Semantic Detection of Phishing Email

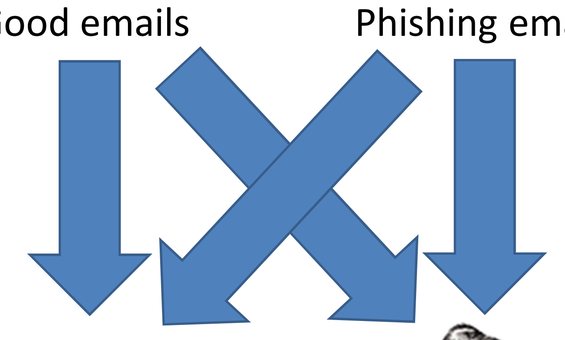
Courtney Falk, PhD Candidate, Information Security

Abstract

This work investigates whether or not the semantic representation of an email's content is more useful than the surface features of the text in classifying an email as a phishing attack email or not. A series of experiments were conducted using machine learning binary classifiers to measure the performance of the competing approaches. The conclusion is that semantic information is just as good if not better in every case than text surface features.

Good emails

Phishing emails



OST

Manual processing

Train & test

Train & test

Train & test

Phishing is a cybersecurity attack that relies on social engineering [1]. The attacker sends a message to the intended victim and tries to convince him/her to open an attached file, click on a link, or some other action that will complete the attack. Phishing continues to be an effective method for attackers because it relies on a human user, and humans are often undertrained and unaware about the kinds of cybersecurity threats they'll face.

Using the Ontological Semantics Technology (OST) [2] approach to natural language processing (NLP) provided the semantic structures that would be input for the experiments. Fifty-six emails were manually processed to serve as a training corpus. Then the meaning-based machine learning (MBML) approach was followed to produce semantic machine learning (ML) features [3].

EXAMPLE: *"The quick brown fox jumps over the lazy dog."*

Surface N-Grams:
{"the" : 2"}, {"quick" : 1"}, {"brown" : 1},
{"fox" : 1}, {"jump" : 1}, {"over" : 1},
{"lazy" : 1}, {"dog" : 1}

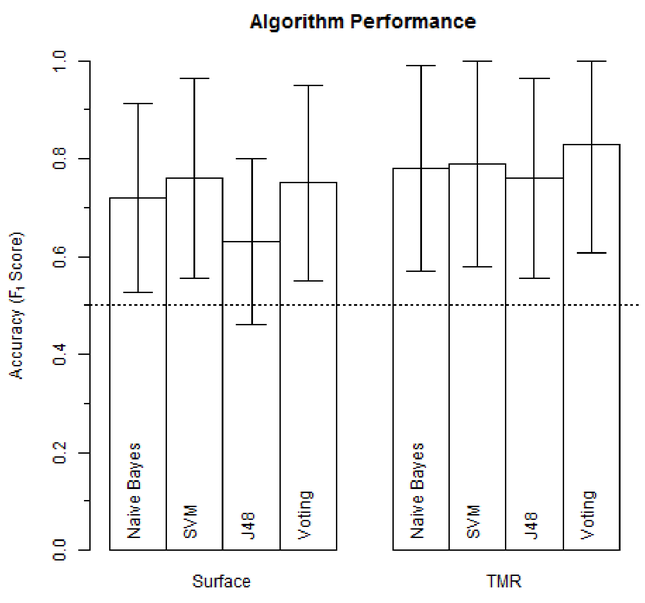
Text Meaning Representation:

```
(JUMP-3
  (AGENT (VALUE (FOX-7
    (HAS-COLOR (VALUE (BROWN)))
    (HAS-SPEED (VALUE (>0.5)))
  )))
  (PATIENT (VALUE (DOG-13
    (AGENT-OF (VALUE (LAZE-4)))
  ))))
```

TMR N-Grams:
{"JUMP|AGENT|FOX" : 1}, {"FOX|HAS-COLOR|BROWN" : 1},
{"FOX|HAS-SPEED|>0.5" : 1}, {"JUMP|PATIENT|DOG" : 1},
{"DOG|AGENT-OF|LAZE" : 1}

Experiments were done by training binary classifiers using three contrasting algorithms: Naïve Bayes, SVM, and J48 (C4.5). The Weka machine learning (ML) suite was used to train and test all the models [4]. Cross validation (K=3) ensured that the models were not overfitting the data.

The semantic ML features definitely perform better than random. And although the semantic approach generally performs better than its unigram version. But the large confidence intervals (95%) mean that this conclusion can't be reached statistically. This is due to the relatively small sample size (56).



References:

1. APWG. "Origins of the Word 'Phishing'." [Online] http://docs.apwg.org/word_phish.html
2. S. Nirenburg and V. Raskin. 2004. *Ontological Semantics*. MIT Press.
3. C. Falk and L. Stuart. "Meaning-Based Machine Learning." FLAIRS-29. 2016.
4. M. Hall et al. 2009. "The Weka Data Mining Software: An Update," *SIGKDD Explorations*, 11(1).