Comparative Analysis of Phishing Email Classifiers

B565 Data Mining

Ashwin Venkatakrishnan, Prateek Giridhar and Prinston Rebello

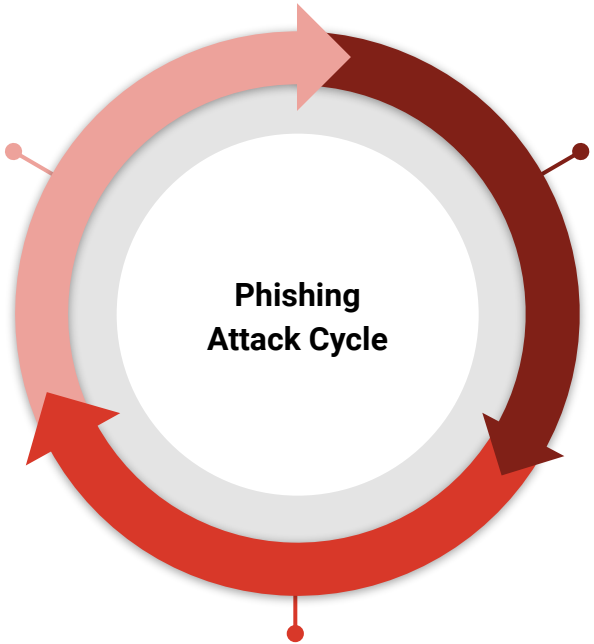
Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Introduction



Reel in Catch

- **70%** malicious links and files not blocked by network protections

- **15%** malicious links and files not blocked by endpoint protections

- **84%** employees took the bait within 10 minutes of receiving malicious email

- **13%** targeted employees reported the phishing attempt

Choose the Bait

- Financial alerts and updates

- Organization-wide announcements and updates

- User-specific alerts

- Promotions, deals and once-in-a-lifetime offers

Set the Hook

*data from CISA's Phishing Infographic: https://www.cisa.gov/sites/default/files/2023-02/phishing-infographic-508c.pdf

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Dataset

Phishing Email Detection is a labeled dataset that consists of:

- 18650 emails - 11322 safe, 7328 phishing
- Complete email body - suitable for text processing and NLP

	Unnamed: 0	Email Text	Email Type
11941	11942	\r\nÂ \r\nDear \r\n Homeowner,Â ...	Phishing Email
6325	6325	want to make more money ? plx order confirmati...	Phishing Email
2448	2448	ReliaQuote - Save Up To 70% On Life Insurance1...	Phishing Email
14849	14850	re : f / u to dr . kaminski @ enron from iris ...	Safe Email
12040	12041	the momentum investor burke , tiger team techn...	Phishing Email

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Exploratory Data Analysis

Dataset Features:
- Features include Index, Email Text, and Email Type.

Data Cleaning:
- Removed 16 records with NaN values in 'Email Text.'
- Eliminated 533 rows with 'empty' values.
- Resulted in 18,100 records remaining.

Text Processing:
- Applied text processing techniques.
- Removed stop words for improved data quality.

Decided to proceed with Random Forest, SVC & XGBoost for classification.

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Approach 1: Direct Classification

- Pre-process
  - Remove “empty” records
  - Undersample
- Split into train and test set
- Vectorize and classify using:
  - Random Forest Classifier
  - SVC
  - XGB Classifier

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Approach 2: Classification using Similarity Feature

- **Preprocessing:**
  - Data Cleansing
  - TFIDF
  - Cosine Similarity
- **Extracting Feature and exporting Modified Dataset**

Email Text	Email Type	Email Text New	Phishing Similarity

- **Test and Train Split** - 33%, 67%
- **Models:**
  - Random Forest Classifier
  - SVC
  - XGB Classifier

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Approach 3: Classification using NLP Techniques

- Feature Extraction
  - Extract URLs from the Email Text and check if it's legit using Google's Safe Browsing API.
  - Checked for lexical errors in the Email Text using Language Tool API. Calculated lexical error scores based on total no. of words in the Email body and appended it to every record.
  - Performed sentimental and emotion analysis on the text and extracted the scores for each Email body.
  - Resultant DataFrame after extraction:

ıdex	Email Text	Email Type	Lexical Errors	URL Verification	Email Text New	scores	compound	Sentiment	TB_score	TB_sentiment	Emotion
0	re : 6 . 1100 , disc : uniformitarianism , re ...	Safe Email	25.65217391304348	NA	disc uniformitarianism sex lang dick hud...	{'neg': 0.031, 'neu': 0.812, 'pos': 0.157, 'co...	0.9795	Positive	(0.17013888888888887, 0.506712962962963)	0.170139	positive
1	the other side of * galicismos * * galicismo *...	Safe Email	28.57142857142857	NA	the other side of galicismos galicismo is ...	{'neg': 0.0, 'neu': 0.966, 'pos': 0.034, 'comp...	0.3612	Positive	(0.009375000000000001, 0.084375)	0.009375	negative

Approach 3: Classification using NLP Techniques

- Preprocessing for Classification :
  - Encoded all binary attributes.
  - Vectorization of Email Text attribute
  - Dropped irrelevant attributes
  - Resultant DataFrame:

Email Text	Lexical Errors	URL Verification	compound	Sentiment	TB_sentiment	Emotion	Email Type

- Test & Train Split: 33%, 67%
- Models :
  - Random Forest Classifier
  - SVC
  - XGBoost
  - ANN

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Approach 4: Classification using NLP and Similarity Features

- Preprocessing:
  - Imported the Rich Processed dataset used in Approach 3
  - Appended the similarity scores calculated in Approach 2 to the imported DataFrame
- Extracting Feature and exporting Modified Dataset

Email Text	Email Type	Lexical Errors	URL Verification	scores	compound	Sentiment	TB_score	TB_sentiment	Emotion	Email Text New	Phishing Similarity

- Test and Train Split - 33%, 67%
- Models:
  - Random Forest Classifier
  - SVC
  - XGB Classifier

Agenda

- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Observations

Accuracy Analysis Table

Approach	Random Forest	SVM	XGBoost
Direct	94.5%	49.9%	94.8%
Similarity	94.3%	63.3%	93.14%
NLP	94.58%	95.77%	95.60%
NLP + Similarity	95.46%	93.8%	94%

Agenda

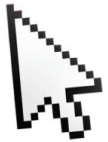
- Introduction to Phishing
- Dataset Overview
- EDA
- Our Work
  - Approach 1: Direct Classification
  - Approach 2: Classification using Similarity as a Feature
  - Approach 3: Classification using NLP Techniques
  - Approach 4: Classification using NLP and Similarity Features
- Observations
- Current Progress

Current Progress

- EDA
- Preprocessing and Feature Extraction
- Classifiers - Training and Testing
- Improve accuracy on ANN
- Test Models against Real Time Email Data

Stop! Don't click that link!

www.phishinglink.com




Thank
You