Exploiting Spammer's Tactics of Obfuscations for Better Corporate Level Spam Filtering

Vipul Sharma, Steve Lewis
Proofpoint, Inc

Agenda

- Machine Learning 101
- Problem
  - Definition
  - Strategies
- Solution
  - Existing
  - Proposal
- Validation
  - Why our system works better
  - Overall improvement in blocking spam
- Conclusions

How does MLX work?

- Machine learning is the study of making computers learn; the goal is to make computers improve their performance through experience.

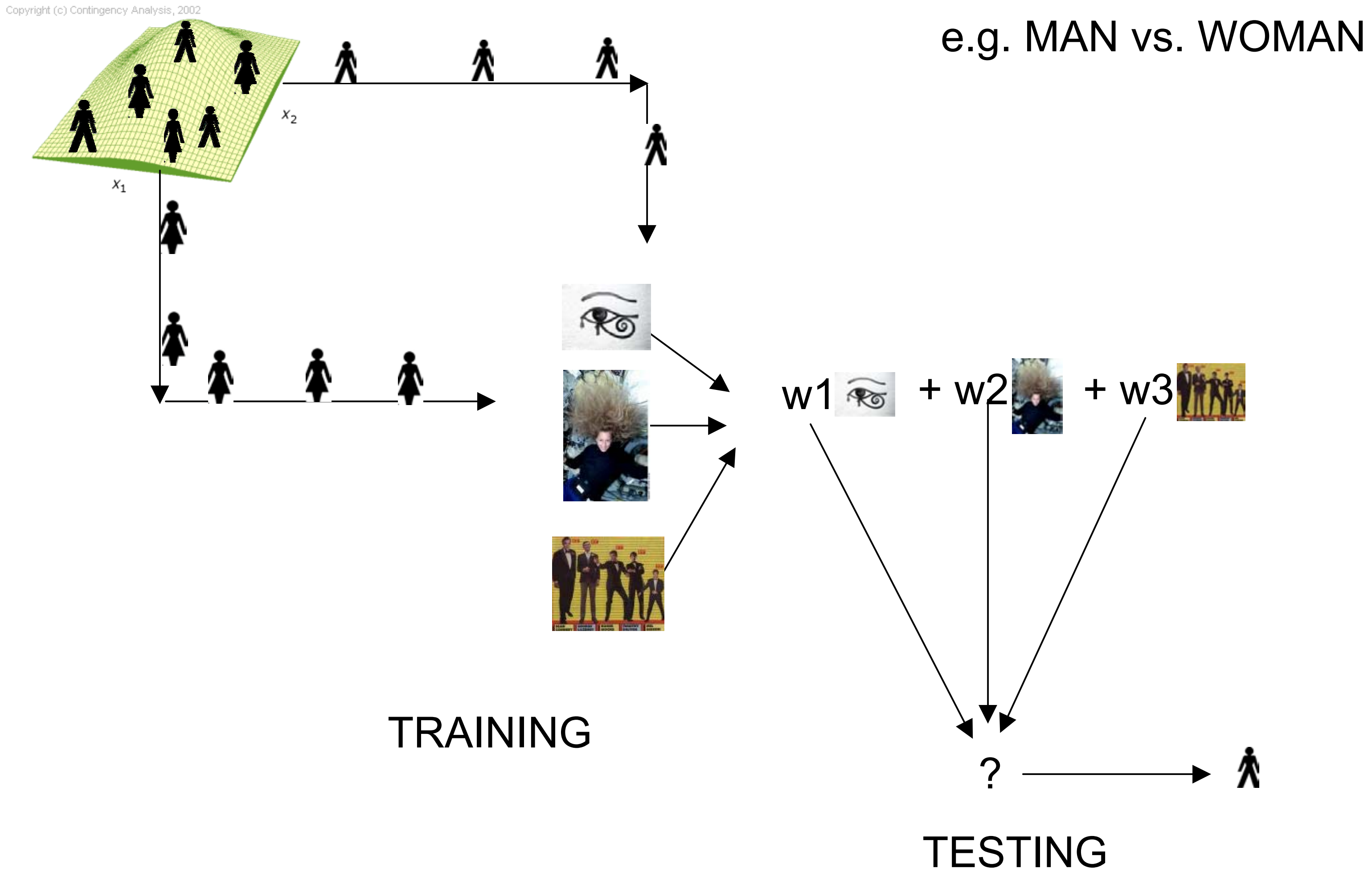
Environment
(Problem Space)
(Uniform Data)
(Feature Set)
(Class of task)

Learning Algorithm
(Logistic Regression)

Performance **P**
(Effectiveness)

