



# Bayesian Spam Classification Applied to Phishing E-Mail

*As phishing becomes more devious, advanced techniques are needed which can identify phishing e-mail faster and more accurately.*

## CONTENTS

Introducing Phishing	2
Identifying Phishing E-mails	3
Applying Bayesian Filters to Phishing	3
- Evaluating text	3
- Including phishing indicators	4
- Training from datasets	5
- Applying the filter	5
- Testing the filter	6
Applying Bayesian Phishing	7
- Blocking phish faster	7
Conclusion	7

## Introducing Phishing

“Phishing” is the term for an e-mail scam that spoofs legitimate companies in an attempt to defraud people of personal information such as logins, passwords, credit card numbers, bank account information and social security numbers. For example, an e-mail may appear to come from PayPal claiming that the recipient’s account information must be verified because it may have been compromised by a third party. However, when the recipient provides the account information for verification, the information is really sent to a phisher, who is then able to access the person’s account. The term phishing was coined because the phishers are “fishing” for personal information.

Phishing e-mails are sent to both consumers and companies, trying to gain either personal information from an individual or confidential information about an enterprise. In phishing e-mail messages, the senders must gain the trust of the recipients to convince them to divulge information. The phishers attempt to gain credibility through mimicking or “spoofing” a legitimate company through methods such as using the same logos and color scheme, changing the “from” field to appear to come from someone in the spoofed company, and adding some legitimate links in the e-mail. Below is an example of phishing e-mail.

