Spam Detection Through Sliding Windowing of E-mail Headers

Francisco Salcedo-Campos, Jesus Diaz-Verdejo , Pedro Garcia-Teodoro

Dpt. of Signal Theory, Telematics and Communications
ETSIIT - CITIC - University of Granada
C/ Periodista Daniel Saucedo Aranda, s/n - 18071 – Granada (Spain)
Phone: +34-958 242304 - Fax: +34-958 240831
Email: fjsalc@ugr.es , jedv@ugr.es , pgteodor@ugr.es

© FJSC and JEDV 2011

Index

Introduction

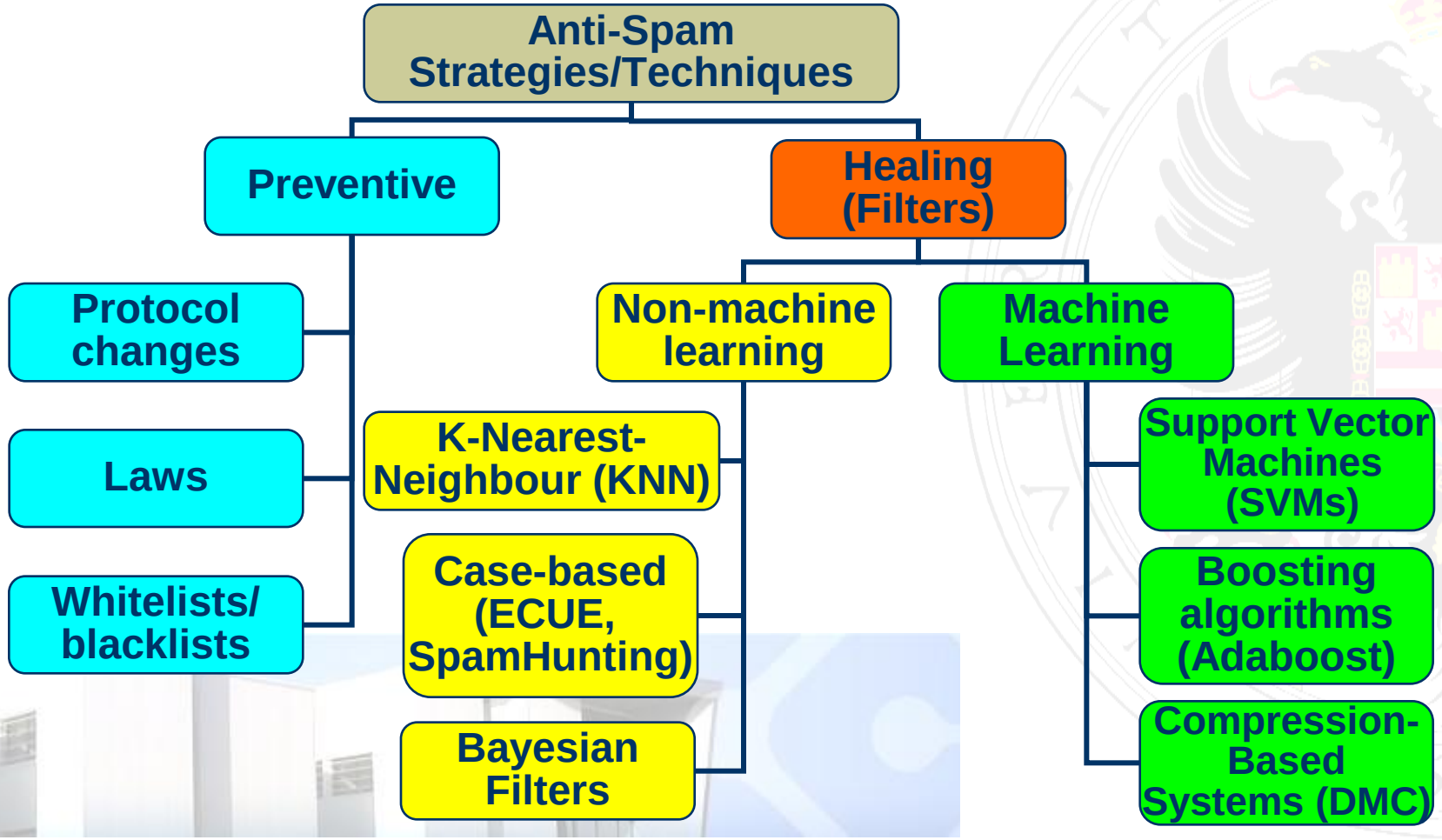
- Spam is a big problem:
  - Time expended by users deleting spam
  - Computer memory wasted
  - Internet Service Providers bandwidth wasted

