CMPE-255 Project Proposal - Detection of Phishing/Suspicious Emails

Rishitha Bandi, Shreya Goyal, Shashank Raghuvanshi, Praveen Kumar

September 24, 2020

Contents

1 Abstract

This study focuses on proposing a model for phishing email detection using a set of features from the different email parts. Features will be extracted from the EML[1] file using the Java program and then will be classified using the J48 algorithm. These features will be assigned weight according to the importance of each feature. Different classifiers will be used to classify emails based on the features extracted from the email.

The goal is to find the best suitable classifier for the phishing email detection with greatest accuracy.

2 Motivation

Phishing is an online attack where phishers used to send an email that seems to be legitimate but it is used to steal personal identity information and financial account credentials. A phishing attack can happen at different levels like a Web page, Email, SMS, or instant voice call. Phishers try to deceive customers by sending an email that contains a link for the webpage. The email contains instructions to go to a fake webpage which seems to be legitimate but originally it ends up stealing personal data.

Over time there has been a considerable amount of increase in online transactions and with that frequency of phishing email have also increased in our inbox.

Training programs organized can help in mitigating phishing emails by raising awareness among users. However, training programs have to be made continuous for knowledge retention otherwise it is for a few days only.

This analysis focuses on enhancing the accuracy of phishing email classifiers.

3 Literature Survey

We found a lot of recent interesting research around the problem of detecting Phishing Emails, many of which vary in the way they approach the problem in notable ways compared to each other. Here we are listing a brief description of some of them.

1. [PS10] Focused mainly on content-based filtering model using a statistical classifier, and along with the usual set of features (that are common) came up with few new features to look out for classification. For extracting topic features a new "latent Class Topic Model(CLTOM)" is used, which finds word clusters which are more specific for the distinction of phishing emails from legitimate emails rather than being general purpose. A model based on the Markov chain is used for sequential analysis of text and external links. A method to detect *Hidden Salting* is also developed, which is essentially a trick used in phishing emails to cause emails to appear normal to

human eye, but contains text which makes automated message processing hard.

While this approach was able to reduce rates of phishing attempts to less than 1%. The author has not commented on the efficiency of the approach as multiple statistical models are being used for feature extraction, however the algorithm seems to be robust, and less dependent on manual selection of features.

2. One of the research [Jon13] proposed the use of lexical analysis of URL in the content of emails for classification, building upon their earlier work [Jon11]. The main focus of the research is to enhance the classification accuracy of existing anti-phishing filters, by supplementing it with their approach. The proposed lexical URL analysis also considered relative positions of tokens to increase precision, and each token is assigned a real value based on how likely it is to be in a phishing email.

3. [Hos15] Focussed mainly on the preprocessing step and extracted 23 hybrid features based on both content and header of the emails. The feature extraction method reduced the noise in the data because of missing values and had reasonable default values for each missing value. The authors did a comparative analysis on a set of classification algorithms with different methods of feature extraction along with their own. The results of the study indicated that the Random Forest, J48 and PART algorithm performed well, with PART having the lowest false positive rate.

4. [VR15] Used n-gram analysis of Message-ID field of email header to extract useful features for classification. The authors argued that even if the Message-ID field is an optional attribute, almost 99% of legitimate emails have that field, which forces the creators of phishing emails to include that, to avoid suspicion. Spoofing Message-Id requires advanced technical knowledge[S08] so most of the phishing emails don't do that, which makes Message-Id field a valuable attribute for classification. [Hos15] Also used Message-ID field to compare the domain name with the sender's domain name as one of the features.

4 Methodology

We are using[Hos15] research as base for our project, and plan to tweak our methodology based on other researches as we make further progress.

The model for classification of a suspicious email from intimidating emails exerts the knowledge discovery process and data mining classifiers. It extracts features and feeds to the model for the detection of fraudulent emails.

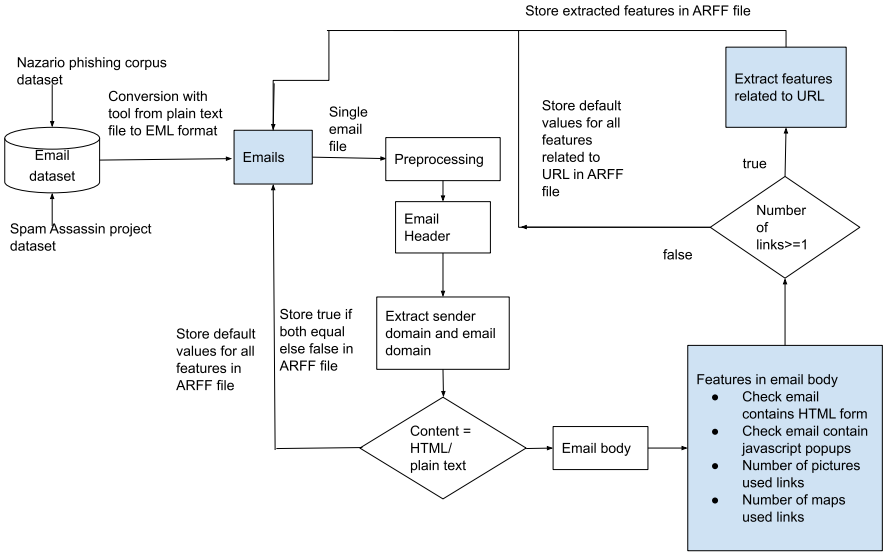


Figure 1: Suspicious Email model

4.1 Data Preprocessing and extraction of features

Data preprocessing and transformation is the key step in data mining. In this step, knowledge is extracted from the data with the features which would affect the model in classification. The dataset of emails is in the form of a plain text file with thousands of records. With the help of tools, it is converted into an EML[1] format for each email. Now that we have data in the format of EML[1], it should be transformed into ARFF[2](Attribute-Relation File Format) file. With the help of a program, the features are extracted from the EML[1] format file and converted into an ARFF[2] file. Features are extracted from each email from four sections i.e Email header, Email body, email sender, URLs in the email.