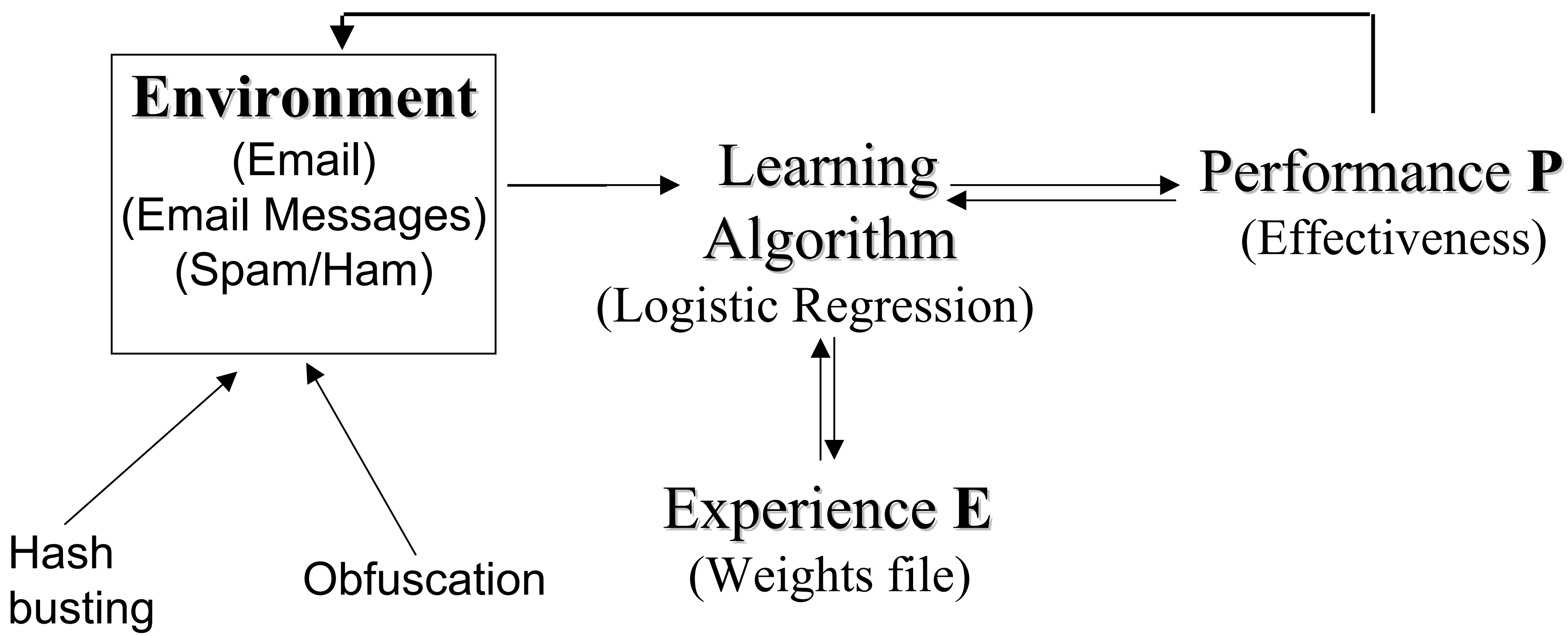
Experience **E**
(Weights file)