- Illustrative figures:
  - In 2008 more than 90% of e-mail were abusive[1]
  - Companies lost more than $70 billion just in workers' productivity only in USA during 2007[2]

[1] Source: Messaging Anti-Abuse Working Group (MAAG)
[2] Source: Nucleus Research

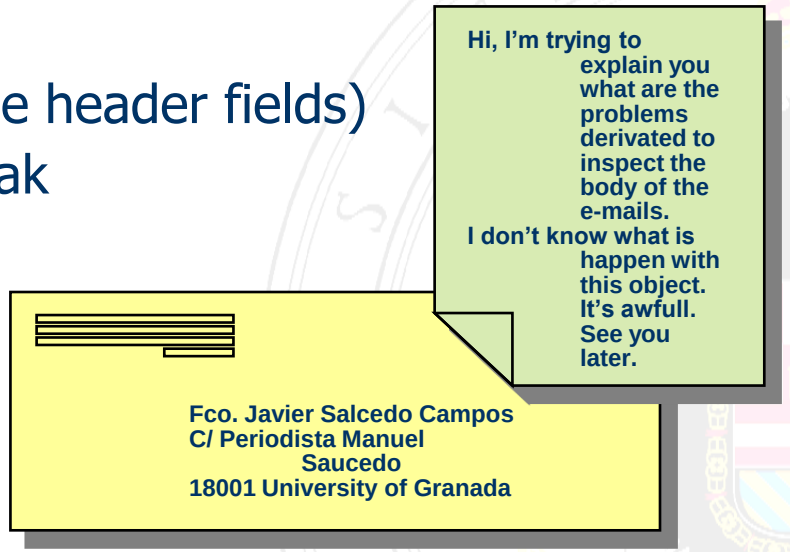
Previous Work on Spam Avoidance

- A wide variety of solutions to fight against spam



Previous Work on Spam Avoidance

- Most of anti-spam methods:
  - Tokenization-based
  - Use the body of e-mails (some header fields)
  - Language-dependent and weak
    - M0n€y instead money
  - Violates users' privacy (illegal in many countries)



- The technique proposed:
  - Based on a new perspective: e-mail headers are interpreted as a digital signal
  - No inspection of the body is needed (avoids privacy and legal concerns)
  - Independent of the language or body encryption used in e-mails.