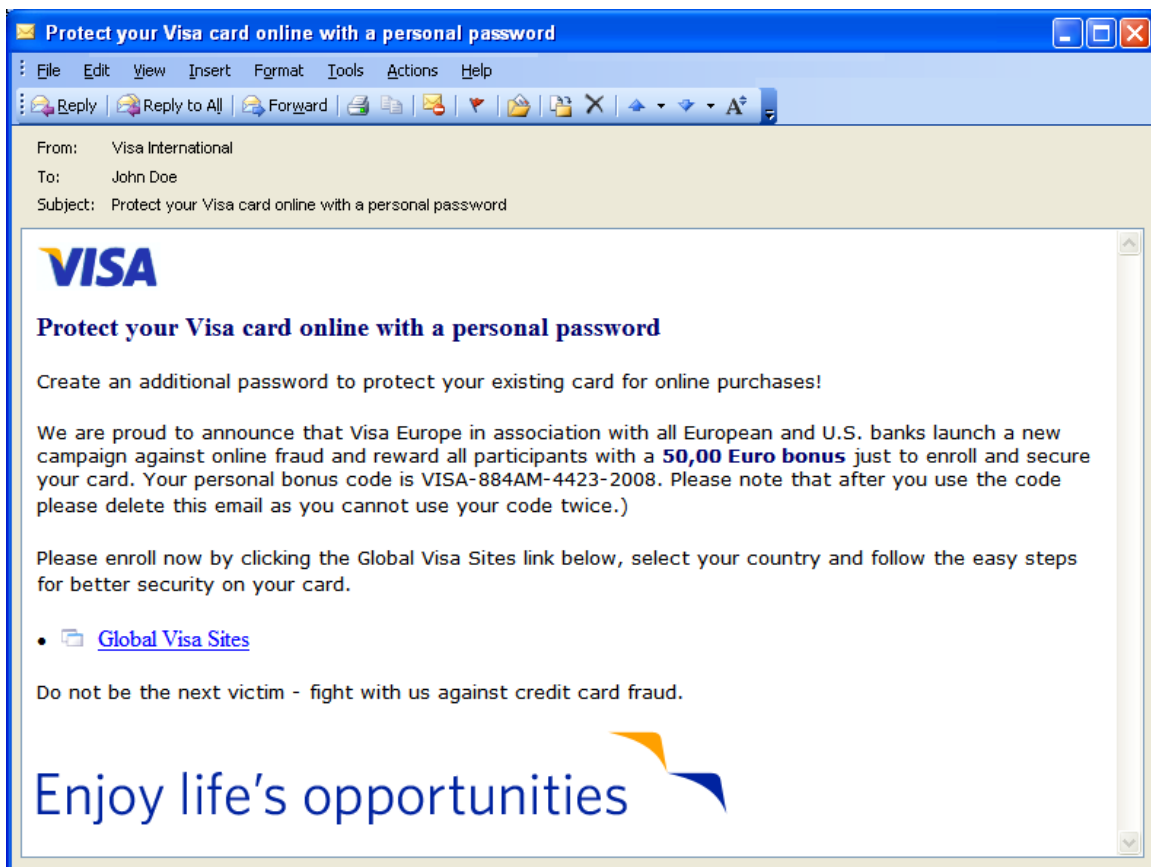


Figure 1 – Consumer Phishing

Once credibility has been established, the phishing e-mail will present a plausible premise that requests the recipient to act. For example, the e-mail may claim that the recipient's account information is outdated, a credit card has expired, or the account has been randomly selected for verification. The request is framed in an urgent situation requiring a quick response. There are numerous approaches and each tries to create a scenario that would convince the recipients that they must provide the requested information. To collect this information, the phishing e-mail generally provides a link to a phishing Web site. Phishing Web pages have fields for collecting the information the phisher is seeking and whatever else seems relevant to the scam. More recently, phishers have also utilized VoIP technology to place phone numbers in a phishing e-mail as the primary form of contact. This type of attack is sometimes known as "vishing". Be it phishing or vishing the results are same, the information collected can be used for financial gain (credit cards, account numbers), identity theft, or even access to a company's private or proprietary information.

## Identifying Phishing E-mails

Not only must phishing e-mails be caught, but they must also be specifically categorized as phishing e-mails. They cannot just be grouped in e-mail junk boxes along with spam. Phishing e-mails can do substantial damage and recipients are easily fooled because they are designed to look like good e-mail. People expect to receive correspondence from their banks and are likely to believe that the phishing e-mails are valid requests for information.

SonicWALL® has posted a Phishing IQ Test on its Web site testing whether people can correctly identify phishing e-mails (<http://www.sonicwall.com/phishing>). This test has been taken by over 1 million people. The results show that the participants misidentified the e-mails 22 percent of the time, phishing e-mails were misidentified as legitimate at a rate of 14 percent and legitimate emails were misidentified as phishing at a rate of 37 percent.

Phishing e-mails need to be classified as phishing to prevent users from believing that the e-mails are legitimate correspondence that were accidentally labeled as spam. SonicWALL found that when phishing emails were placed in a general junk folder, these e-mails were "unjunked," or put back in the inbox, up to ten percent of the time, whereas spam e-mails are unjunked at a rate which is significantly less than one quarter of one percent.

Although it is important that phishing e-mails be labeled as phishing to communicate to users that these are bad correspondence, it is then crucial to avoid any false positives, which misidentify good e-mails as phishing. Users will expect the phishing e-mails to resemble legitimate correspondence and will most likely not be able to identify when good e-mails, possibly requiring important transactions, have been mislabeled as phishing. An effective phishing filter will target features specific to phishing e-mails, identifying these e-mails as phishing while letting good e-mail pass through.

## Applying Bayesian Filters to Phishing

Bayesian spam filtering has a well established history as an anti-spam weapon. However, these filters are less than effective at identifying phishing emails. Spam emails are generally a sales pitch aimed at promoting a product or service. While phishing emails are designed to look like legitimate transactional correspondence and almost always work to disguise their true source. To accurately catch phishing emails, Bayesian filters must be specifically designed for that purpose.

### Evaluating text

Naïve Bayesian filtering can be applied to identifying phishing e-mails. The probability that an e-mail is Phishing,  $P(\mathbf{F}|\mathbf{E})$ , can be calculated using Bayes' Rule:

$$P(F | E) = \frac{P(E | F) * P(F)}{P(E)}$$

A similar calculation can be made for non-phishing. These calculations include the unknown but fixed value **P(E)**. This value can be ignored because it is simply a scaling factor and only relative values and not absolute values are needed. Thus the equation would be reduced to the following:

$$P(F | E) \propto P(E | F) * P(F)$$

We consider each e-mail to be a set of words and features  $W_0 \dots W_n$ . To determine **P(E|F)** in the calculation above, the product of the conditional probabilities for each word  $W_i$  must be calculated. Naïve Bayes assumes that each word appears independently and the probability that an email is phishing is simply the product of the probabilities for each word.

$$P(E|F) = P(W_0|F) * P(W_1|F) * \dots * P(W_n|F) = \prod_i P(W_i|F)$$

Inserting this into the calculation for **P(F|E)**:

$$P(F|E) \propto \prod_i P(W_i|F) * P(F)$$

Some phishing e-mails contain language that would not be used in legitimate transactional correspondence, making it easy to catch with simple Naïve Bayesian filtering applied to the words in the e-mail. However, the majority of phishing e-mails are sophisticated enough that they will not be identified as phishing merely based on an analysis of the e-mail text. Other features specific to phishing must be included as part of the analysis.

