Exploiting Redundancy in Natural Language to Penetrate Bayesian Spam Filters

Christoph Karlberger,
Günther Bayler,
Christopher Kruegel,
& Engin Kirda

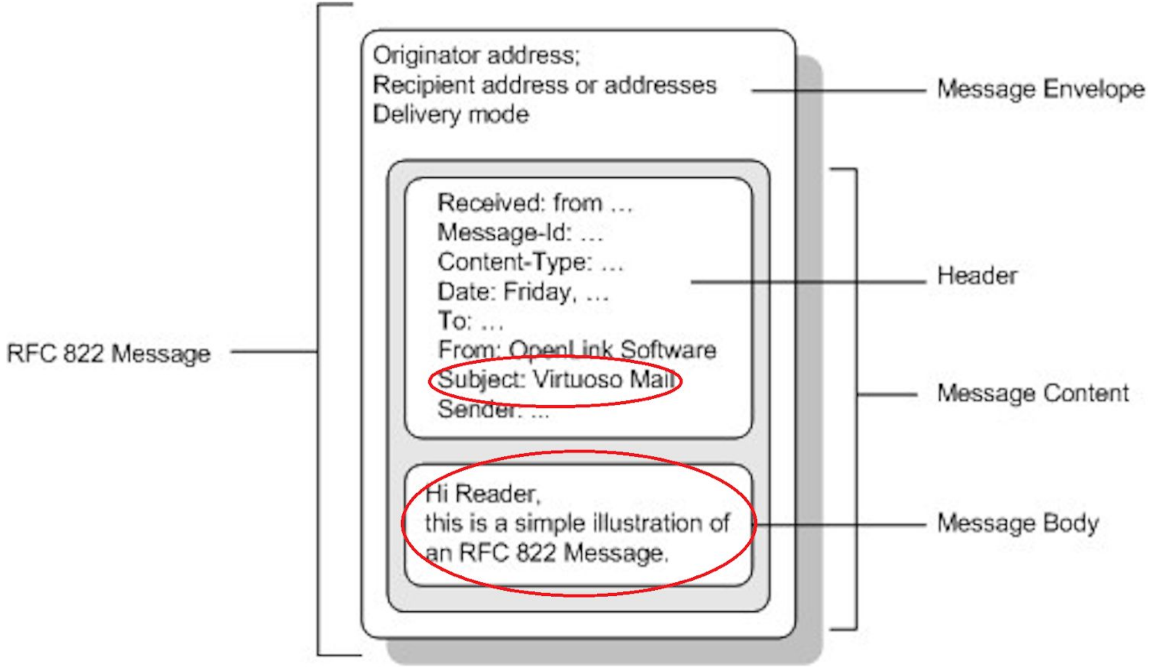
WOOT '07: Proceedings of the first USENIX workshop on Offensive Technologies

Chris Li,
Amy Min,
Claire Wang,
& Jack Steilberg

Problem statement

Summary

What is in an email?



What is a Bayesian spam filter?

How does a Bayesian spam filter work?

Calculating the probabilities for individual words

$$P_{spam}(token) = \frac{\frac{n_{spam}(token)}{n_{spam}}}{\frac{n_{spam}(token)}{n_{spam}} + \frac{n_{ham}(token)}{n_{ham}}}$$

Ham means not spam

Training a Bayesian spam filter

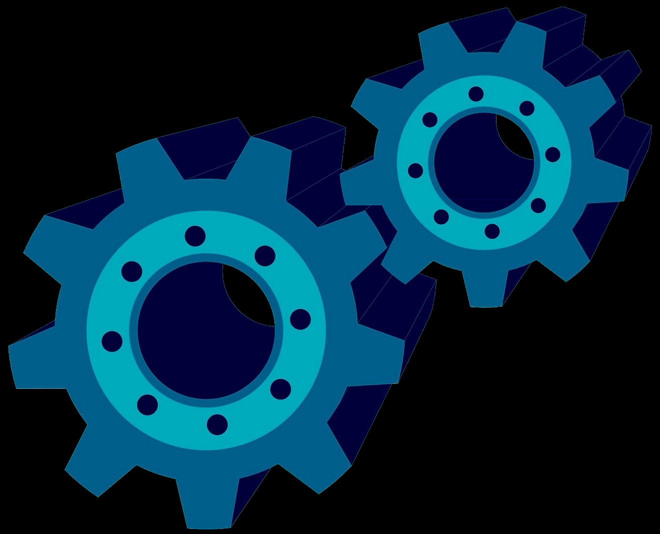
1. Tokenize emails
2. Analyze messages

Training a Bayesian filter

2. Analyze messages

Formula derived from Bayes' theorem combining individual probabilities

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1 - p_1)(1 - p_2) \cdots (1 - p_N)}$$



How it Works

Typical attacks: Appending filler words

1. Random word attack
2. Common word attack
3. Common word + uncommon in spam attack

Alternate attack: Substitution

	Synsets	Hypernym sets	If no synonym sets
Car:	“an automobile with four wheels”	“motor vehicle”	a → @
	“a motor vehicle with four wheels”	“automobile”	i → l (lower case L)
	“a cabin for transporting people”		

