**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 each approach in specific paper
  + Explain where their focus was and where my focus concentrates on
* 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 mark at 93.50%.

**Introduction**

Phishing, in the simplest of terms, is the practice of sending in fraudulent messages under the guise of being a reputable source with the intention of tricking the recipients of the phishing messages into revealing personal information (credit card information, login information, etc.) to the attacker. By performing these types of attacks, attackers can yield access to various amounts of information that are beneficial to their purposes. Phishing, as an attack, is classified as a social engineering attack which utilizes psychological manipulation into tricking users into a security mistake. While phishing is a broad range of attacks, phishing can be classified into four common versions of the attack. The four common versions of the attack are:

1. Spear Phishing – This version of the phishing attack has a designated target as its victim and crafts the attack into having higher likelihood of success. The spearing analogy originates from the fact that you can target a specific fish with a spear compared to a fishing pole.
2. Whaling – This version of the phishing attack is a subset of spear phishing by targeting a person who is high up in a company or in a position of power. The whaling analogy arises from the fact that a whale is a target that is even bigger than a fish.
3. Smishing – This version of the phishing attack is distributed through text messaging or short message service (SMS) which has a clickable link within the message or phone number to call back to.
4. Email Phishing – The most common form of the phishing attack, this version distributes the attack to any email addresses within their possession and try to harvest personal information from the victims through clickable links and social engineering tactics.

With all these options (and many more) as potential avenues for attacks, that underlies the reason why phishing attacks have increased in both number and volume. Research has found that about 90% of companies/organizations have been targeted by these types of attacks in some capacity. Since the people within these institutions are frequently targeted because they are high profile candidates to the attackers, which is why, along with human education and preparation for these attacks, a technical approach can be utilized as well in the protection of these phishing attacks.

**Related Work**

Various other research studies within the field of phishing emails vary in the approach made to the analyzation of the phishing attacks. A popular approach found within this field of study is the research on phishing websites and how these websites differ from the typical legitimate email. This method of analysis does not place a significant level of importance on the efforts of social engineering and rather focuses on the information that can be revealed from the phishing websites domain and the analysis of responses and other key features. In a much smaller subsection of phishing analysis, there can be found analysis on phishing emails. The analysis on phishing emails varies from how the analysis performed in this study persists as the methodology used within that research focused on analysis technique mixing the analysis of URL and few key words that appeared within the email. The methodology that will be used within this research focuses more on the social engineering aspect of phishing attacks which checks multiple key words in addition to the types of word used to engineer these attacks.

The research performed by Rawal, Rawal, Shaheen, and Malik focuses primarily on the detection of phishing emails through machine learning techniques and data processing. The methodology behind their approach to detection of phishing emails lies in extracting key features from the emails which then leads to a dataset that machine learning techniques can be performed on. The primary features that were extracted from the emails can be categorized into the following three categories:

1. Link-Based
2. Tag-Based
3. Word-Based

From these three displayed categories, a dataset consisting of nine fields was created and passed to the machine learning algorithms for analysis. The research and methodology that is being proposed within this analysis differs by placing more emphasis on the word-based approach and creating a larger dataset. From there, it focuses on various words and the natural language of the emails to the end goal of producing a dataset that holds key words and the percentages of parts of speech in the emails.

The research performed by Mahajan and Siddavatam focuses on the detection of phishing websites by pulling out features that can be found within the website itself. Some of the features that can be detailed in the research are whether a URL is redirected and various other features that can be extracted from a URL. By performing this type of analysis on the phishing websites, the researchers observe the behavior of websites and determine what type of behavior and responses from websites illicit phishing qualities of websites or legitimate types of websites. The research handles phishing emails and the natural language that attackers employ to coerce victims into clicking on links within the email. This practice will allow for the analysis of phishing emails without the opening of links that occur when checking for certain attributes of URL features.

The research performed by Krishna et al. also places a primary focus on phishing websites compared to phishing emails for the primary analysis. By placing a larger focus on the websites and the behavior of the website compared to the word choice used within the phishing email, their research aimed at the identification of phishing websites without the factor of social engineering in the attacks. As attackers develop more and more nuanced ways of deceiving people, an underlying factor beneath all the attacks will be convincing the user to land on the website and give their credentials. The research that was performed within this paper focuses on those manipulation tactics of the attackers in the phishing emails. By placing more emphasis on the word choice within the emails compared to the websites of the email, the tricks/trends that the attackers use become more apparent.

**Proposed System**

The preprocessing system of this methodology makes use of Python and the NLP library that exists in the python libraries (NLTK). Through use of multiple libraries in the Python language, a dataset that was suitable for supervised machine learning analysis was implemented. The utilized models for the analysis of the dataset used techniques, such as k-fold cross validation and other techniques geared toward their respective model, to perform analysis and infer which emails were phishing emails and ham emails. The methodology behind each level of the process in the proposed system is explored in more detail in the following sections of the **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 must 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 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 number 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.” An example of a few words that were checked for in the emails were the following:

Account – phishers are trying to harvest the credentials of users, therefore, “account” appears often in a phishing email.

Alert – phishers are aiming to rattle a person into performing actions without doing the proper inspections of a legitimate email, therefore, stating “alert” raises likelihood of error.

Confidential – often, phishers make statements stating that confidential credentials have been exposed and a password reset is required which would lead to the stealing of the victim’s information

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**

Support Vector Machine – Support Vector Machine is a supervised machine learning model that uses classification algorithms for two group classification problems. Once a support vector machine has been given a dataset to learn on, it can predict what an unlabeled set of data should be by dividing regions within its learning set.

Decision Trees – Decision Trees is a supervised machine learning model that were used to categorize classes based on how previous questions were answered in the tree. The Decision Tree learning model is trained on a dataset and the questions are made to classify the types given in the dataset.

K-Nearest Neighbors – The K-Nearest-Neighbors model is a supervised machine learning model that categorize an unknown data point depending on the neighboring data points surrounding the unknown data point.

Linear Regression – Linear Regression depicts the relationship between variables using a straight line to predict unknown data points within the dataset provided to the model. Linear Regression is a supervised machine learning model that categorizes data points according to the linear line.

**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 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**

**Bibliography (Works Cited)**

Alkhalil, Zainab, et al. “Phishing Attacks: A Recent Comprehensive Study and a New Anatomy.” *Frontiers*, Frontiers, 9 Mar. 2021, https://www.frontiersin.org/articles/10.3389/fcomp.2021.563060/full.

Krishna V, A., A, A., Jose, B., Anilkumar, K., & Lee, O. T. (2021, August 2). *Phishing detection using machine learning based URL Analysis: A Survey*. International Journal of Engineering Research & Technology. Retrieved September 29, 2022, from https://www.ijert.org/phishing-detection-using-machine-learning-based-url-analysis-a-survey

Mahajan, R., & Siddavatam, I. (2018, October). *Phishing website detection using machine learning algorithms - IJCA*. Phishing Website Detection using Machine Learning Algorithms. Retrieved September 29, 2022, from https://www.ijcaonline.org/archives/volume181/number23/mahajan-2018-ijca-918026.pdf

Rawal, Srishti, et al. “Phishing Detection in e-Mails Using Machine Learning.” *Research Gate*, Oct. 2017, https://www.researchgate.net/publication/320257918\_Phishing\_Detection\_in\_E-mails\_using\_Machine\_Learning.