### Including phishing indicators

When applying Bayesian filtering to phishing, various phishing indicators must be included in addition to text analysis to generate an effective result. Some indicators of phishing include:

- Links based on raw IP addresses instead of domain names
- @ symbols used in URLs
- Null characters in hostnames
- Illegal or double-redirects of URLs
- Other methods of obscuring and encoding URLs
- Inconsistent contact points (URLs and phone numbers within the e-mail are from different sources)
- Non-standard ports (using ports other than 80).
- Similar, but illegitimate domain names (for example, paypalverify.net instead of paypal.com)

The value of a phishing feature in discriminating phishing from non-phishing can be determined by examining the Odds Ratio for that feature:

$$OddsRatio = \frac{P(W_i | phishing)}{P(W_i | non - phishing)}$$

Table 1 provides examples of features and their associated odds ratio showing the odds of that feature appearing in phishing e-mail over that feature appearing in a non-phishing e-mail.

Token	Description	Odds Ratio
SUBJ_verify	word “verify” appears in the subject	368.7
PHRASE_verify_your_identity	phrase “verify your identity” is found in the email	206.44
SUBJ_debit	word “debit” appears in the subject	165.7
SUBJ_bank	word “bank” appears in the subject	159.2
URL_:4903	URL is directed to port 4903	194.4
WORD_suspension	word “suspension” is found in the email	142.59
URL_cit	URL contains “cit” for Citibank phishing	139.8
URL_:87	URL is directed to port 87	139.8
LINK_INCONSISTENT	text of HTML link differs from the actual target	50.40

**Table 1 – Odds Ratio Features and Values**

Based on the examples in this table, having the word “verify” in the subject is the strongest indicator of phishing. When the odds ratios of different tokens are considered together, the probability that a particular e-mail is phishing can be effectively estimated. Because text is not a strong indicator of phishing, these more intelligent features need to be layered as part of the Bayesian phishing analysis.

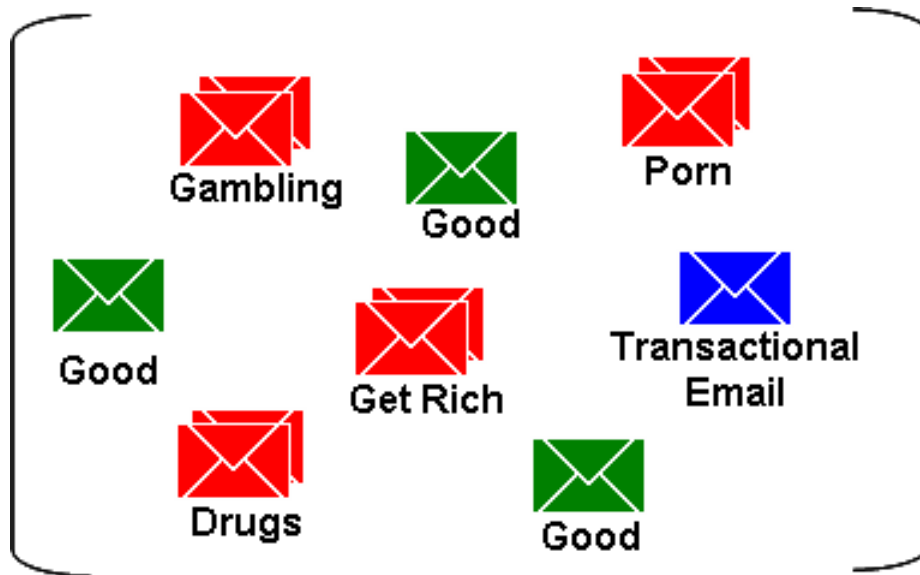
### **Training from datasets**

Once the features have been selected, the Bayesian phishing filter must be trained to determine the probability that a particular occurrence of that feature is an indicator of phishing. For example, if non-standard ports are selected as a feature, the Bayesian phishing filter must be trained to determine the probability that a specific non-standard port, for example port 5880, is an indicator of phishing.