Static Parameterization of Headers

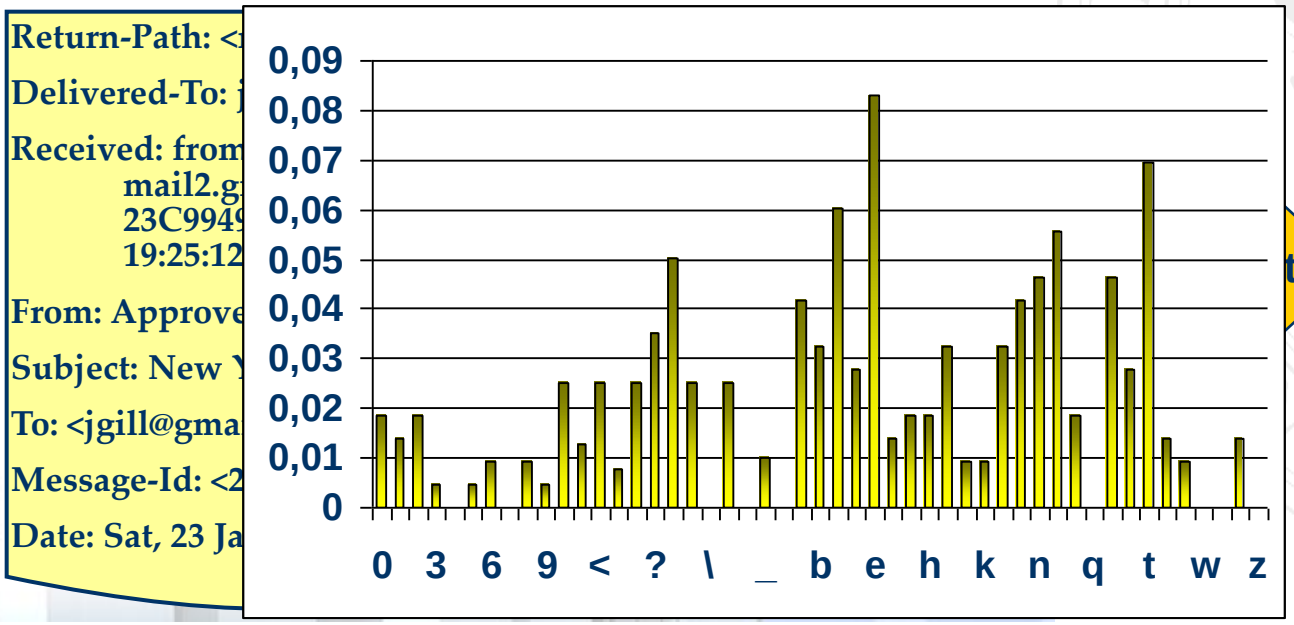
Headers can be described as a sequence of elements of a vocabulary

$$H = h_1 h_2 ... h_T \quad h_i \in V$$

$$V = \{h_i \mid 1 \le i \le K\}$$

$$f_i = \frac{count(v_i, H)}{\sum_{j=1}^{K} count(v_j, H)}$$

$$\vec{p} = \chi(H) = < f_1, f_2, ..., f_K >$$



Segmental Approach for Spam Detection

- Sliding Windowing of E-mail headers
  - Header are considered as digital signals codified in ASCII

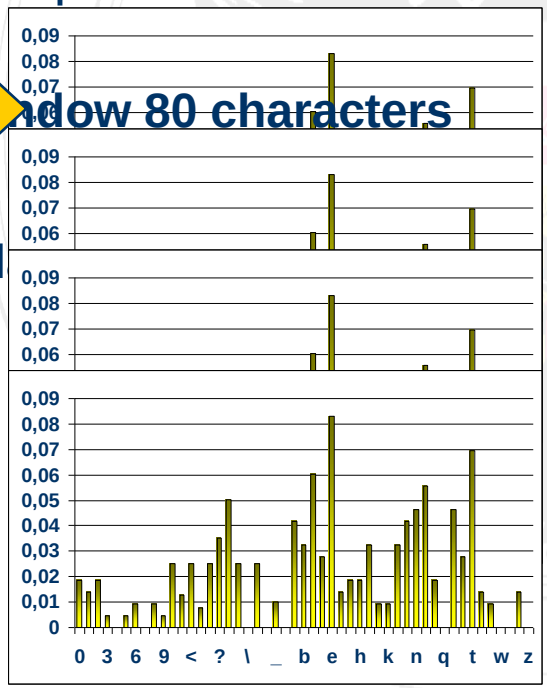
$$H = h_1 h_2 ... h_T \quad h_i \in V$$

  - Segmentation: windows of W characters displaced Δ

Return-Path: morukin@andin.org\nDelivered
-To: jgill@gmail.com\nReceived: from amyski
tchen.net (unknown [87.241.68.49]) by ma .g
mail.com (Postfix) with SMTP id 23C9949 5
for <jgill@gmail.com>; Sat, 23 Jan 2010 19:25.1
2 +0100 (MET)\nFrom: Approved Store <moru
kin@andin.org> \nSubject: New Year Sales\n
To: <jgill@gmail.com>\nMessage-Id: 20100123
182514.23C9949815@mail2.gmail.com\nDate:S
at, 23 Jan 2010 19:25:12 +0100\n

count ndow 80 characters

displ



Segmental Approach for Spam Detection

Sliding Windowing of E-mail headers

- Improving the dynamics of the system: adding 1st and 2nd derivatives of f^k in each window i

