DETECTION OF PHISHING WEBSITES USING GENERATIVE ADVERSARIAL NETWORK

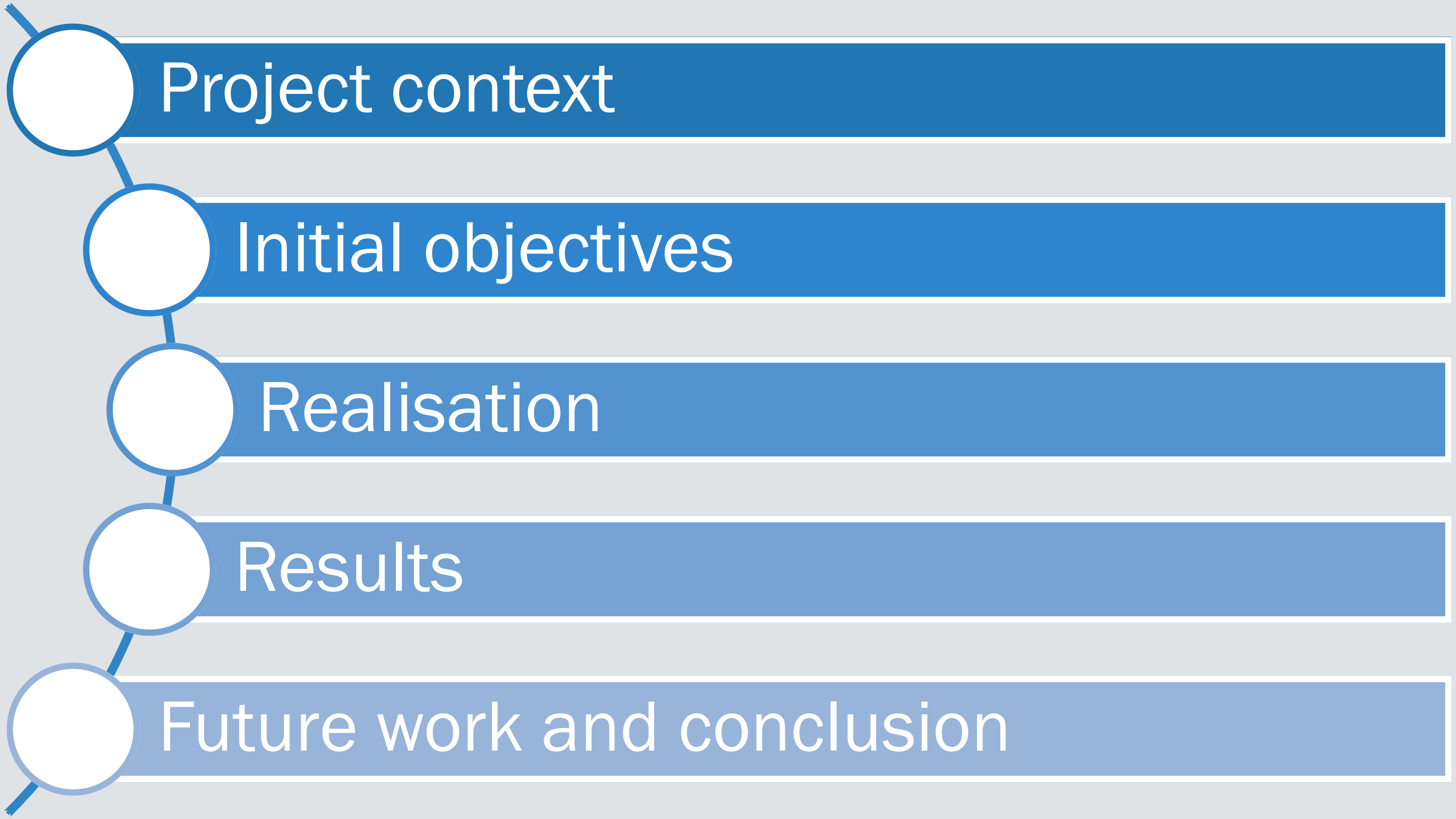
Pierrick Robic--Butez

Thu Yein Win

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES CENTRE VAL DE LOIRE

UNIVERSITY OF GLOUCESTERSHIRE at Cheltenham and Gloucester

Plan



Project context

Initial objectives

Realisation

Results