Automating Substitution Attacks

1. Identify all words with high spam probability
2. Find a synonym set with a lower spam probability
3. Replace words in the email with one of the synonym sets
4. Test altered email against spam filter

1. Identifying all words with high spam probability

Training spam filters with spam and ham emails:

1. Find the spam probability of every word
2. Use a *substitution threshold*

2. Finding sets of words with similar meaning

1. Find synonym sets using **WordNet**
    a. If none found, use *exchange threshold* for doing e.g. a → @
2. Give WordNet the role of the word using **LingPipe NLP** package
3. Use **SenseLearner** to choose the synset closest semantically to the original term

3. Replacing words in the email

Two methods of selecting from the set of synonym sets found:

1. Random
2. Minimum spam probability

Results

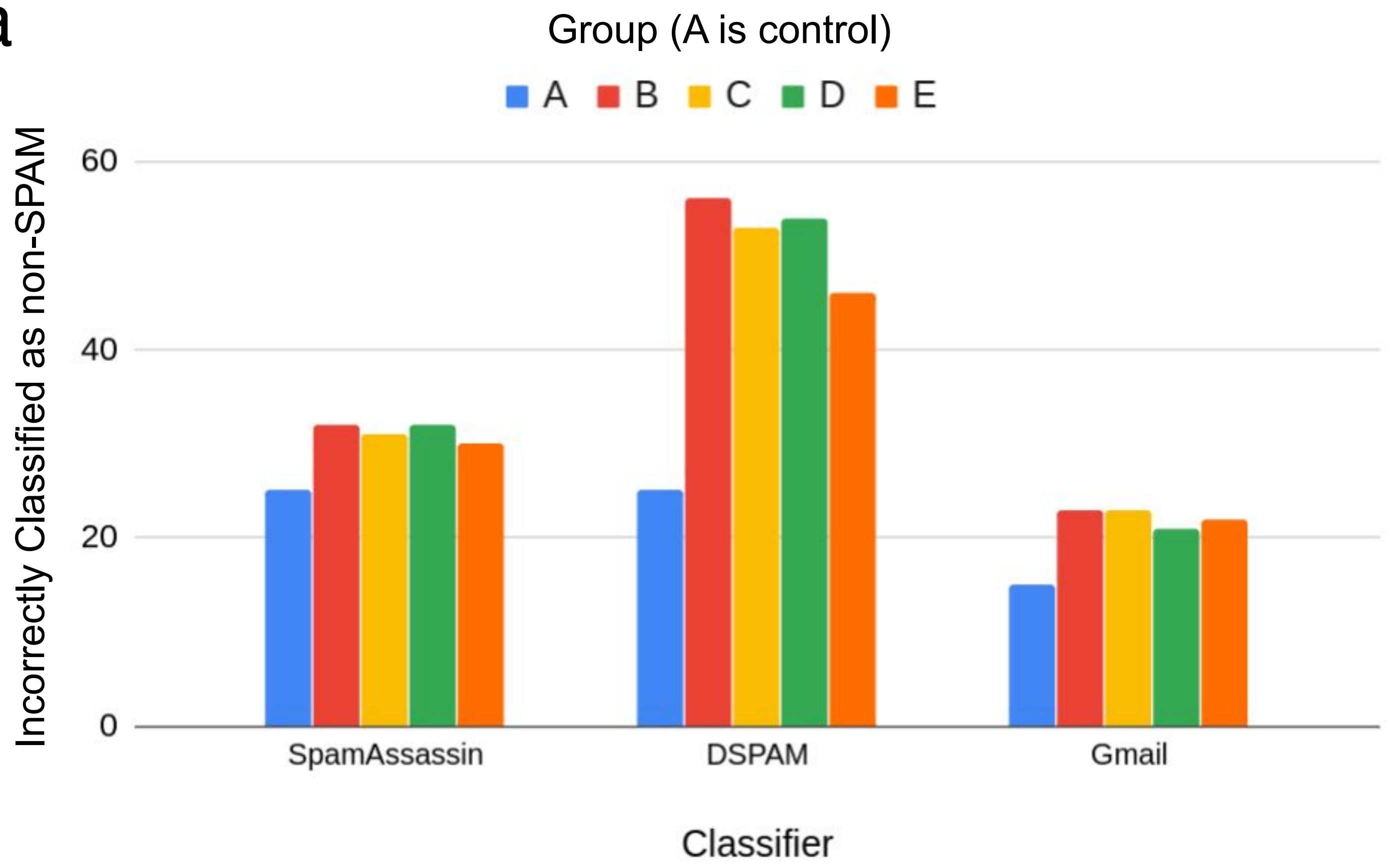
Evaluation

- Results were evaluated with three different spam filters
  - SpamAssassin 3.1.4
  - DSPAM 3.8.0
  - Gmail
- Spam emails obtained from Bruce Guenter’s SPAM archive