Training/Testing

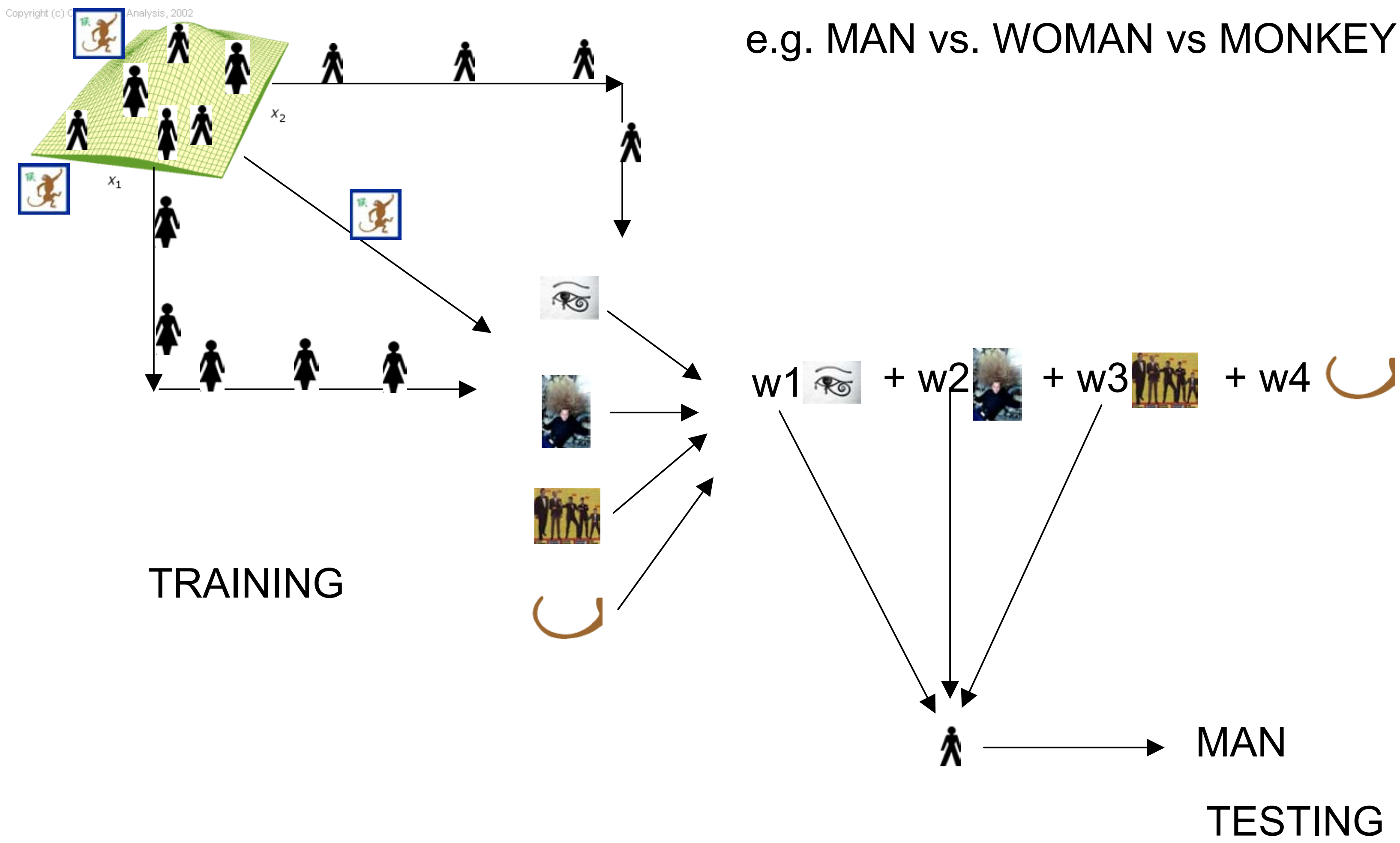


Spam → Adversarial

- Spam is a special problem of ML



Training/Testing



Spam Features

Header Information:	**Date, Subject, MTA, MUA, content types, etc.**
Body:	**Words, phrases, URL etc.**
Meta:	**Boolean combinations of other features**
Surrounding Technologies:	**Image features, IP, obfuscations, etc**

Deceiving Content based Spam Filters using Text Obfuscation

- Come play your favorite ca$in0 games online right now.
- Come play your favorite ca#ino games online right now.
- Come play your favorite caniso games online right now.
- Come play your favorite c$in0 games online right now.
- Come play your favorite cassiino games online right now.
- Come play your favorite ca $ ino games online right now.

proofpoint

Types of Text Obfuscation

- Come play your favorite ca$in0 games online right now.

  Substitution

- Come play your favorite caniso games online right now.

  Shuffling

- Come play your favorite casno games online right now.

  Deletion

- Come play your favorite cassiino games online right now.

  Addition

- Come play your favorite ca s ino games online right now.

  Segmentation

- Come play your favorite ($iino0 games online right now.

  Combination

How to Counter V!@gr@@?

- Come play your favorite c$in0 games online right now

Deobfuscation →

- Come play your favorite casino games online right now

- Come play your favorite c$in0 games online right now

Detection →

- Come play your favorite c$in0 games online right now

proofpoint

Advantages / Disadvantages

- Deobfuscation (Lee et al. CEAS 2005)
  - HMM
    - Accurate (97%),
    - Very Slow (240 letters/sec) on English letters **(Bad for corporate level spam filters)**
- Identification
  - Regular Expressions
    - Inaccurate
    - Expensive to maintain
  - Edit distance (Oliver et al. Spam Conference 2005)
    - Less Accurate (75%) **(Bad for corporate level spam filters)**
    - Cheap / Faster

proofpoint

What is a Good Solution?

Accurate (~95%)

&

Fast (near real time)

&

Computationally Inexpensive (minimal overhead)

&

Easy to Maintain

proofpoint

Obfuscation Detection Model

- A machine learning based detection system
- Benchmarked several supervised multivariate classification techniques
- Uses a domain knowledge of ~800 hand collected frequently obfuscated words (FOW)
- Auxiliary classifier that can be easily integrated with base classifier
- Fast, accurate and easy to maintain

Frequently Obfuscated Words (FOW)

- Come play your favorite ca$$iino games online right now

- we offer real, genuine degrees, that include bachelo-rs's, ma|ster's, mba, and do,ctorate degrees. they are fully verifiable

WHAT WOULD SPAMMERS WANT TO HIDE?