Extracted features which are saved in ARFF[2] format is our dataset divided into training and testing the model using 15 fold cross-validation. Based on the performance metrics results are evaluated for the data mining classifiers.

Cross-validation is used for the improvement of the performance of the classifiers. When we consider 15 fold cross-validation, 15% of the data is used for testing and the remaining 85% of data is used for training of the model. This process is repeated for all the 15 groups of the dataset. The average of 15 fold performance metrics is considered.

[1]EML: Ecological metadata language
[2]ARFF: Attribute relation file format

4.2 Email Classifiers

Data mining algorithms are used for the classification of the emails as legitimate and suspicious. Training data is fed to the classifiers and weight for the features are attained and applied to the test data for classification.

For the suspicious email classification, we will be using following classifiers:

1. **Logistic regression**
2. **Naive Bayes:** Classifier based on Naive Bayes theorem. It works on the assumption that the presence of each feature in a class is independent of the presence of other features in class.
3. **SVM:** Support vector machine is based on the principle of creating a line in 2D or hyperplane in multi-dimension to separate data into classes.
4. **Decision Tree:** It creates the model in the form of a decision tree and works on the basis of breaking down the dataset into smaller subsets.
5. **Random forest:** It works on the principle of decision trees which are created by selecting the random subset of decision trees. It selects the model by voting from different decision trees.

4.3 Performance Evaluation

To evaluate the model for the classification of emails into legitimate and suspicious for potential effectiveness, the following metrics are calculated.

Suppose NS is the number of suspicious emails, NL is the number of legitimate emails that are detected correctly. Ns is the number of suspicious emails detected legitimate and Nl is the number of legitimate emails detected as suspicious. S is the total number of suspicious emails, L is the total number of legitimate emails.

1. **True positive (TP):** The number of suspicious emails correctly detected as suspicious.

   $TP = NS/S$

2. **True Negative (TN):** The number of legitimate emails correctly detected as legitimate.

   $TN = NL/L$

3. **False positive (FP):** The number of legitimate emails incorrectly detected as suspicious.

   $FP = Nl/L$

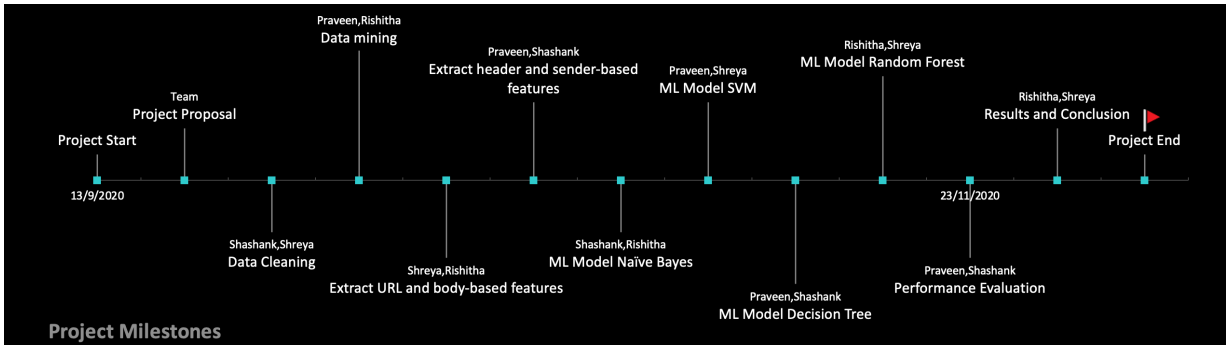
4. **False Negative (FN):** The number of suspicious emails incorrectly detected as legitimate.

   $FN = Ns/S$

5. $Precision = TP/(TP + FP)$
6. $Sensitivity = TP/(TP + FN)$
7. $Accuracy = (TP + FN)/TP + FP + TN + FN$

With the above metrics calculated the performance of the hypothesis is measured and the best suitable classifier is identified.

5 Deliverables



References

[S08] Pasupatheeswaran S. "Email 'Message-IDs' helpful for forensic analysis?" In: *Australian Digital Forensics Conference* (2008). DOI: 10.4225/75/57b2735e40cbe.

[PS10] André Bergholza; Jan De Beerb; Sebastian Glahna; Marie-Francine Moensb; Gerhard Paaßa and Siehyun Strobela. "New filtering approaches for phishing email". In: *Journal of Computer Security* (2010). DOI: 10.3233/JCS-2010-0371.

[Jon11] Mahmoud Khonji; Youssef Iraqi; Andrew Jones. "Lexical URL analysis for discriminating phishing and legitimate websites". In: *The 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)* (2011). DOI: 10.1145/2030376.2030389.

[Jon13] Mahmoud Khonji; Youssef Iraqi; Andrew Jones. "Enhancing Phishing E-Mail Classifiers: A Lexical URL Analysis Approach". In: *International Journal for Information Security Research (IJISR)* (2013).

[Hos15] Sami Smadi; Nauman Aslam; Li Zhang; Rafe Alasem; M.A Hossain. "Detection of phishing emails using data mining algorithms". In: *Software Knowledge, Information Management and Applications (SKIMA), IEEE* (2015). DOI: 10.1109/SKIMA.2015.7399985.

[VR15] Rakesh Verma and Nirmala Rai. "Phish-IDetector: Message-Id Based Automatic Phishing Detection". In: *International Conference on Security and Cryptography(SECRYPT)* (2015). DOI: 10.5220/0005574304270434.