On-line Spam Filter Fusion

Thomas Lynam & Gordon Cormack

originally presented at SIGIR 2006

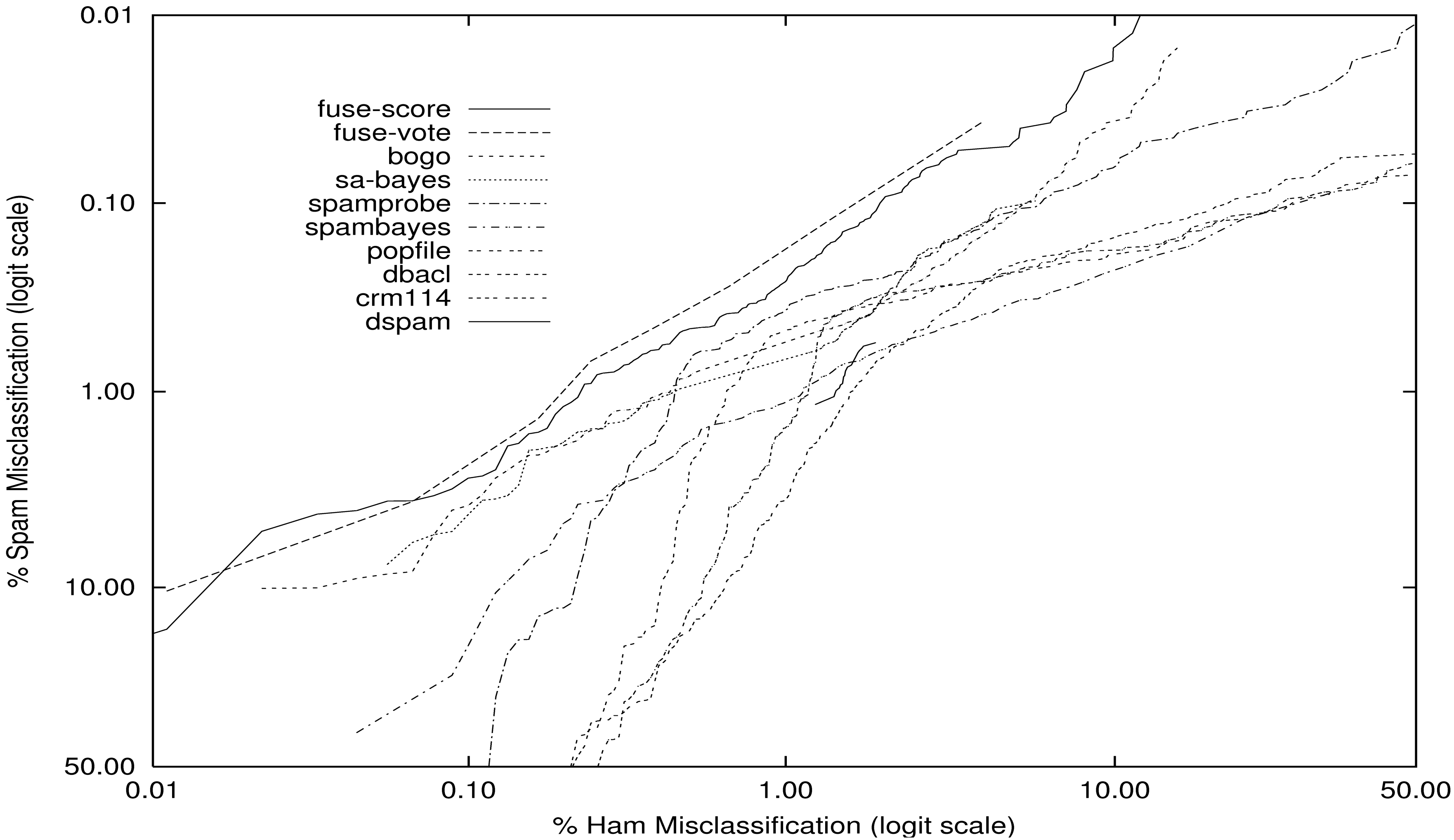
On-line vs Batch Classification

- Batch Hard Classifier
  - separate training and test data sets
  - Given ham/spam classification of training set
  - Compute ham/spam class for each message
- On-line Soft Classifier
  - Chronological sequence
  - Compute *spamminess* for each in sequence
    - ham/spam class by comparing to fixed threshold
  - Given ham/spam classifcation afterwards
    - Immediate, correct feedback (idealized user)

