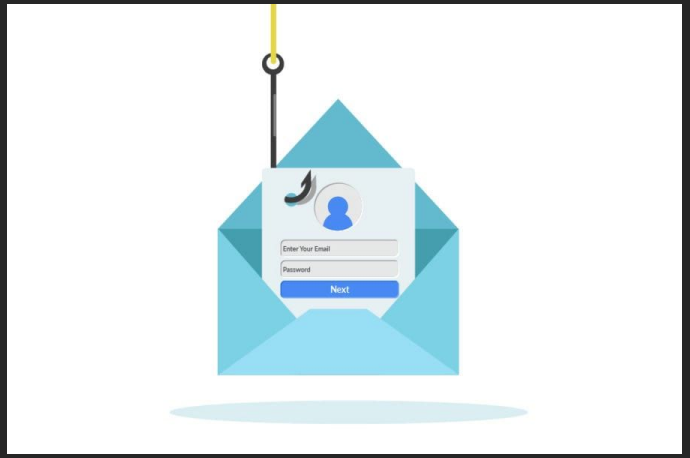


INFO-I513 Usable AI

Phishing Email Detection

Sharanbasav and Sumanth

Problem Statement

Scam emails have become a common occurrence in today's culture. The goal of this project is to create a system that uses machine learning and natural language processing techniques to determine whether or not an email is trustworthy.

For this task we built a machine learning classifier that can calculate the phishing probability of an email. The model input consist of features and attributes of a specific email, and desired output is “phishing” or “not phishing”.

Dataset

The Data for this Project was Borrowed from the Authors of the below cited Paper

Verma, R. M., Zeng, V., & Faridi, H. (2019). Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2605–2607. Presented at the London, United Kingdom. doi:10.1145/3319535.3363267

The author manually collected every email that are phishing and legit and stored them as text file (.txt extension) with and without the email headers.

Data Preparation

- The Initial data was that we obtained was as text files
- 500 legit and 4000 Phishing text files
- Text Files convert to Dataframe with email Parser Library

```
Return-Path: <user@domain>
X-Original-To: user@domain
Delivered-To: user@domain
Received: from domain.com (domain.com [10.5.6.7])
        by domain.com (Postfix) with ESMTP id F1DC646929
        for <user@domain>; Thu,  8 Jun 2015 05:10:11 -0400 (EDT)
Received: from fr-155.domain.com (unknown [211.115.206.155])
        by domain.com (Postfix) with ESMTP id C844C6CCF43
        for <user@domain>; Thu,  8 Jun 2015 05:10:12 -0400 (EDT)
Received: (qmail 17149 invoked by uid 531); 7 Jun 2015 18:54:29 +0900
Date: 7 Jun 2015 18:54:29 +0900
Message-ID: <user@domain>
To: user@domain
Subject: eBay One Time Offer: Become a Power Seller!
From: user@domain <user@domain>
Content-Type: text/html
Status:
X-Status:
X-Keywords:

Dear Customer,
Currently we are trying to upgrade our onlinebanking methods. All accounts have been temporarily suspended untill ea
ch person completes our secure online form. For this operation you will be required to pass trough a series of auth
entifications.
To begin upgrading your account please click the link below.

<<link>>
Please note:
If we don't receive your account verification within 72 hours from you, we will further lock down your account unti
ll we will be able to contact you by e-mail or phone.

© Copyright 1998 - 2006, The domain.com Union At The organization of Chicago, Inc. All Rights Reserved
```

