Appendix B: Experimental Procedure

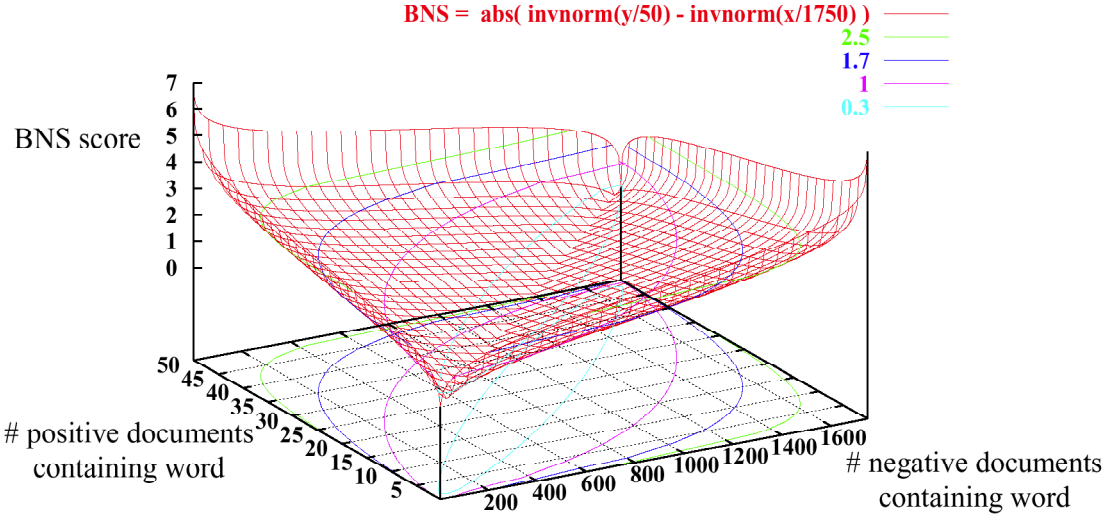
The following pseudo-code details the experimental procedure for collecting and processing the data:

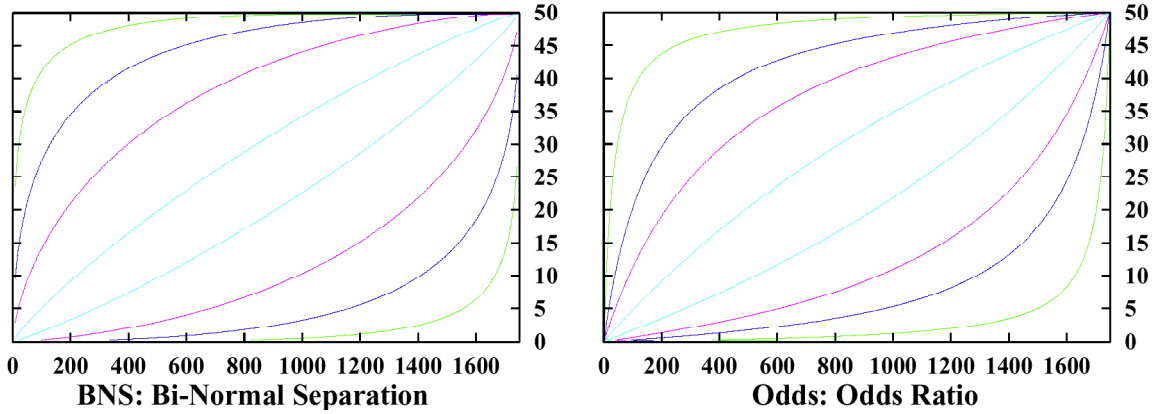
```
for each dataset d:
|  for each class c of the dataset d, using c as the positive class and the others as the negative class:
|     for N=5 random trials:
|     |  for each random splits of the dataset for 4-fold stratified cross-validation:
|     |     for each feature selection metric:
|     |     |  for each number of features to select—10,20,50,100,200,500,1000,2000:
|     |     |     select top features via the metric using only the training set
|     |     |     train a SVM classifier on the training set split
|     |     |     measure performance on the testing set split
|     |     |  end
|     |     end
|     |  end
|     |  record the 4-fold cross-validation scores for  accuracy, precision, recall, F-measure
|     |  (also record the maximum attained over any number of features)
|     end
|     determine the average scores over all N=5 trials
|     record these under the unique key: dataset, class, feature selection metric, number of features
|     (record the 'average maximum' attained with key: dataset, class, feature selection metric)
|  end
end
macro-average the performance measures over all 229 classification tasks
record the results under the key: feature selection metric, number of features
for each goal—accuracy, precision, recall, F-measure:
|  for each of the 229 tasks:
|     determine the best 'average maximum' performance attained by any metric & number of features
|  end
|  for each tolerance level t from 1% to 10%:
|     for each feature selection metric:
|     |  determine percentage of problems on which the metric ties for best within t% tolerance
|     end
|  end
end
```