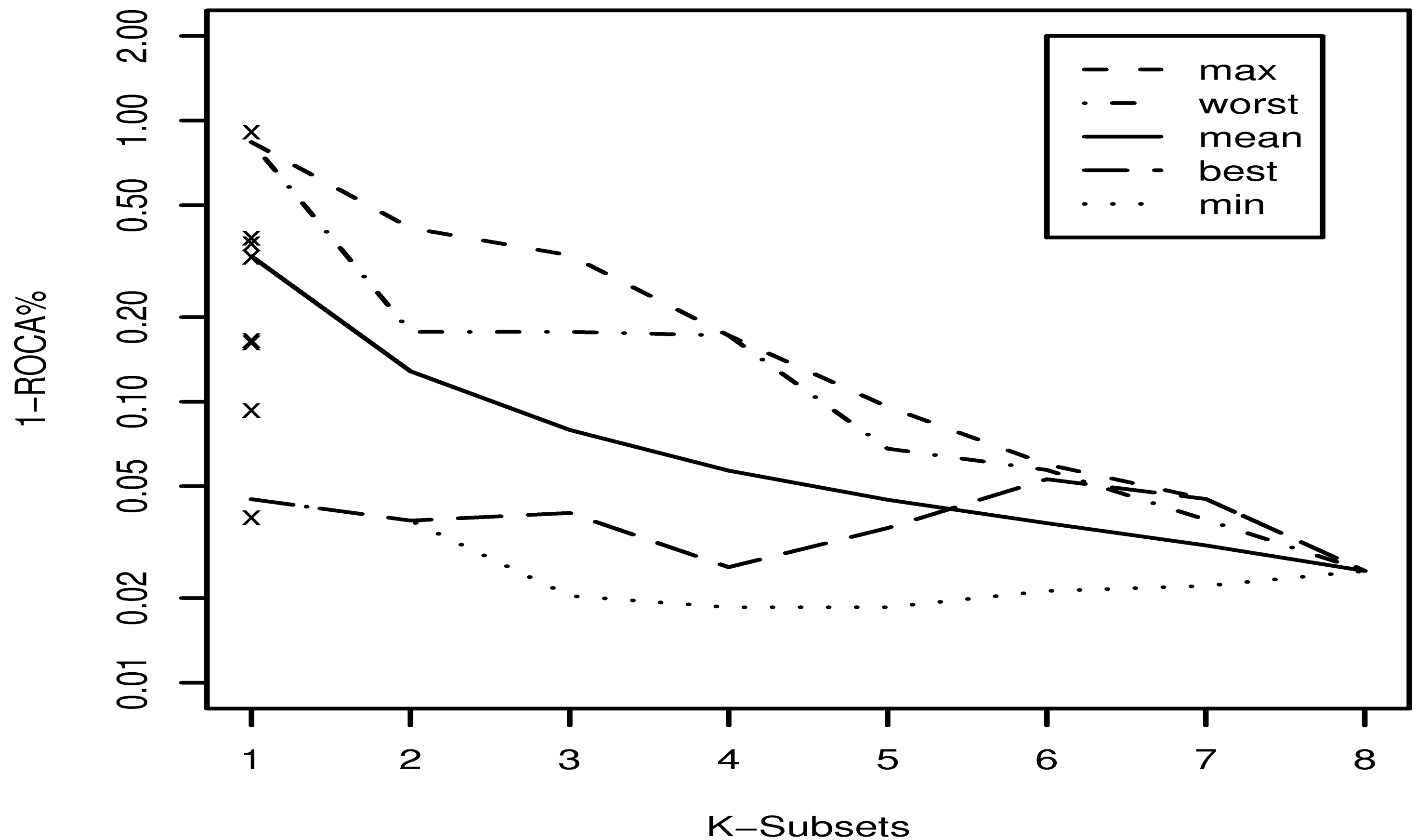
Measures of Success & Failure

- ROC Curve
- ROC Area *above* the curve (as percentage)
- Ham & spam misclassification rates
  - Sm(%) when threshold set for Hm(%) = .1
- 95% confidence intervals
  - For ROC area (logit transformed)
  - For difference between ROC areas (logit trans)
    - Significant result: difference interval excludes 0

Pilot Test ROC
(Mr X corpus)



Pilot Tests K Subsets
(Mr X corpus)



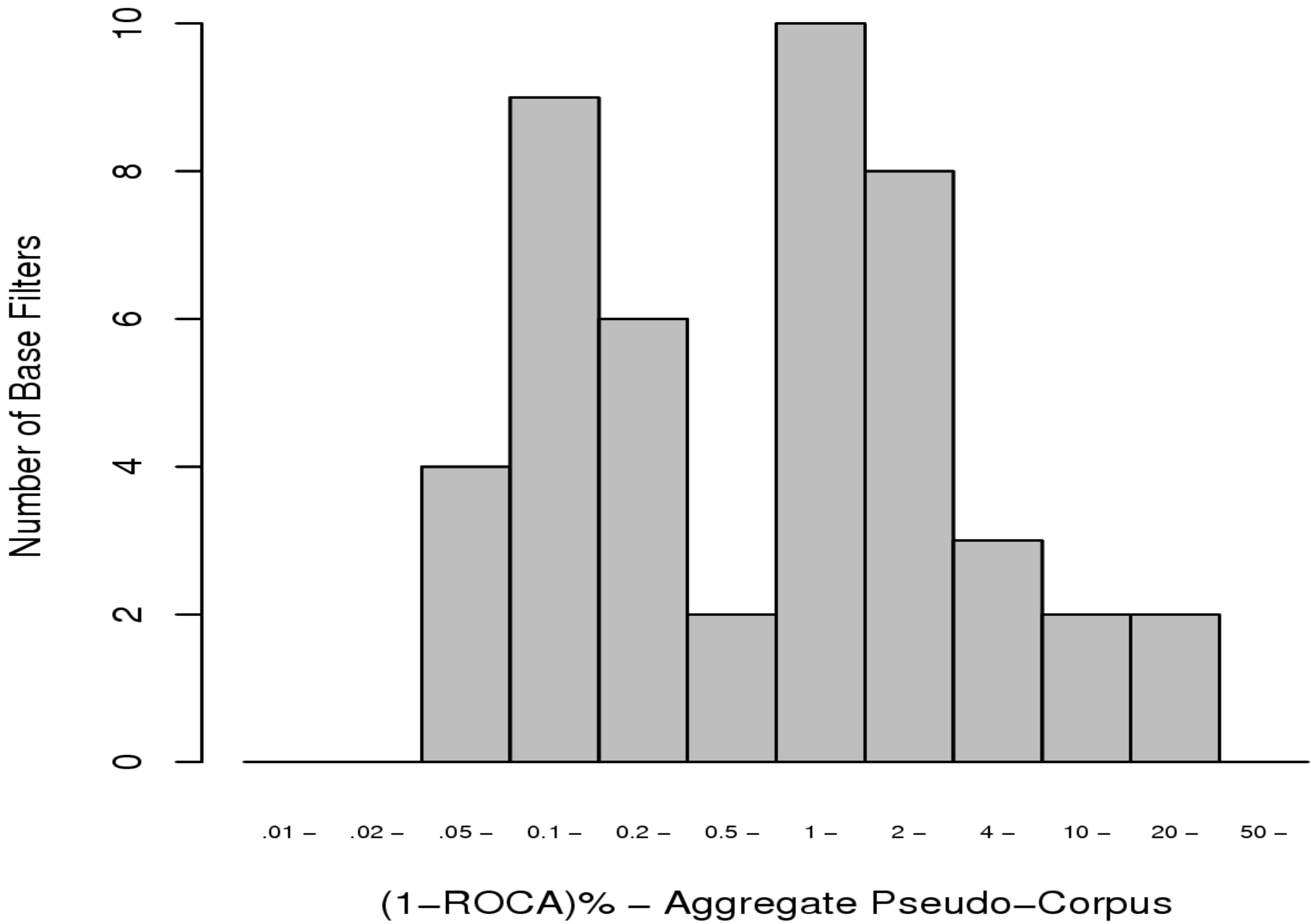
TREC 2005 SPAM TRACK

- 4 corpora
  - 1 public, 3 private
- submit runs on public corpus
- submit filter to be run on private corpora
- 53 runs (different filters)
- 17 different organizations represented

TREC Spam Track Corpora

	Ham	Spam	Total
Mr X	9038	40048	49086
S B	6231	775	7006
T M	150685	19516	170201
Full	39399	52790	92189
Aggregate	205253	113129	318482

TREC Filter Performance Distribution



Fusion Methods

- Best System (Baseline)
- Voting
- SumScore
- Log-odds Averaging
- SVM
- Logistic Regression

Log-odds Averaging

- 53 unknown systems
  - unknown min/max scores.
  - linear/nonlinear scoring
- How to normalize scores?

