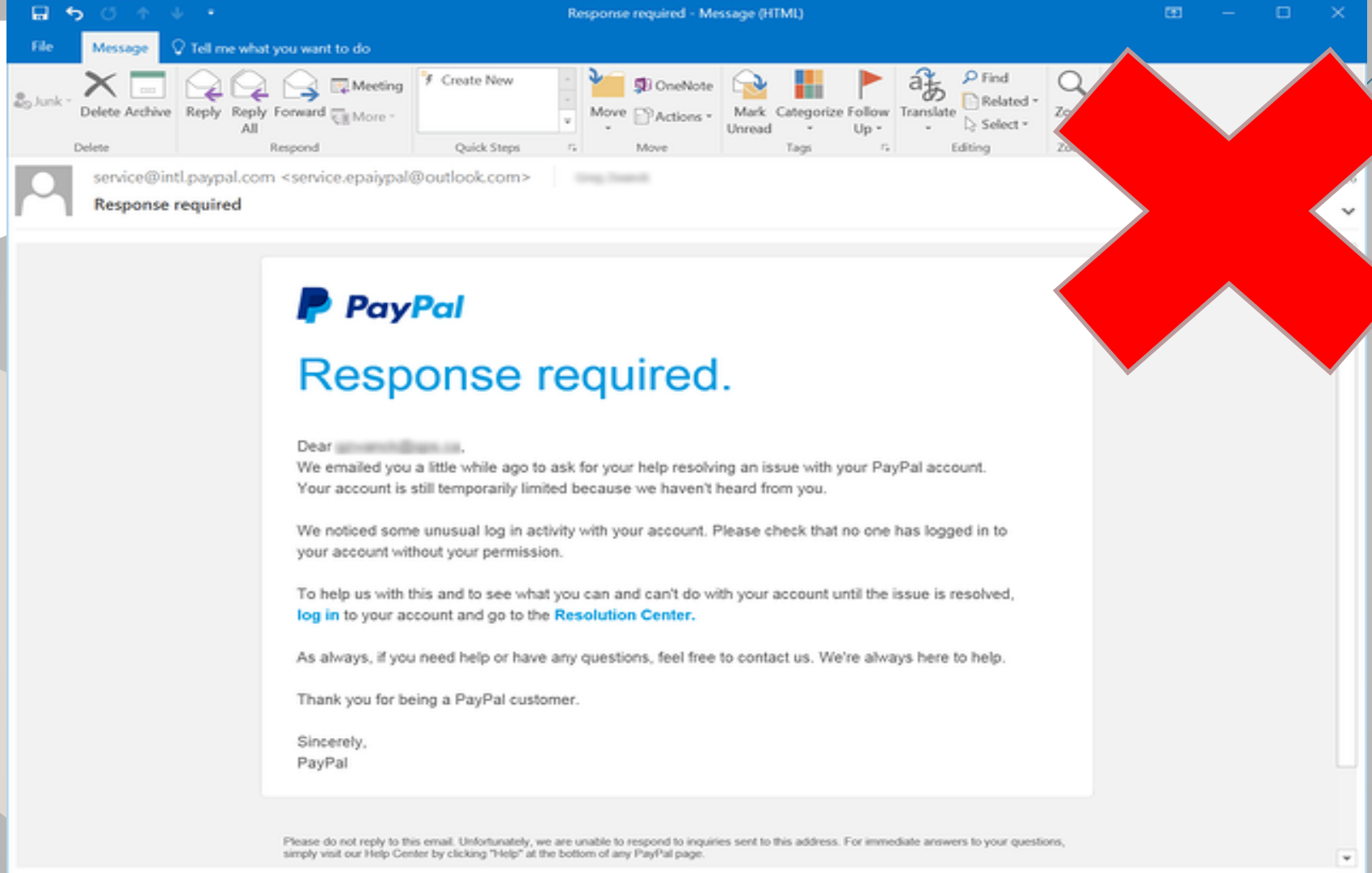
Data Acquisition and Corpus Creation for Security-Related Domain

Sanja Seljan, Nevenka Tolj, Ivan Dunđer
sanja.seljan@ffzg.unizg.hr, ntolj@ffzg.unizg.hr ivandunder@gmail.com

Faculty of Humanities and Social Sciences - University of Zagreb, Information and Communication Sciences


Response required - Message (HTML)
File
Message
Tell me what you want to do
Junk
Delete
Archive
Reply
Reply All
Forward
Meeting
More
Delete
Respond
Create New
Quick Steps
Move
OneNote
Actions
Move
Mark Unread
Categorize
Follow Up
Tags
Translate
Find
Related
Select
Editing
service@intl.paypal.com <service.epaiypal@outlook.com>
Response required
PayPal
Response required.
Dear
We emailed you a little while ago to ask for your help resolving an issue with your PayPal account.
Your account is still temporarily limited because we haven't heard from you.
We noticed some unusual log in activity with your account. Please check that no one has logged in to your account without your permission.
To help us with this and to see what you can and can't do with your account until the issue is resolved, log in to your account and go to the Resolution Center.
As always, if you need help or have any questions, feel free to contact us. We're always here to help.
Thank you for being a PayPal customer.
Sincerely,
PayPal
Please do not reply to this email. Unfortunately, we are unable to respond to inquiries sent to this address. For immediate answers to your questions, simply visit our Help Center by clicking "Help" at the bottom of any PayPal page.

