ICCST 2025 conference in San Antonio (Texas)

Features Analysis of all Email Components for Phishing Detection

Armand Florent Tsafack Piugie*+, Mathieu Valois*, Philippe Chauvat*, Emmanuel Giguet+ and **Christophe Rosenberger+**

*téïcée company, 14760 Bretteville-sur-Odon, France

+Université de Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

{ftsafack,mvalois,pchauvat}@teicee.com,
emmanuel.giguet@unicaen.fr,
christophe.rosenberger@ensicaen.fr

téïcée

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

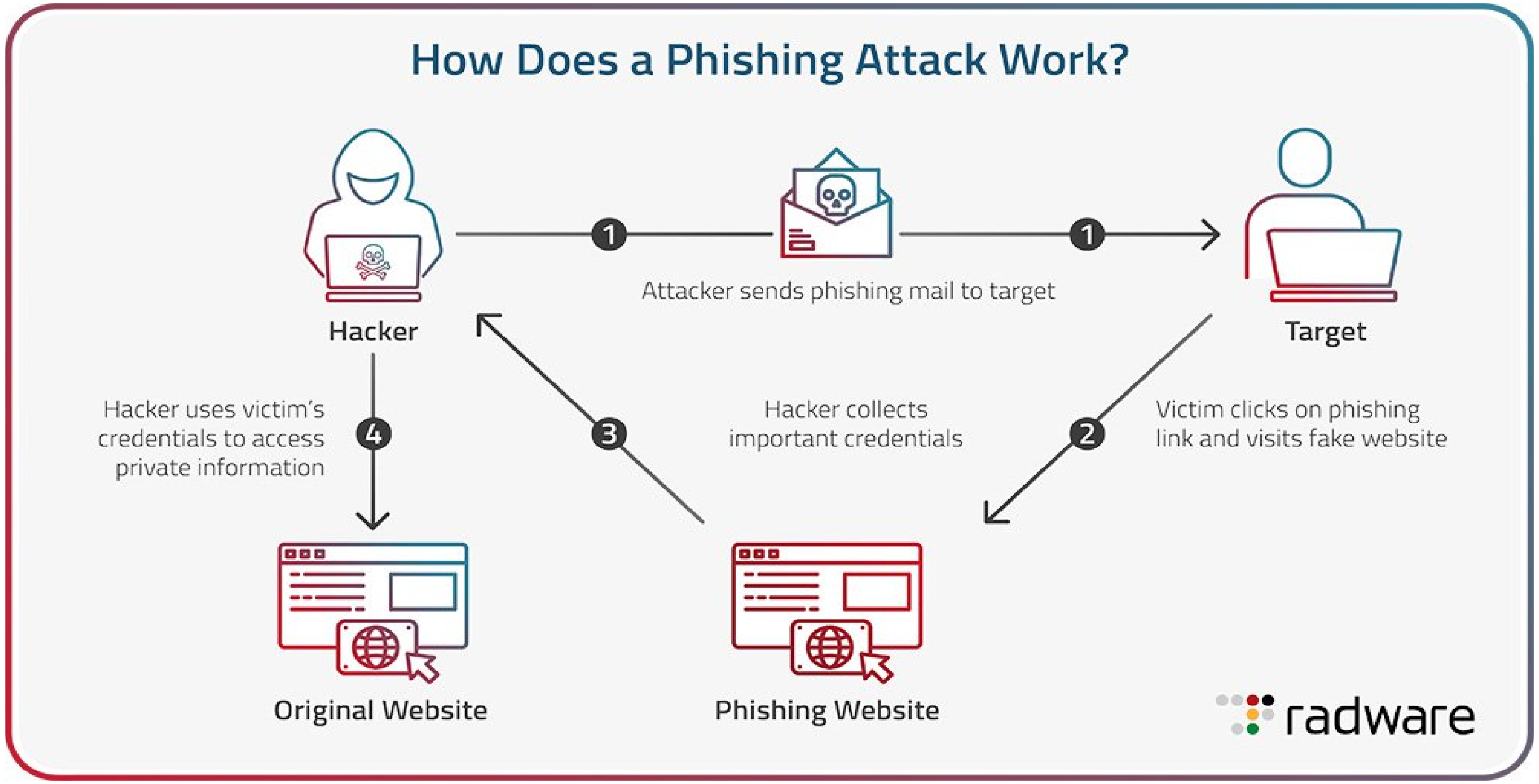
cnrs

Outline

1. Introduction : context

Phishing : a threat for organizations

- **~91% of cyber attacks** *[TEHTRIS, 2024]* .
- **~50% of organizations** *[TEHTRIS,2024]* .
- **~4,65M USD** *[L.Ana, 2024]* .