$$L_n = \log\left(\frac{\left|\{ i<n \mid s_i \leq s_n \text{ and ith message is spam}\}\right| + \epsilon}{\left|\{ i<n \mid s_i \geq s_n \text{ and ith message is ham }\}\right| + \epsilon}\right)$$

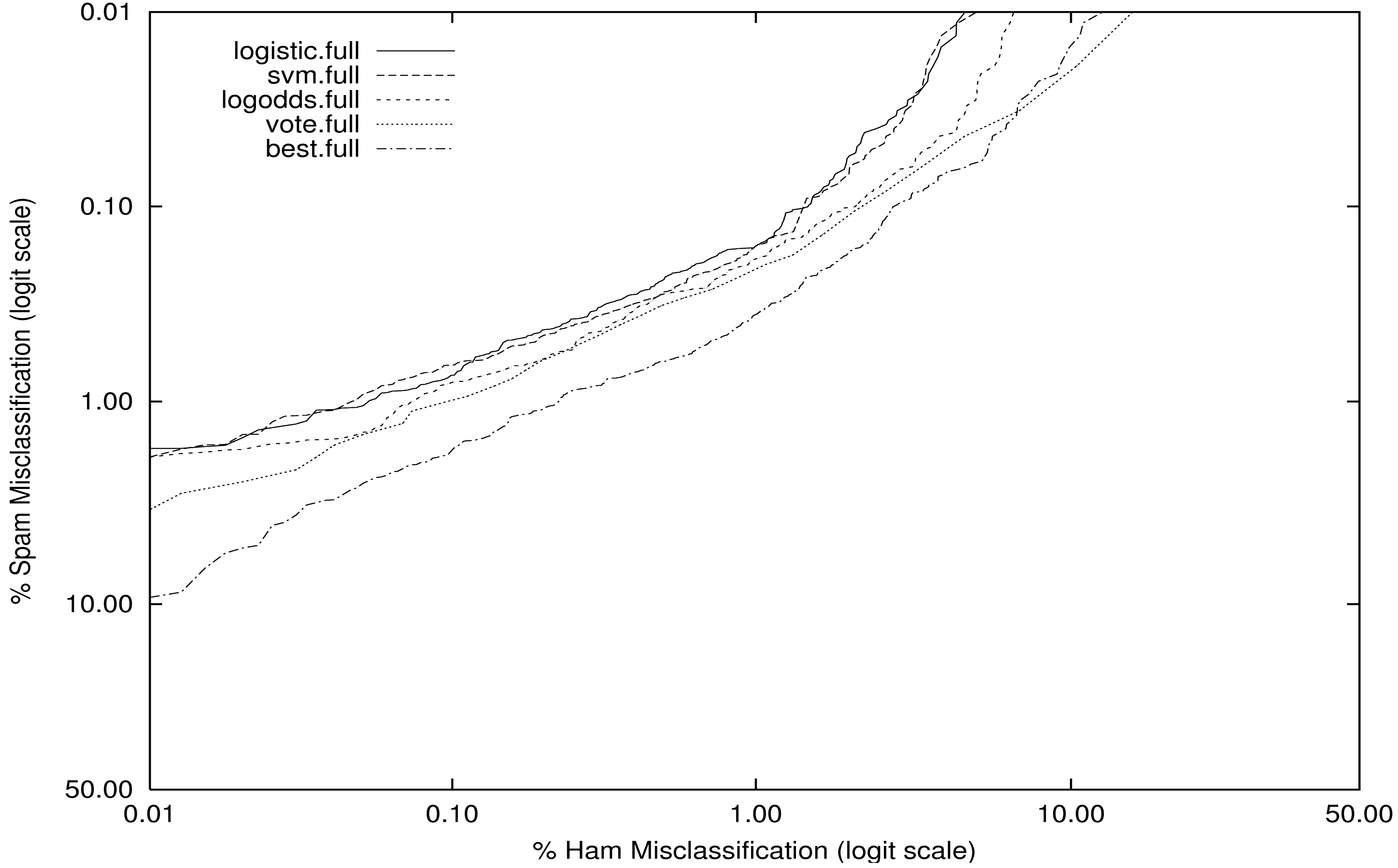
SVM Fusion

- SVM^{light}
  - default kernel and parameters
  - log-odds averaging used as features
- training set sizes of
  0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000
- output used as spamminess score

Logistic Regression

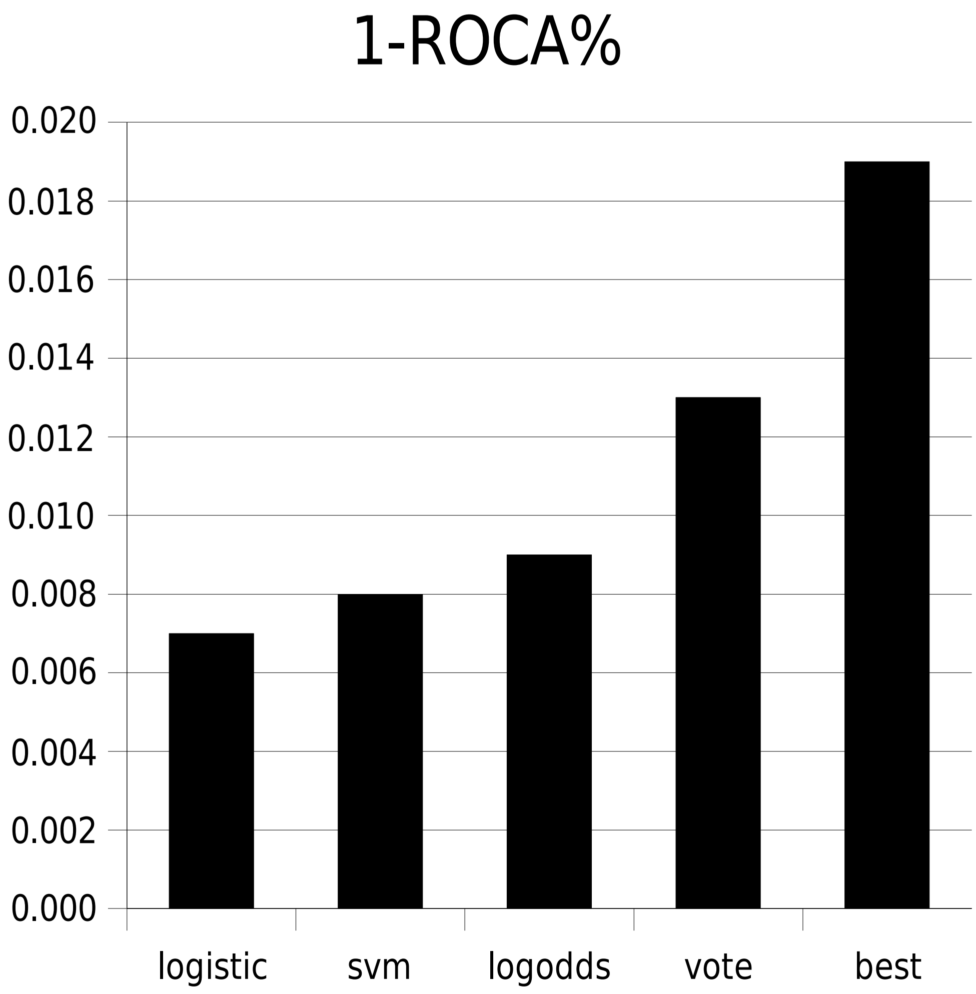
- LR-TRIRLS logistic regression package
- weights predict prior classification
- Negative weights considered over-fitting
- initial weight equal 1/number of filters
- training set sizes of
  0, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 2100, 4100, 9100, 19100, 39100, 69100, 99100, 129100, 159100.
- weighted average uses as spamminess score

