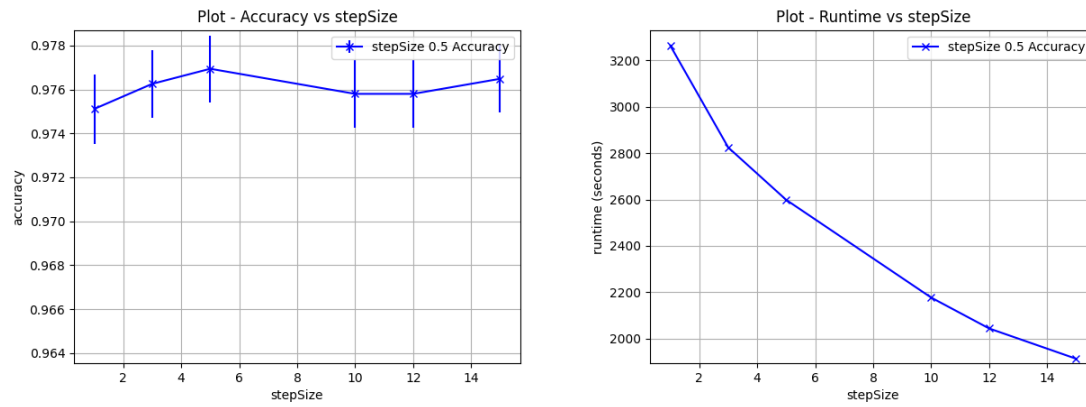*Step 1: Sweep number of words by mutual information*
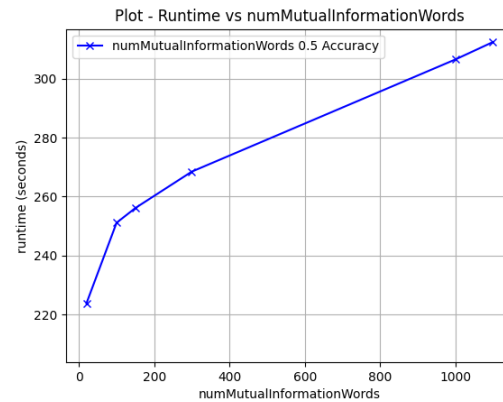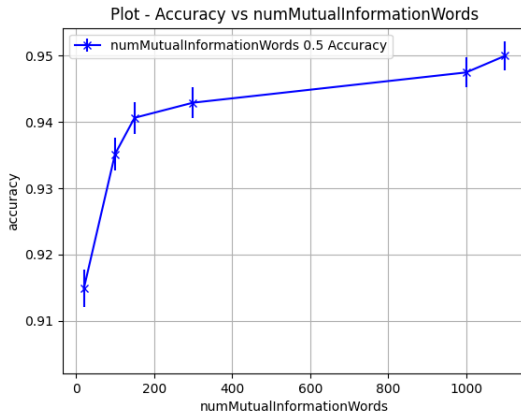
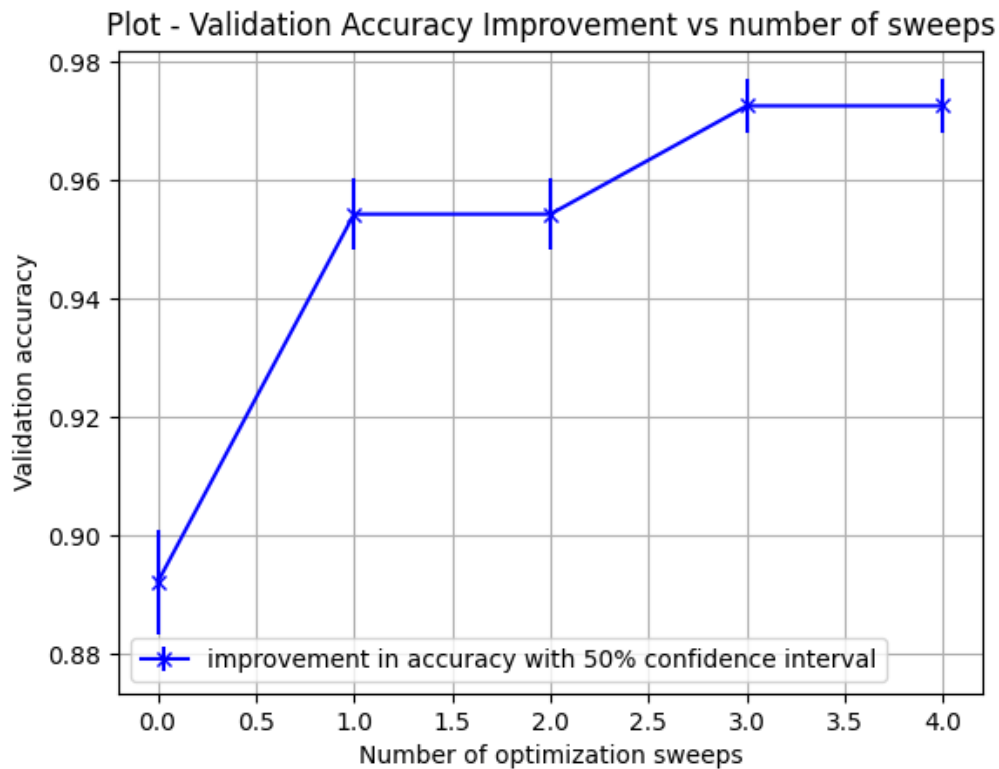*Step 2: Sweep number of words by frequency*
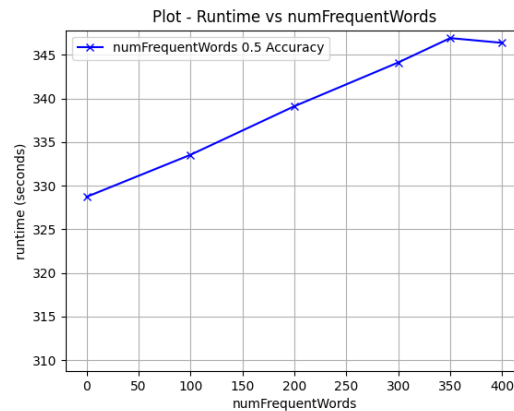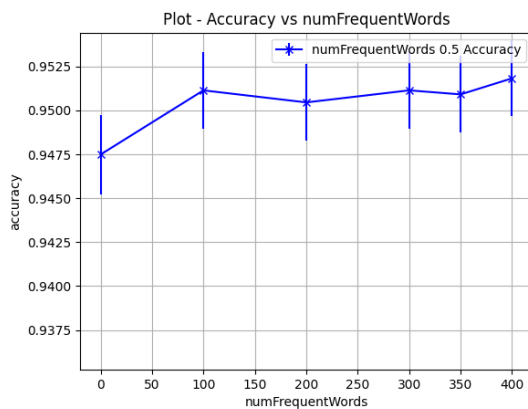
Step 3: Sweep convergence

