CERIAS

The Center for Education and Research in Information Assurance and Security

Text-based Approaches to Detect Phishing Attacks

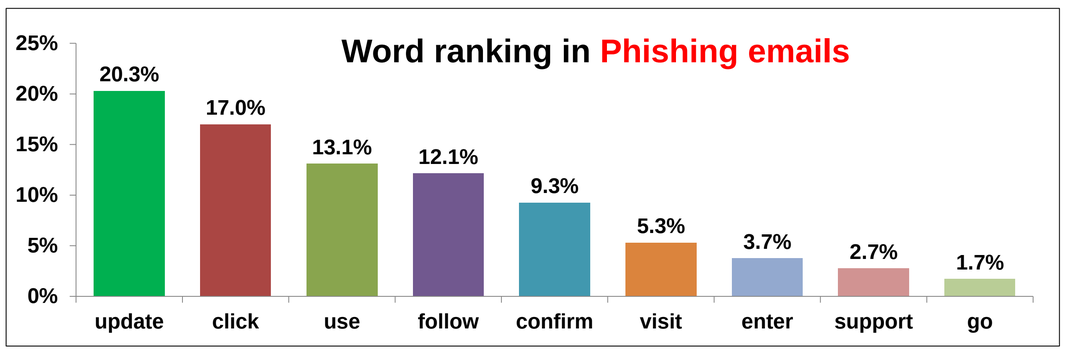
Students: Gilchan Park & Lauren Stuart / Advisors: Julia M. Taylor & Victor Raskin

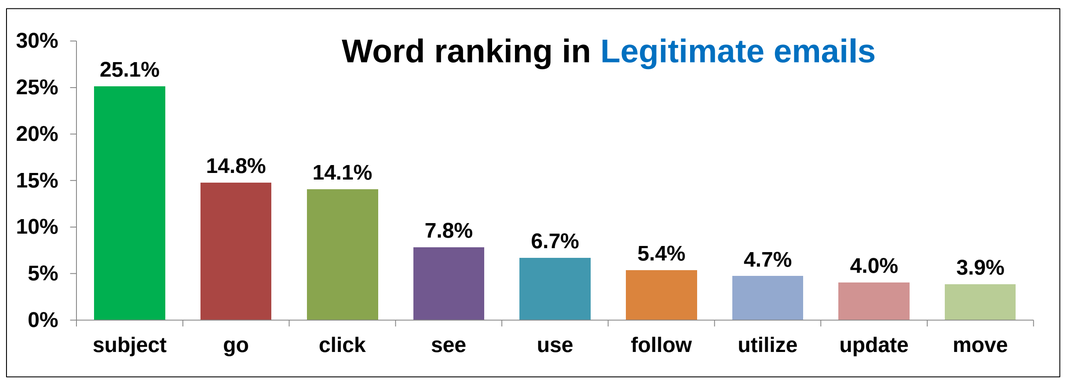
Abstract

The purpose of the first research is to report on an experiment into text-based phishing detection. The developed algorithm uses previously published work on the, so-called PhishNet-NLP, a content based phishing detection system. In particular, this research aims to analyze the keywords that lead used to do some actions in email texts. The algorithm produced the considerable results in filtering out malicious emails (TPR); however, the rate of text falsely identified as phishing (FPR) needed to be addressed. To solve the FPR problem, tradeoff between TPR and FPR was performed to reduce the FPR while minimizing the decrease in the phishing detection accuracy.

The second research's aim is to compare the results of computer and human ability to detect phishing attempts. Two series of experiments were conducted, one for machine and the other one for humans, using the same dataset, and both were asked to categorize the emails into phishing or legitimate. The results prove that machine and human subjects differ in classification of phishing emails. This comparison suggests that humans intelligence to detect some types of phishing emails that machine could not recognize needs to be semantically computerized so as to ameliorate the machine's phishing detection ability.

Text-based Phishing Detection





The process of the text analysis

1. Set the list of keywords.
2. Parse the email text into words assigning parts of speech (POS) to them.
3. Compare the parsed tokens with the synonym sets of keywords.
4. The scores of found keywords are calculated.
5. The maximum score determines whether the email is phishing or not.

< Examples of keywords>

Word / POS	Sentence (from phishing emails)
update / verb	Please **update** and verify your information by signing in your account below.
update / noun	A recent review of your transaction history determined that we require an **update** of your account in order to provide you with secure services.
click / verb	you must **click** the link below and then complete all steps from the following page as we try to verify your identity.
clicking / present participle	please contact Pay Pal by visiting the Help Center and **clicking** "Contact Us".

Results

- **TPR**: ***83.5%*** & **FPR**: ***14.9%*** (the testing data: # 4558 phishing emails, # 7944 legitimate emails)
- Trade-off between TPR and FPR in order to reduce FPR (the words that increase FPR but not TPR were removed).
- Trade-off resulted in about ***5.4%*** decrease in FPR, and about ***1.5%*** decrease in TPR.

Comparing Machine and Human Ability to Detect Phishing Emails

[Example of **Machine**-friendly substitution]

- Some part of phishing email -
eBay sent this **point** ~~message~~ to Jose Nazario (jnazario). Your ~~registered~~ **automated** ~~name~~ **construction** is included to show this message ~~originated~~ **automated** from eBay.

[Example of **Human**-friendly substitution]

- Some part of phishing email -
eBay sent this **communication** ~~message~~ to Jose Nazario (jnazario). Your ~~registered~~ **personal** ~~name~~ **sign** is included to show this message ~~originated~~ **started** from eBay.

Word ranking in **phishing** emails
73. registered
84. originated
137. message
234. name

Word ranking in **legitimate** emails
31. personal
77. communication
84. automated
138. point
234. construction
357. started
531. sign

Objective

- The objective of this research is to see how well machine and human can identify phishing emails, and to compare their abilities..

Experiments

(* phishing emails classified as legitimate by machine)

Emails \ Tester	Machine (Weka - SVM classifier)		Human subjects (Amazon Mechanical Turk)
Legitimate	# 12 Original	# 13 Machine-friendly	# 12 Human-friendly
Phishing	# 12 Original	# 13 Machine-friendly	# 12 Human-friendly
* Misclassified	# 12 Original	# 13 Machine-friendly	# 12 Human-friendly

- Modification of the testing emails with word ranking lists by machine.
  - *Machine-friendly substitution* : keyword substitution with its counterpart in the opposite list based on the word ranking.
  - *Human-friendly substitution* : keyword substitution with its counterpart in the in the opposite list based on the word sense.

Results

- Legitimate emails - machine and humans similarly performed (Humans classified # 9 as legitimate, machine # 10 for machine-friendly substitution).
- Phishing emails - machine outperformed humans (Humans classified # 8 as legitimate for original phishing emails).
- Misclassified emails - humans outperformed machine (Humans correctly identified over # 8, but machine correctly identified at most # 4).

CERIAS®

PURDUE UNIVERSITY