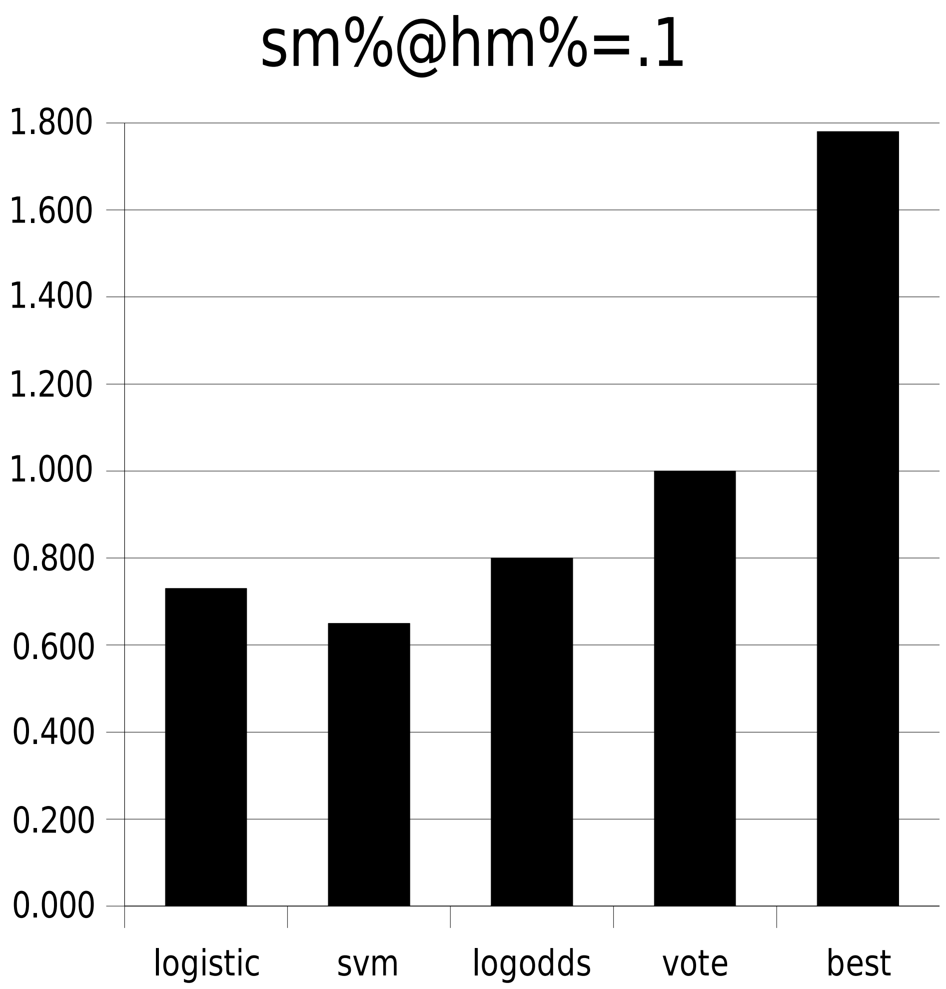
ROC
(Full Corpus)