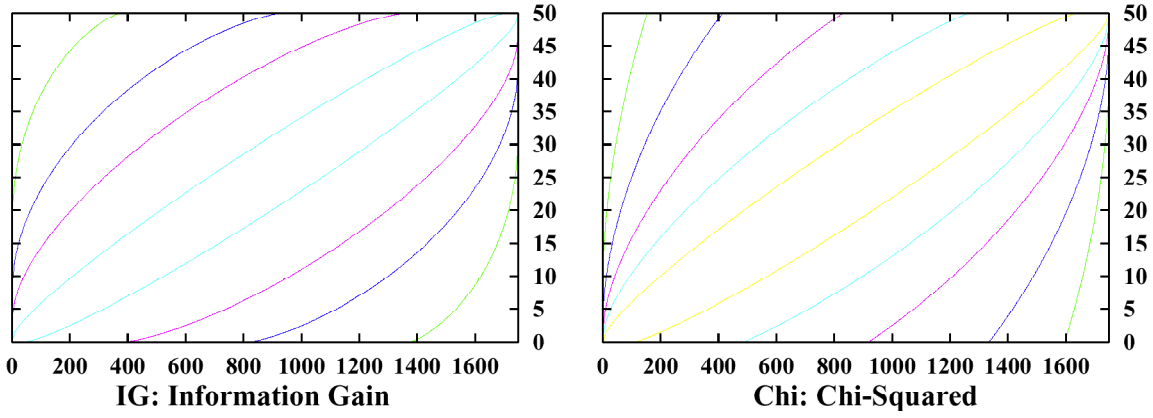
This resulted in nearly half a million invocations of the learning algorithm, run on over a hundred different machines. We performed additional variations of this procedure for several side studies.

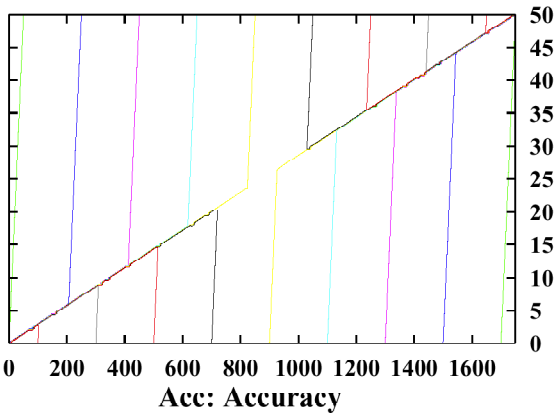
Appendix C: Graphical Analysis of Feature Selection Metrics

This section contains color graphs illustrating the shape of each feature selection metric. Each graph shows a set of isocline contours, as a topographic map. For reference, compare the BNS isoclines to its three-dimensional plot below, which includes contours both on the surface and projected down on the plane below. Like colored isoclines indicate the same value within a plot (but are not comparable between plots). Any feature in the top left or bottom right corners would be a perfect predictor, and so each metric scores highest in these corners, except DFreq, which gives highest value to the most frequent features, i.e. the top right.