Phishing attack steps - https://www.radware.com/cyberpedia/bot-attacks/phishing-attack/

Social engineering attack vectors :

- Phishing Email
- Spear phishing
- Smishing
- Vishing
- Quishing

1. Introduction : problem and goals

Problem :
Attackers are refining their techniques to increase their profitability (by also using AI).

Main goal :
Design a protection system with automatic detection and human moderation.

Specifically :
- Using AI and NLP techniques for phishing email detection,
- Collaborative approach : Combine automated analysis results and user reports, with IT/CISCO moderators validating the email's status.

2. State of the art

Table I: Review of Phishing Email Detection methods.

Paper	Method	Dataset	Email part used	Feature extraction	Metric
[4]	NB, SVM, DTC, LR, AdaBoost and RF	IWSPA-AP	Text body	TF-IDF and Doc2Vec	Accuracy (73.1%)
[5]	NB, SVM, DTC, KNN, LR and RF	IWSPA-AP	Text body	TF-IDF, SVD and NMF	F1 (50.9%)
[6]	NB, LR and SVM	IWSPA-AP	Text body	TFIDF	Accuracy (94.3%)
[7]	NLP model	IWSPA-AP	Text body	fastText	F1 (96.8)%
[8]	SVM, KNN, NB, AdaBoost, RF, GBdt, LR, DTC, NN, RNN	SpamAssassin, Enron1, SMS-Spam-Collection	Text body	-	Accuracy (99.5%)
[9]	RF, DT, GNB, MNB, LR, KNN, Bagg., Boost., SVM, DL, AutoSk, TPOT	Email Benchmark Dataset	header + body	header + body + TFIDF	Accuracy (99.9%)
[10]	KNN attribute graph	Symantec	header and body	header and body features	F1 (90%), FPR (0.01)
[11]	SVM	Nazario and SpamAssassin	header and body	DES, SBW, URLD, URLS, URLIP, US, UD, DMID	Accuracy (97.2%)

2. State of the art : ML Methods

Table I: Review of Phishing Email Detection methods.

Paper	Method	Dataset	Email part used	Feature extraction	Metric
[4]	NB, SVM, DTC, LR, AdaBoost and RF	IWSPA-AP	Text body	TF-IDF and Doc2Vec	Accuracy (73.1%)
[5]	NB, SVM, DTC, KNN, LR and RF	IWSPA-AP	Text body	TF-IDF, SVD and NMF	F1 (50.9%)
[6]	NB, LR and SVM	IWSPA-AP	Text body	TFIDF	Accuracy (94.3%)
[7]	NLP model	IWSPA-AP	Text body	fastText	F1 (96.8)%
[8]	SVM, KNN, NB, AdaBoost, RF, GBdt, LR, DTC, NN, RNN	SpamAssassin, Enron1, SMS-Spam-Collection	Text body	-	Accuracy (99.5%)
[9]	RF, DT, GNB, MNB, LR, KNN, Bagg., Boost., SVM, DL, AutoSk, TPOT	Email Benchmark Dataset	header + body	header + body + TFIDF	Accuracy (99.9%)
[10]	KNN attribute graph	Symantec	header and body	header and body features	F1 (90%), FPR (0.01)
[11]	SVM	Nazario and SpamAssassin	header and body	DES, SBW, URLD, URLS, URLIP, US, UD, DMID	Accuracy (97.2%)

IWSPA Challenge :

- **Best performance** achieved by [7] with an **NLP model** : F1 score of 96.8%.
- NB, LR, and SVM performed well, with an accuracy of 94.3% [6].
- Most ML models were **trained on the message body** of the emails.

[4] N. Unnithan, H. N B, V. Ravi, S. Kp, and S. Sundarakrishna, “Detecting phishing e-mail using machine learning techniques cen-securenlp,” 2018.
[5] N. Harikrishnan, R. Vinayakumar, and K. Soman, “A machine learning approach towards phishing email detection,” in Proceedings of the anti- phishing pilot at ACM international workshop on security and privacy analytics (IWSPA AP), vol. 2013, 2018, pp. 455–468.
[6] N. Unnithan, H. N B, A. Soman, V. Ravi, and S. Kp, “Machine learning based phishing e-mail detection security-cen@amrita,” 03 2018.