Bank transfer

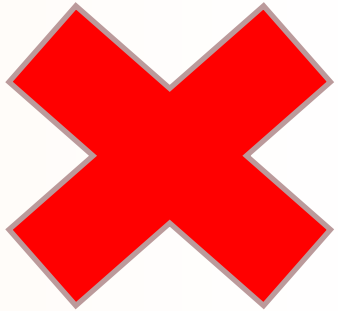
Pošiljatelj:
Primatelj:
Kopija:
Predmet: RE: Bankovni transfer

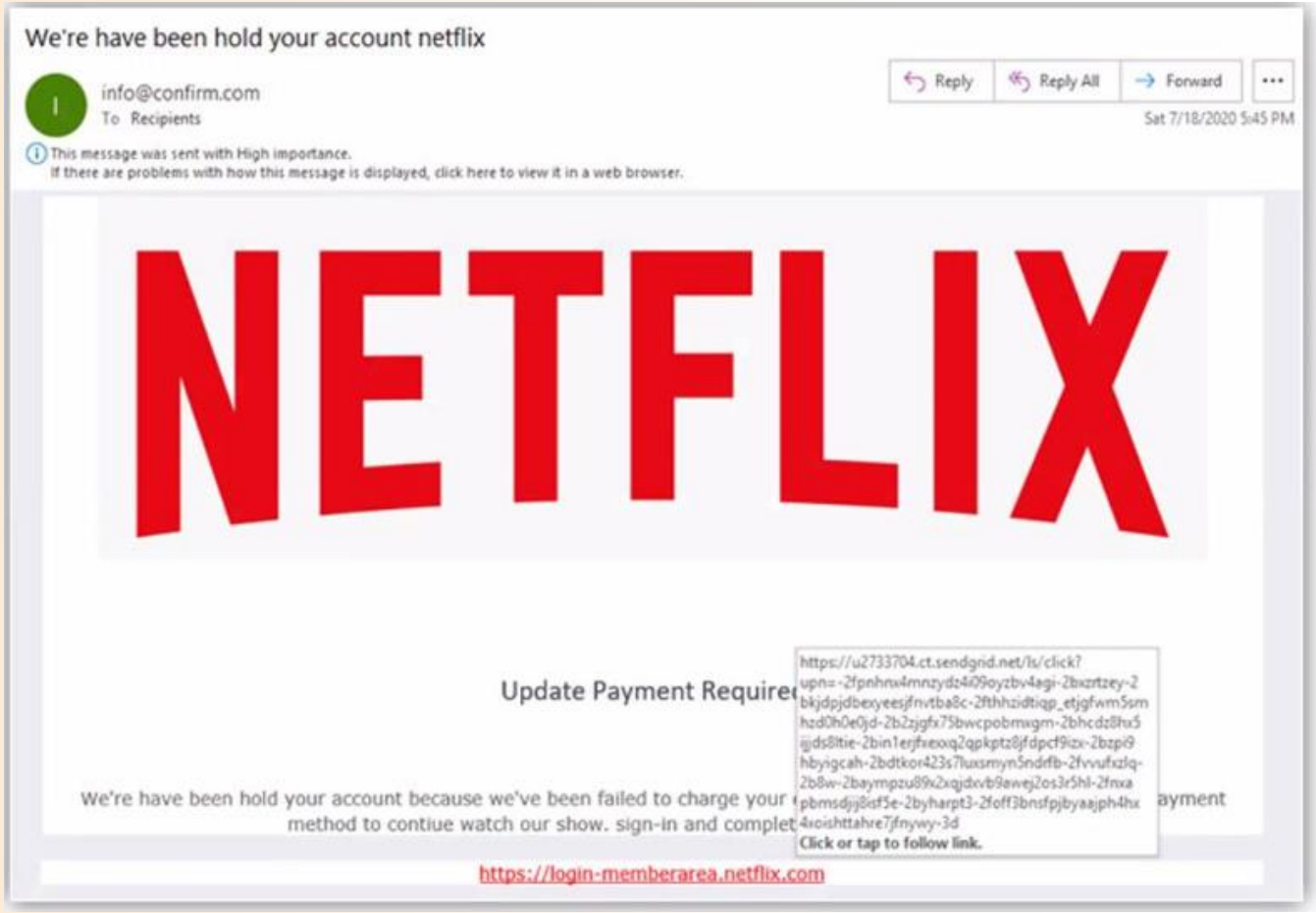
isplatiti 9 700 eura u banku kako slijedi:

Valuta: EUR
Naziv : Angnetha Sjoberg
banke Naziv: Nordea BANK PLC
adresa :Smalandsgatan 17,10571 Stockholm Sweden Swift code: NDEASESS
IBAN: SE9630000000006712014841
Account no : 6712014841

Molimo da prijenos odmah i recite mi kada ste gotovi.