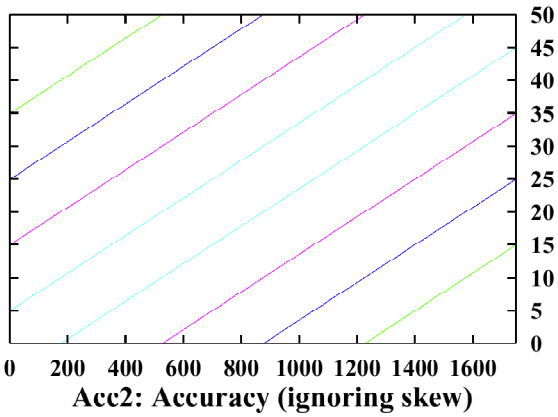




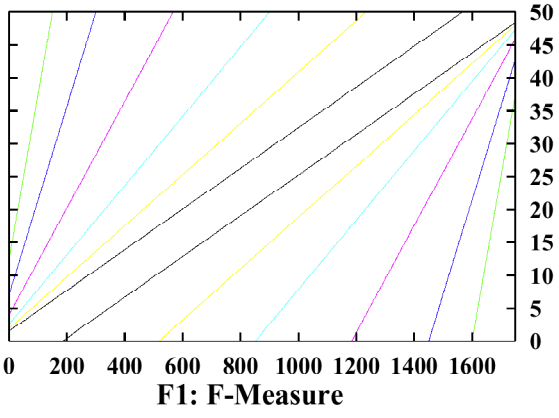




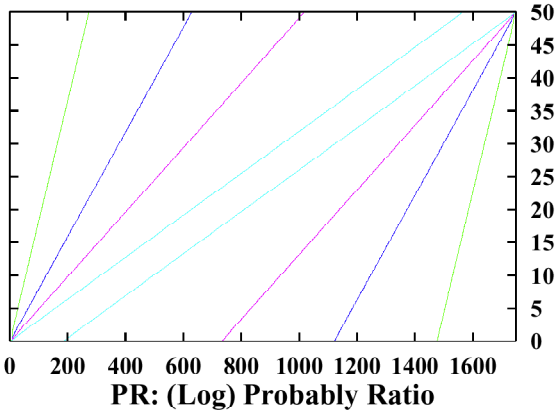
Acc: Accuracy



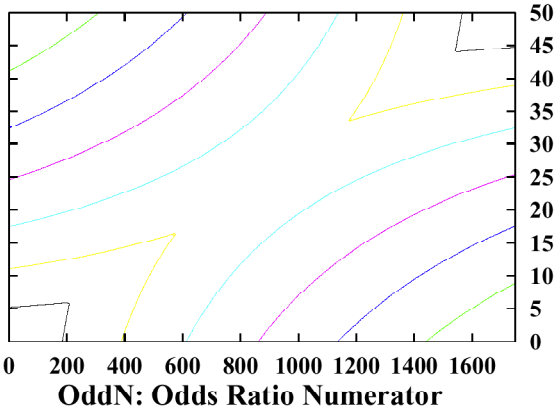
Acc2: Accuracy (ignoring skew)



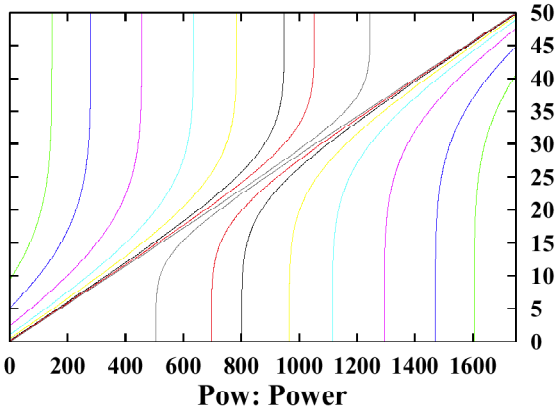
F1: F-Measure



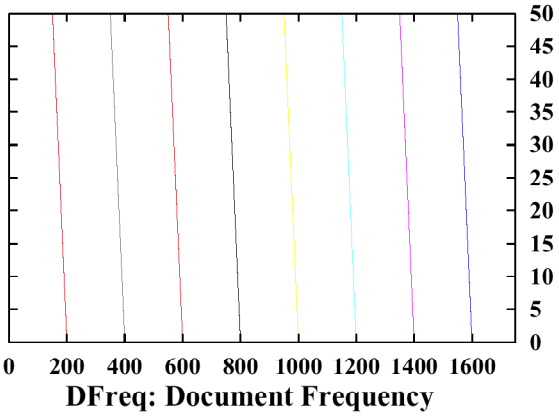
PR: (Log) Probably Ratio



OddN: Odds Ratio Numerator



Pow: Power



DFreq: Document Frequency

Rand: Random (varies)