2. State of the art : ML Methods

Table I: Review of Phishing Email Detection methods.

Paper	Method	Dataset	Email part used	Feature extraction	Metric
[4]	NB, SVM, DTC, LR, AdaBoost and RF	IWSPA-AP	Text body	TF-IDF and Doc2Vec	Accuracy (73.1%)
[5]	NB, SVM, DTC, KNN, LR and RF	IWSPA-AP	Text body	TF-IDF, SVD and NMF	F1 (50.9%)
[6]	NB, LR and SVM	IWSPA-AP	Text body	TFIDF	Accuracy (94.3%)
[7]	NLP model	IWSPA-AP	Text body	fastText	F1 (96.8)%
[8]	SVM, KNN, NB, AdaBoost, RF, GBdt, LR, DTC, NN, RNN	SpamAssassin, Enron1, SMS-Spam-Collection	Text body	-	Accuracy (99.5%)
[9]	RF, DT, GNB, MNB, LR, KNN, Bagg., Boost., SVM, DL, AutoSk, TPOT	Email Benchmark Dataset	header + body	header + body + TFIDF	Accuracy (99.9%)
[10]	KNN attribute graph	Symantec	header and body	header and body features	F1 (90%), FPR (0.01)
[11]	SVM	Nazario and SpamAssassin	header and body	DES, SBW, URLD, URLS, URLIP, US, UD, DMID	Accuracy (97.2%)

- Aassal and al. [9] has **very good performance** by analysing headers, body, and text content with **ML Models** : 99.9% Acc.
- Email Benchmark Dataset Characteristics :
  - Not publicly available, even for research purposes
  - Balanced, labeled as phishing (10500) and legitimate (10500)

[9] A. El Aassal, S. Baki, A. Das, and R. M. Verma, “An in-depth benchmarking and evaluation of phishing detection research for security needs,” Ieee Access, vol. 8, pp. 22 170–22 192, 2020

2. State of the art : Databases

Dataset	Parts of emails used	Size	ham	spam
SpamAssasin	Body + Header		2 551	501
SMS Spam Collection	Body	5 572	4 825	747
WestPac	Body	659 673	613 048	46 525
IWSPA-AP-2018	Body + Header	10 306	9 174	1 132
CSDMC 2010 SPAM	Body + Header	4 327	2 949	1 378
TREC corpus	Body		39 399	52 790
Enron	Body	500 000	43 000	50 000
Enron1	Body	6 447	3 228	3 219

IWSPA-AP Database [22] :

- Emails are labeled as **phishing** or **legitimate**.
- Imbalanced class ratio.
- Emails are written in English.
- Available for research purposes.

[22] A. E. Aassal, L. Moraes, S. Baki, A. Das, and R. Verma, “Anti-phishing pilot at acm iwspa 2018: Evaluating performance with new metrics for unbalanced datasets,” Proc. IWSPA-AP Anti Phishing Shared Task Pilot 4th ACM IWSPA, pp. 2–10, 2018.

3. Methodology

- Work on an imbalanced dataset
  - Using an existing dataset : IWSPA dataset
  - Creating a new dataset with French emails : téïcée dataset

- Using an existing taxonomy of feature based on headers, body and text content [9]
  - New feature extraction techniques and preparation

- Designing a new ML pipeline for training and evaluating ML Models
  - Resampling techniques to handle imbalanced datasets
  - Fusion Model to keep the best performance

- Evaluating on two datasets

[9] A. El Aassal, S. Baki, A. Das, and R. M. Verma, “An in-depth benchmarking and evaluation of phishing detection research for security needs,” Ieee Access, vol. 8, pp. 22 170–22 192, 2020

3. Methodology

Considering all components of the email : headers and message body content

eBay.com
aw-confirm@ebay.com

To username@domain.com

WARNING-IFRAME-Your account at eBay has been suspended

email headers

Need Help?

Dear eBay User,

We regret to inform you, that we had to block your eBay account
because we have been notified that your account may have been compromised by outside parties.

Our terms and conditions you agreed to state that your account must always be under your control or those you designate at all times. We have noticed some activity related to your account that indicates that other parties may have access and or control of your information in your account.