S postovanjem,




We're have been hold your account netflix
info@confirm.com
To Recipients
Reply
Reply All
Forward
Sat 7/18/2020 5:45 PM
This message was sent with High importance.
If there are problems with how this message is displayed, click here to view it in a web browser.
NETFLIX
Update Payment Require
https://u2733704.ct.sendgrid.net/ls/click?
upn=-2fpnhnx4mnzydz4i09oyzbv4agi-2bxzrtzey-2
bkjdpjdbexyeesjfnvtba8c-2fthhzidtiqp_etjgfwm5sm
hzd0h0e0jd-2b2zjgfx75bwcpobmxgm-2bhcdz8hx5
ijds8itie-2bin1erjfxeoxq2qpkptz8jfdpcf9izx-2bzpi9
hbyigcah-2bdtkor423s7luxsmyn5ndrfb-2fvvufxzlq-
2b8w-2baympzu89x2xqjdxvb9awej2os3r5hl-2fnxa
pbmsdjyj8isf5e-2byharpt3-2foff3bnsfpjbyaajph4hx
4xoishttahre7jfnywy-3d
Click or tap to follow link.
We're have been hold your account because we've been failed to charge your
method to contiue watch our show. sign-in and complet
ayment
https://login-memberarea.netflix.com

*** VIRUS ***URGENT ORDER CONFIRMATION

To undisclosed-recipients:

Urgent order confirmati...
370 KB

Good Day,

We would like to get your discount price for the order attached file.

Due to the urgency, Can you Kindly issue us your proforma invoice with your best price and terms of payment equally spelled out.

Waiting urgently for your prompt action.

Best Regards

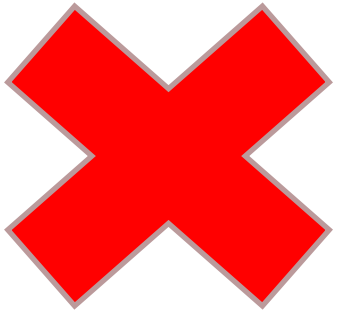
Giovanna Lorenzo
Sales Manager

MEDICAL
EQUIPMENT


From:
Amazon <management@mazoncanada.ca> on behalf of
not an Amazon email address
(note the missing A in Amazon)
To: @sheridanc.on.ca
Cc:
Subject: Suspension
amazon.com
Dear Client,
Generic non-personalized greeting
We have sent you this e-mail, because we have strong reason to belive, your account has been used by someone else.In order to prevent any fraudulent activity from occurring we are required to open an investigation into this matter. We've locked your Amazon account, and you have 36 hours to verify it, or we have the right to terminate it.
To confirm your identity with us click the link bellow:
https://www.amazon.com/exec/obidos/sign-in.html
Hovering over the link reveals it points to a non-Amazon site - "http://redirect.kereskedj.com"
Sincerely,
The Amazon Associates Team
© 1996-2013, Amazon.com, Inc. or its affiliates

Pošiljatelj: "Raiffeisenbank." <support@we-are.team>

Subject: Neki podaci na vašem računu nedostaje ili nije ispravan !



Poštovani kupci,

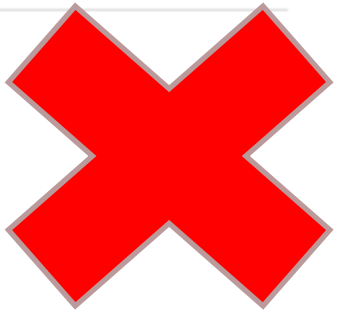
Neki podaci na vašem računu nedostaje ili nije ispravan
moramo povremeno provjeravamo podatke o računu Jamčimo da naši korisnici mogu koristiti naše usluge propisno.

