



Data Security: Machine Learning and Regex Matching Based Phishing Detection System Development for a System That Performs SMTP Phishing as an Attack

Umutcan Sevdı - 19011091

İsmet Güngör - 19011100

Semih Yazıcı - 19011087

Oğuzhan Ercan - 18011054

Yıldız Technical University Computer
Engineering Department

Information Systems Security Term Project

Should We Trust our eMails?

Common Email Spam File Types



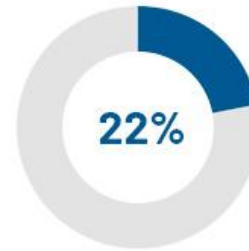
94% of malware is delivered via email



Office doc files

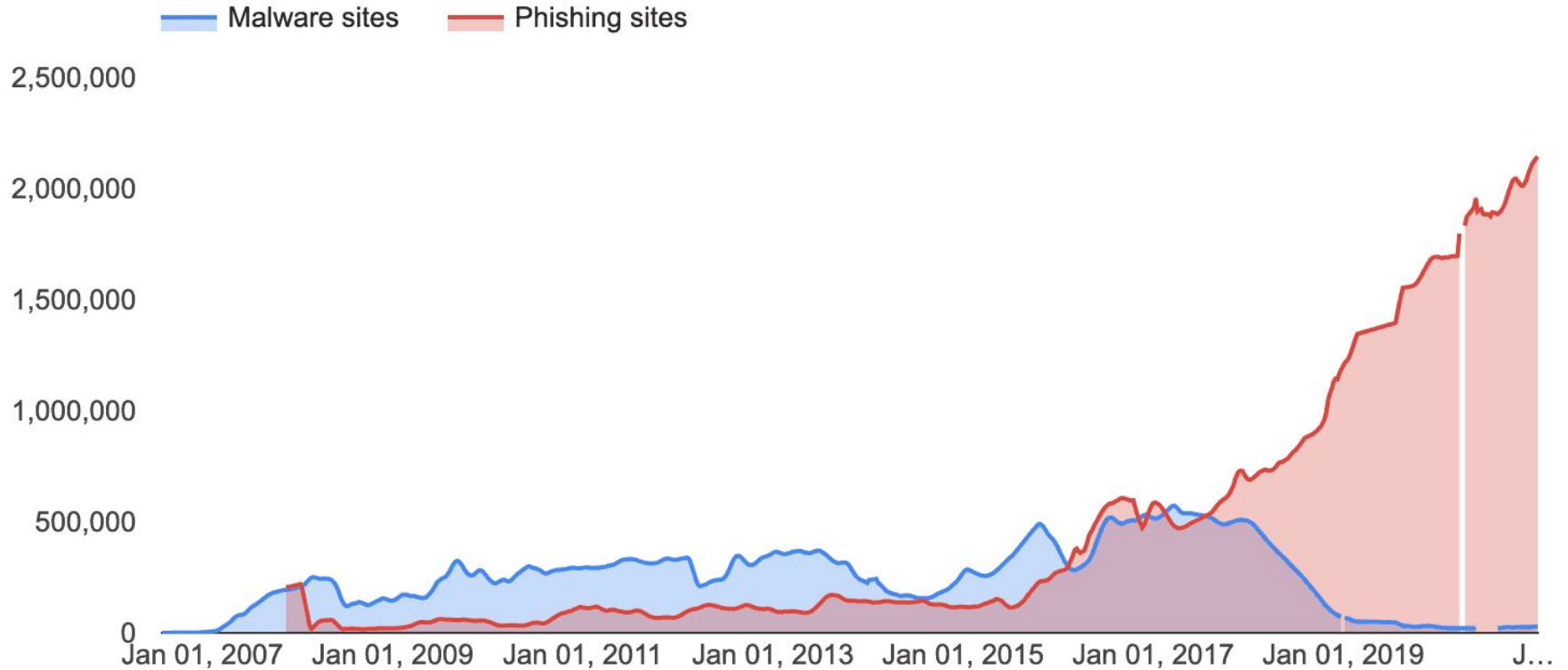


Windows apps



Other

Phishing is More Dangerous than the Malware



What is eMail Phishing

YTU < > 192.168.185.212 e-Devle...