	file_name	text	Phish	Date	From	Subject	To	body
0	289.txt	Return-Path: <user@domain>\nX-Original-To: use...	1	Mon, 18 Jun 2015 03:02:34 +0100	"eBay"<user@domain>	Account On-hold: Please confirm your eBay info...	undisclosed-recipients:;	Message sent trough \n eBay S...
1	262.txt	Return-Path: <user@domain>\nX-Original-To: use...	1	Mon, 31 Jul 2015 21:14:15 -0100	"eBay Priority Protection" <user@domain>	Alert eBay Unpaid Item Strike Received	user@domain	\n\n\n<!--\nstyle2 {color: #0000CC}\n--><\n\n ...
2	276.txt	Return-Path: <user@domain>\nX-Original-To: use...	1	Sun, 17 Jun 2015 10:41:44 -0500	"Jenn Crabtree" <user@domain>	Aknowledge The Receipt Of the Mail	undisclosed-recipients:;	A Computer Database Maintainance is currently ...
3	29.txt	Return-Path: <user@domain>\nX-Original-To: use...	1	Wed, 12 Jul 2015 22:00:42 +0900	"E-gold"<user@domain>	Notification of limited account access	undisclosed-recipients:;	We recently reviewed your e-gold account, and ...
4	15.txt	Return-Path: <user@domain>\nX-Original-To: use...	1	Sat, 3 Mar 2015 22:15:17 +0800	"Simon John Rubias Dela Cruz" <user@domain>	KEEPING TRACK OF YOUR USAGE.	undisclosed-recipients:;	Your web mail quota has exceeded the set quota...

Feature Engineering

- Frequency of top 5 words in Phishing emails
- Frequency of top 5 words in legit emails
- Frequency of uppercase letters
- Frequency of punctuations
- Frequency of stop words
- Datetime to hours and minute

Data Dictionary

The following headers were used for the model:

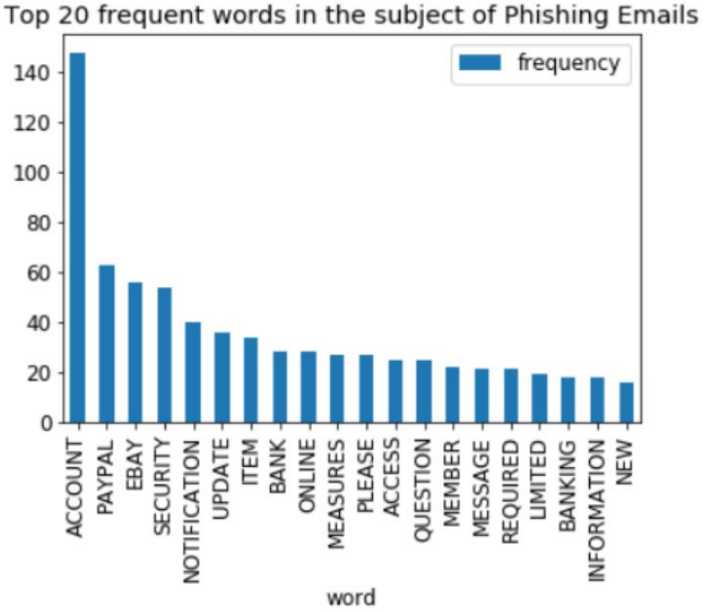
- file_name: This shows the text file name from where the data was extracted from.
- From: This displays who the message is from, however, this can be easily forged and can be the least reliable.
- Subject: This is what the sender placed as a topic of the email content.
- Date: This shows the date and time the email message was composed.
- To: This shows to whom the message was addressed but may not contain the recipient's address.
- body: This is the actual content of the email itself, written by the sender.
- Phish: This Indicated if the email was Phishing or Legit

Data Cleaning

- Removal of stop words
- Removal of non english words
- Removal of Punctuation
- Removal of non Alphanumeric Characters
- Removal of Blank spaces
- Stemming