Ažurirajte odmah vaše podatke za nastavak uživate u svim prednostima svog računa.
Ako ne ažurirati svoje podatke u roku od 2 dana, hoće ograničiti upotrebu računa

Ažuriranje podataka

https://royalbioenergy.com/wp-admin/w8hyr.php

Pošiljatelj: Olt Director <Olt.Director@tgie.ro>
Poslano: 20. veljače 2020. 10:44
Primatelj: no-reply@microsoft.net
Predmet: Vaš račun za e-poštu treba odmah potvrditi



MICROSOFT VAŽNA OBAVIJEST

Vaš račun za e-poštu treba odmah **potvrditi** ili će vaš račun za e-poštu biti obustavljen ako nije potvrđen sada.

https://ismcadmissions.wixsite.com/mysite

Hvala na razumijevanju

Microsoftov tim za provjeru

Hrvatska pošta

Dragi kupče,

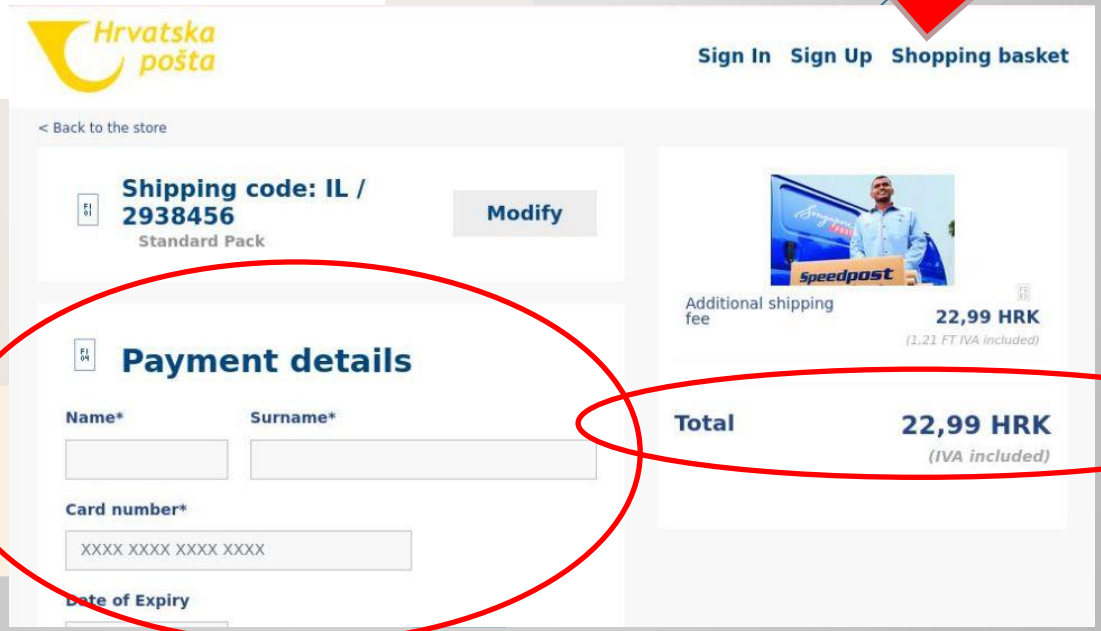
Obavještavamo vas da vaša pošiljka čeka dostavu.

Potvrdite uplatu od 22,99 kuna na donjoj poveznici.

Napomena: postupak provjere mora se obaviti u sljedećih 02 dan

Kliknite donju poveznicu :

https://www.posta.hr/payment



MZ

Thu 9/3/2020 11:01 AM

Ministarstvo zdravlja Hrvatska <covid-19@zdravlje.gov.hr>

Besplatna distribucija zaštitne opreme Covid-19 (Ministarstvo zdravlja)

Prima undisclosed-recipients:

Važnost ove poruke: Visoko.
Ako se pojavljuju problemi vezani uz način prikaza ove poruke, kliknite ovdje da biste je pogledali u web-pregledniku.

covid-19_besplatna zaštitna oprema-pdf.zip
Datoteka – .zip

Republika Hrvatska
Ministarstvo zdravlja

Poštovani građani

Sukladno članku 3. Zakona o unutarnjem zakonodavstvu od 25-4-2020 (Vlada Hrvatske 42 / A / 25-4-2020) "Hitne mjere za sprječavanje i ograničenje širenja koronavirusa".

Mi, Ministarstvo zdravstva RH, u suradnji s hrvatskom vladom želimo besplatno podijeliti maske za lice Covid-19, respirator, ispitne strojeve i drugu zaštitnu opremu covid-19 svim hrvatskim registriranim tvrtkama, a koordinira Središnje vijeće Zdr
priloženi obrazac i provjerite je li na tom obrascu napisan točan broj zaposlenika i adresa tvrtke.

Ispunite priloženi obrazac i pošaljite nam kopiju prije zatvaranja danas i veselimo se vašem brzom odgovoru.