e-Devlet Kapısı Kimlik Doğrulama Sistemi

Giriş Yapılacak Adres www.turkiye.gov.tr
Giriş Yapılacak Uygulama e-Devlet Kapısı

[e-Devlet Şifresi](#) [Mobil İmza](#) [e-İmza](#) [T.C. Kimlik Kartı](#) [İnternet Bankacılığı](#)

T.C. Kimlik Numaranızı ve e-Devlet Şifrenizi kullanarak kimliğiniz doğrulandıktan sonra işleminize kaldığınız yerden devam edebilirsiniz. [e-Devlet Şifresi Nedir, Nasıl Alınır?](#)

* T.C. Kimlik No

[Sanal Klavye](#) [Yazarken Gizle](#)

* e-Devlet Şifresi

[Sanal Klavye](#) [Şifremi Unuttum](#)


* e-Devlet [Şifrenizi unutmanız durumunda doğruladığınız cep telefonunuzdan yenileme işlemi yapabilirsiniz.](#)

[< İptal Et](#) [Giriş Yap >](#)

© 2022, Ankara - Tüm Hakları Saklıdır [Gizlilik ve Güvenlik](#) [Hızlı Çözüm Merkezi](#)

Bir menü görüntüle

Email Phishing in 2022



22 percent of data breaches are a result of phishing.^[1]



3.4 billion scam or phishing emails are sent each day.^[3]



Microsoft is the most impersonated brand in phishing attacks.^[4]



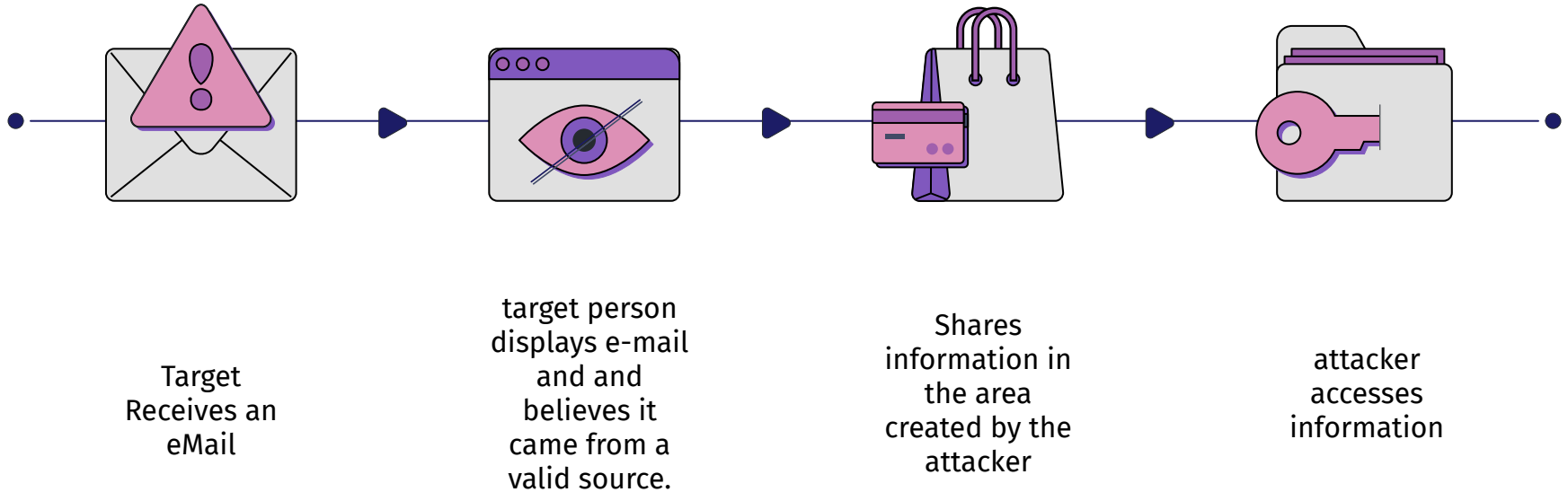
Phishing is the most common type of cybercrime.^[2]