First order:

$$d_i^k = \frac{(f_i^k - f_{i-1}^k)\cdot W}{\Delta}$$

Second order:

$$a_i^k = \frac{(d_i^k - d_{i-1}^k)\cdot W}{\Delta}$$

Thus a header Q is:

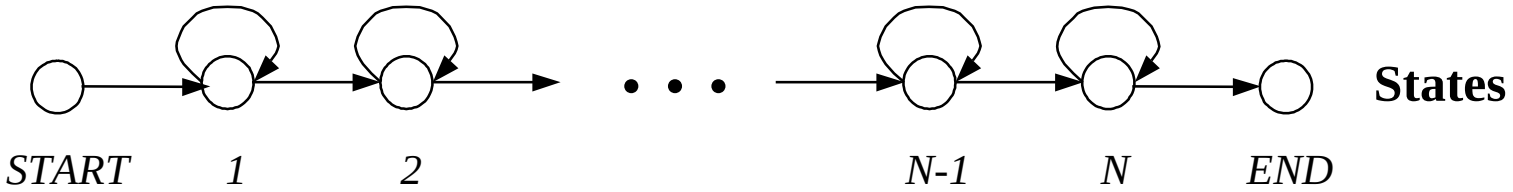
$$Q = \vec{q}_1\vec{q}_2...\vec{q}_R, \quad \vec{q}_i =< f_i^1, f_i^2, ... f_i^K, d_i^1, d_i^2, ..., d_i^K, a_i^1, a_i^2, ..., a_i^K >$$

Segmental-based HMM-based E-mail Detection

- Hidden Markov Models (HMMs) have been successfully used to model processes of a sequential nature (speech recognition)
- Evaluation $P(\lambda_i/Q)$ probabilities where λ_i is an HMM model and Q the observed header (Bayes' rule)

$$class(Q) = \arg\max_i \{P(Q \mid \lambda_i) \cdot P(\lambda_i)\}$$

- Two initial models $\Lambda = \{\lambda_S, \lambda_L\}$
- Determine topology and windows width W and displacement Δ



Segmental-based HMM-based E-mail Detection

- First experiment: determine *W* and *Δ* using 3 states models
  - Best CA value obtained *W=25* and *Δ=10*

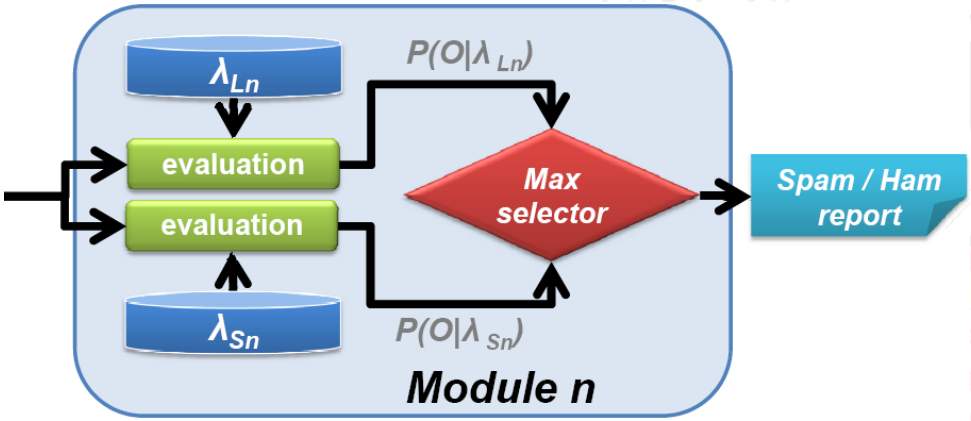
- Determining the number of states of HMMs
  - According to the topology and the length of headers (24 to 2124 segments)
  - Solution: Use more than one model for spam and ham depending on the size of the header