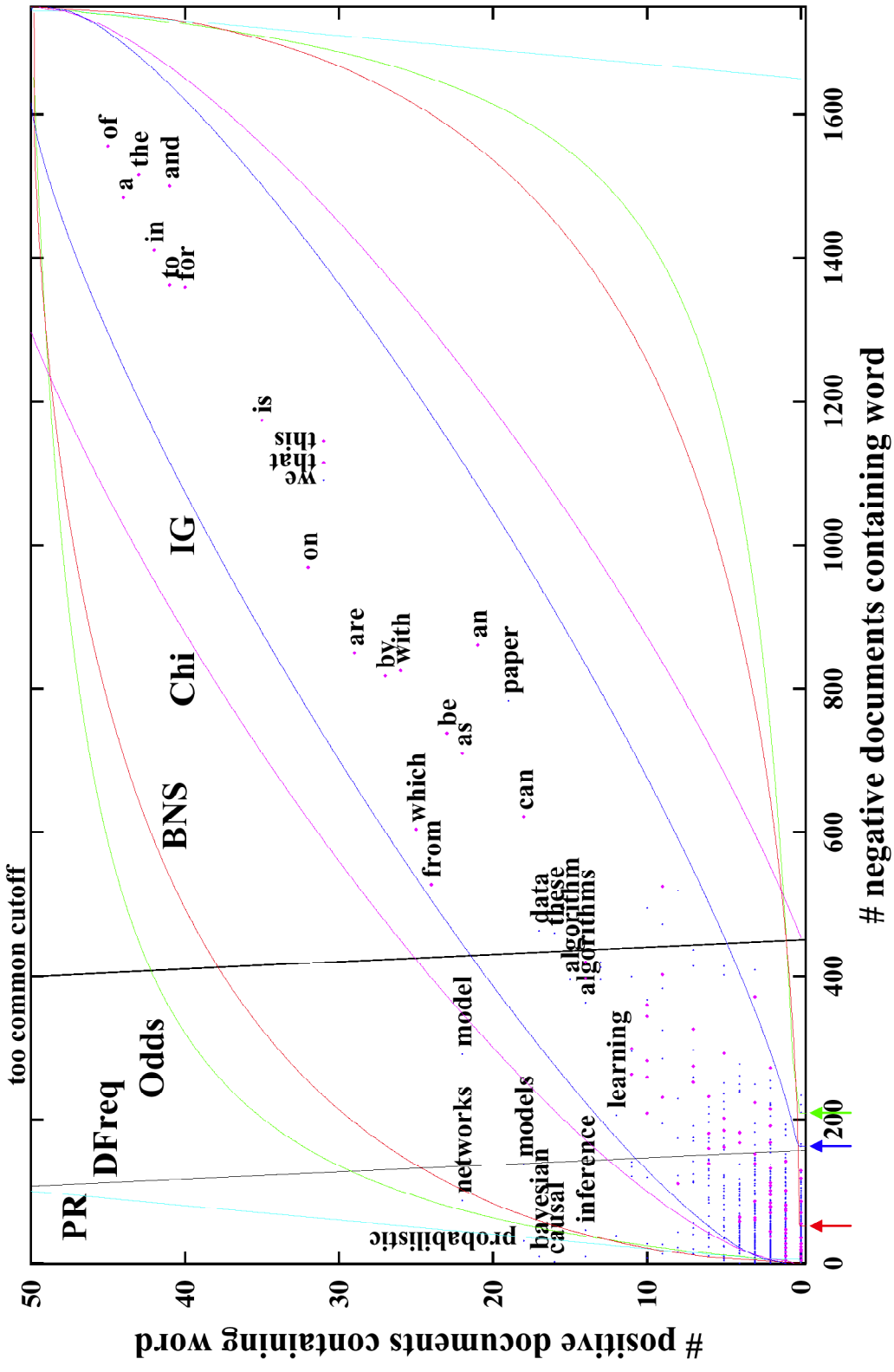


Figure 13. Color version of Figure 2. Note that BNS selects many more negative features.

Appendix D: Additional Color Graphs of Results

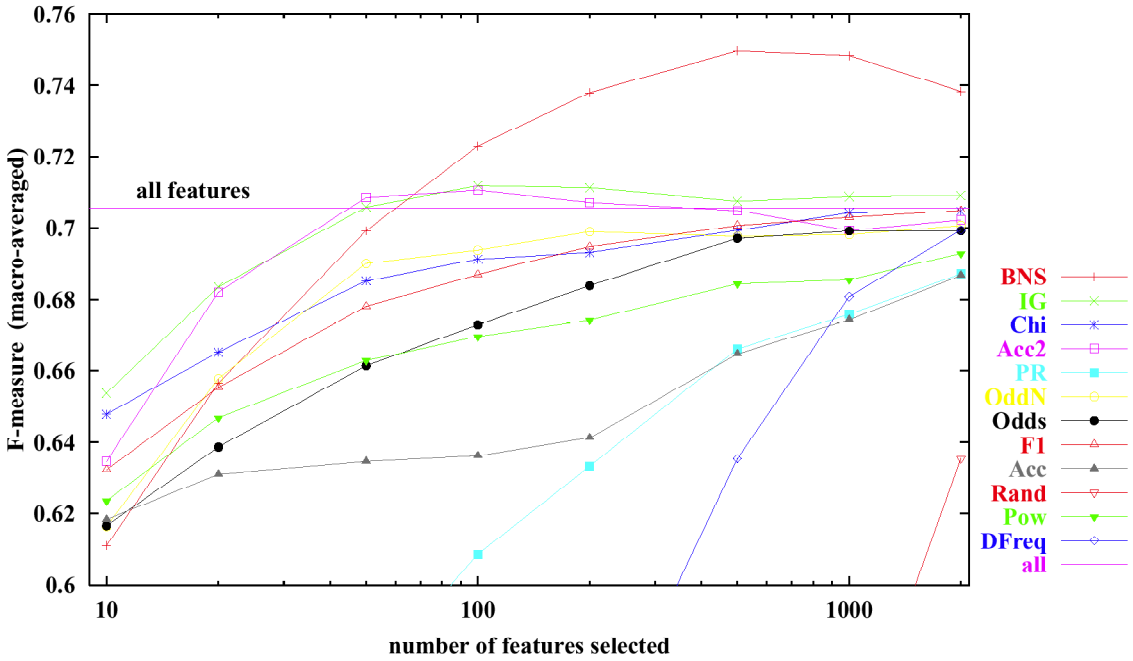


Figure 14. Color zoomed version of Figure 4. F-measure macro-averaged over repeated trials on 229 binary text classification tasks, as we vary the number of features selected for each feature selection metric.