Exploratory Data Analysis

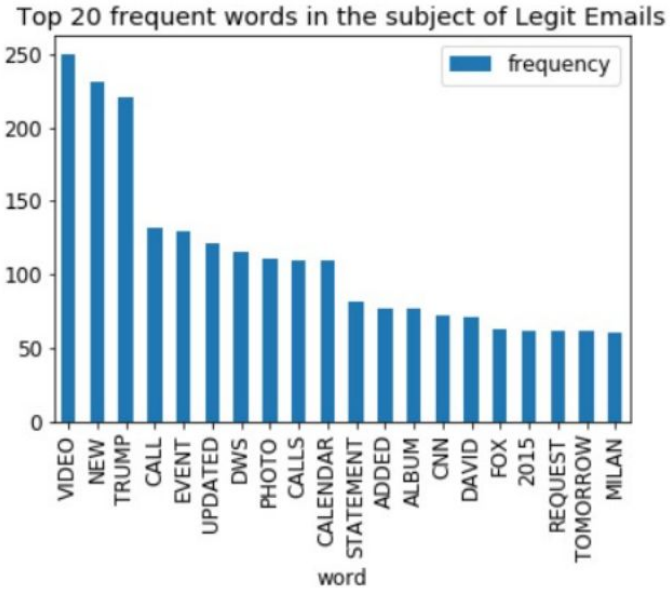
The Bar charts below show the top 20 frequent words that appear in the **Subject** of the **Phishing** emails.



Phishing Mails Word Cloud

Exploratory Data Analysis

The Bar charts below show the top 20 frequent words that appear in the **Subject** of the **Legit** emails.



Legit Mails Word Cloud

Modeling & Results

Logistic Regression model

	precision	recall	f1-score	support
Phish	0.99	0.99	0.99	1008
Legit	0.94	0.89	0.91	125
accuracy			0.98	1133

Naive bayes model (Baseline)

Precision = 0.54 **Recall** = 0.97

	precision	recall	f1-score	support
Phish	1.00	0.89	0.94	1008
Legit	0.53	0.97	0.69	125
accuracy			0.90	1133

Random Forest model

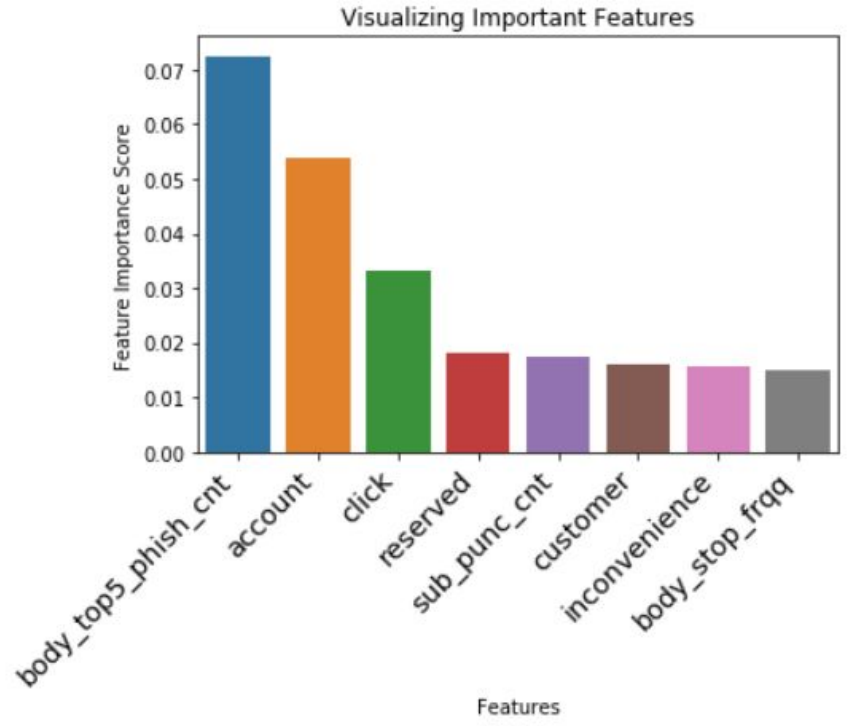
Precision = 0.98 **Recall** = 0.84

	precision	recall	f1-score	support
Phish	0.98	1.00	0.99	1008
Legit	0.97	0.83	0.90	125
accuracy			0.98	1133

Modeling & Results

Feature importance

feature	importance
body_top5_phish_cnt	0.072467
account	0.053950
click	0.033112
reserved	0.018205
sub_punc_cnt	0.017354
customer	0.016023
inconvenience	0.015846
body_stop_frqq	0.014996



Future Work

- Use TF-IDF and check model performance
- Perform sentiment analysis on the text data
- Tree based boosting models like XgBoost and Neural Networks


Q
A