PHISHING DETECTION SYSTEM

GROUP MEMBERS:
Kevin John
Neeraj Sachi
Mohammed Midhlaj V
Rumaisa Shajahan

GUIDED BY:
Ms. Asha Raj

CONTENTS

1. Introduction
2. Problem Statement
3. Abstract
4. Current System
5. Proposed System
6. Objectives
7. Datasets
8. Architecture
9. Modules
10. Methodology
11. Conclusion
12. Reference

INTRODUCTION

Phishing is a cybercrime tactic used to deceive individuals and organizations into revealing sensitive information. This presentation explores the use of machine learning for phishing detection.

PROBLEM STATEMENT

To develop an extension for detecting phishing websites and emails that can accurately identify and flag potentially malicious websites and emails in real-time.

ABSTRACT

Phishing attacks continue to be a prevalent threat to individuals and organizations, compromising sensitive information and causing financial and reputational damage. To address this growing concern, we propose an innovative browser extension and email filtering system. The extension seamlessly integrates with web browsers, offering a robust defense mechanism against phishing attempts. Users are alerted and prevented from accessing deceptive websites, reducing their vulnerability to phishing attacks. Detecting these malicious sites requires an approach that combines machine learning, heuristics, and data analysis. Key techniques involve analyzing website content, URL characteristics, and user behavior to identify anomalies and patterns associated with phishing attempts. Machine learning models, such as random forest and decision trees, play a crucial role in automating the detection process. In addition to web protection, our solution extends to email security. The integrated email filtering system scans inbound messages, using advanced content analysis and URL inspection to identify phishing attempts. Suspicious emails are flagged and quarantined, ensuring that users are shielded from fraudulent links and potentially harmful attachments.

CURRENT SYSTEM

- **Blacklist-Based Detection:**
  - Maintains a list of known phishing URLs or domains.
  - Ineffective against zero-day phishing attacks and frequently changing domains.
- **High False Positive Rates:**
  - Current systems often generate false positives, causing inconvenience to legitimate users.
- **Dependency on User Reports:**
  - Often relies on user-generated reports of suspected phishing sites, leading to delayed responses.

PROPOSED SYSTEM

- Utilizes machine learning algorithms like XGBoost and Random Forest .
- Integrates real-time analysis of user behavior and website content.
- Incorporates deep URL analysis.

OBJECTIVES

The primary objective of the extension is to protect individuals and organizations from the significant financial, reputational, and security risks associated with phishing attacks by identifying and blocking fraudulent websites and educating users about the threat.

- Protection Against Fraud
- Data Security
- Prevent unauthorised access
- Reduced Cybersecurity risks
- Early Threat Detection