- Buy cheapest Vi@gr@, Ci@#lis, mbian on!!ine

- Re|^ian@nce your m0rtg@g3 today. Click here

proofpoint

Environment

- Problem Space → To detect variations of FOW
- Dataset
  - 67,907 hand collected obfuscated words
  - 250,000 valid words, parsed from ham messages
  - 12,000 commonly used valid word as dictionary
  - 727 frequently obfuscated words (FOWs)
- Class of Task → Obfuscated | Valid

Feature Set

- Feature Set
  - A: Count of non-alphanumeric characters (~!@#$ ..)
  - B: Count of Numeric letters (not on boundary)(01234 ..)
    - DEC009988
    - M0r1gage
  - C: Length of the word
  - D: Dictionary presence of the word {0,1}
  - E: Similarity between FOW and the word (0-1)

Similarity Metric

- A common technique of obfuscation is *shuffling* for e.g. *mtograge*
- Levenshtein Distance, Jaro Winkler metric match
  - *L(mortgage, mortal) = 4*
  - *L(mortgage, mrogtgae)=6*
- Other metric are sensitive towards ordered variations
- We need a metric that neglects order of letters

proofpoint

Similarity Metric

- L is the list of FOW; l_i ? L *(Viagra, Mortgage ..)*
- b_i = length(l_i)
- m be any test word Vi!@gra
  - Filtered word m' = Vigra
- b_m = length(m')
- b_{im} = common letters (b_i, $b_{m'}$)
- $S_{im} = max_i(b_{im}/(b_i + b_{m'} - b_{im}))$

proofpoint

Similarity Metric Example

- $L = \{ \text{Viagra, mortgage} \}$; $m = \textit{mrogtgae}$
- $b_{viagra} = 6$; $b_{mortgage} = 8$; $b_{mrogtgae} = 6$
- $b_{viagra,\ mrogtgae} = 3$; $b_{mortgage,\ mrogtgae} = 8$
- $S_{viagra,\ mrogtgae} = 3/(6+8-3) = 0.27$
- $S_{mortgage,\ mrogtgae} = 8/(8+8-8) = 1$
- $S = \max(1, 0.27) = 1$
- $S_{\text{mortgage, mortal}} = 5/8+6-5 = 0.55$

Preprocessing

- Discretization (Fayyad & Irani's MDL method)
  - Converted numeric features to nominal features
  - Increase classification accuracy for certain classifiers
  - Certain classifier works only for nominal features
- Cutoff bins are calculated such that the entropy of the model is minimized