2021 was the most expensive year for data breaches in 17 years.^[3]



How Mail Phishing Attack Occurs



Components of Our Program

Phishing Web Server

Mercury is the smallest planet



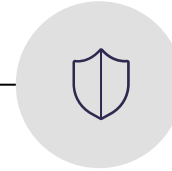
Mail Sender

Venus has a beautiful name



Traffic Analyser

Regex Based SMTP Analyser



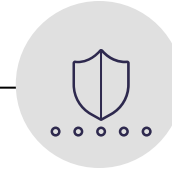
Mail Hog

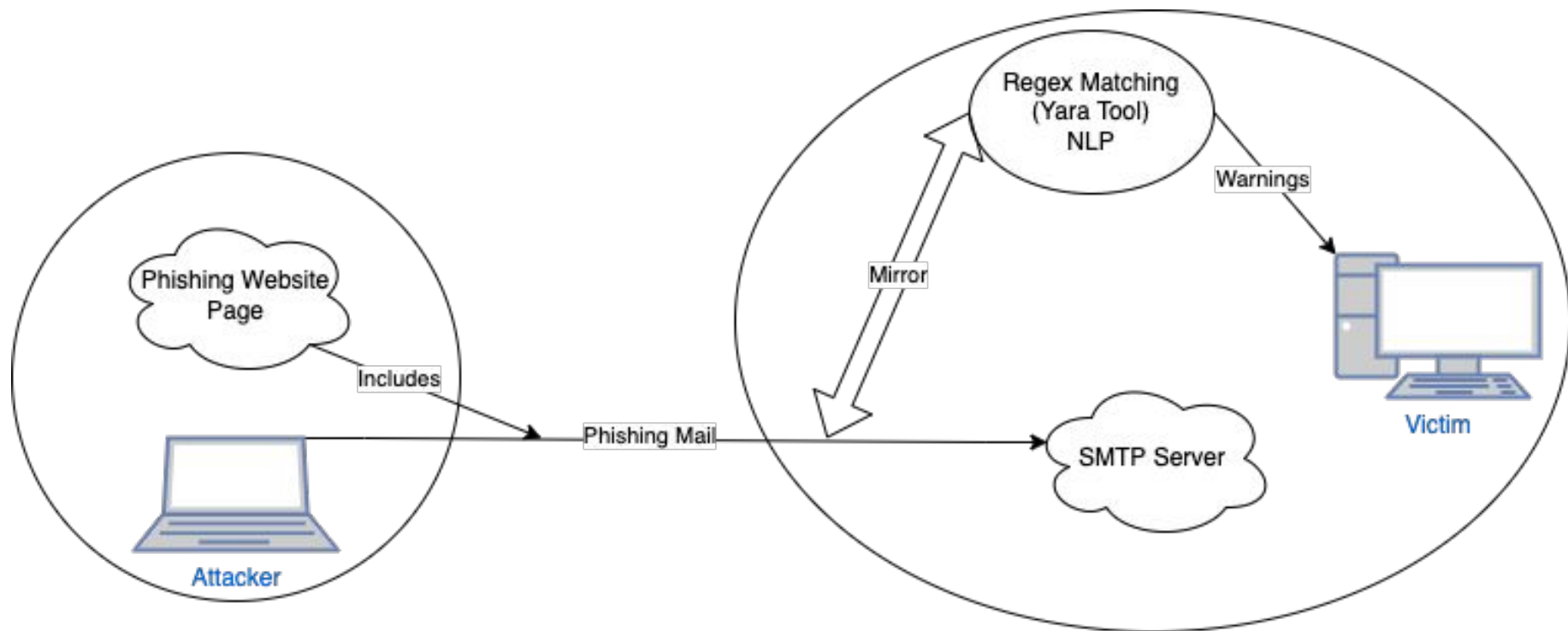
SMTP Server



NLP Analyser

NLP Based SMTP Analyser





Cybersecurity Infographics



01

Phishing Server

A web server written in Go which imitates edevlet.gov.tr while stealing critical information



02

SMTP Traffic Analyzer

A Go program that is mirroring traffic that is transferring through SMTP Server.