DATASET

Source:
https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset/data

DATASET

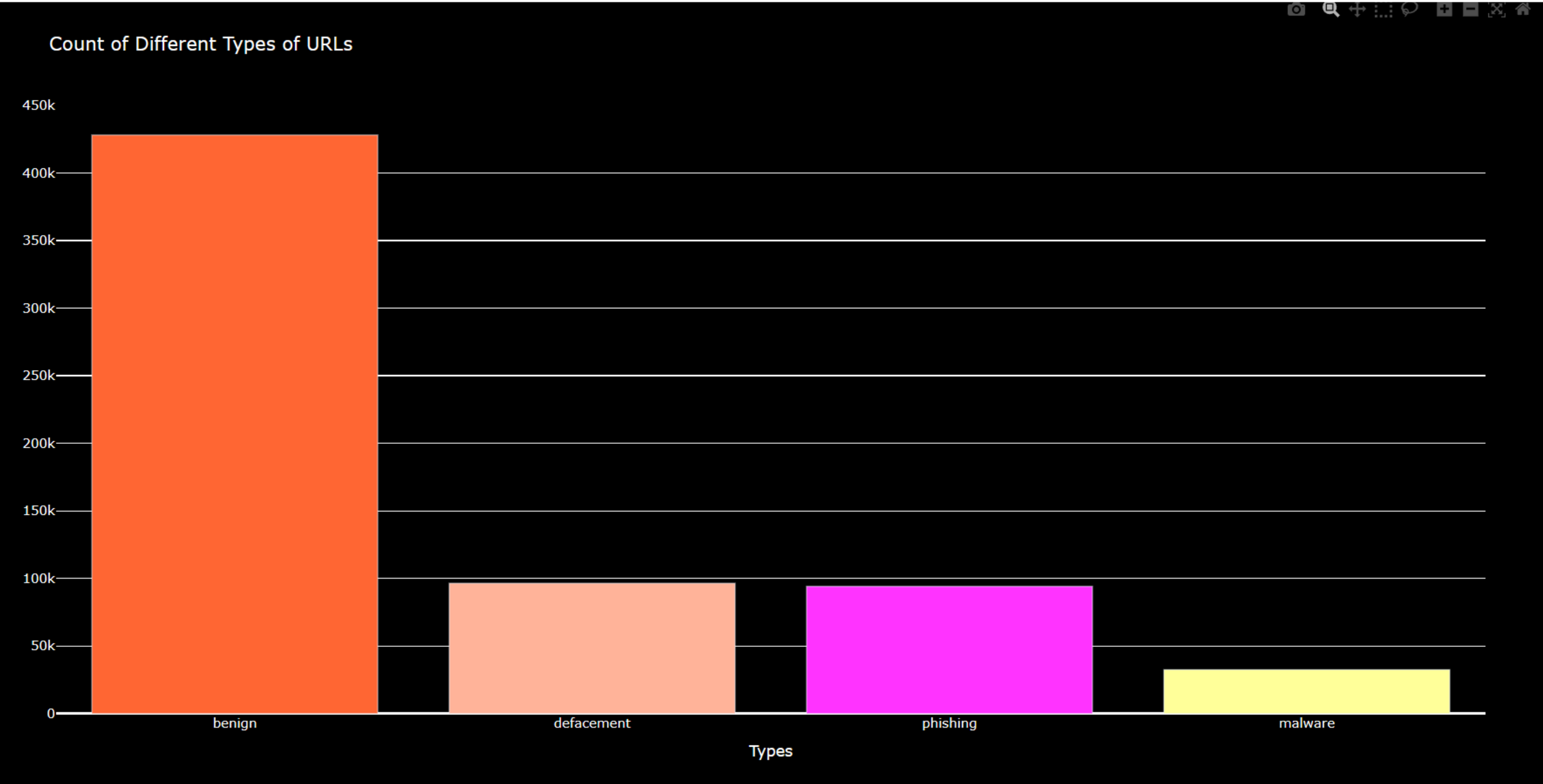
Raw Dataset

- br-icloud.com.br,phishing
- mp3raid.com/music/krizz_kaliko.html,benign
- bopsecrets.org/rexroth/cr/1.htm,benign
- http://www.garage-pirenne.be/index.php?option=com_content&view=article&id=70&vsig70_0=15,defacement
- myspace.com/video/vid/30602581,benign
- http://www.lebensmittel-ueberwachung.de/index.php/aktuelles.1,defacement
- http://www.szabadmunkaero.hu/cimoldal.html?start=12,defacement
- http://buzzfil.net/m/show-art/ils-etaient-loin-de-s-imaginer-que-le-hibou-allait-faire-ceci-quand-ils-filmaient-2.html,benign
- espn.go.com/nba/player/_/id/3457/brandon-rush,benign
- http://www.vnic.co/khach-hang.html,defacement

DATASET

Raw Dataset

- http://www.824555.com/app/member/SportOption.php?uid=guest&langx=gb,malware
- http://www.raci.it/component/user/reset.html,defacement
- https://docs.google.com/spreadsheet/viewform?
- formkey=dGg2Z1lCUHlSdjllTVNRUW50TFIzSkE6MQ,phishing
- psychology.wikia.com/wiki/Phonemes,benign
- info.centriq.com/content/fivereasons,benign
- lowery,benign infinitysw.com/,benign strawberrycreekgardens.com/,benign
- peopleaz.org/firstname/Fowlkes/6,benign
- orensamtraktrip.wordpress.com/,benign
- retajconsultancy.com,phishing
- alexa.com/whatshot?q=james+arness,benign