proofpoint

Entropy/Information Gain

Entropy → measure of randomness ~ prediction

Low Entropy

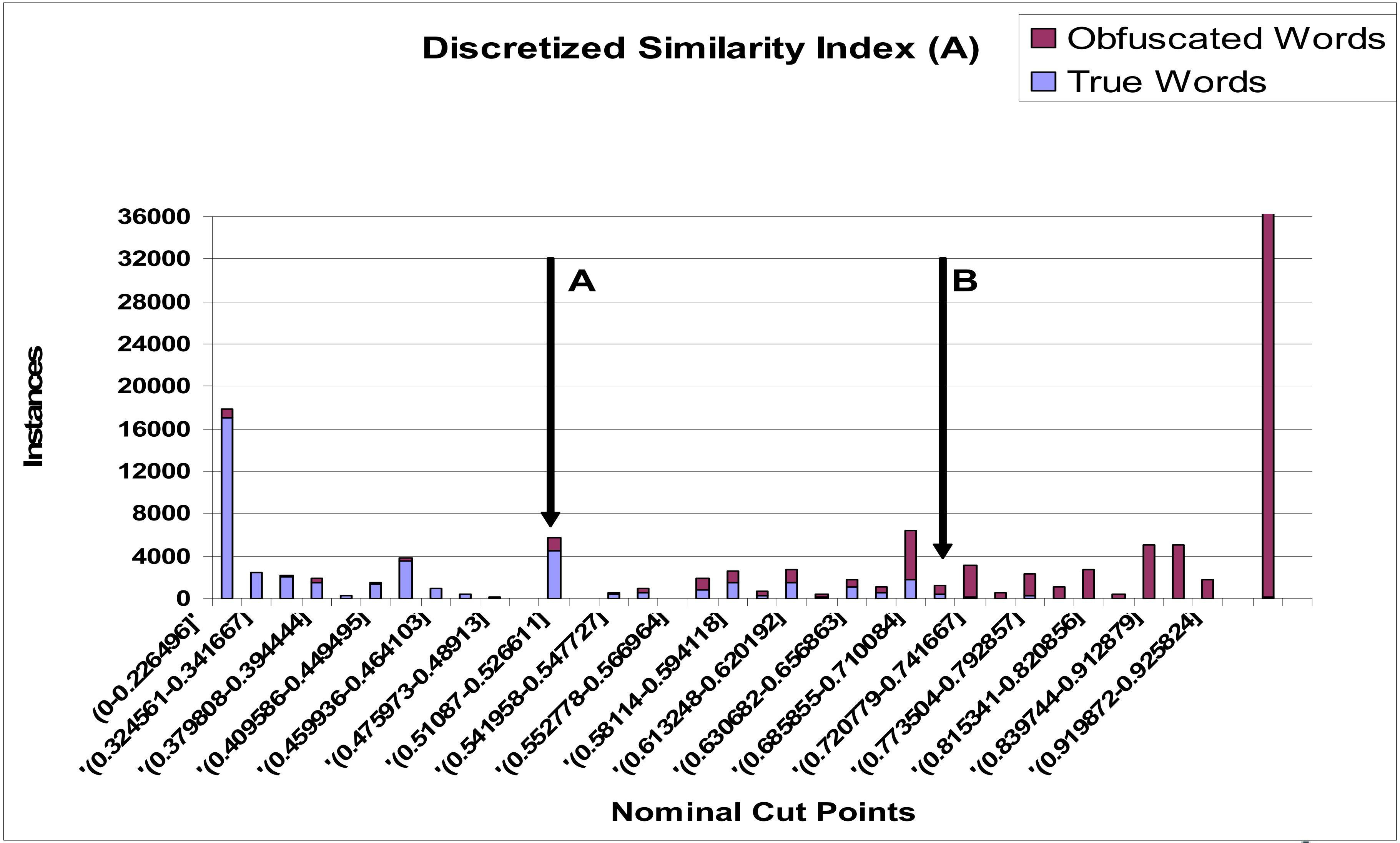
More Predictive

High Entropy

Less Predictive

Images taken from eb.cecs.pdx.edu/~york/cs510w04/infogain09.pdf

Discretization



Feature Generation using Constructor Functions

- General Constructor Functions generated 13 features
  - Set of operators used +, > , <, =, !=, &, |
  - Cut-points generated via discretization used as ranges
  - Use beam search
  - Heuristics used → maximize Information Gain

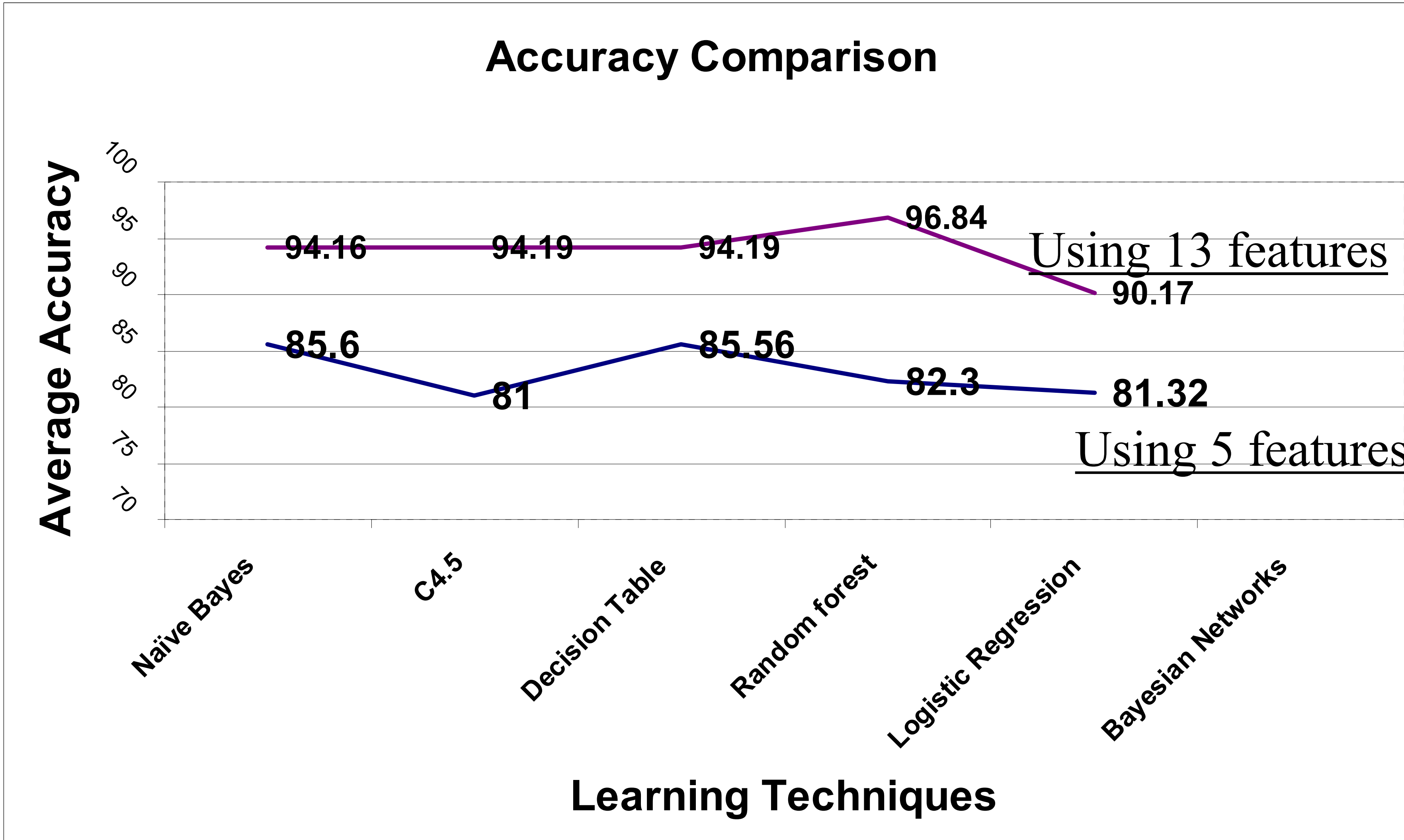
- E > 0.710084 and A > 1 (V!a-gra)
  - If similarity index > 0.710084 and number of non alpha numeric character > 1 ➔ strong representation of obfuscated class