Please be aware that until we can verify your identity no further access to your account will be allowed.As a result,Your access to bid or buy on eBay has been restricted.To start using your eBay account fully,Please uptake and verify your information by clicking below

http://signin.ebay.com/aw-c gi/eBayISAPI.dll?Verify

Regards,

eBay Member Service

Please Do Not Reply To This E-mail As You Will Not Receive A Response

email body

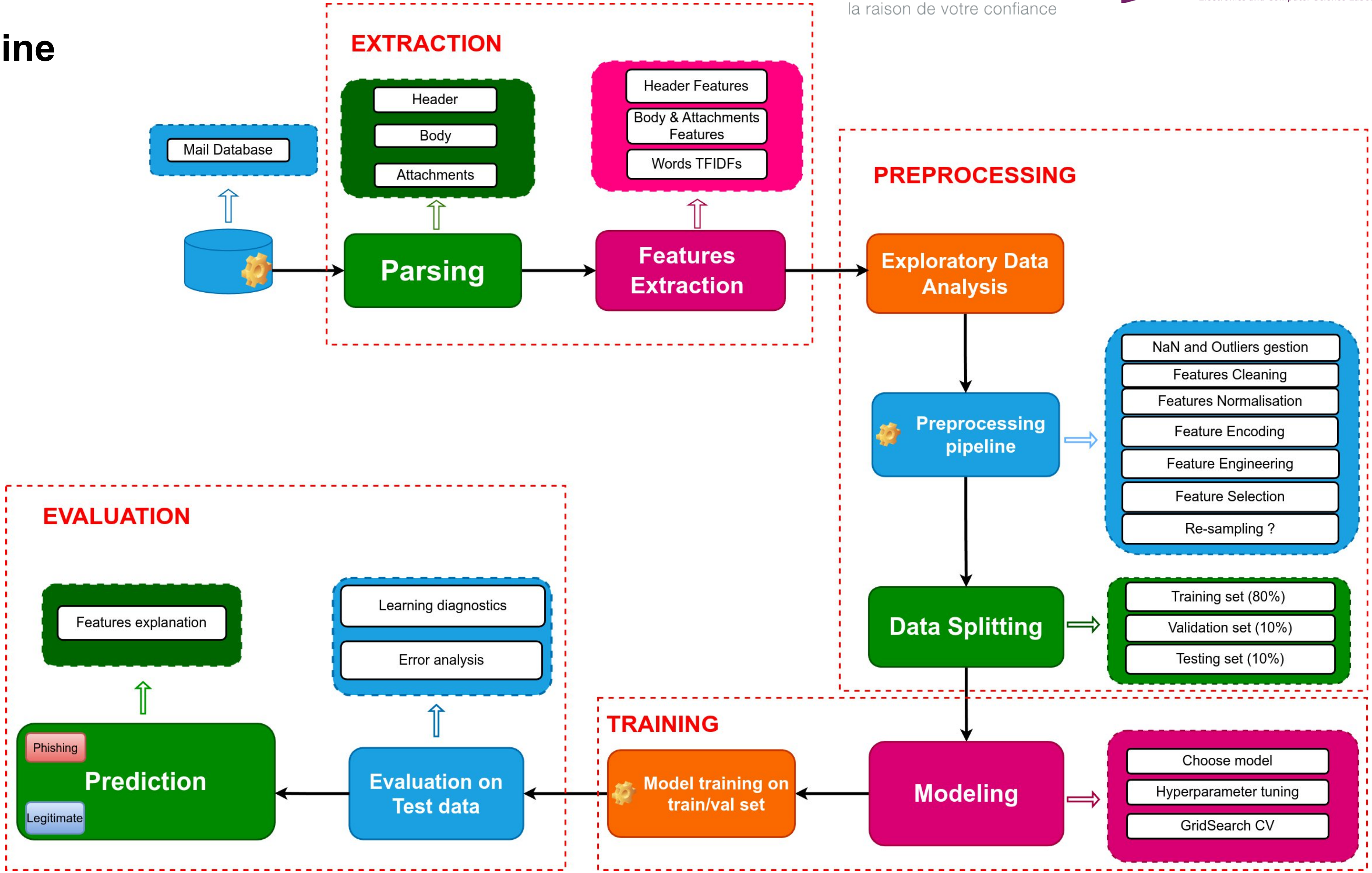
```
From nobody Sun Jan 25 13:47:53 2015
Return-Path: <aw-confirm@ebay.com>
X-Original-To: username@domain.com
Delivered-To: username@domain.com
Received: from mail2.domain.com (mail2.domain.com [192.168.64.8])
 by spanky.domain.com (Postfix) with ESMTP id 6B4A5539307
 for <username@domain.com>; Sun, 28 Nov 2004 17:24:56 -0500 (EST)
Received: from 220.124.143.10 (unknown [220.124.143.10])
 by mail2.domain.com (Postfix) with SMTP id A06CD18A0BF
 for <username@domain.com>; Sun, 28 Nov 2004 06:10:42 -0500 (EST)
Received: from 65.167.151.237 by ; Sun, 28 Nov 2004 15:32:17 -0100
Message-ID: <JDXSYRKPDISQIRUBXXMCHBC@goldenmail.ru>
From: "eBay.com" <aw-confirm@ebay.com>
Reply-To: "eBay.com" <aw-confirm@ebay.com>
To: username@domain.com
Old-Subject: Your account at eBay has been suspended
Date: Sun, 28 Nov 2004 10:28:17 -0600
X-Mailer: midwife biceps around 3
MIME-Version: 1.0
Content-Type: multipart/alternative;
    boundary="--=====767125884059=_"
X-Priority: 1
X-MSMail-Priority: High
X-YAVR: IFRAME
Subject: WARNING-IFRAME-Your account at eBay has been suspended
Status: RO
X-Status:
X-Keywords:
X-UID: 5
```