To train a Bayesian spam filter, only two datasets are required: a dataset of spam and a dataset of good e-mail. To apply Bayesian filtering to phishing, two additional datasets are needed: legitimate transactional e-mail and phishing e-mail. A large set of legitimate transactional e-mail is needed because this set of e-mail most resembles phishing e-mails and the filter must have numerous examples of legitimate transactional email to help ensure a low false positive rate.

### **Applying the filter**

After the filter has been trained on sufficiently large datasets, it can be applied to incoming e-mails. First, Bayesian detection methods are applied, which causes three types of messages to emerge: good e-mail, spam e-mail, and transactional e-mail which includes both legitimate and phishing transactional e-mail.



**Figure 2 – Sorting e-mail by type**

Once the transactional e-mails have been separated, the intelligent phishing features are applied to make a phishing judgment. This will separate the phishing emails from the legitimate transactional e-mails.

### Testing the filter

SonicWALL trained a Bayesian phishing filter using the methods discussed above and then tested the filter for effectiveness. The test e-mails were passed through the Bayesian phishing filter and gave the following results.

	Phishing set	Legitimate transactional set	Good set
Set Count	2,193	1,177	12,978
False Negative	582 (27%)		
False Positive		0 (0%)	2 (0%)

**Table 2 – Bayesian Phishing Test Results**

The methods used were set to guarantee a very low false positive rate; zero percent of the legitimate transactional e-mails were misidentified as phishing. As a result of ensuring a low false positive rate, some phishing e-mails were not identified as phishing: there was a twenty-seven percent false negative rate. The over all result was seventy-three percent of phishing emails were caught while all legitimate transactional e-mails were delivered showing that Bayesian methods successfully identify phishing.

When the Bayesian phishing filter was tested, other e-mail filtering methods were not used to ensure a true test of the filter's effectiveness. Methods that were not used included:

- Reputation services
- Authentication
- White-listing

- Real-time black lists
- Real-time phishing link lists

When Bayesian phish filtering is combined with these other techniques, SonicWALL has been able to achieve an effectiveness rate of over 98 percent at identifying phishing e-mails with a 0 (zero) percent false positive rate.

## Applying Bayesian Phish Filtering

Today most anti-spam and e-mail security products have some type of phishing protection. Nearly all of these products rely principally on some form of URL detection and similar techniques noted previously or by using their spam filtering techniques. As we have noted, using anti-spam techniques on a phishing e-mail is insufficient because spam and phishing e-mails are truly different in their purpose, construction and social engineering techniques. Further, Bayesian phishing detection techniques can dramatically reduce the chances of anyone in your organization from being exposed to a phishing e-mail.

### **Blocking phish faster**

Blocking phishing e-mails through real-time black lists, reputation services and other related methods require that the phishing e-mail first be received by someone and then analyzed to ensure that the message is in fact phishing. Only then can the “phishing” designation for that e-mail be communicated to other parties (customers). In other words someone has to be exposed to the phishing threat before it can be identified and stopped.

With Bayesian phish filtering, a phishing message can be determined as phishing “on-the-spot” the first time it is seen as part of the analysis done by SonicWALL Email Security and SonicWALL Anti-Spam products and services. In addition, that phishing designation can be communicated to other parties via the SonicWALL GRID Network. As a result, no one has to be exposed to the threat because it is blocked the first time it is seen.

## Conclusion

Effective e-mail security solutions must protect users from all types of e-mail threats. However, each threat must be treated differently to address its unique properties. Bayesian filtering has been successfully combating spam. However, this same filter cannot be applied to phishing. Instead, a Bayesian phishing filter can be created through layering different features which are strong indicators of phishing. Test results show that Bayesian rules can be effectively applied to identifying phishing. This identification allows phishing e-mails to be uniquely categorized. Even if recipients can not recognize the e-mails as phishing, the e-mails can be securely removed from the inbox and labeled appropriately, helping to keep users safe.

©2008 SonicWALL is a registered trademark of SonicWALL, Inc. Other product names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Specifications and descriptions subject to change without notice.