Future work and conclusion

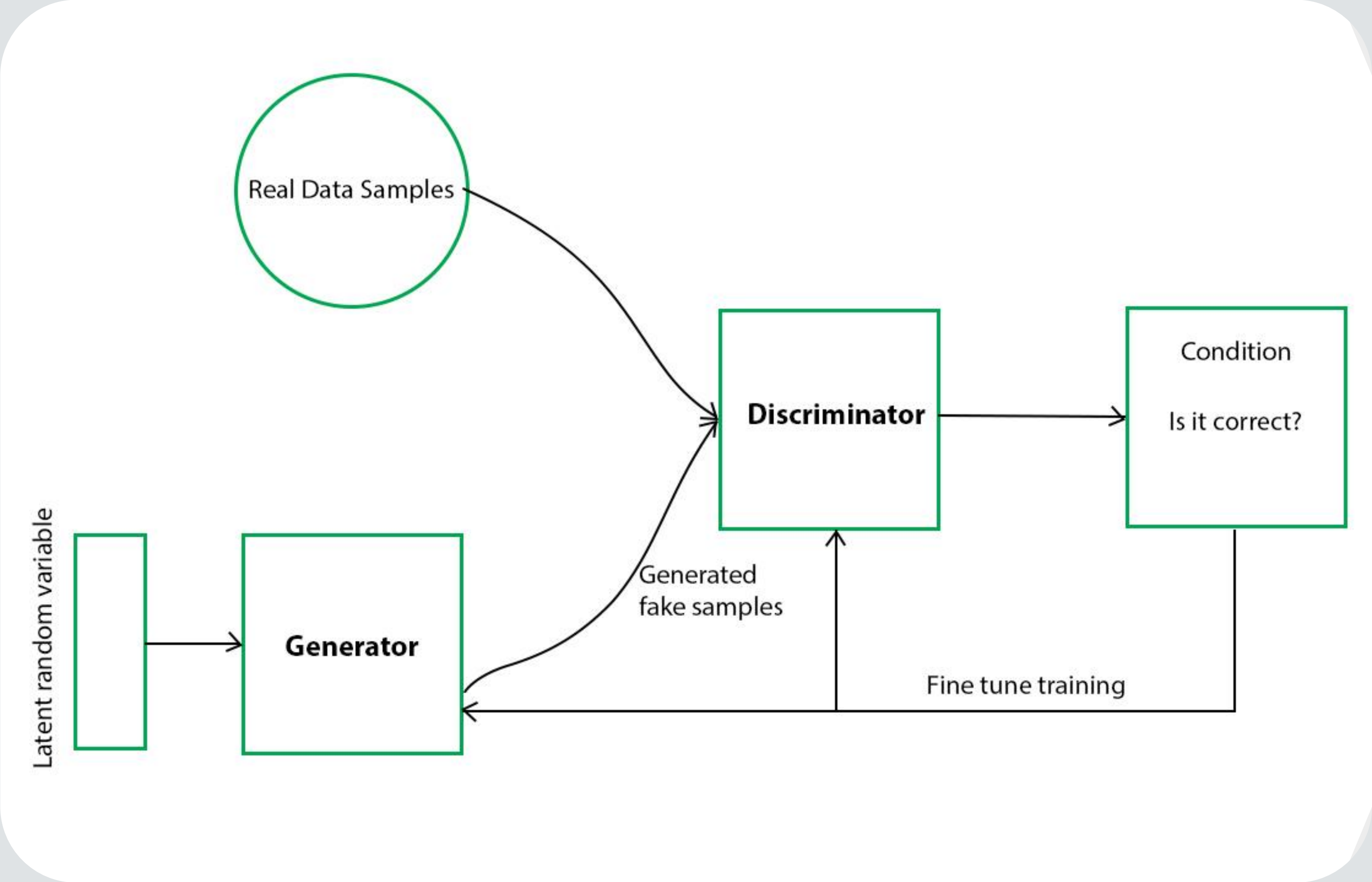
PROJECT CONTEXT

Phishing

- *“Phishing is the fraudulent attempt to obtain sensitive information by disguising oneself as a trustworthy entity in an electronic communication”*
- One of the most popular type of attack
  - *Easy*
  - *Not expensive*
- In 2018: ~800,000 unique phishing websites were detected