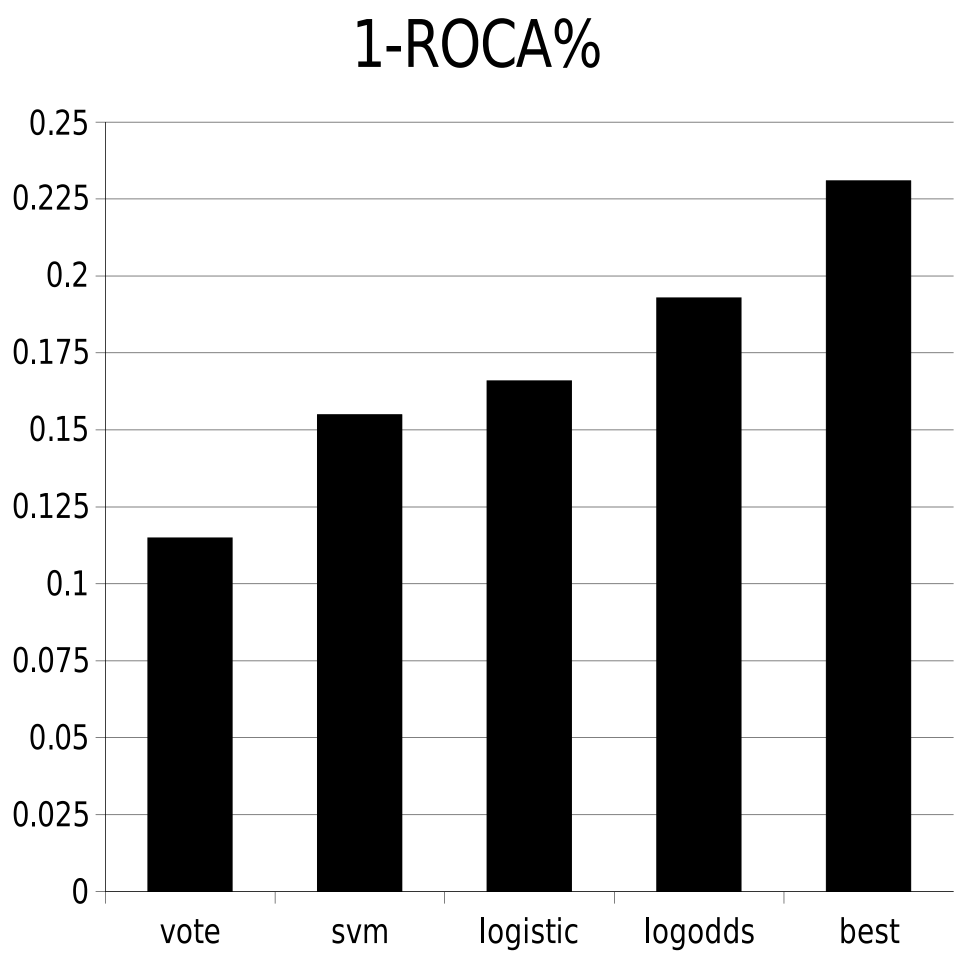
University of Waterloo

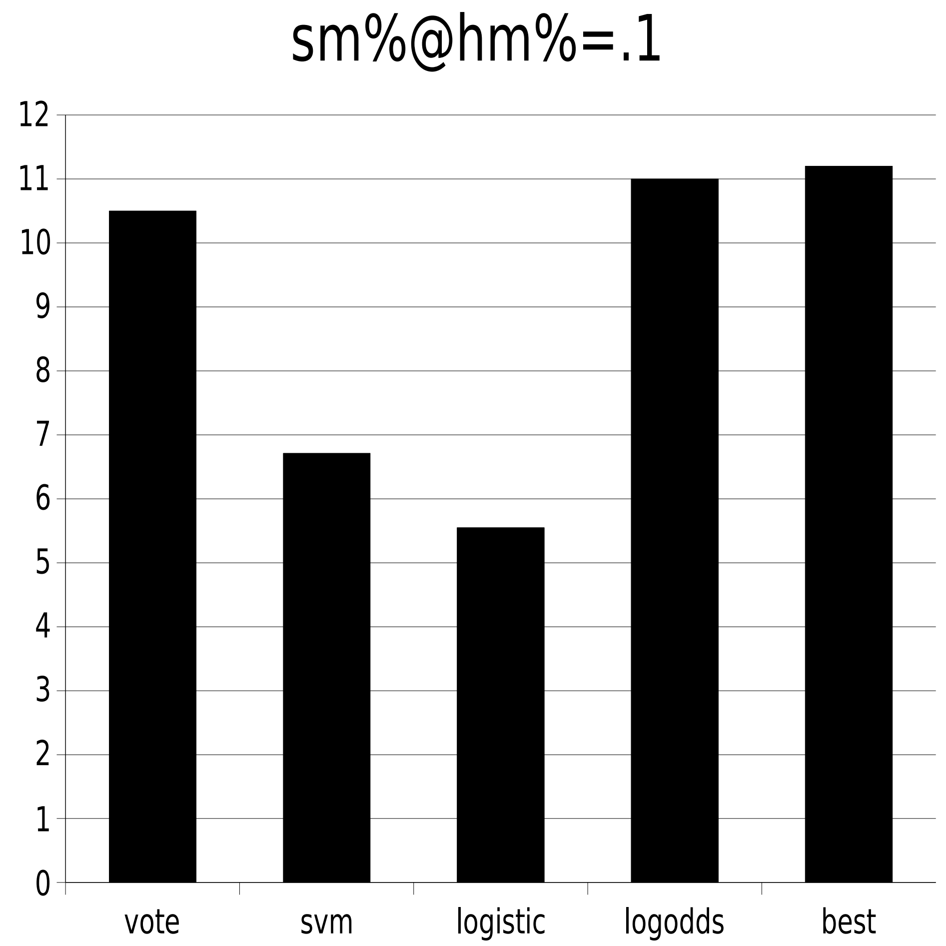
Full Results



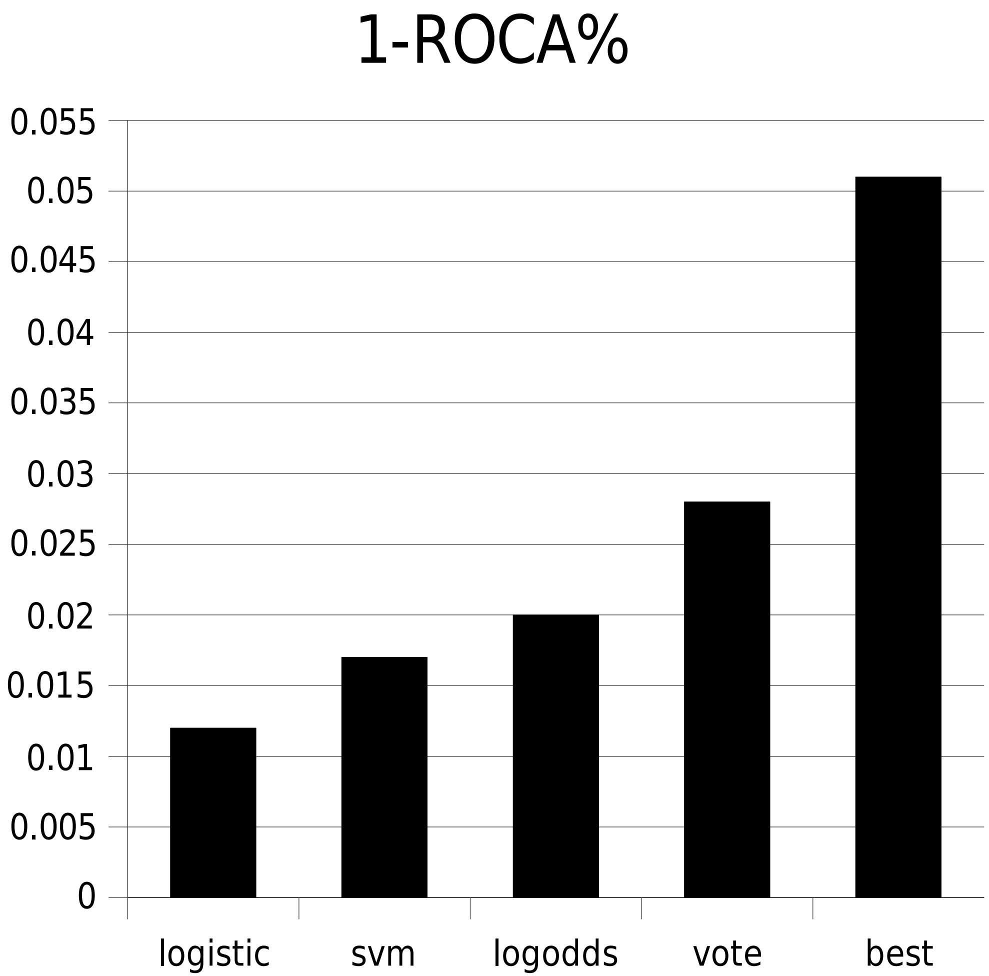