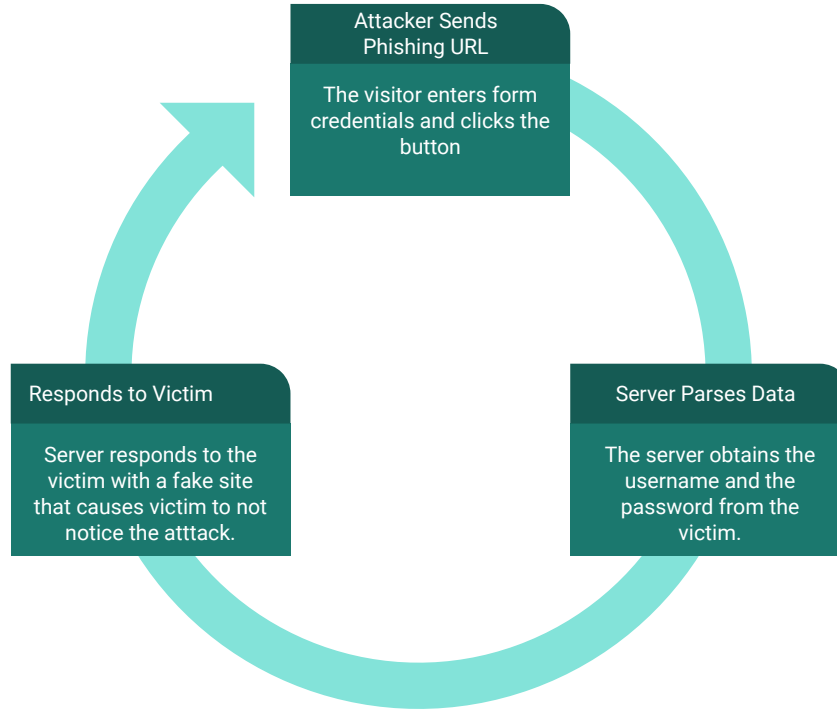
03

NLP Based Phishing Detection

A Deep Learning Model Consist on Long Short Term Memory (RNN)

Phishing Mail Server

We developed a Phishing Web Server in Go programming language. This web server sends web pages that imitates the edevlet.gov.tr. However unlike the original web page, this version does not encrypt the form messages. Also the target URL is replaced with the phishing site. When the victim sends the form data, it will be parsed by the program and the username and the password values are revealed. Phishing site responds to the form data with a dashboard that looks like the victim has entered to the website. This server URL will be used in the phishing mail.



SMTP Traffic Analyzer



Go Program

System runs on a program written in Go that actively listens SMTP server's port and read active traffic..



Pcap & Gopacket

To listen ongoing traffic Pcap library used. Packet source generated in live-mod and Reached packet details with gopacket.



Plaintext Transfer to Python Script

Parsed mail body transferred through TCP to the Python script which is containing NLP model.

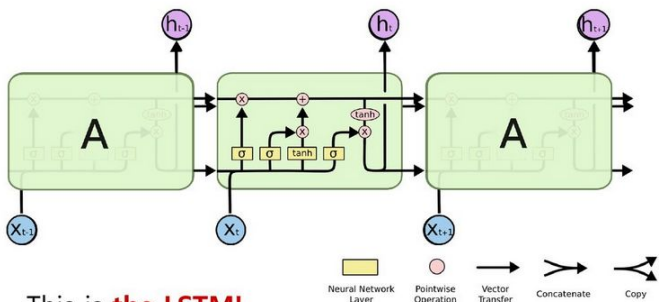


Yara Tool

Yara rules generated for regex matching to detect possible phishing keywords.

NLP Based Phishing Detection

Long Short Term Memory



This is **the LSTM!**

NLP based phishing detection system consists of 3 different components.

1

Text Preprocessor: used for cleaning and preprocessing text by removing punctuation and lowercasing, removing stop words, and optionally applying stemming or lemmatization.

2

Language Model: Long Short Term Memory, a type of recurrent neural network (Kind of RNN) well-suited for modeling natural language.

3

Attention and Classifier Model: The attention weighs the input sequences, and the classifier predicts based on the weighted input whether phishing or innocent.