Segmental-based HMM-based E-mail Detection

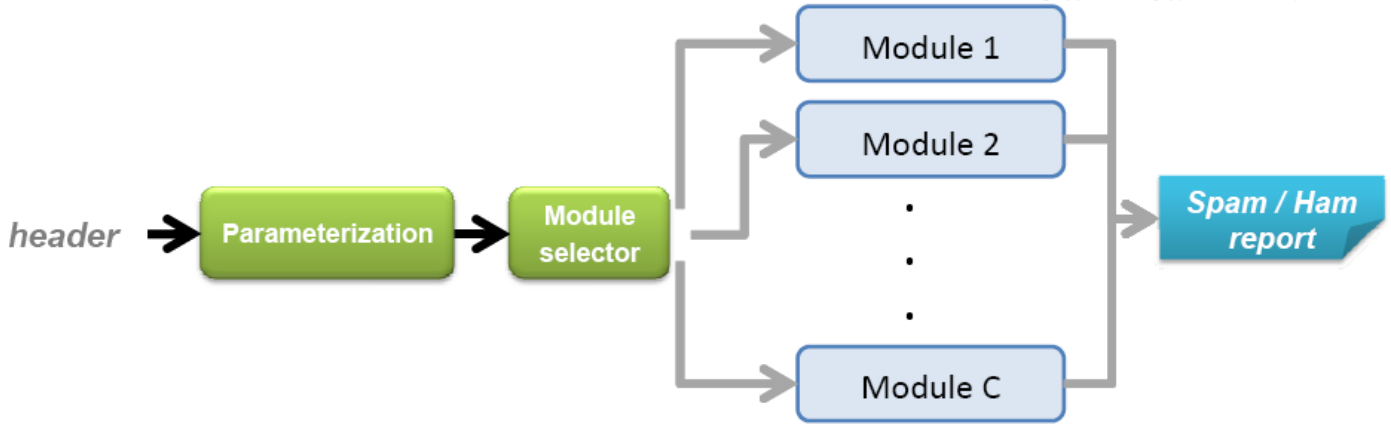
Segmental-based HMM-based E-mail Detection

- Module assignement

$$j = \arg\min_{1 \le i \le C} \left| N_i - \frac{R}{5 \cdot \Delta} \right|$$



- 19 modules SpamAssassin



Experimental Results

- Framework
  - SpamAssassin public corpus (9351 e-mails with 2.90 ham/spam ratio)
  - 10-fold stratifed cross-validation to increase the confidence in the experimental findings
  - Measures to compare results