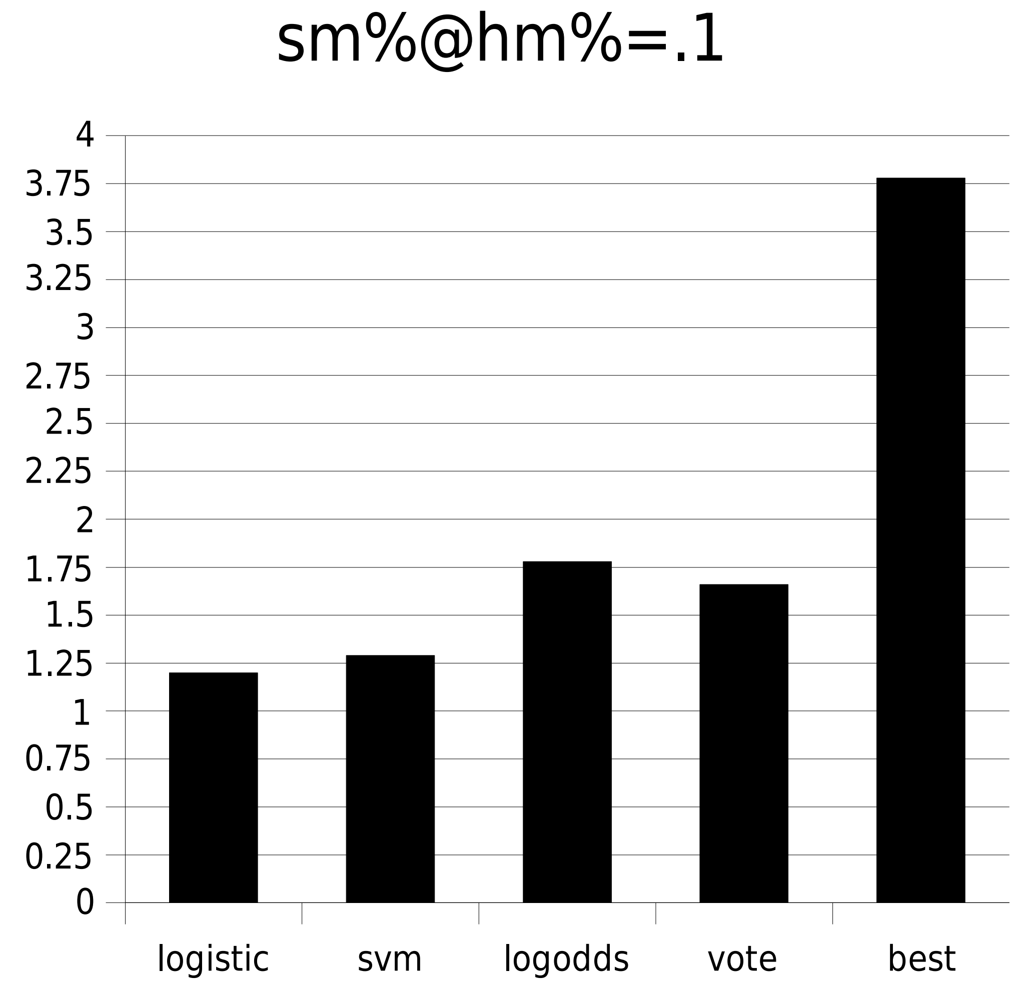
S B Results


1-ROCA%
0.25
0.225
0.2
0.175
0.15
0.125
0.1
0.075
0.05
0.025
0
vote
svm
logistic
logodds
best