Evaluation

- HTML stripped from messages
- Manually corrected pre-existing word-alternation based filter attacks
  - E.g. “he==llo” => “hello”

Data


Group (A is control)
A
B
C
D
E
Incorrectly Classified as non-SPAM
60
40
20
0
SpamAssassin
DSPAM
Gmail
Classifier

Data (uglier)

Mail set	Substitution threshold	Exchange threshold	Replacement strategy	Mails not recognized as spam by		
				SpamAssassin 3.1.4	DSPAM 3.8.0	Gmail
Test Set A	100%	100%	-	25	25	15
Test Set B	60%	95%	minimum	32	56	23
Test Set C	60%	100%	minimum	31	53	23
Test Set D	60%	100%	random	32	54	21
Test Set E	80%	100%	minimum	30	46	22

Limitations

- Substitution was not always able to find a good word to use
    - Instead do character exchanges, but those do not usually fool spam filters
- Sometimes word substitutions do not make sense to a human
- Spam often has bad grammar which makes substitution more difficult

Later Research

Mostly ways to counter the attack proposed in our paper

Enhanced Topic-based Vector Space Model for semantics-aware spam filtering [2]

2012

Igor Santos, Carlos Laorden, Borja Sanz, and Pablo G. Bringas

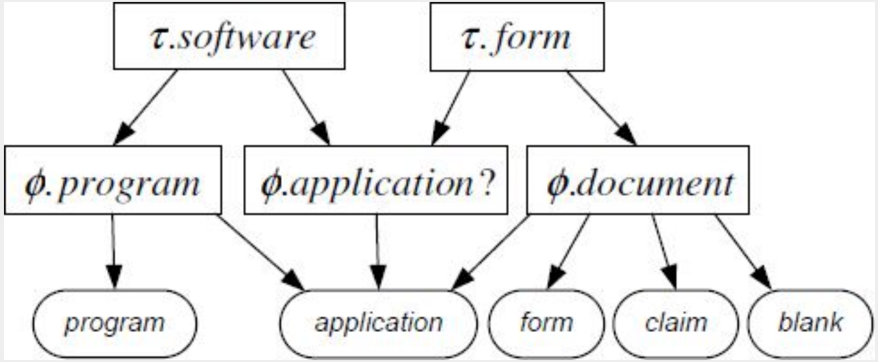
VSM

- Models natural language
- Used in information retrieval
- Treats words as independent

eTVSM

- Accounts for meaning
- Topics → interpretations → terms

[3]



2012 - eTVSM

- Represented emails with eVTSM
- Trained machine learning classifiers
- Successfully identified many spam messages

Evasion-Robust Classification on Binary Domains [4]

2018

Bo Li and Yevgeniy Vorobeychik

- Our paper was an evasion attack
  - Intelligent adversary
- And had a binary feature space

2018 - Evasion-Robust Classification

- Authors created 2 frameworks
  - General
    - Mixed-integer linear programming
    - Accounts for feature cross-substitution attacks
  - RAD
    - Algorithm for retraining with arbitrary attack models & classifiers

- And tested them
  - Filtering spam
  - Identifying handwritten numbers

Opportunities to do similar research

NEU SecLab - practical security

- Security applications of program analysis
- Web & mobile security
- Malware
- Botnets

Basic knowledge of security is helpful

https://seclab.ccs.neu.edu/ ek@ccs.neu.edu

FACULTY

Engin Kirda
Professor

William Robertson
Associate Professor

Conclusion

- Spam emails are a serious concern and major annoyance
- Bayesian spam filters are an important technology for removing spam
- They are not perfect and can be fooled by substitution
  - Replacing suspicious words with more innocuous ones
  - This can be used to improve filters in the future
- This shows we need more improvements to filter spam

References

[1] Christoph Karlberger, Günther Bayler, Christopher Kruegel, and Engin Kirda. 2007. Exploiting redundancy in natural language to penetrate Bayesian spam filters. *WOOT '07: Proceedings of the first USENIX workshop on Offensive Technologies*, Article 9 (2007), 7 pages.

[2] Igor Santos, Carlos Laorden, Borja Sanz, and Pablo G. Bringas. 2011. Enhanced Topic-based Vector Space Model for semantics-aware spam filtering. *Expert Systems with Applications* 39, 1 (Jan. 2012), 437-444. DOI: https://doi.org/10.1016/j.eswa.2011.07.034

[3] Ahmed Awad, Artem Polyvyanyy, and Mathias Weske. 2008. Semantic Querying of Business Process Models. *12th International IEEE Enterprise Distributed Object Computing Conference* (2008), 85-94. DOI: https://doi.org/10.1109/EDOC.2008.11

[4] Bo Li and Yevgeniy Vorobeychik. 2018. Evasion-Robust Classification on Binary Domains. *ACM Trans. Knowl. Discov. Data*. 12, 4, Article 50 (June 2018), 32 pages. DOI: https://doi.org/10.1145/3186282