Sve popunjene obrasce pošaljite na ovu e-mail adresu: covid-19@zdravlje.gov.hr

pozdravi

Prof. dr. sc. Vili Beroš
Ministarstvo zdravlja

REPUBLIKA HRVATSKA MINISTARSTVO ZDRAVLJA
Ksaver 200a, 10000, Zagreb, Croatia
Telephone: (+385)1 4670 555

*** VIRUS ***RE: ZAHTJEV ZA PONUDU (Sveučilište u Zagrebu) EUI894/BU466

To undisclosed-recipients:

If there are problems with how this message is displayed, click here to view it in a web browser.
Click here to download pictures. To help protect your privacy, Outlook prevented automatic download of some pictures in this message.

University of Zagreb

Pozdrav sa Sveučilišta u Zagrebu,

Pažnja:

U skladu s dobrim preporukama Vaše tvrtke, mi smo Sveučilište u Zagrebu pod vodstvom profesora Damira Borasa.

Potrebna nam je vaša cijena za naš proračun za 2022. godinu (u prilogu).

Pošaljite svoju ponudu na ili prije roka 5. ožujka 2022.

ESN ZAGREB

GEOF OSGL

Erasmus+

UNIVERSITY OF ZAGREB
FACULTY OF SCIENCE
Department of Biology

Sveučilište u Zagrebu
Trg Republike Hrvatske 14
10000 Zagreb, Croatia
Telephone: +385 1 4564 111
Fax: +385 1 4830 602

Ulica kralja Zvonimira 8
10000 Zagreb, Croatia
Telephone: +385 1 4698 100
Fax: +385 1 4698 141

NIN: 36612267447
E-mail: unizginfo@unizg.hr

Outline

I. Introduction & motivation
II. Research
    - Dataset acquision and corpus creation
        - NLP analysis
        - Context analysis
    - ML analysis
        - Classification
        - Category prediction
III. Conclusion

I Introduction

- Digital corpora are crucial for the study of phenomena in human languages, as they enable researchers access to a vast collection of textual data, which can be employed for building machine learning models that are applicable to several tasks
  - specialy crafted corpora can also be used for teaching purposes in various courses in higher education
- for the Croatian language there is an evident **lack of recent research in the field of phishing detection**, and especially in **research of corpora** that are essential for conducting analyses and for building models
  - high-quality and **domain-specific** textual corpora is of great importance

Phishing

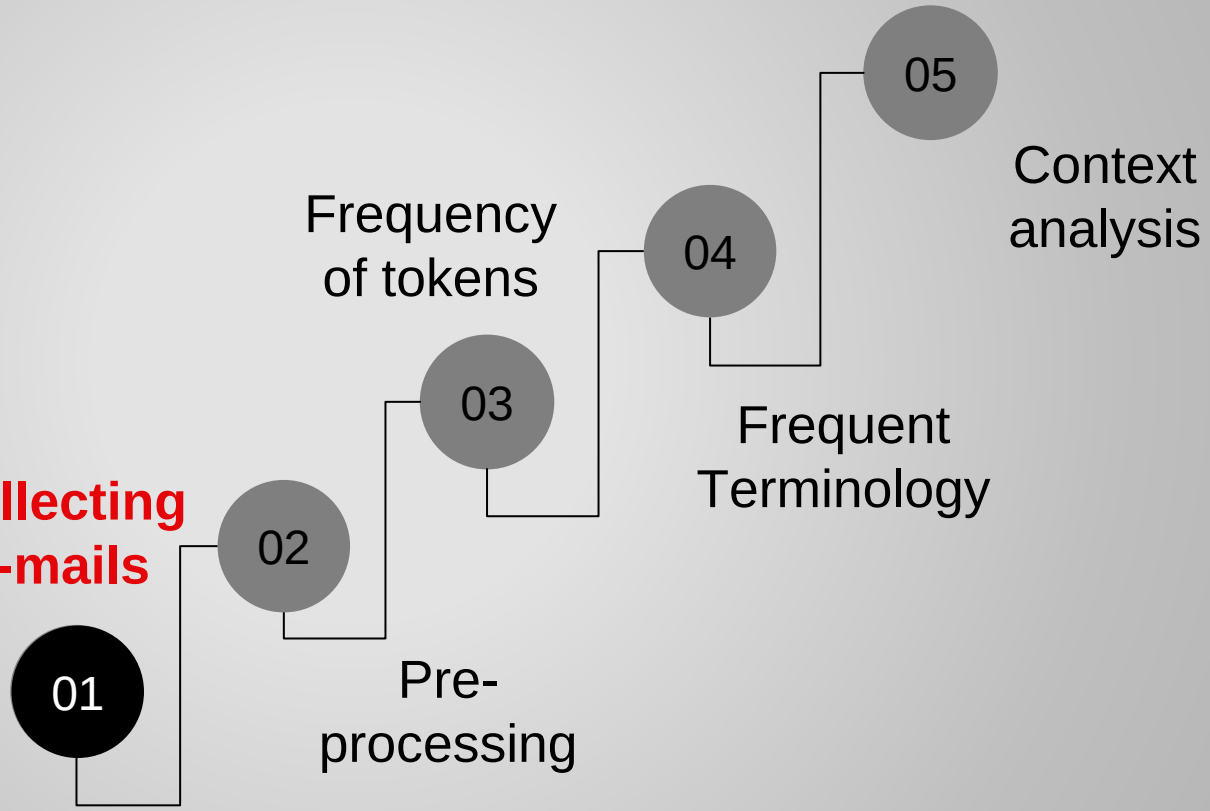
Phishing attack	a type of social engineering in which an attacker poses as a reliable entity in order to deceive a victim into disclosing sensitive information
Phishing detection	the effort to recognize and stop phishing attacks on people and business enterprises, and as such represents a crucial endeavor in the field of information and cyber security
Phishing datasets	purposefully crafted and prepared in order to account for linguistic and technical nuances