proofpoint

Learning Model

- Various multivariate classification techniques were compared using Weka
- 10-fold cross validation was used for accuracy estimation
- Accuracy was compared on both feature set
  - 5 basic feature before preprocessing
  - 13 generated features after preprocessing

Obfuscation Detection Accuracy Comparison



proofpoint

Detection

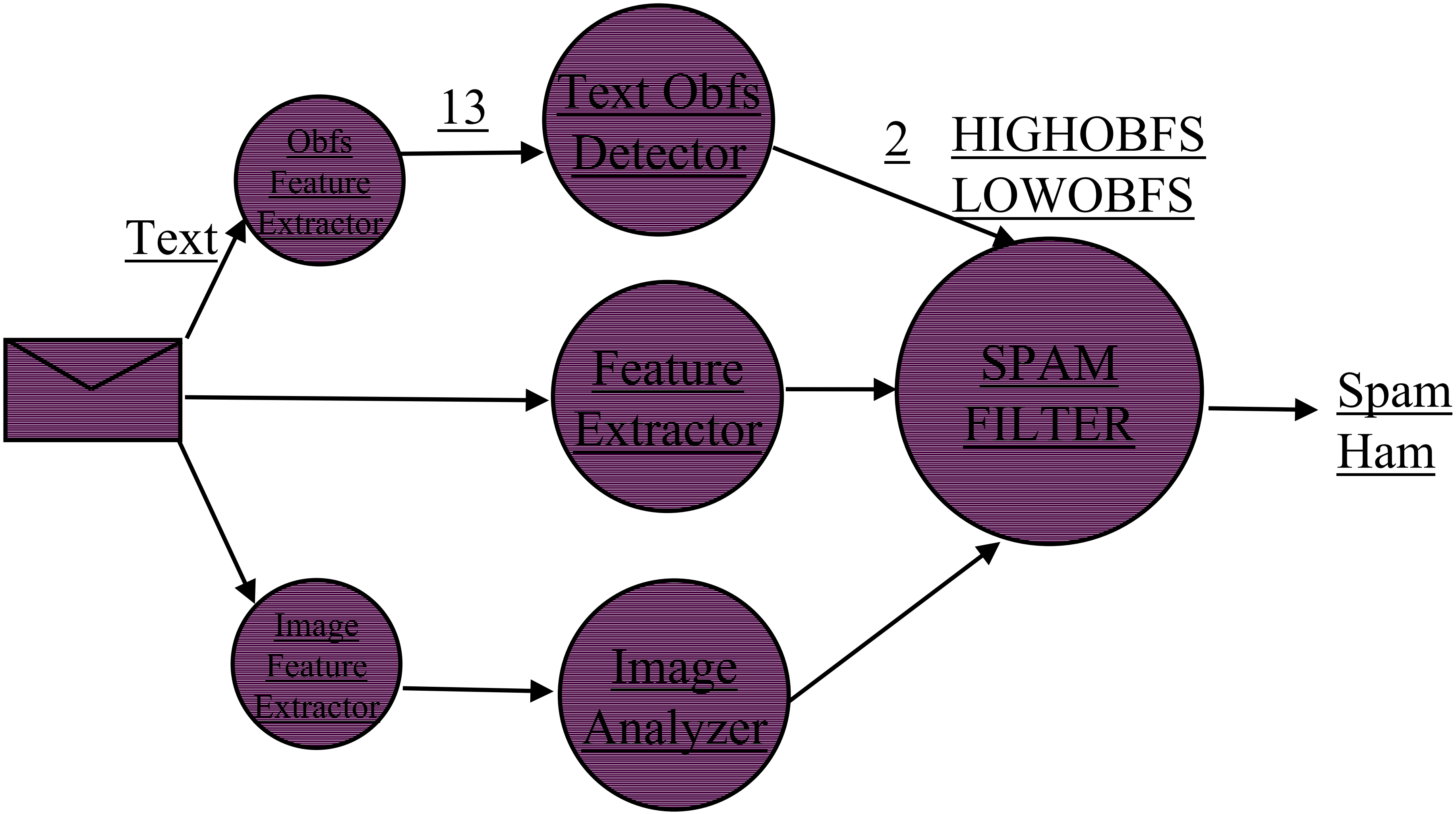
- Training produces weights for all the 13 feature
- Any given word will be converted in the form of a feature vector
- A score for each word is calculated using the weights and the logistic function
- If Score > 0.5 ➔ Obfuscated
- If Score < 0.5 ➔ True