3. Methodology

Proposed ML Pipeline

Pipeline :

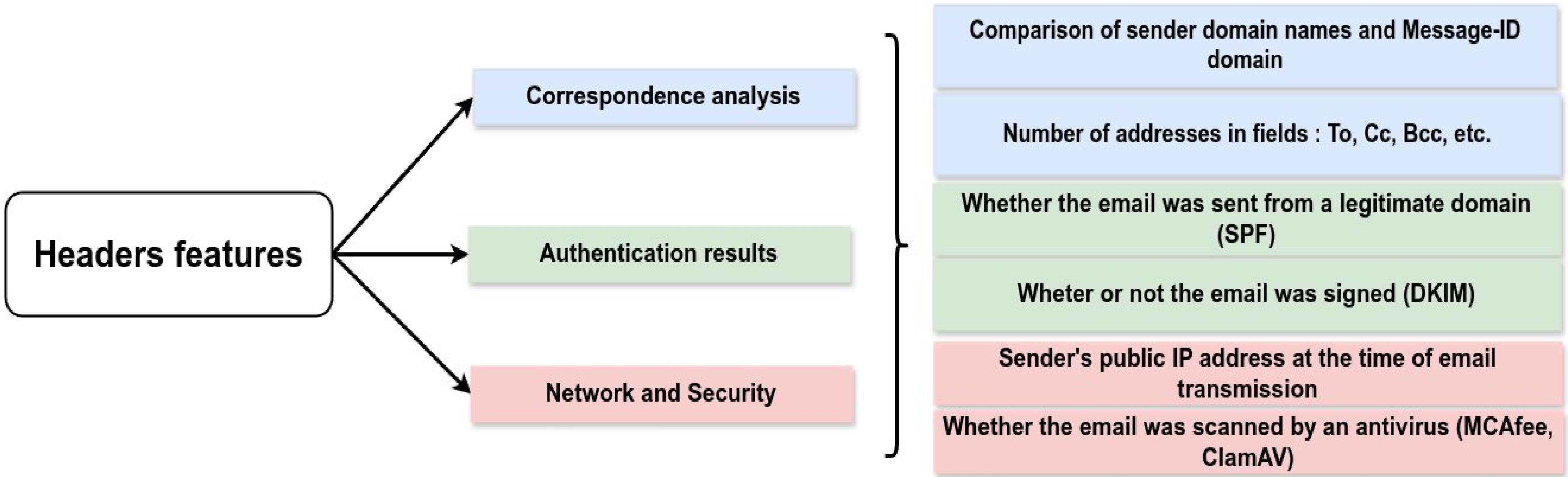
- Extraction
- Preprocessing
- Training
- Evaluation