Generative Adversarial Network (GAN)

- Machine learning system invented by Ian Goodfellow et al in 2014
- Two neural networks contest each other in a game



Generative Adversarial Network (GAN)

- Often used in image processing
  - *Detection of aggressive Prostate Cancer* by Simon Kohl *et al.* in 2017
  - *Text to Photo-realistic Image Synthesis* by Han Zhang *et al.* in 2016

- But not often for cybersecurity

INITIAL OBJECTIVES

Initial objectives

Classic method

- Create a database of phishing websites
- Compare to this database to check if a website is phishing or not
- Based on report by users

Detection with GAN

- Train a GAN to detect phishing websites
- Real time method
- Can detect a newly created phishing website

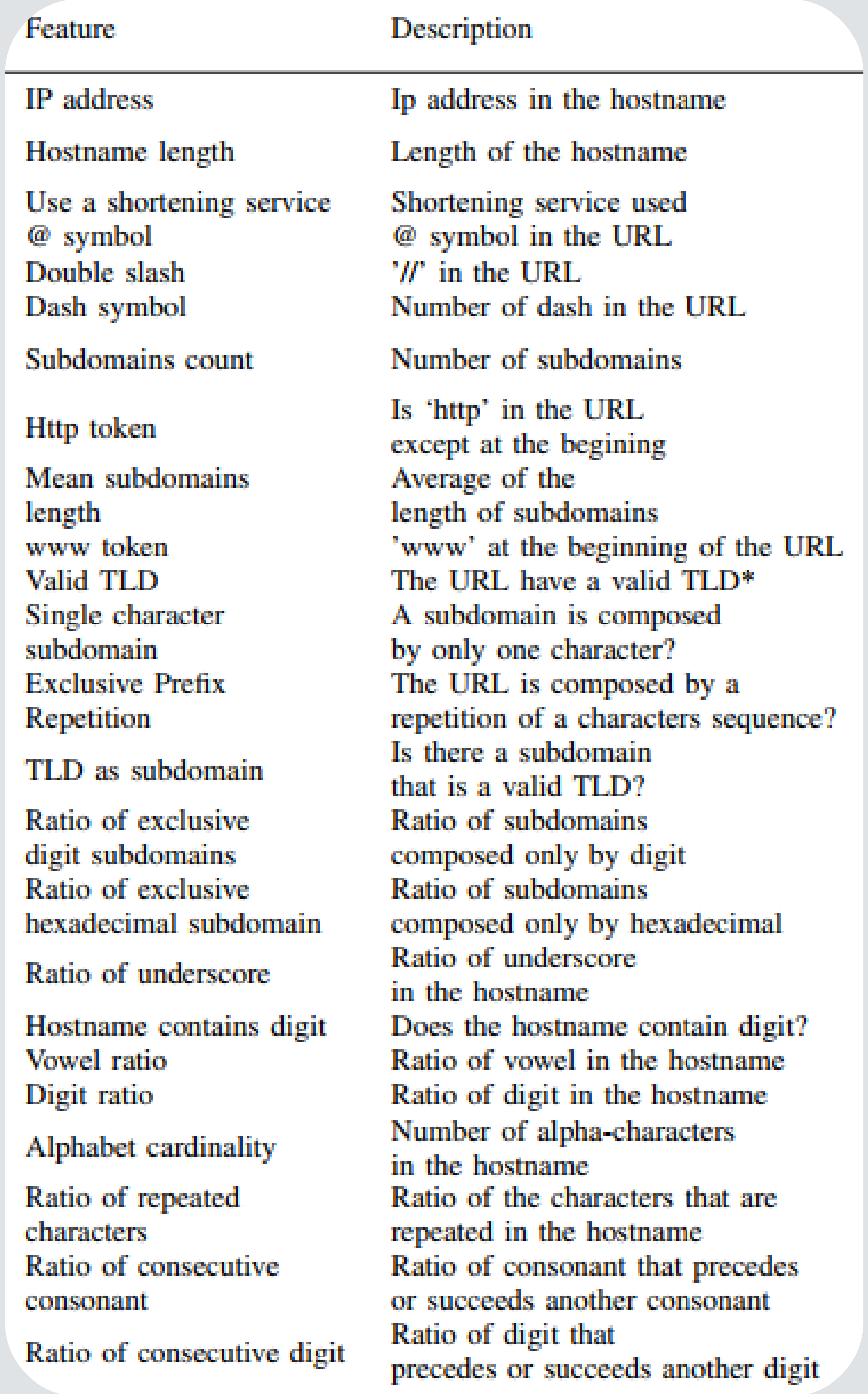
REALISATION

Features extraction

- Based on the work of R. M. Mohammad *et al.* and S. Schüppen *et al.*
- 46 features are extracted for each website
  - 3 categories: structural, based on source code, from third party sources
  - 2 types of value possible: discrete or continuous

Structural features

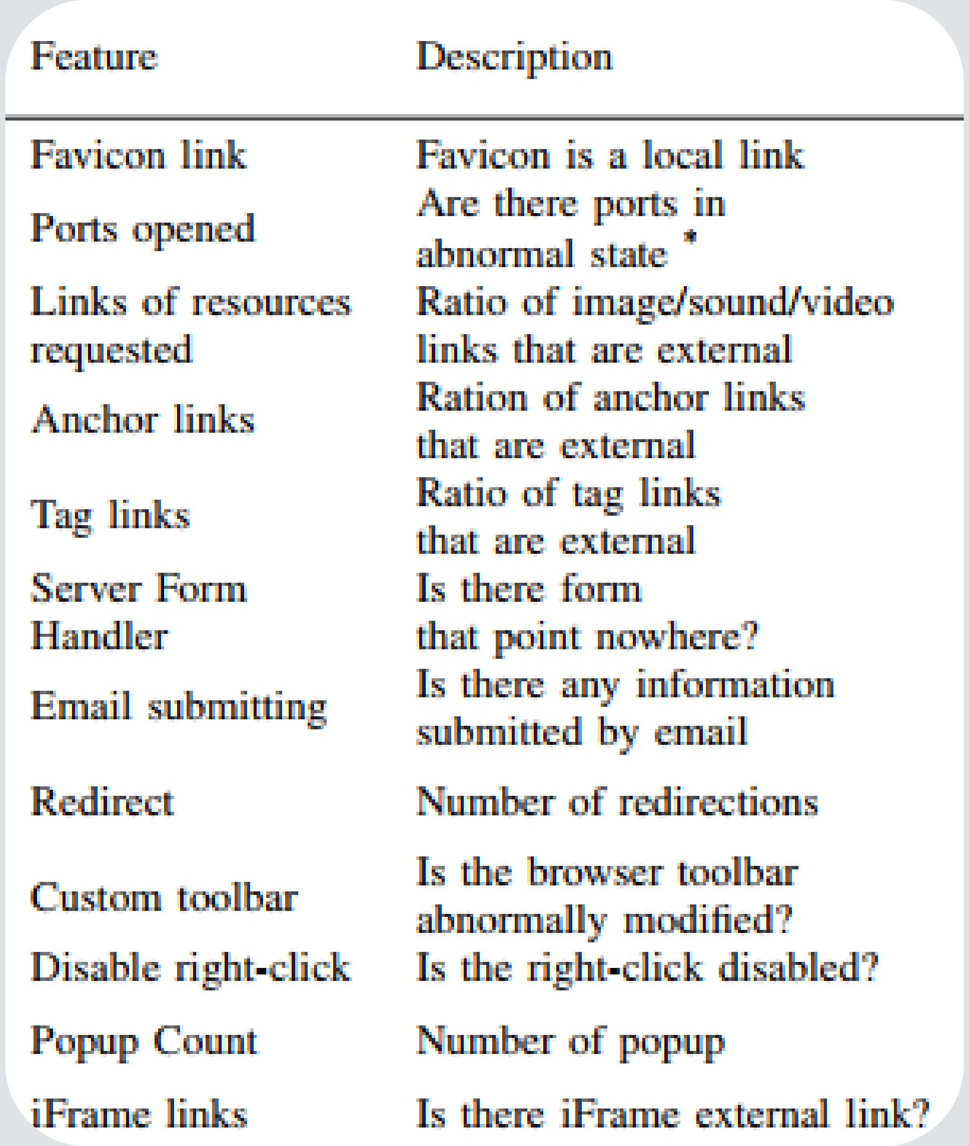
These features are extracted from the URL of the website