Count of Different Types of URLs
450k
400k
350k
300k
250k
200k
150k
100k
50k
0
benign
defacement
phishing
malware
Types

DATASET(BEFORE PREPROCESSING)

```
url,type
br-icloud.com.br,phishing
mp3raid.com/music/krizz_kaliko.html,benign
bopsecrets.org/rexroth/cr/1.htm,benign
http://www.garage-pirenne.be/index.php?option=com_content&view=article&id=70&vsig70_0=15,defacement
http://adventure-nicaragua.net/index.php?option=com_mailto&tmpl=component&link=aHR0cDovL2FkdmVudHVyZS1uaWNhcmFndWEubmV0L2luZGV4LnBocD9
http://buzzfil.net/m/show-art/ils-etaient-loin-de-s-imaginer-que-le-hibou-allait-faire-ceci-quand-ils-filmaient-2.html,benign
espn.go.com/nba/player/_/id/3457/brandon-rush,benign
yourbittorrent.com/?q=anthony-hamilton-soulife,benign
http://www.pashminaonline.com/pure-pashminas,defacement
allmusic.com/album/crazy-from-the-heat-r16990,benign
corporationwiki.com/Ohio/Columbus/frank-s-benson-P3333917.aspx,benign
http://www.ikenmijnkunst.nl/index.php/exposities/exposities-2006,defacement
myspace.com/video/vid/30602581,benign
http://www.lebensmittel-ueberwachung.de/index.php/aktuelles.1,defacement
http://www.szabadmunkaero.hu/cimoldal.html?start=12,defacement
http://larcadelcarnevale.com/catalogo/palloncini,defacement
quickfacts.census.gov/qfd/maps/iowa_map.html,benign
nugget.ca/ArticleDisplay.aspx?archive=true&e=1160966,benign
uk.linkedin.com/pub/steve-rubenstein/8/718/755,benign
http://www.vnic.co/khach-hang.html,defacement
baseball-reference.com/players/h/harrige01.shtml,benign
signin.eby.de.zukruygxctzmmqi.civpro.co.za,phishing
192.com/atoz/people/oakley/patrick/,benign
nytimes.com/1998/03/29/style/cuttings-oh-that-brazen-raucous-glorious-hibiscus.html,benign
```

PREPROCESSING METHODOLOGY

Loading the Dataset:

- The dataset is loaded using the **pd.read_csv** function from the Pandas library

Handling Missing Values:

- The presence of missing values is checked using **urls_data.isnull().sum()**
- Missing values are replaced with 0 using **urls_data.fillna(0, inplace=True)**

URL Cleaning:

- The 'www.' prefix is removed from the 'url' column using **urls_data['url'] = urls_data['url'].replace('www.', '', regex=True).**
-

Mapping Types of Websites:

- The 'type' column is mapped to numerical values using **urls_data["url_type"] = urls_data["type"].replace(...)**

DATASET(AFTER PREPROCESSING)