Step 4: Sweep step size

1.0 points - The final plot you produce in step 4, showing how validation accuracy improves throughout the entire optimization run.

I was slightly confused as to what a point on the X axis should represent. The assignment asks for "the number of optimization sweeps done
to that point on the X-axis". A single hyperparameter sweep represents trying 6 distinct values for a single parameters, and there were 4 parameters.

There are 5 points on the X axis. X=0 represents the initial model, without any optimization. Every successive point on the X axis represents a single hyperparameter sweep with 6 distinct values.

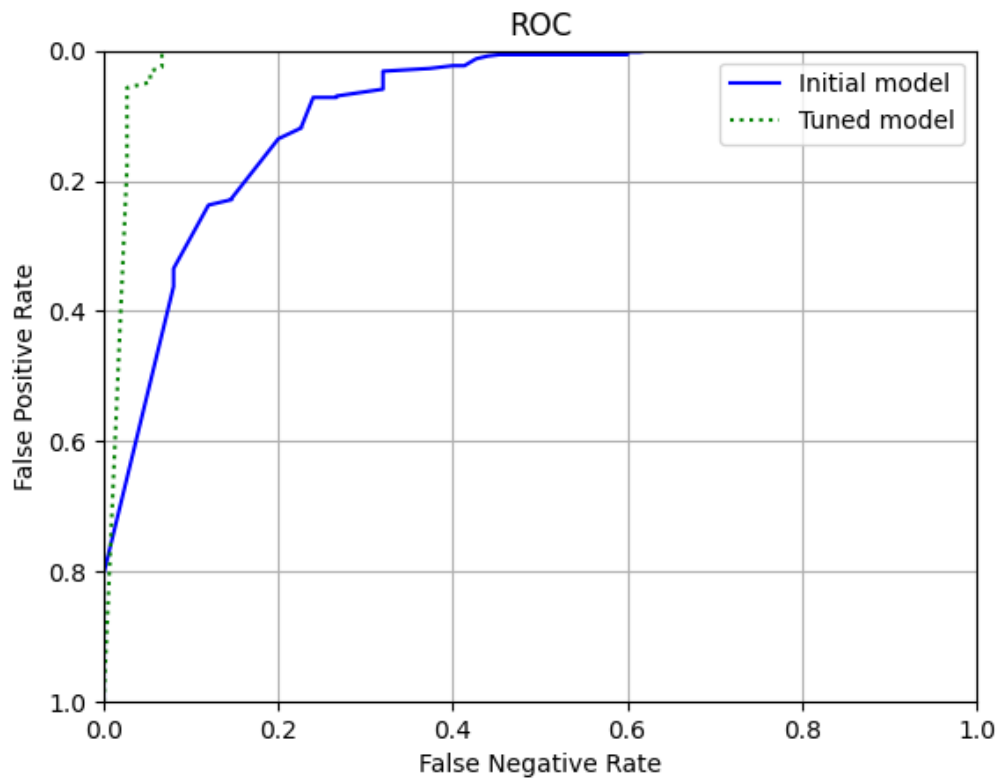1.0 points - The final hyperparameters you learned.

'stepSize': 15.0

'convergence': 0.0001

'numFrequentWords': 0

'numMutualInformationWords': 1000
1.0 points - The ROC plot you produced on the test set (comparing your best model to the initial model).

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

The initial model results in a test set accuracy of 91.40 with 50% 2 sided bounds (90.66%, 92.21%) and the optimized model results in a test accuracy of 99.08% with 50% 2 sided bounds (0.9881%, 99.35%). Performance at low false positive rate is much more optimized compared to the initial model, with the new model obtaining 0 false positive rate with as little at (6.66%) false negative rate. The initial model would only give 0 false positives at (62.66%) false negatives. The optimized model is better than the initial model because of a larger AUC. Operating points of interest - 0% FPR (False positive rate), 10% FPR, and 50% FPR all result in lower false negatives in the optimized model than the original model.