proofpoint

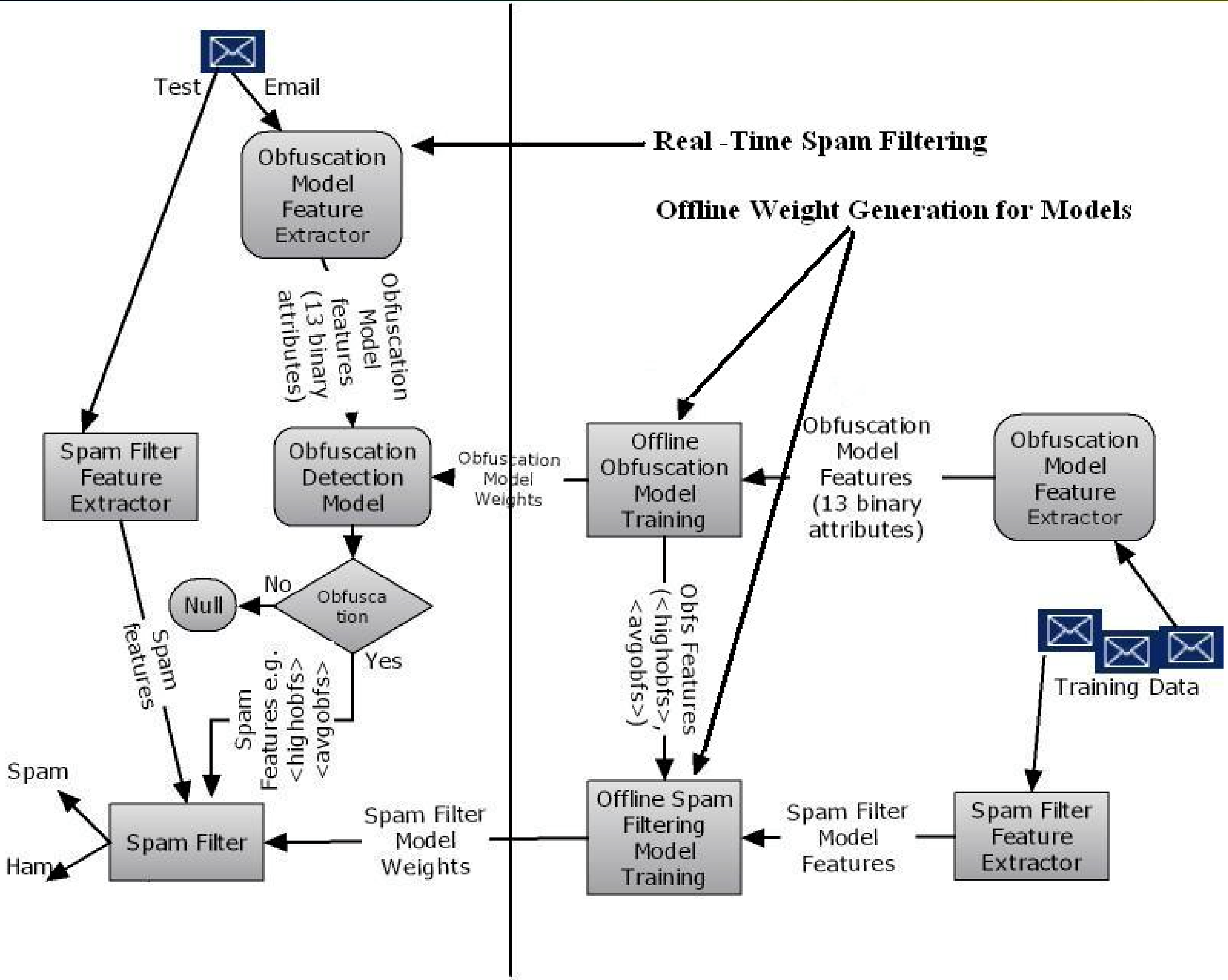
Integration with Base Classifier

- Used to generate features for base classifier
- Weight of the spam filter feature a confidence of obfuscation (Score of the term)
- Logistic regression scores each term between 0-1
- Score(term) > 0.5 → obfuscated
- Score(term) > 0.9 → HIGHOBFS
- 0.7 < Score(term) < 0.9 → LOWOBFS
- The weight of HIGHOBFS, LOWOBFS is determined during base classifier training