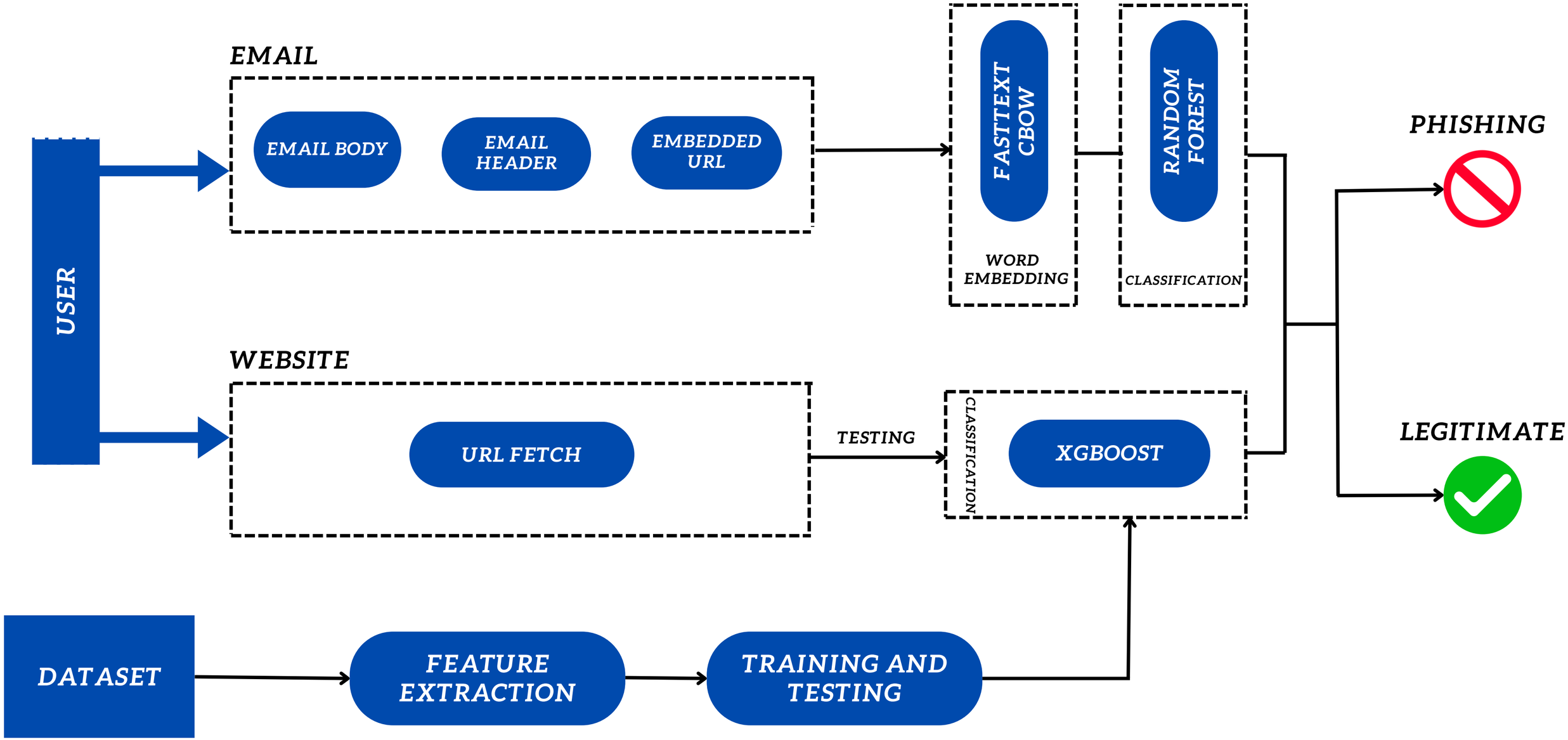
```
url,type,url_type,url_len
br-icloud.com.br,phishing,2,16
mp3raid.com/music/krizz_kaliko.html,benign,0,35
bopsecrets.org/rexroth/cr/1.htm,benign,0,31
http://garage-pirenne.be/index.php?option=com_content&view=article&id=70&vsig70_0=15,defacement,1,77
http://adventure-nicaragua.net/index.php?option=com_mailto&tmpl=component&link=aHR0cDovL2FkdmVudHVyZS1uaWNhcmFndWEubmV0L2luZGV4LnBocD9
http://buzzfil.net/m/show-art/ils-etaient-loin-de-s-imaginer-que-le-hibou-allait-faire-ceci-quand-ils-filmaient-2.html,benign,0,111
espn.go.com/nba/player/_/id/3457/brandon-rush,benign,0,45
yourbittorrent.com/?q=anthony-hamilton-soulife,benign,0,46
http://pashminaonline.com/pure-pashminas,defacement,1,33
allmusic.com/album/crazy-from-the-heat-r16990,benign,0,45
corporationwiki.com/Ohio/Columbus/frank-s-benson-P3333917.aspx,benign,0,62
http://ikenmijnkunst.nl/index.php/exposities/exposities-2006,defacement,1,53
myspace.com/video/vid/30602581,benign,0,30
http://lebensmittel-ueberwachung.de/index.php/aktuelles.1,defacement,1,50
http://szabadmunkaero.hu/cimoldal.html?start=12,defacement,1,40
http://larcadelcarnevale.com/catalogo/palloncini,defacement,1,41
quickfacts.census.gov/qfd/maps/iowa_map.html,benign,0,44
nugget.ca/ArticleDisplay.aspx?archive=true&e=1160966,benign,0,52
uk.linkedin.com/pub/steve-rubenstein/8/718/755,benign,0,46
http://vnic.co/khach-hang.html,defacement,1,23
baseball-reference.com/players/h/harrige01.shtml,benign,0,48
signin.eby.de.zukruygxctzmmqi.civpro.co.za,phishing,2,42
192.com/atoz/people/oakley/patrick/,benign,0,35
nytimes.com/1998/03/29/style/cuttings-oh-that-brazen-raucous-glorious-hibiscus.html,benign,0,83
```