1. True Positives (TP)

$$TP = \frac{n_{S \to S}}{N_S} \cdot 100$$

2. False Positives (FP)

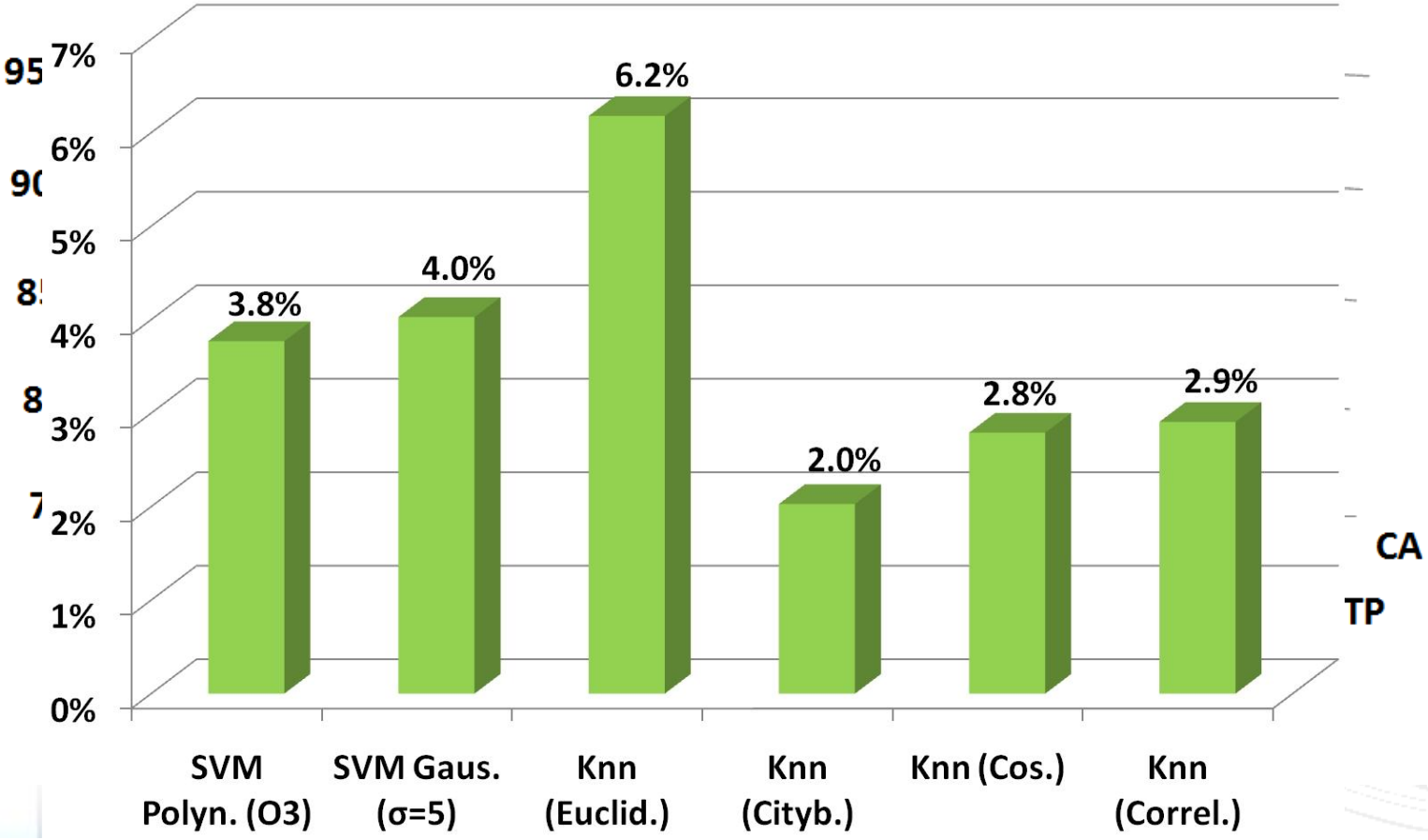
$$FP = \frac{n_{L \to S}}{N_L} \cdot 100$$