sm%@hm%=.1
12
11
10
9
8
7
6
5
4
3
2
1
0
vote
svm
logistic
logodds
best

Aggregate Results



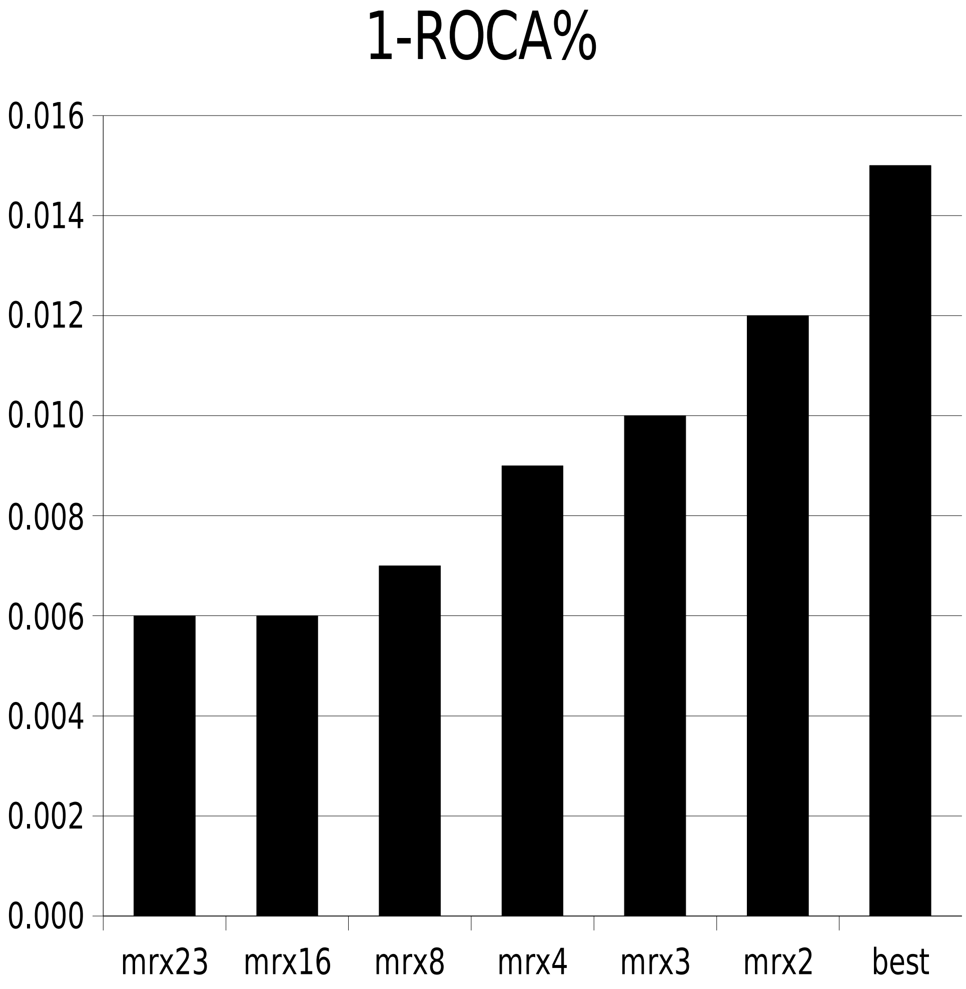


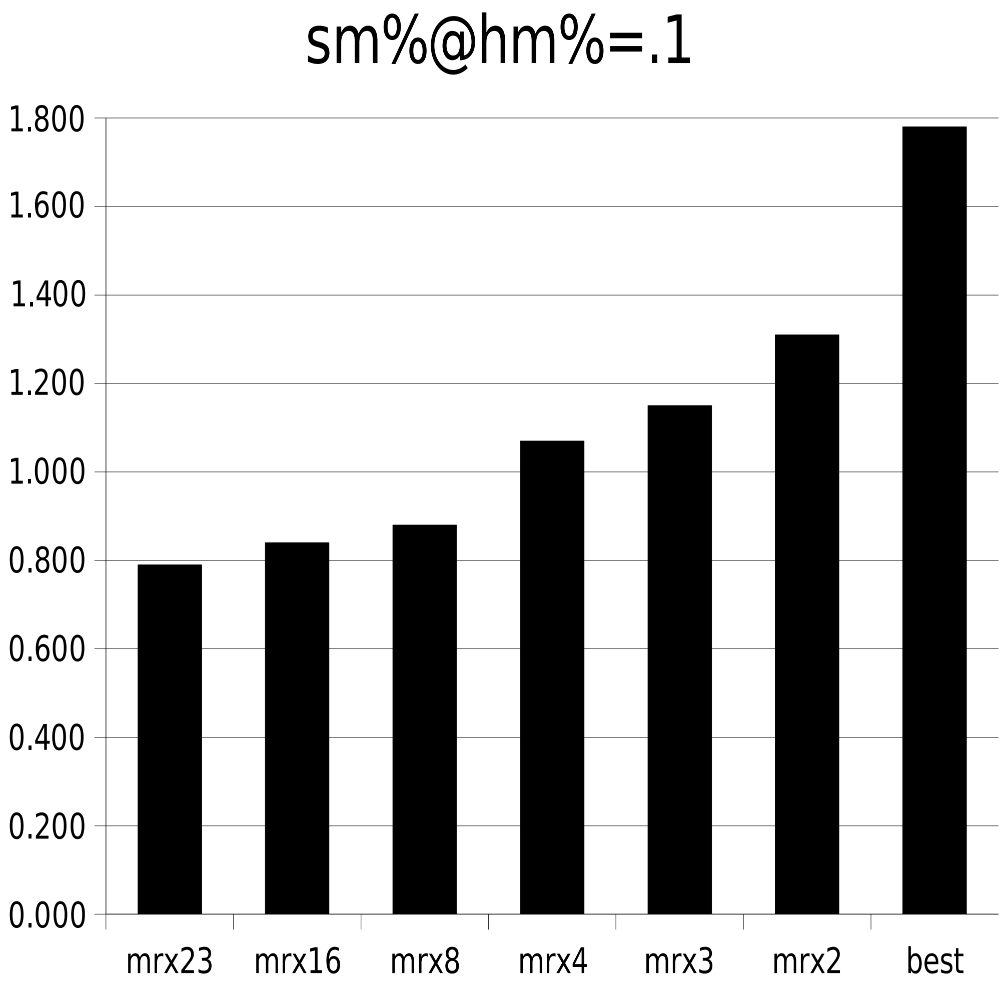
Subset Experiment

- logistic regression subset selection
  - eliminate smallest filter weight
  - recompute logistic-regression weight
  - repeat
- train on Mr X and S B corpora
- subset size of

