**Abstract**

**Introduction**

* Introduce the topic of phishing
  + What it means and why is this type of attack employed
  + The different types of phishing
    - i.e spearing, whaling, etc.
* Report the losses of phishing attacks of victims and who was targeted
* Introduce the methodology that will be used to detect the phishing emails
  + Pre-processing: Python NLP
    - Explain what pre-processing is
  + Machine Learning: various algorithms
    - Explain what machine learning is

**Related Work**

* Explain how a lot of research has been done on phishing websites but far less on phishing emails specifically
  + Explain their approach to performing the analysis on phishing websites
* Explain how the research that has been performed on phishing emails differs from the research that I intend to perform
  + My method of preprocessing will place more emphasis in the amount that focuses on NLP than other research performed in the field

**Proposed System**

* Explain the how the data comes from being in raw emails to a dataset
  + Explain why certain choices were made
    - i.e focus on speed over accuracy or why more words were chosen
* Create a visual that explains the process simpler:
  + Emails -> Email Classification -> NLP analysis -> Feature Extraction -> Dataset creation -> Results and Machine Learning Analysis

**Features**

* Words
  + Explain which words were chosen and why they were chosen
    - Checked the frequency of which keywords show in phishing emails
  + Explain how the NLP analysis helped this analysis
    - Explain the process of stemming
* Part-of-speech Analysis
  + Explain what POS analysis and how it was performed
  + Explain why it was performed
    - The objective of phishing emails is to cause action thus action verbs are important

**Classifiers**

* Explain what each classifier that was used in machine learning analysis
* Explain if different types of models were used for the same classifier, then why

**Dataset**

* List the total number of features of the dataset
* List the total number of phishing emails
* List the total number of ham emails
* List the total number of emails

**Results**

* Explain what k-cross fold validation is
* Explain what classifies as a false positive result
* Explain what classifies as a false negative result
* Explain what classifies as correctly classified result (ham and phish)
* Explain results of each classifier & display them

**Conclusion**

* Summarize all that was done to get features
* Summarize results of all findings
* Talk about what these mean as a result
* Where room for improvement lies

**Abstract**

Communication by email is one of the most popular and commonly used by people today. Emails can be utilized in various forms and because of the versatility of email communication, sensitive and confidential information can be distributed through emails. Phishing is a method of obtaining this sensitive/confidential information from emails by tricking people into revealing this information. Phishing is a social engineering technique that is deployed to obtain sensitive/confidential information from victims by tricking the victim into believing that malicious host is a reputable host. Phishing attacks have characteristics that allow for easier distinguishing between safe emails and malicious emails since attackers cannot take the identity of the reputable source and they all have a certain goal when employing their attacks. For this research, the main approach of detecting these phishing emails was through classification and extracting features from the emails themselves to create the dataset. Then, machine learning algorithms would use the dataset to make predictions on whether an email was safe (ham) or phishing. Highest results for the algorithms currently marks at 93.50%.

**Introduction**

**Related Work**

**Proposed System**

The creation of the dataset takes place in a multiple phase process. First, phishing emails and ham emails had to be collected from online resources for the analysis of the project. The phishing emails (.eml files) were collected from the phishing corpus and the ham emails were found in a resource on Kaggle that listed all the ham emails (.txt files). From there, the emails have to be within the same format for proper analysis to be performed, therefore, the ham emails were converted into the same format that the phishing emails consisted of. The files are now prepared to be fed into a preprocessor and have various features extracted.

The preprocessor extracts various features from the emails but the first feature that is given to each of the emails is whether an email is a phishing email or a ham email. These labels are given to each email depending on which source (in the directory) the email came from and labels it accordingly. Then, using Python’s BeautifulSoup library and string manipulation, emails are cleaned of extra text that would influence the speed and efficiency of extracting features from the emails. Sequentially, using Python’s NLTK library, each email was broken up word-by-word and features were extracted according to those listed in **Features**.

Once every email has been analyzed, all of the features that have been extracted are exported into a csv file that can be used for machine learning analysis. From there, machine learning models and results are explored using the classifiers found in **Classifiers** and results are detailed in those found in **Results**.

**Features**

Word-based Features – Word-based features consist of checking the amount of times a certain word appears in an email. Words that have the same meaning but are written in different forms were stemmed to allow for a smaller csv file to be generated. For example, the word “run” and “running” were considered the same word since they both contain the stem word of “run.” The words that were checked for in the emails were the following:

# of Account –

# of Alert –

# of Confidential –

# of Fraudulent –

# of Indefinite –

# of Information –

# of Notification –

# of Password –

# of Key Action Verbs –

POS Features – Parts-of-Speech (POS) consist of the percentages of how much each part of speech appears in the email. The parts of speech that were checked for were the number of times the following appeared in the emails:

* Nouns
* Verbs
* Adjectives
* Adverbs

**Classifiers**

**Dataset**

The analysis of the project focused on a dataset that consisted of both ham and phishing emails. There was a total of 7,017 emails analyzed within this dataset. Of those 7,017 emails, 4,466 (64%) of those emails were classified as phishing. The remaining 2,551 (36%) of those emails were classified as ham emails. From all of these emails, there were a total of twenty features extracted from each of the emails.

**Results**

A false positive, in the context of this research, describes an email that was a ham email but was classified as a phishing email. A false negative, in the context of this research, describes an email that was a phishing email but was classified as a ham email. Correctly classified emails will be predicted correctly based from the label given to them in the dataset.

The results of each classifier can be seen in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Accuracy (%)** | **False Positive (%)** | **False Negative (%)** |
| Support Vector Machines | 90.5 | Yet to be implemented | Yet to be implemented |
| Decision Trees | 93.5 | Yet to be implemented | Yet to be implemented |
| K-Nearest Neighbors | 92.0 | Yet to be implemented | Yet to be implemented |
| Linear Regression | 90.6 | Yet to be implemented | Yet to be implemented |
|  |  |  |  |

**Conclusion**