3. Classification Accuracy (CA)

$$CA = \frac{n_{S \to S} + n_{L \to L}}{N_L + N_S} \cdot 100$$

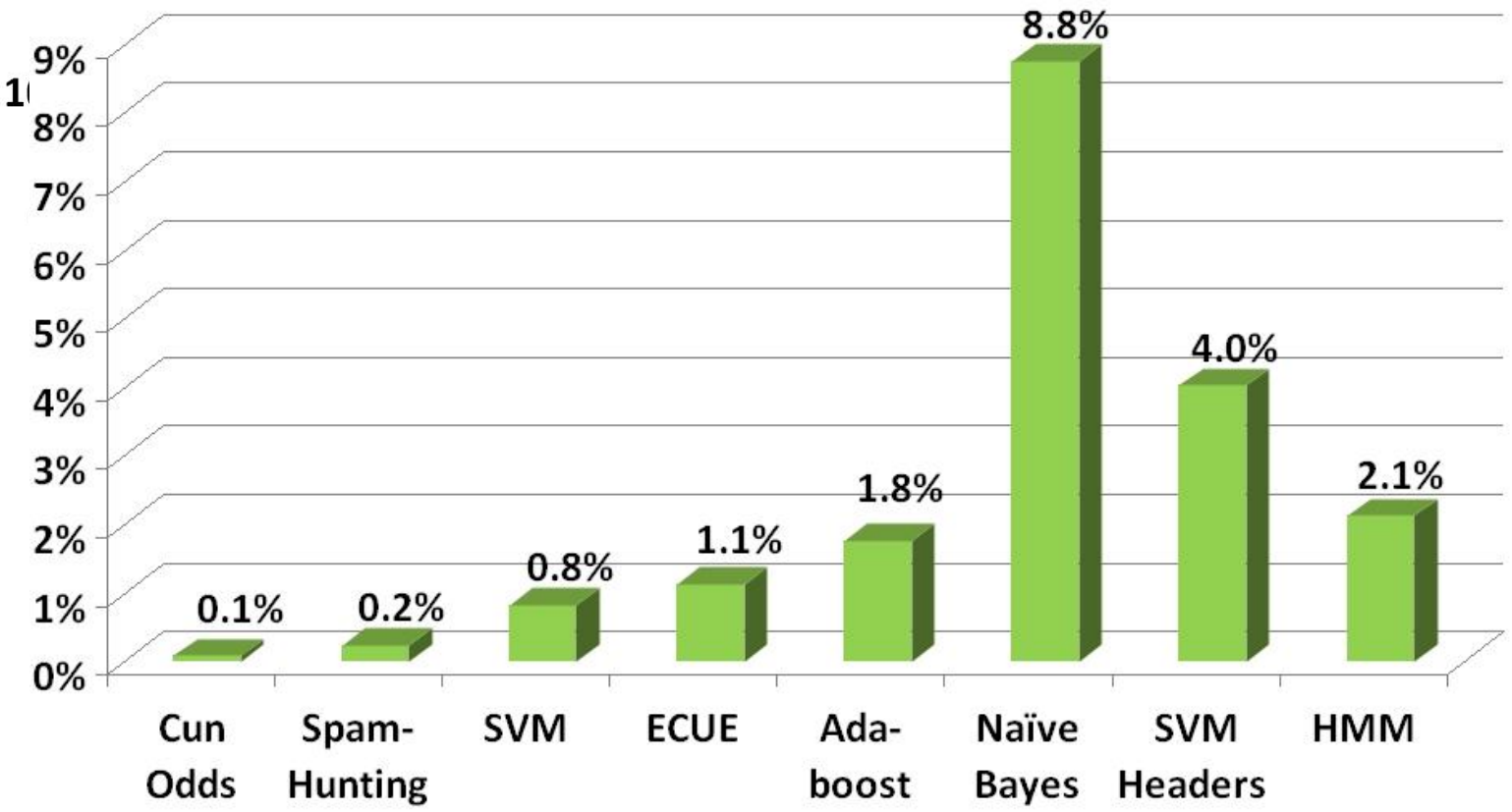
Experimental Results

Results using static parameterization



Experimental Results

Last results using CEAS08 corpus: TP=98.4% CA= 98.6% FP=0.4%



Conclusions

- A novel method for spam detection is proposed:
  a) Based only in the inspection of the headers
  b) Consider the header as a signal and parameterize and process it accordingly
  c) Uses HMMs as basic detector
  d) Preserves the privacy of users
  e) Independent of the users' language and body encryption
- Similar or better results than other methods like SVM, Naïve Bayes, etc.

Thanks for your attention

Spam Detection Through Sliding Windowing of E-mail Headers

Francisco Salcedo-Campos, Jesus Diaz-Verdejo , Pedro Garcia-Teodoro

Dpt. of Signal Theory, Telematics and Communications
ETSIIT - CITIC - University of Granada
C/ Periodista Daniel Saucedo Aranda, s/n - 18071 – Granada (Spain)
Phone: +34-958 242304 - Fax: +34-958 240831
Email: fjsalc@ugr.es , jedv@ugr.es , pgteodor@ugr.es

© FJSC and JEDV 2011