DATASET(AFTER FEATURE EXTRACTION)

url	type	url_type	url_len	pri_domain	letters_count	digits_count	special_chars_	shortened	abnormal_url	secure_http	have_ip
br-icloud.com.br	phishing	2	16	br-icloud.com.br	13	0	3	0	0	0	0
mp3raid.com/mu	benign	0	35	mp3raid.com	29	1	5	0	0	0	0
bopsecrets.org/r	benign	0	31	bopsecrets.org	25	1	5	0	0	0	0
http://garage-pire	defacement	1	77	garage-pirenne.b	60	7	17	0	1	0	0
http://adventure-	defacement	1	228	adventure-nicara	199	22	14	0	1	0	0
http://buzzfil.net/	benign	0	111	buzzfil.net	93	1	24	0	1	0	0
espn.go.com/nba	benign	0	45	espn.go.com	31	4	10	0	0	0	0
yourbittorrent.co	benign	0	46	yourbittorrent.co	40	0	6	0	0	0	0
http://pashminao	defacement	1	33	pashminaonline.	34	0	6	0	1	0	0
allmusic.com/alb	benign	0	45	allmusic.com	33	5	7	0	0	0	0
corporationwiki.c	benign	0	62	corporationwiki.c	47	7	8	0	0	0	0
http://ikenmijnku	defacement	1	53	ikenmijnkunst.nl	47	4	9	0	1	0	0
myspace.com/vi	benign	0	30	myspace.com	18	8	4	0	0	0	0
http://lebensmitte	defacement	1	50	lebensmittel-ueb	47	1	9	0	1	0	0
http://szabadmu	defacement	1	40	szabadmunkaer	37	2	8	0	1	0	0

DATASET(AFTER FEATURE EXTRACTION)

http://szabadmur	defacement	1	40	szabadmunkaerc	37	2	8	0	1	0	0
http://larcadelcar	defacement	1	41	larcadelcarneval	42	0	6	0	1	0	0
quickfacts.censu	benign	0	44	quickfacts.censu	37	0	7	0	0	0	0
nugget.ca/Article	benign	0	52	nugget.ca	38	7	7	0	0	0	0
uk.linkedin.com/	benign	0	46	uk.linkedin.com	31	7	8	0	0	0	0
http://vnic.co/kha	defacement	1	23	vnic.co	23	0	7	0	1	0	0
baseball-referen	benign	0	48	baseball-referen	40	2	6	0	0	0	0
signin.eby.de.zul	phishing	2	42	signin.eby.de.zul	36	0	6	0	0	0	0
192.com/atoz/pe	benign	0	35	192.com	26	3	6	0	0	0	0
nytimes.com/199	benign	0	83	nytimes.com	62	8	13	0	0	0	0
escholarship.org	benign	0	33	escholarship.org	24	5	4	0	0	0	0
songfacts.com/d	benign	0	33	songfacts.com	23	5	5	0	0	0	0
casamanana.org	benign	0	30	casamanana.org	26	0	4	0	0	0	0
http://hollywoodli	benign	0	78	hollywoodlife.cor	58	12	15	0	1	0	0
http://marketingb	phishing	2	60	marketingbyinter	43	17	7	0	1	0	0
en.wikipedia.org/	benign	0	34	en.wikipedia.org	29	0	5	0	0	0	0
soaps.sheknows	benign	0	81	soaps.sheknows	62	5	14	0	0	0	0


