Crossroads Classic Analytics Challenge

Team: Highwaywintercrow

Andrew Huang
Shiue-Yuan Chuang
Indiana University Bloomington

February 18, 2022

Overview

1. Team Members
2. Idea and Strategy
3. What is Phishing-type Email
4. Model Structure
5. Results

Strategies

- The CCAC data has no labels
- We assume the CCAC data were not sampled from current public datasets
- To make our models valuable,
  we need our model enable to predict future phishing email without having future email as training data.

Strategies

- What could be anticipated from future phishing emails?
- Phishing email will innovate
  - Not just try to classify emails to patterns in old datasets
  - **We try to design the strategies that can fit in this scenario: predict new phishing emails with only old emails data**

Strategies

Scenerio

With only old data as training data in hands,
we want to predict new phishing emails.
That is, use data(A) as train data to predict data(B)

Features

- Invariant Features: URLs, Email Address
- Variant Features: Email Body

Test Strategies

We test this strategies on different public datasets,
such as *lingspam* and *spamham*

What is Phishing-type Email

- Total 4898 emails
- Initial Idea:
  - "Bad Guys" send phishing emails: Detect Email Address
  - "Phishing" needs URLs: Detect URLs

Senders, Receivers

- Where do they come from ?
- 53% emails were sent from the same address
- 78% emails were sent from 'ccac.sales'

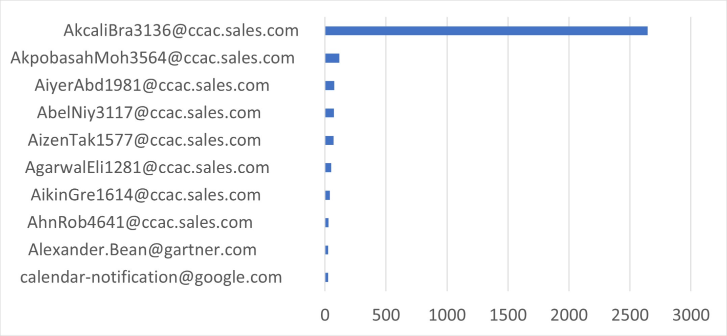


Figure: Top 10 Senders

Senders, Receivers

- Who receives them ?
- 30% emails were sent to the same five address

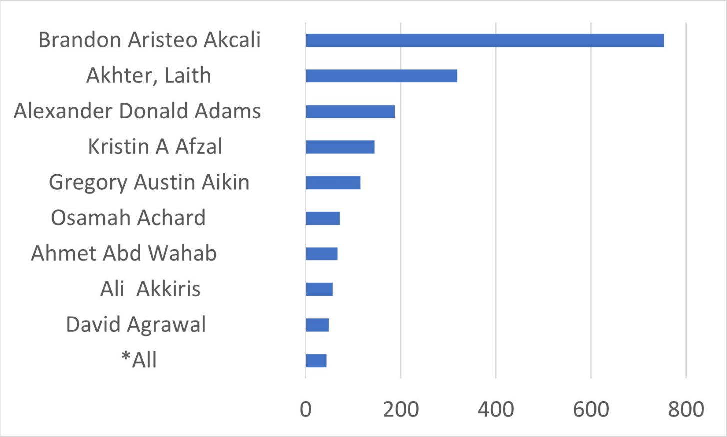


Figure: Top 10 Receivers

Senders, Receivers

- Who are likely targets ?
- We consider these emails are mainly (or pretend as) internal emails in CCAC organization.
- The Senders and Receivers cluster in few groups.

Senders, Receivers

- **Conditional on LIMITED TIME,**
  We would NOT consider Email Address as the Most Priority features to check

Phishing URLs

- Total 4898 emails

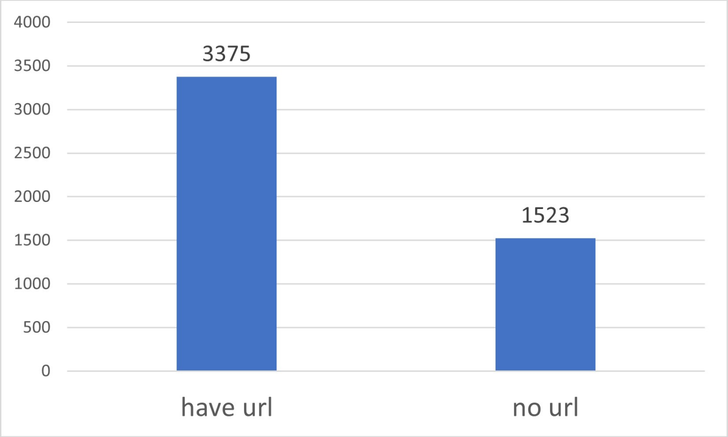
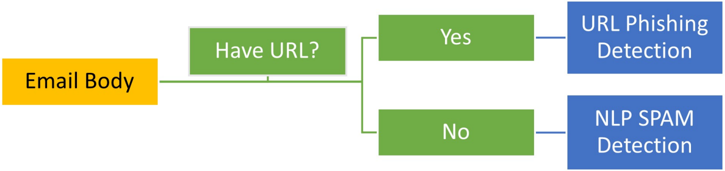


Figure: Emails contain urls in Body

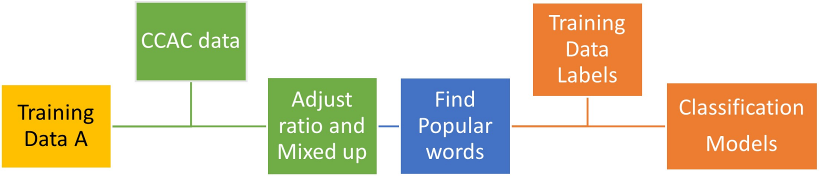
What is Phishing-type Email

- Total 4898 emails
- 69% emails contain urls, 21% do not
- We turn out to classify email into 2 approaches:
  - If the email contains NO URLs, go to SPAM model
  - If the email contains URLs, go to URL Phishing Model



SPAM Model

- For emails contain NO URLs,
  we use NLP spam model to determine if it is phishing email
- **Make train and test data have similar distribution after tokenization**
- voting results from RNN, LSTM, GRU models

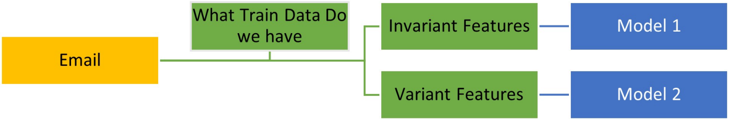


Phishing URL Model

- For emails contain URLs,
  we use rule-base method to determine if it is phishing email
- Phishing Websites Features suggested by Mohammad et al. (2015)
- Consider following email address features:
  - URL contents:
    - having IP
    - URL length
    - shorten URL address
    - number of subdomain
    - having , ’-’, double-slash
  - Registration information
    - is redirected
    - registered date
    - active status

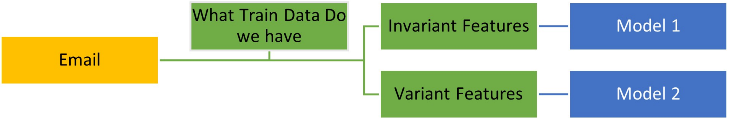
Results

- We have 0.64 in final accuracy
- Not a perfect score,
  but we CAN apply this STRATEGY to future data



Results

- Can this method be generalized to other phishing-type problems?
- Sure it can!
- We need to define the invariant and variant features between train data and the new emails



Thank You