3. Methodology

EXTRACTION

Physical analysis of email headers



Number of headers features extracted: **68**

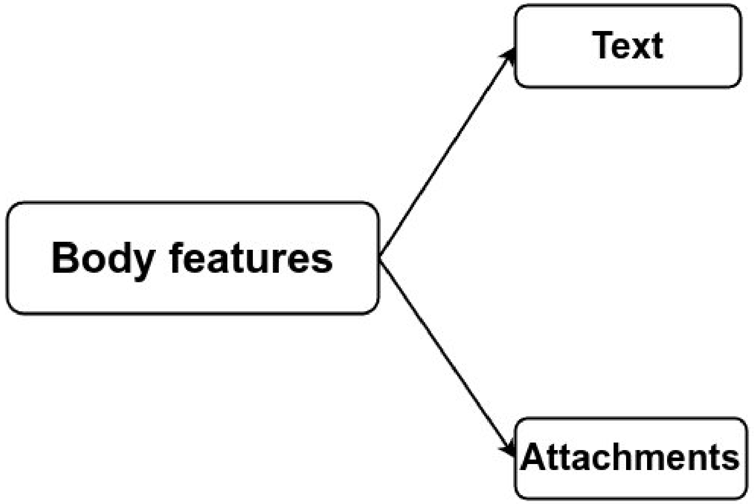
3. Methodology

la raison de votre confiance

Electronics and Computer Science Laboratory

EXTRACTION

Physical and semantic analysis of email body



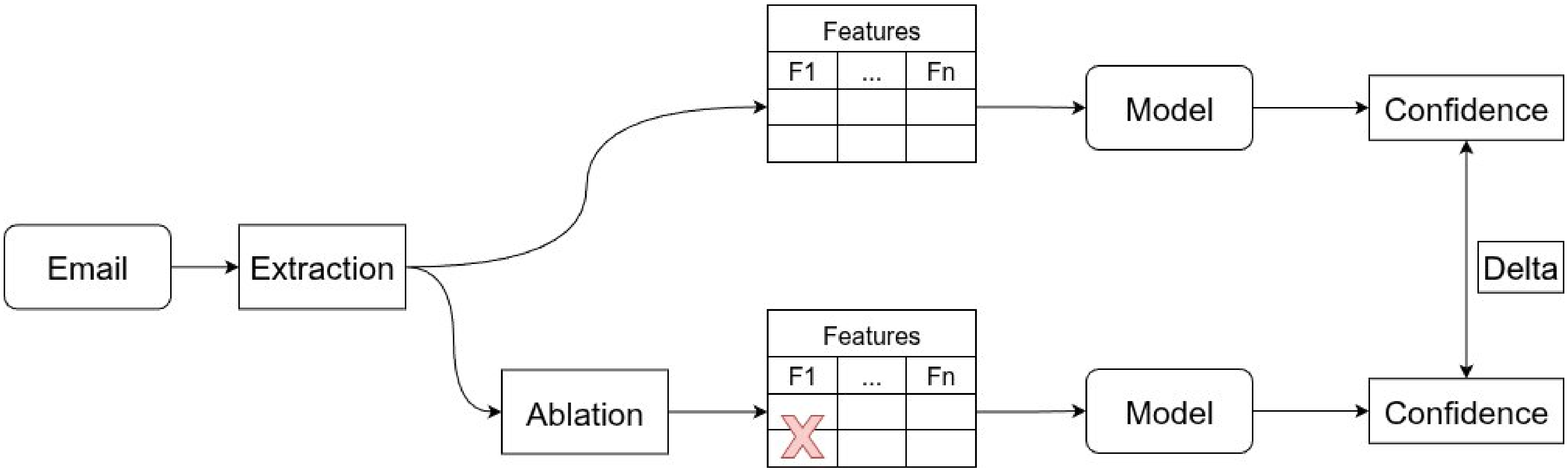
content type region	Number of email parts with a declared content-type
	Number of parts with an identical content-type
charset region	Number of parts with a declared charset
	Number of parts with an identical charset
content transfert encoding	Number of parts with a declared content transfer encoding
	Number of parts with an identical content transfert encoding
vocabulary	Number of words in the body
	Number of characters in the body
html part	Number of opening and closing tags
	Number of clickable links
semantic	Word frequency in term of TF-IDF (semantic analysis)

Number of body features extracted: **76**

Number of TFIDF features : **10 000**

3. Methodology

Explaining phishing detection



Delta quantifies the confidence difference before and after feature ablation.