II Research - NLP

Set A
originally Croatian

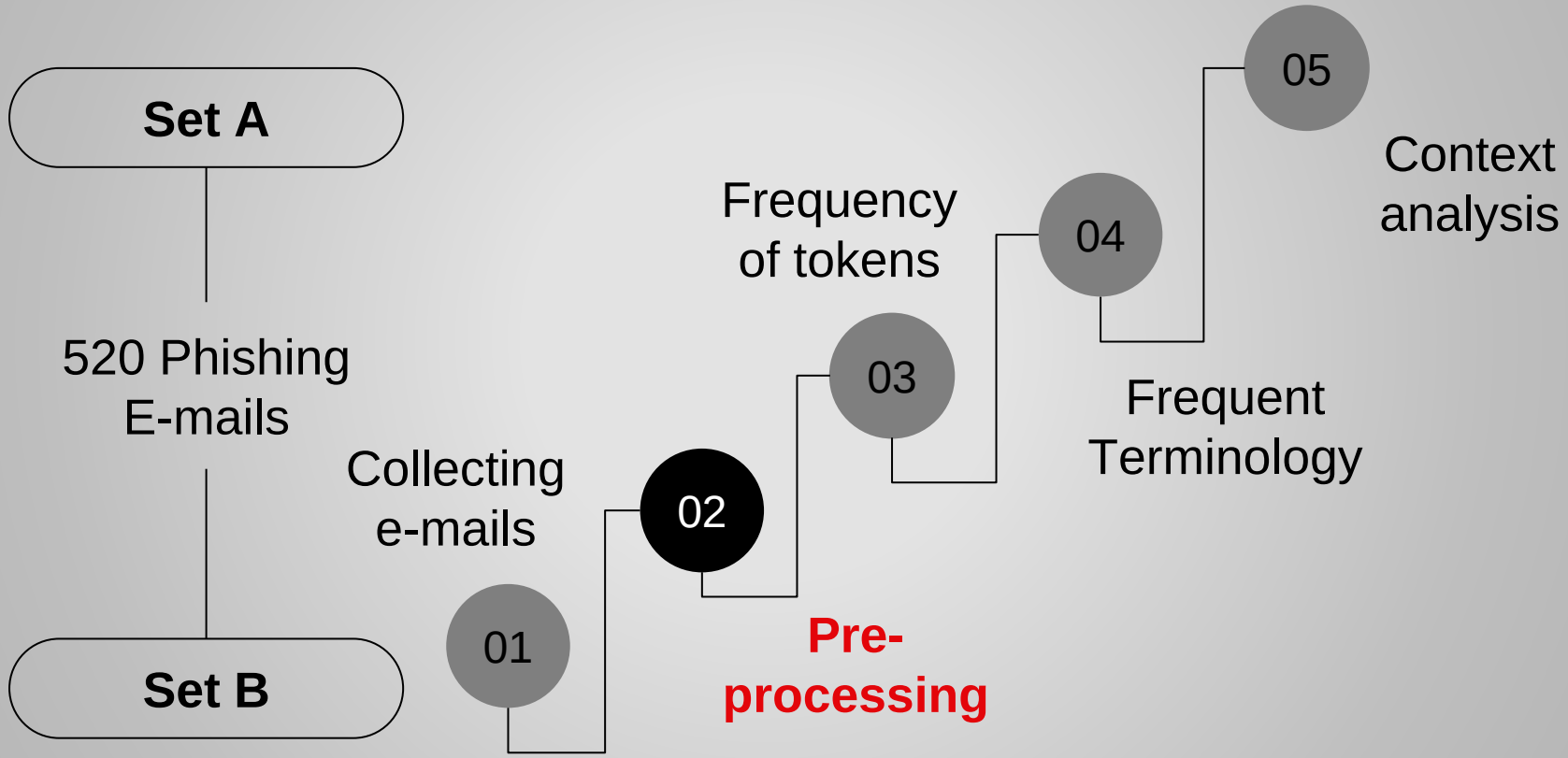
520 Phishing
E-mails

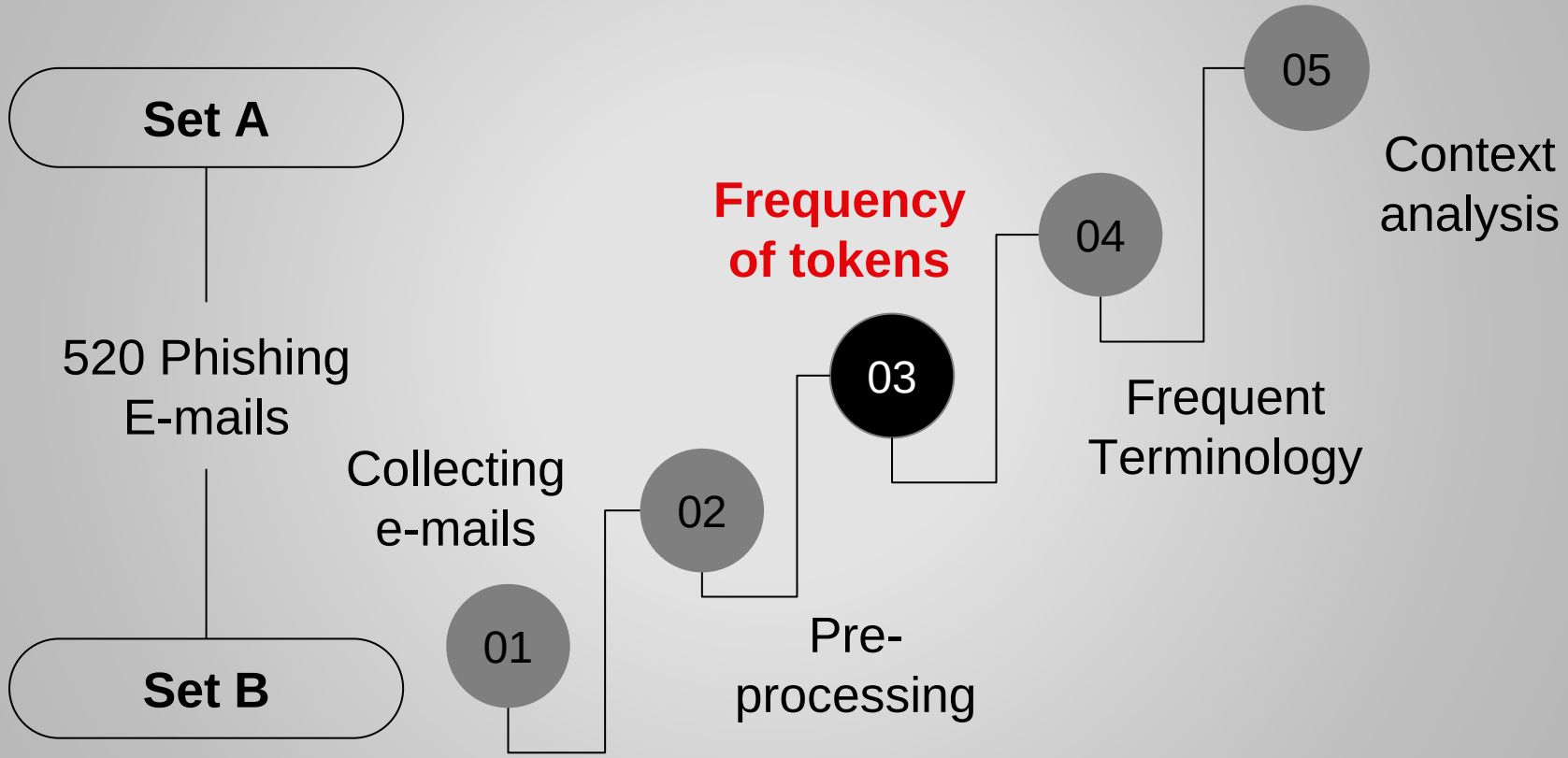
Set B
originally in English
and translated into
Croatian



Collecting e-mails

- dataset was acquired and crafted by extracting data that derived from several personal official e-mail accounts
  - **Set A** contains **260 e-mails originally written in Croatian** (not necessarily grammatically correct)
  - **Set B** contains **260 e-mails originally written in English and translated automatically into Croatian** with Google Translate and without any post-editing


Set A
520 Phishing E-mails
Set B
01
Collecting e-mails
02
Pre-processing
03
Frequency of tokens
04
Frequent Terminology
05
Context analysis


Set A
520 Phishing E-mails
Set B
01
Collecting e-mails
02
Pre-processing
Frequency of tokens
03
04
Frequent Terminology
05
Context analysis