EMAIL
EMAIL BODY
EMAIL HEADER
EMBEDDED URL
USER
WEBSITE
URL FETCH
TESTING
FASTTEXT CBOW
WORD EMBEDDING
RANDOM FOREST
CLASSIFICATION
CLASSIFICATION
XGBOOST
PHISHING
LEGITIMATE
DATASET
FEATURE EXTRACTION
TRAINING AND TESTING

MODULES

Module 1:

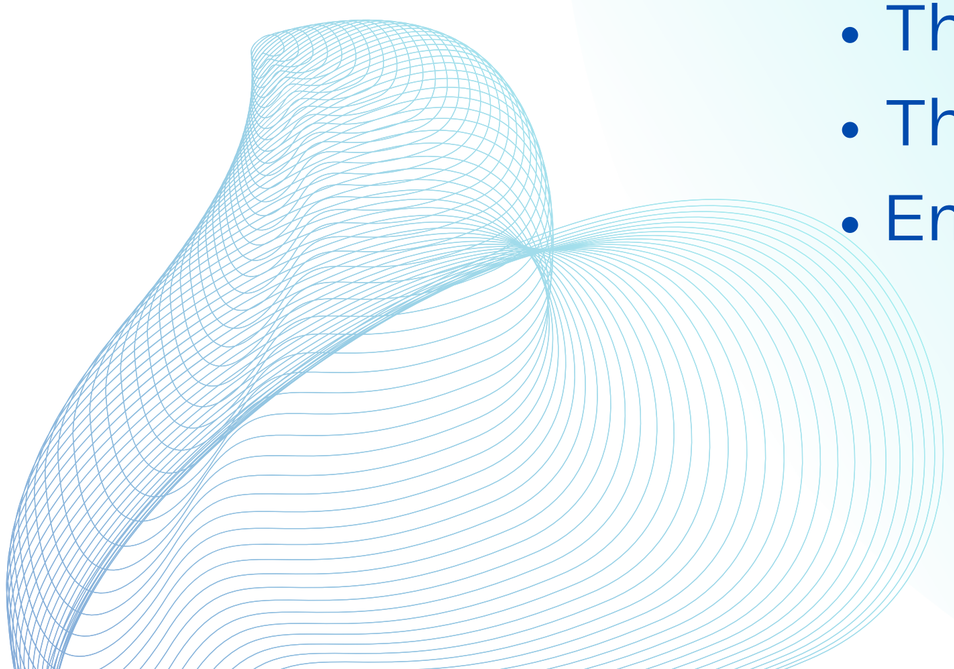
Feature Extraction:

–In Website Detection Extract

- url

–In Email Detection Extract

- The header of the email
- The body of the email
- Embedded urls



MODULES

Module 2:

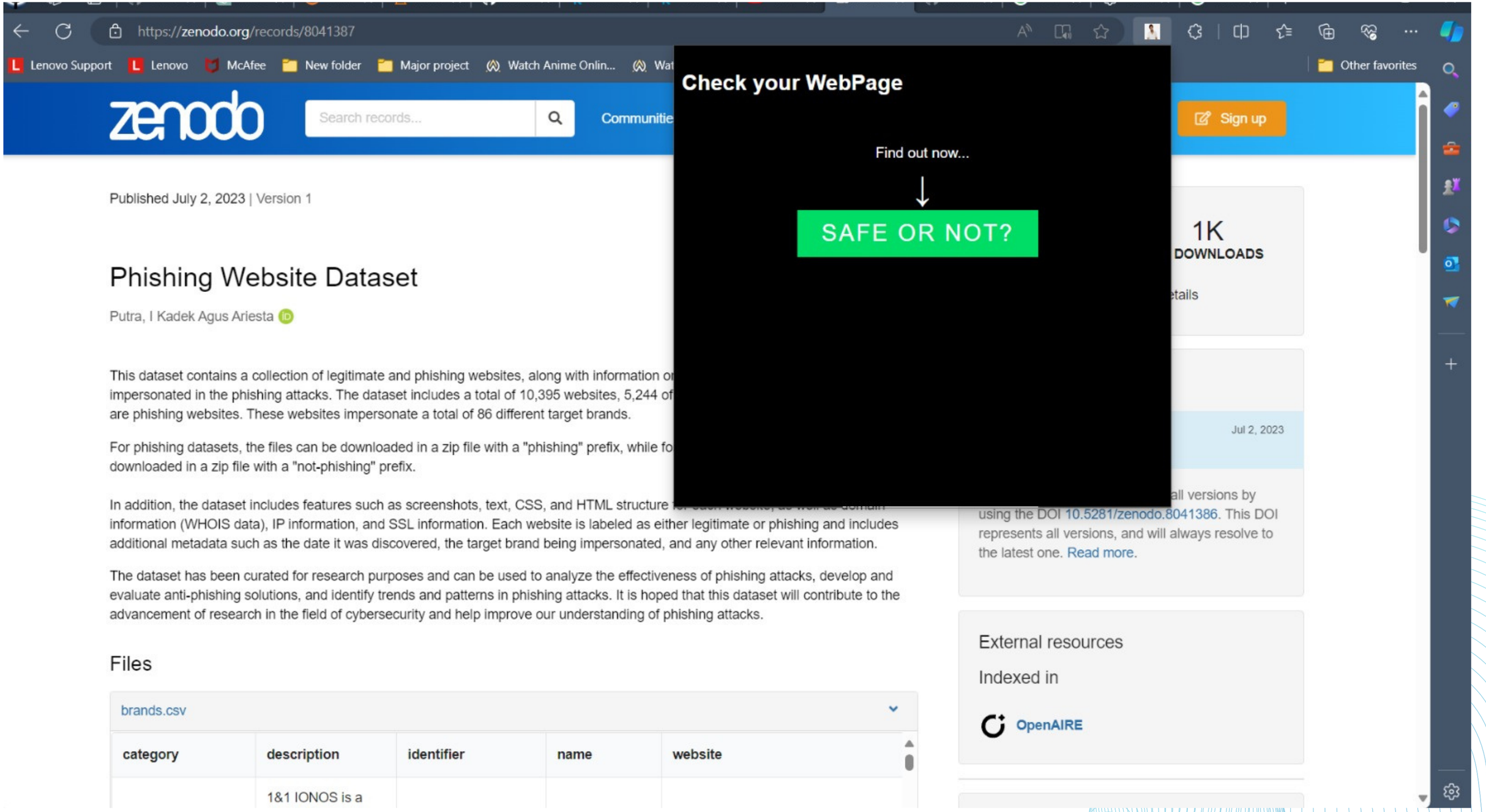
Word Embedding:

- Word embedding is done using FastText-CBOW

Module 3:

Prediction Analysis:

- Website and Email content- Random Forest Algorithm.
- Embedded URLs and website URL-URL Analysis


https://zenodo.org/records/8041387
Lenovo Support
Lenovo
McAfee
New folder
Major project
Watch Anime Onlin...
Other favorites
zenodo
Search records...
Communitie
Sign up
Check your WebPage
Find out now...
SAFE OR NOT?
Published July 2, 2023 | Version 1
Phishing Website Dataset
Putra, I Kadek Agus Ariesta
This dataset contains a collection of legitimate and phishing websites, along with information o
impersonated in the phishing attacks. The dataset includes a total of 10,395 websites, 5,244 of
are phishing websites. These websites impersonate a total of 86 different target brands.
For phishing datasets, the files can be downloaded in a zip file with a "phishing" prefix, while fo
downloaded in a zip file with a "not-phishing" prefix.
In addition, the dataset includes features such as screenshots, text, CSS, and HTML structure
information (WHOIS data), IP information, and SSL information. Each website is labeled as either legitimate or phishing and includes
additional metadata such as the date it was discovered, the target brand being impersonated, and any other relevant information.
The dataset has been curated for research purposes and can be used to analyze the effectiveness of phishing attacks, develop and
evaluate anti-phishing solutions, and identify trends and patterns in phishing attacks. It is hoped that this dataset will contribute to the
advancement of research in the field of cybersecurity and help improve our understanding of phishing attacks.
Files
brands.csv
category
description
identifier
name
website
1&1 IONOS is a
1K
DOWNLOADS
etails
Jul 2, 2023
all versions by
using the DOI 10.5281/zenodo.8041386. This DOI
represents all versions, and will always resolve to
the latest one. Read more.
External resources
Indexed in
OpenAIRE

METHODOLOGY

- **Data Collection and Preparation:** Gather a diverse dataset of both phishing and legitimate websites.
- **Feature Engineering:** Create meaningful features that can help distinguish between phishing and legitimate websites
- **Machine Learning Model Selection:** Choose appropriate machine learning algorithms or architectures and split the dataset into training, validation, and testing sets.
- **URL Analysis:** Implement URL analysis algorithms to identify phishing indicators in URLs.