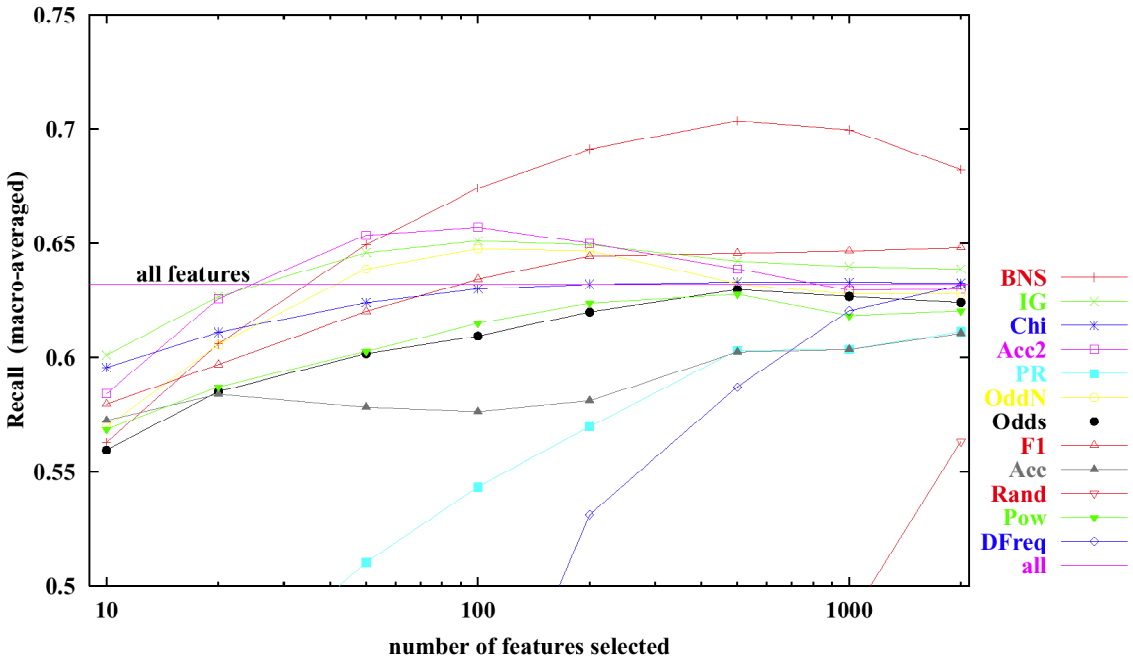


Figure 15. Same as Figure 14, but for recall.

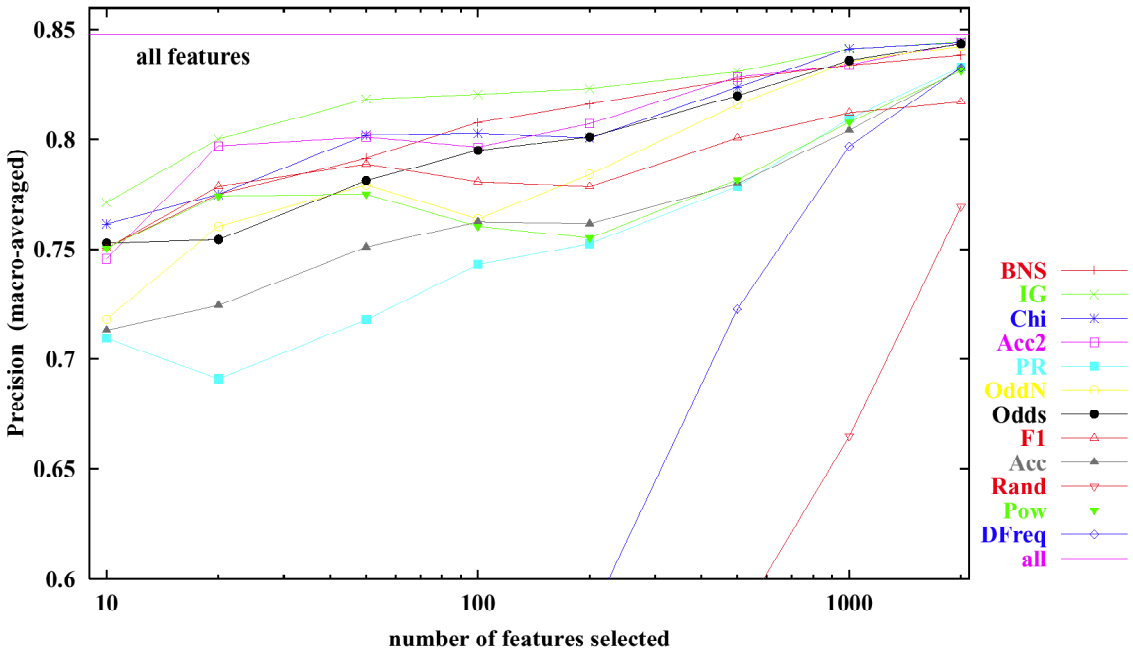


Figure 16. Same as Figure 14, but for precision.