4. Experimental Protocol

DATABASES

IWSPA-AP 2018

- Origin : IWSPA-AP 2018 competition
- With headers
- Number of emails : **4579**
- 2 classes in the dataset :
  - **phishing** : 497
  - **legitimate** : 4082
- Class imbalance ratio : ~8.21
- Text in the body : cleaned
- Text language : english

téïcée company

- Origin : téïcée company
- With headers
- Number for emails : **803**
- 2 classes in the dataset :
  - **spams** : 377
  - **hams** : 426
- Class imbalance ratio : ~1.13
- Text in the body : not cleaned
- Text language : multilingual (french, english, etc.)

4. Experimental Protocol

ML Protocol

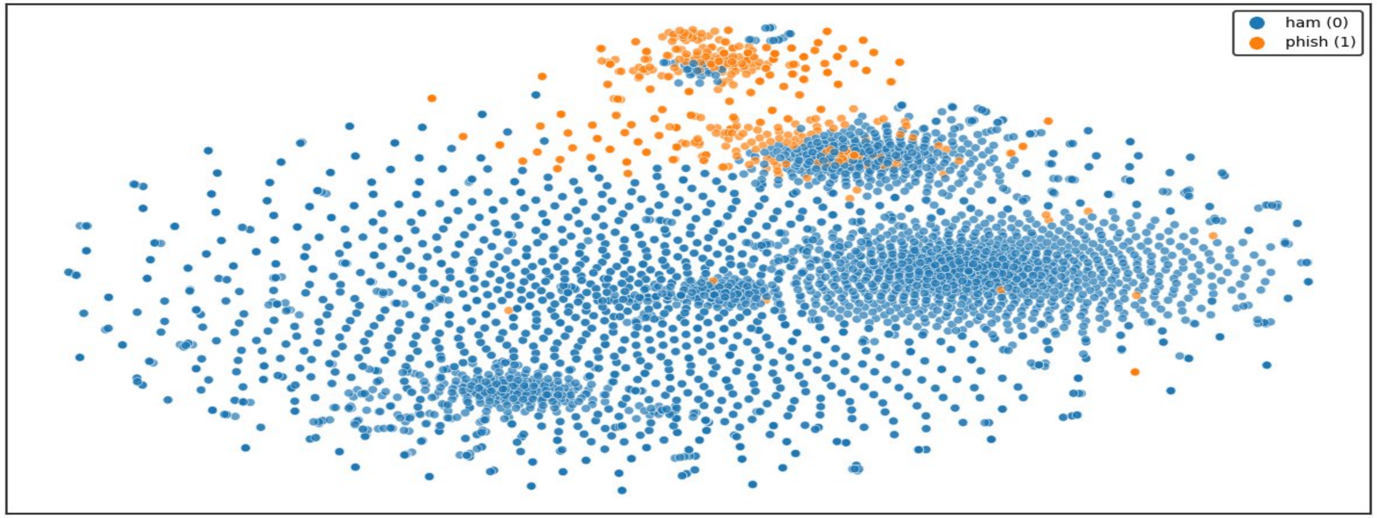
- Datasets:
  - IWSPA-2018 (497 | 4082)
  - téïcée company emails (377 | 426)

- Email Parsing
- Feature Preparation Pipeline:
  - Extraction, Cleaning, Encoding, Normalization, Selection + PCA

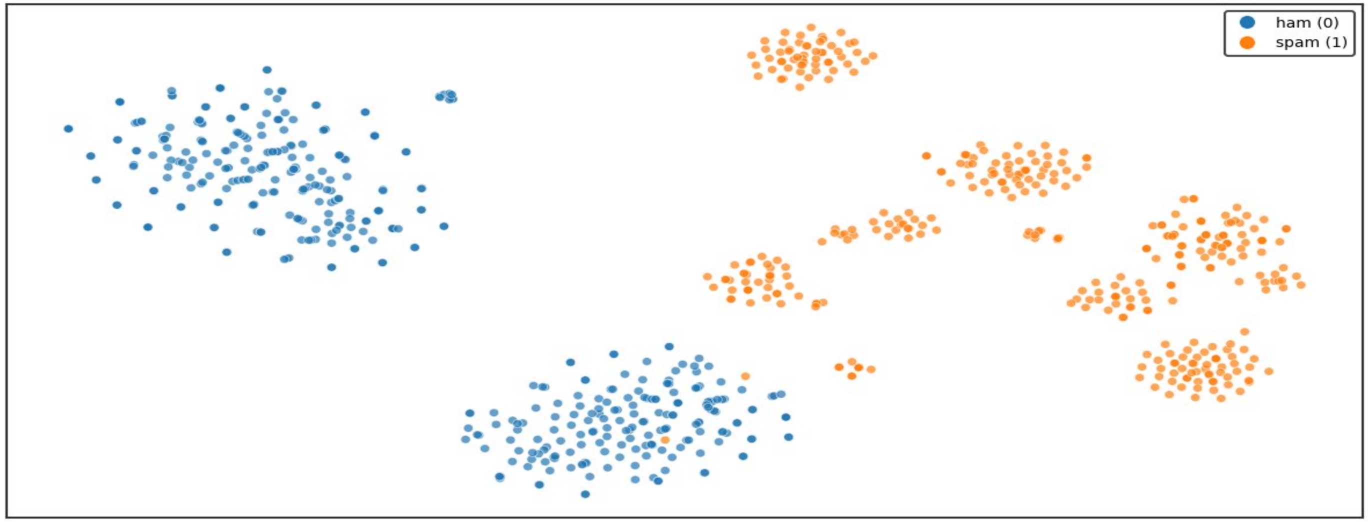
- **Resampling :** SMOTE, ADASYN, RandomUnderSampler, None