CONCLUSION

In conclusion, the proposed Phishing Detection Extension integrates advanced technology, user education, and real-time threat intelligence to provide effective protection against phishing attacks. It prioritizes user awareness, privacy, and continuous improvement to mitigate risks and enhance cybersecurity in the digital age.

REFERENCE

[1]Gururaj Harinahalli Lokesh & Goutham BoreGowda (2020): Phishing website detection based on effective machine learning approach, Journal of Cyber Security Technology. Available: https://doi.org/10.1080/23742917.2020.1813396

[2]ABDUL KARIM, MOBEEN SHAHROZ, KHABIB MUSTOFA,SAMIR BRAHIM BELHAOUARI, AND S. RAMANA KUMAR JOGA,Phishing Detection System Through Hybrid Machine Learning Based on URL,Volume 11,2023, 10.1109/ACCESS.2023.3252366

[3]MARIA SAMEEN, KYUNGHYUN HAN , ANDSEONG OUN HWANG,PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System, VOLUME 8, 2020, 10.1109/ACCESS.2020.2991403

[4]MANUEL SÁNCHEZ-PANIAGUA , EDUARDO FIDALGO FERNÁNDEZ,ENRIQUE ALEGRE ,WESAM AL-NABKI , AND VÍCTOR GONZÁLEZ-CASTRO,Phishing URL Detection: A Real-Case Scenario Through Login URLs,VOLUME 10, 2022,10.1109/ACCESS.2022.3168681

REFERENCE

[5]Ehsan Nowroozi , Senior Member, IEEE, Abhishek , Member, IEEE,Mohammadreza Mohammadi , Member, IEEE, and Mauro Conti ,An Adversarial Attack Analysis on Malicious Advertisement URL Detection Framework,, VOL. 20, NO. 2, JUNE 2023, 10.1109/TNSM.2022.3225217

[6]E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa and C. G. Duque, "The Answer is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering," in IEEE Access, vol. 8, pp. 223529-223547, 2020, doi: 10.1109/ACCESS.2020.3043396.

[7]N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," in IEEE Access, vol. 10, pp. 36429-36463, 2022, doi: 10.1109/ACCESS.2022.3151903.

[8]K. Althobaiti, M. K. Wolters, N. Alsufyani and K. Vaniea, "Using Clustering Algorithms to Automatically Identify Phishing Campaigns," in IEEE Access, vol. 11, pp. 96502-96513, 2023, doi: 10.1109/ACCESS.2023.3310810.

REFERENCE

[9]Mittal, Apurv; Engels, Dr Daniel; Kommanapalli, Harsha; Sivaraman, Ravi; and Chowdhury, Taifur (2022) "Phishing Detection Using Natural Language Processing and Machine Learning," SMU Data Science Review: Vol. 6: No. 2, Article 14. https://scholar.smu.edu/datasciencereview/vol6/iss2/14

[10] Umer Ahmed, Rashid Amin, Hamza Aldabbas, Senthikumar Mohan, Bader Alouffi, Ali Ahmadian,"Cloud-based email phishing attack using machine and deep learning algorithm"Complex Intell. Syst. 9, 3043-3070 (2023). https://doi.org/10.1007/s40747-022-00760

[11] Somesh m, Alwyn R Pais,"Classification of Phishing Email Using Word Embedding and Machine Learning Techniques"2022 vol 11 issue 3 https://doi.org/10.13052/jcsm2245-1439.1131

[12] Yong Fang, Cheng Zhang, Cheng Huang, Liang Liu, Yue Yang,"Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism"IEEE May 9,2019.

THANK YOU