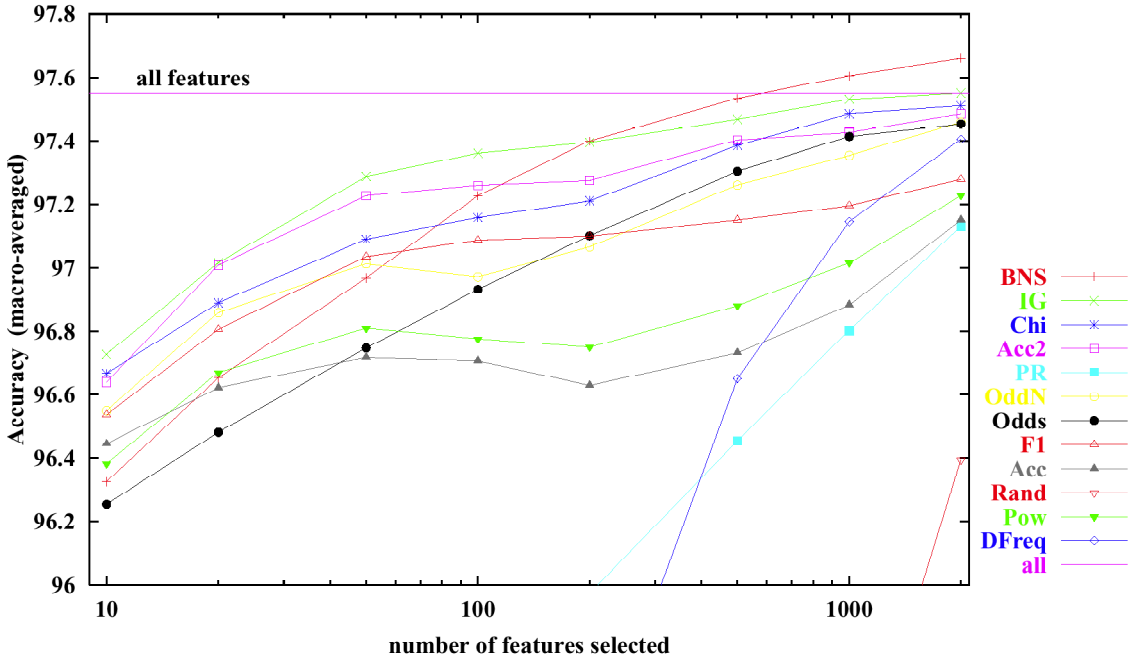


Figure 17. Same as Figure 14, but for accuracy.

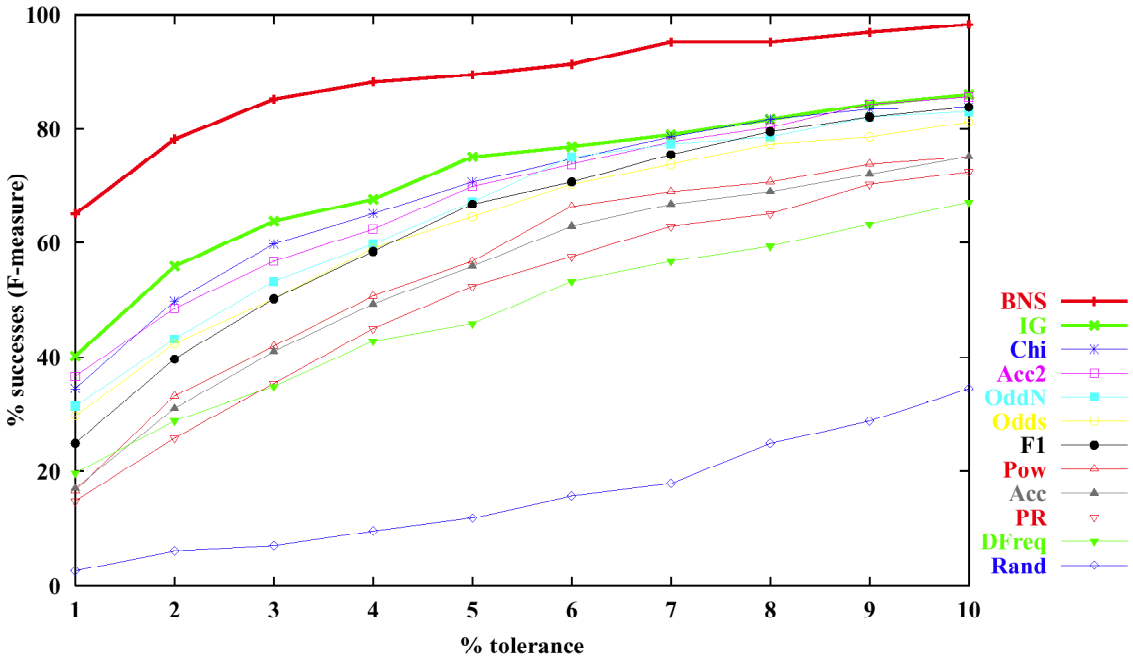


Figure 18. Color version of Figure 9a. Percentage of problems on which each metric scored within x% tolerance of the best F-measure achieved by any metric.