- Train/Val/Test : 80/10/10
- 13 ML models with 4 folds cross validation
- Meta model
- Metrics: Accuracy, FP/N, FP/P.

iwspa database 2D visualisation



téïcée email database 2D visualisation



5. Results

Performance evaluation

Table II. Performance of the best model on the two datasets (training time in second)

Dataset	Acc (%)	FP/N (%)	FN/P (%)	Training
IWSPA-AP	99	0.2	**0.0**	6985
téïcée	100	**0.0**	**0.0**	**95**

- Fusion model that gives best performance ;
- Excellent performance ;
- But it should be mitigated by the fact that datasets are small and models are tuned for each of them.

Generalization capability

Table III. Performance of the best model on the two datasets when training with the other one.

Dataset	Acc (%)	FP/N (%)	FN/P (%)
IWSPA-AP/téïcée	83.6	60.0	**0.5**
téïcée/IWSPA-AP	**89.0**	**4.0**	14.0

- Good generalization capability on IWSPA;

5. Results

Algorithm performance accross resampling method on IWSPA-AP testing set

ML Models	NONE	Random UnderSampling	ADASYN	SMOTE
ADABOOST	99.1	**99.7**	**99.7**	99.6
BGC	99.1	98.0	99.4	**99.5**
BNB	95.8	92.0	91.9	**96.2**
DTC	98.2	97.0	99.0	**99.1**
ETC	98.5	99.0	98.5	**99.5**
GBDT	98.7	**100**	99.5	99.5
GNB	96.1	93.0	**95.7**	95.8
KNC	91.5	**96.0**	95.2	95.1
LR	99.3	99.0	**99.7**	99.6
MNB	**96.3**	93.0	93.9	94.9
RF	99.1	99.0	99.5	**99.7**
SVC	98.0	98.0	98.7	**99.8**
XGB	99.3	**100**	99.6	99.5

- Increase of performances with resampling on IWSPA-AP dataset;
- Best results in average with **SMOTE** approach.
- XGB archieved the highest accuracy (99.3%) without dataset resampling, surpassing **[6]** result.

[6] N. Unnithan, H. N B, A. Soman, V. Ravi, and S. Kp, “Machine learning based phishing e-mail detection security-cen@amrita,” 03 2018.

5. Results

Explaining phishing detection

Table V. Explanation of the decision for 3 emails by making feature ablation.

Email class	prediction	Confidence	Top 3 Features
phishing	phishing	97%	if the body has content in html (10%) number of MTA the email acrossed (10%) number of parts declared as text/html (7%)
ham	ham	100%	brunet index (22%) lexical entropy (17%) normalized token type ratio (6%)
phishing	phishing	59%	blacklisted words in the subject (14%) has DKIM signature (9%) blacklisted words in the body (6%)

For example, in the second email :

- The **brunet index** features contributes 22% of the confidence score
- It measures the vocabulary richness of the text contained in the email.

Conclusion

- Analysis of original raw email messages
- Feature extraction and preparation pipeline
- Evaluation on two datasets
- Feature ablation for detection explainability

Future works

- Research on larger email datasets (more than 40.000 samples) from SignalSpam organization;
- Explore phishing detection approach using LLM training ;

Thanks for your kind attention!

Questions ?