Feature	Description
IP address	Ip address in the hostname
Hostname length	Length of the hostname
Use a shortening service	Shortening service used
@ symbol	@ symbol in the URL
Double slash	'//' in the URL
Dash symbol	Number of dash in the URL
Subdomains count	Number of subdomains
Http token	Is ‘http’ in the URL except at the begining
Mean subdomains length	Average of the length of subdomains
www token	’www’ at the beginning of the URL
Valid TLD	The URL have a valid TLD*
Single character subdomain	A subdomain is composed by only one character?
Exclusive Prefix Repetition	The URL is composed by a repetition of a characters sequence?
TLD as subdomain	Is there a subdomain that is a valid TLD?
Ratio of exclusive digit subdomains	Ratio of subdomains composed only by digit
Ratio of exclusive hexadecimal subdomain	Ratio of subdomains composed only by hexadecimal
Ratio of underscore	Ratio of underscore in the hostname
Hostname contains digit	Does the hostname contain digit?
Vowel ratio	Ratio of vowel in the hostname
Digit ratio	Ratio of digit in the hostname
Alphabet cardinality	Number of alpha-characters in the hostname
Ratio of repeated characters	Ratio of the characters that are repeated in the hostname
Ratio of consecutive consonant	Ratio of consonant that precedes or succeeds another consonant
Ratio of consecutive digit	Ratio of digit that precedes or succeeds another digit

Features based on source code

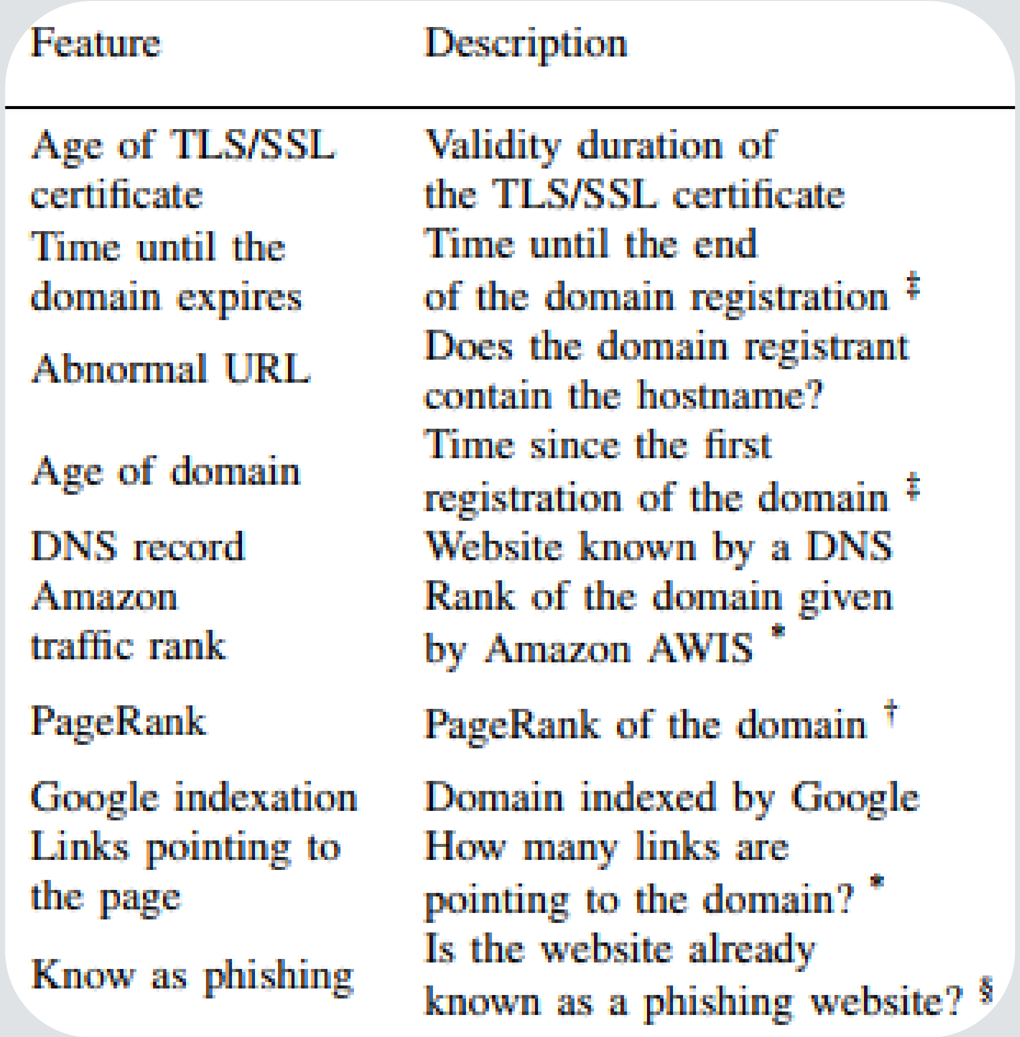
These features are extracted from the HTML source code of the website