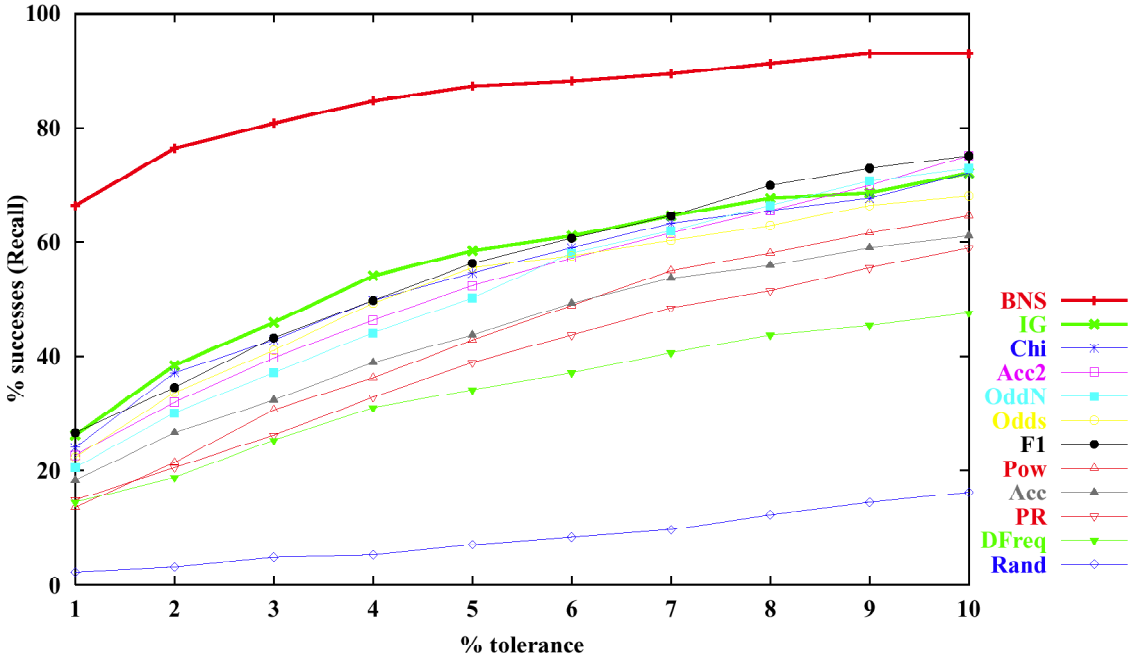


Figure 19. Same as Figure 18, but for recall.

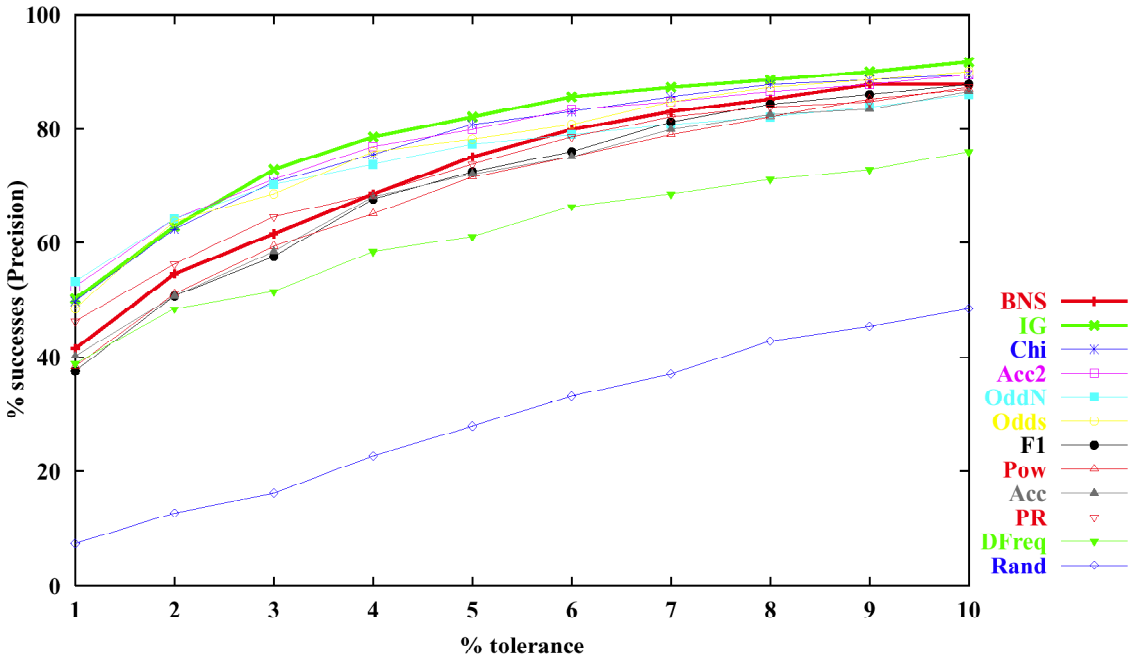


Figure 20. Same as Figure 18, but for precision.