2, 3, 4, 8, 16 ..., largest subset with only postive weights

Training on Mr X Corpus
Results on Full Corpus

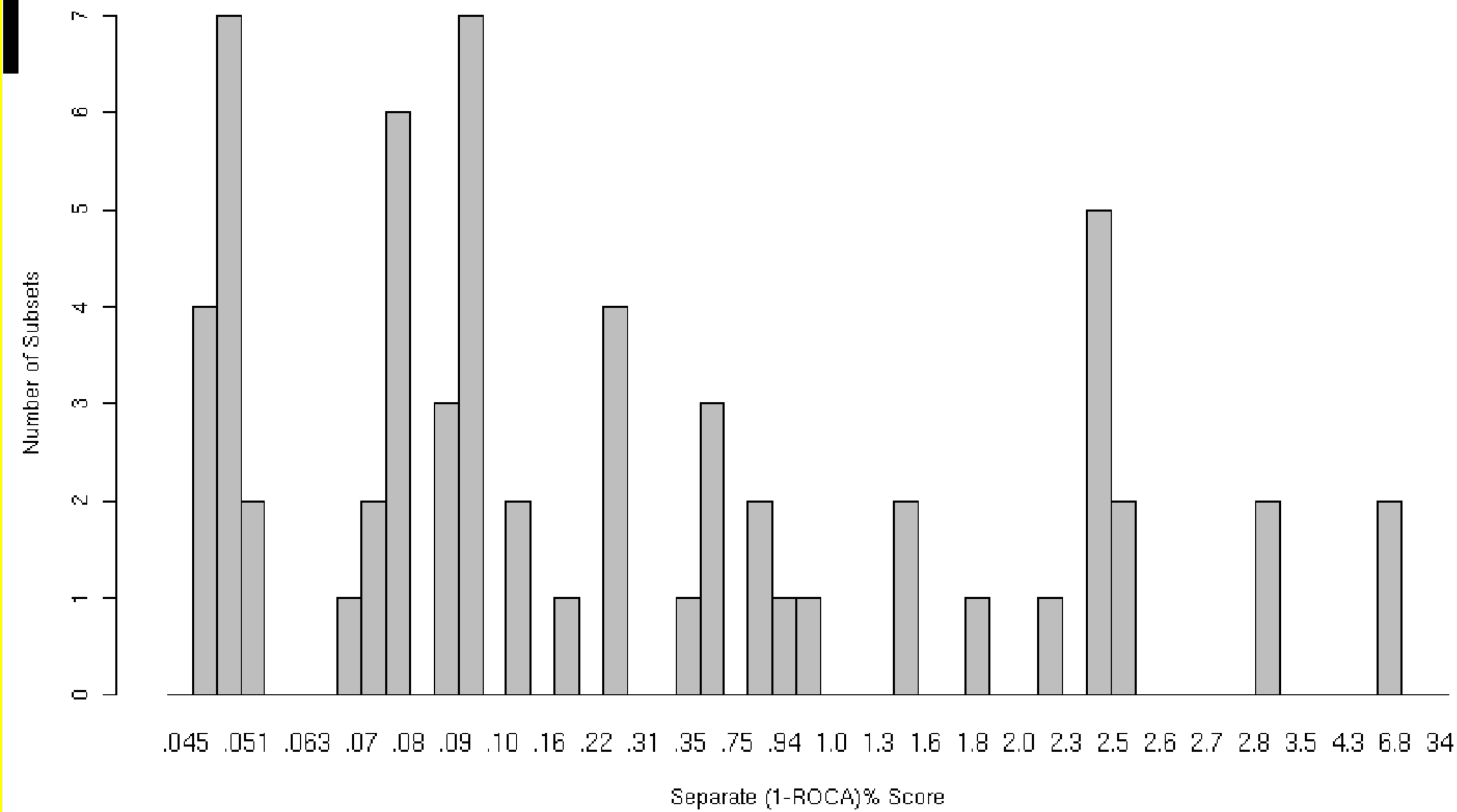




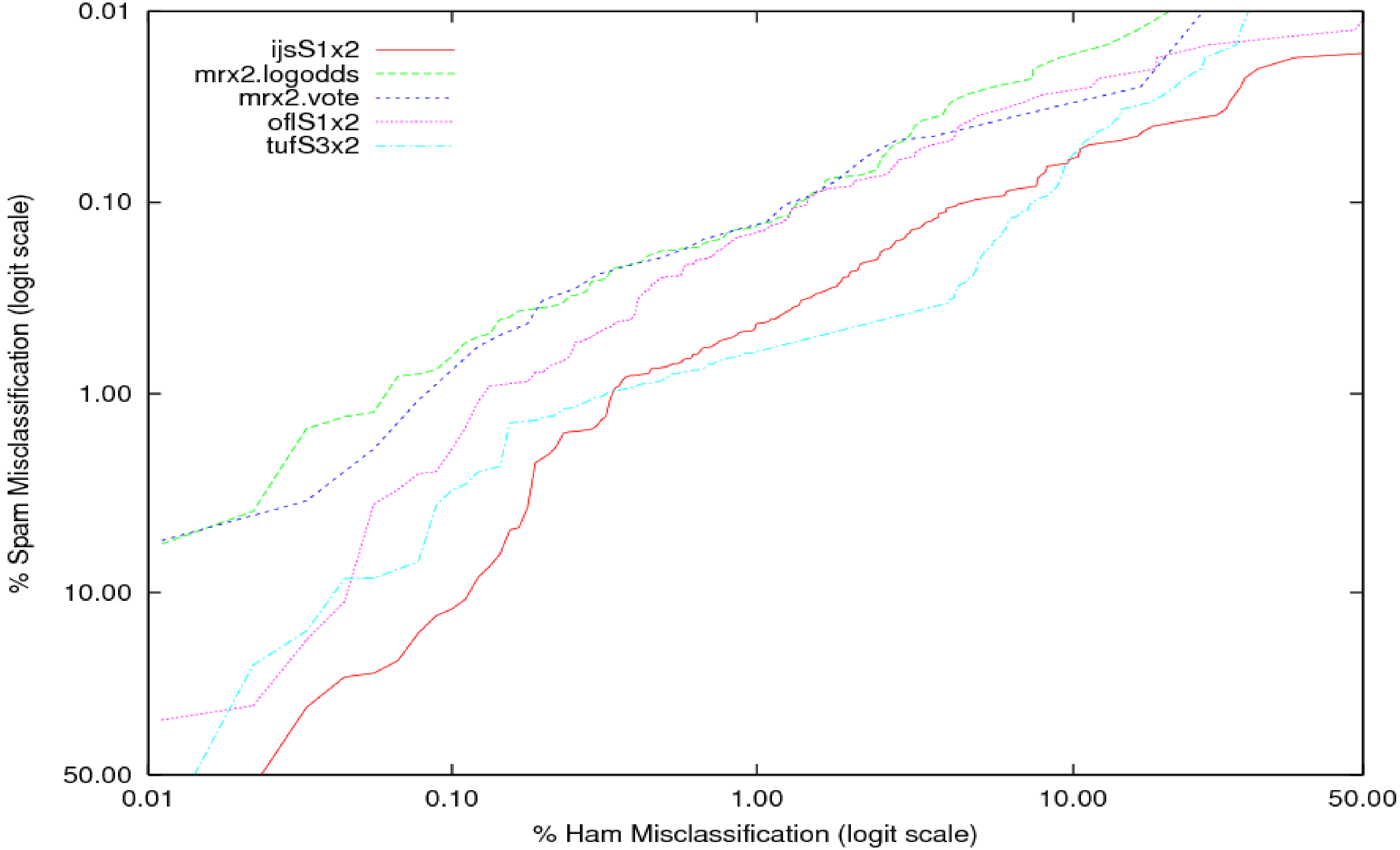
MrX-derived subsets on trec05p-1

Subset	(1-ROCA)%	sm%@hm%=.1
mrx23	.007*** (.006-.009)	.79*** (.62-.99)
mrx16	.007*** (.006-.009)	.84*** (.69-1.02)
mrx8	.009*** (.007-.011)	.88*** (.71-1.08)
mrx4	.012*** (.009-.015)	1.07*** (.82-1.39)
mrx3	.012*** (.010-.016)	1.15*** (.92-1.44)
mrx2	.016 (.012-.021)	1.31** (1.01-1.68)
best	.019 (.015-.023)	1.78 (1.42-2.22)

Base Filter Participation in Subsets
(by Separate Performance)



TREC 06 MrX II Corpus



1-ROCA(%) on Mrx II

- Logodds: 0.196 (.007 - .05)
- Vote: 0.224 (.009 - .05)
- Ofl: 0.363 (.02 - .06)
- Significance
  - Logodds – Ofl $p < .04$ (96% confidence)
  - Vote – Ofl $p < .06$ (94% confidence)

Analysis

- All fusion methods substantially outperformed the best system
- On small corpus SVM and Logistic regression are less effective
- Voting seems more stable
- log-odds essential for other methods
- negative LR weights not always overfitting

Conclusions

- Voting works surprisingly well
- Log-odds averaging works a little better
- Logistic Regression is slightly better
- SVM is the best for large corpus
- 53 filters not feasible
- predicting good small subsets possible

Future Work

- explore meta analysis
- different methods of score normalization
- apply fusion to other areas

Questions?

University of Waterloo

Subset	(1-ROCA)%		sm%@hm%=.1	
mrx23	.007***	.006-.009	.79***	.62-.99
mrx16	.007***	.006-.009	.84***	.69-1.02
mrx8	.009***	.007-.011	.88***	.71-1.08
mrx4	.012***	.009-.015	1.07***	.82-1.39
mrx3	.012***	.010-.016	1.15***	.92-1.44
mrx2	0.02	.012-.021	1.31**	1.01-1.68
best	0.02	.015-.023	1.78	1.42-2.22

SpamAssassin Corpus ROC curves

Mr X Corpus ROC Curves