Feature	Description
Favicon link	Favicon is a local link
Ports opened	Are there ports in abnormal state *
Links of resources requested	Ratio of image/sound/video links that are external
Anchor links	Ration of anchor links that are external
Tag links	Ratio of tag links that are external
Server Form Handler	Is there form that point nowhere?
Email submitting	Is there any information submitted by email
Redirect	Number of redirections
Custom toolbar	Is the browser toolbar abnormally modified?
Disable right-click	Is the right-click disabled?
Popup Count	Number of popup
iFrame links	Is there iFrame external link?

Features from third party sources

These features are extracted from information given by third party



Feature	Description
Age of TLS/SSL certificate	Validity duration of the TLS/SSL certificate
Time until the domain expires	Time until the end of the domain registration [‡]
Abnormal URL	Does the domain registrant contain the hostname?
Age of domain	Time since the first registration of the domain [‡]
DNS record	Website known by a DNS
Amazon traffic rank	Rank of the domain given by Amazon AWIS [*]
PageRank	PageRank of the domain [†]
Google indexation	Domain indexed by Google
Links pointing to the page	How many links are pointing to the domain? [*]
Know as phishing	Is the website already known as a phishing website? [§]

Datasets creation

- 2 datasets were created:
  - Phishing dataset containing 11250 phishing websites with their features
    - Extracted from *phishtank.com*
  - Clean dataset containing 24000 non-phishing websites with their features
    - Extracted from web browser history of a classic workstation and top 25000 of most visited websites in the US

GAN training

- Find the best type of value for each feature
  - *24 features can be extracted as discrete or continuous value*
  - *Each of them was tested and the most efficient type was kept*
- 2 GANs were trained
  - *With clean data as input*
  - *With phishing data as input*

RESULTS

With Clean data

- Good accuracy
- Low false positive rate

	Phishing	Clean
Phishing	1106	107
Clean	110	2297

Accuracy	94,00%
Precision	90,95%
Recall	91,17%
F1-Score	91,06%
False positive rate	4,57%
False negative rate	8,82%

With Phishing data

- Good recall
- High false positive rate
- Low false negative rate

	Phishing	Clean
Phishing	1150	63
Clean	228	2179

Accuracy	91,96%
Precision	83,45%
Recall	94,80%
F1-Score	88,76%
False positive rate	9,47%
False negative rate	5,19%

FUTURE WORK & CONCLUSION

How to improve it?

- Results are quite good
- Combine it to some other detection system
- Add a third class « suspicious »

Thank you!

Some questions?

Repo: github.com/khuzd/phishgan