Integration



Sample Integration

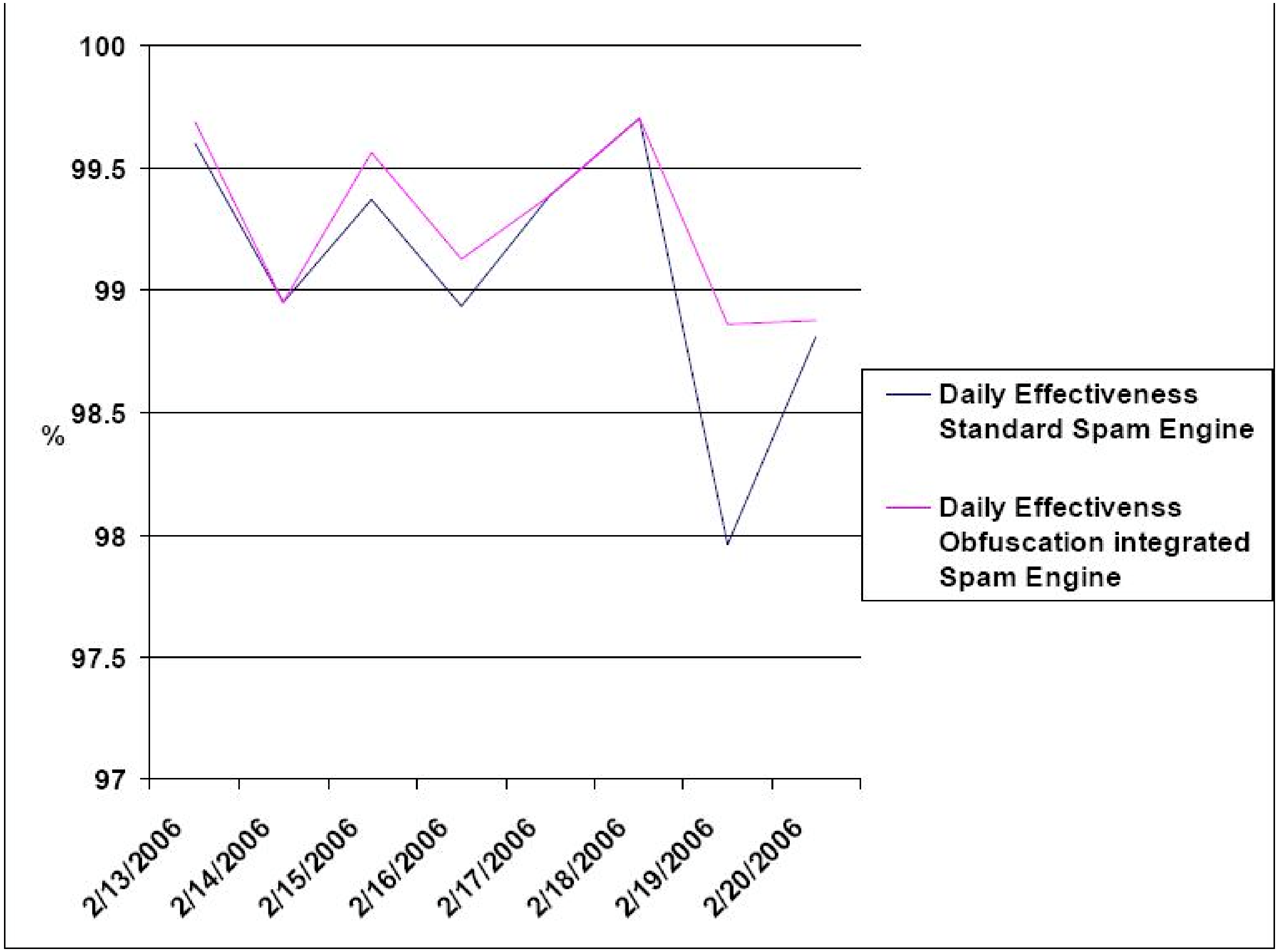


Overall Spam Detection Accuracy

- Test Corpora
  - 400,000 spam messages
  - 112,000 ham messages
- Accuracy improvements
  - FNs decreased by 50%
  - A negligible increase in FP ~ 0%
  - Overall accuracy ~ average increase 0.3%

Overall Spam Detection Accuracy

- Tested on one of the Proofpoint's honeypot



proofpoint

Conclusions

- Obfuscation can be detected with high accuracy
  - Concentrate on FOW
  - Use preprocessing techniques for feature generation
- A very low overhead to spam engine
- Logistic regression achieved highest detection accuracy with lowest false positives
- Similarity Metric should not be weighted around ordered similarity
- We noticed a significant improvement in spam detection accuracy with almost no false positives

Discussions

- Biased towards the FOW list
- Works for all languages
- FOW list do not contain words with length equal or less than 4
- FP rate can be decreased by adding the errors in dictionary
- A interesting method of using supervised classification technique for feature generation

proofpoint

Thank You for Your Time!
Q&A

For information about Proofpoint,

contact us at:

info@proofpoint.com

408-517-4710

www.proofpoint.com

For a FREE 45-day trial of Proofpoint, visit:

www.proofpoint.com/vb2006

proofpoint