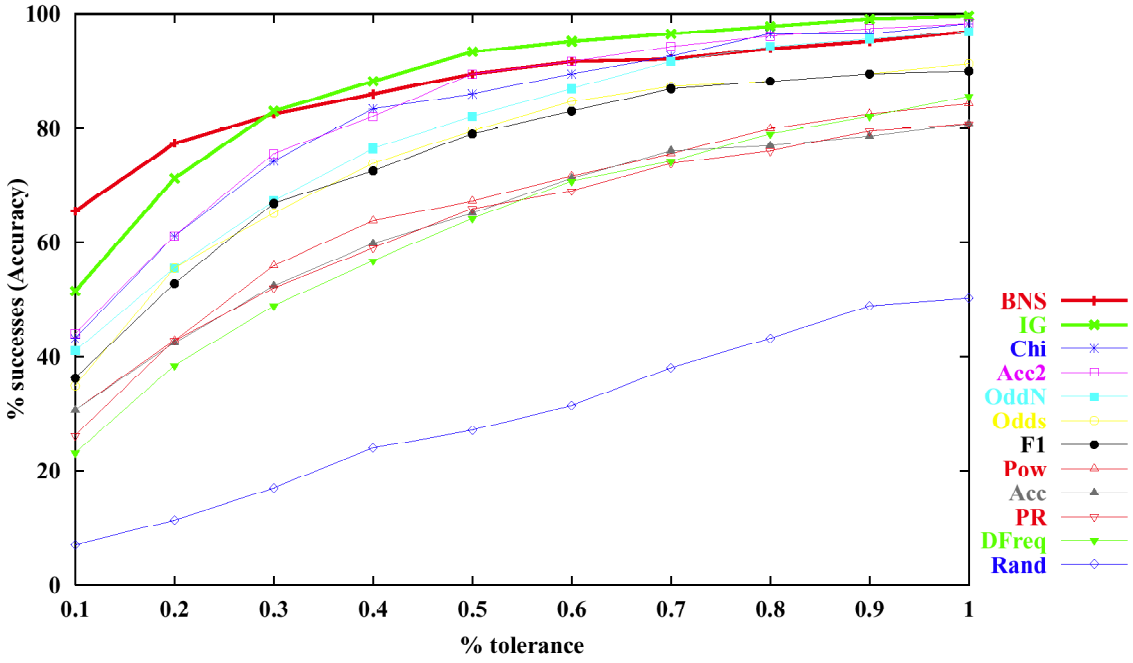


Figure 21. Same as Figure 18, but for accuracy. Note the smaller x-axis scale.

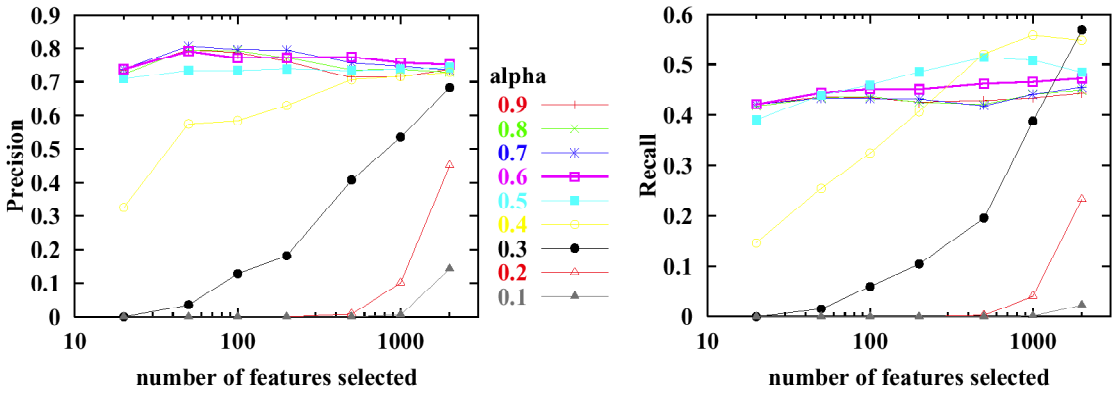


Figure 22. (a) Precision vs. number of features selected, as we systematically vary the preference weight α for positive vs. negative features in BNS. High α prefers positive features. (b) Same, but for Recall. (Note: only the Cora dataset was used, so the results are not comparable to other figures. While α=40% yields a superior F-measure for the 36 classes of the Cora dataset on average, for the entire benchmark of 229 tasks, we were unable to beat the F-measure of unweighted BNS by experimenting with different values of α.)