Frequency of tokens

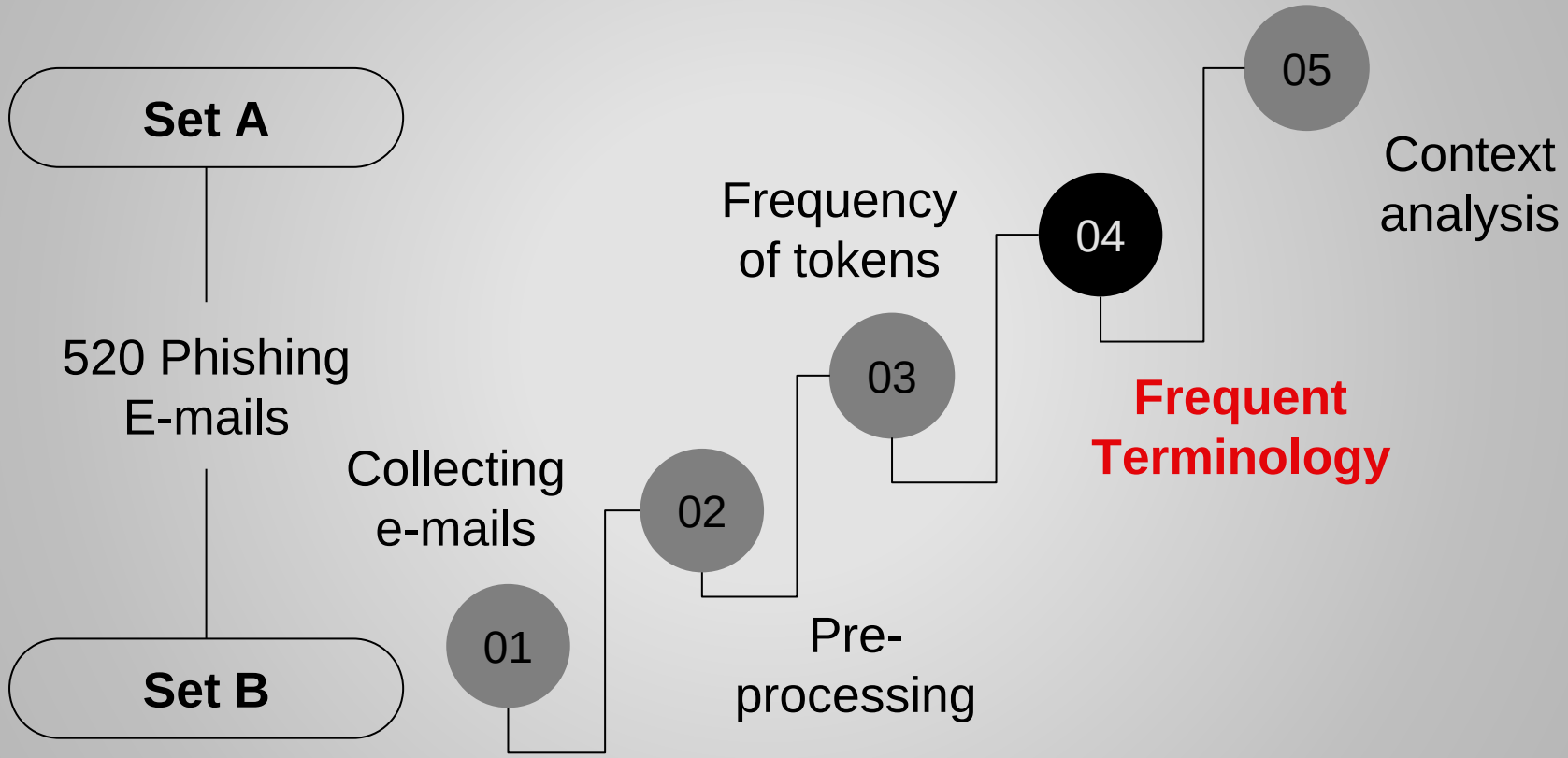
Set A

mostly nouns and vocabulary related to *financial fraud*, valuable data, verbs reflecting communication of *receiving and giving*

Set B

contains mostly *obscene vocabulary,* mentions websites and e-mails as means of communication, *transaction-related data*

Set A		Set B	
Nouns	**Verbs**	**Nouns**	**Verbs**
banka	kontaktirati	djevojka	željeti
fond	moći	e-pošta	tražiti
e-pošta	poslati	stranica	razgovarati
ime	moliti	račun	podijeliti
račun	dobiti	pratnja	poslati
adresa	željeti	usluga	kontakirati
sredstvo	primiti	pozdrav	moliti
broj	odgovoriti	transakcija	pogledati
podatak	pomoći	url	pozivati
pozdrav	dati	mreža	pomoći


Set A
520 Phishing E-mails
Set B
01
Collecting e-mails
02
Pre-processing
03
Frequency of tokens
04
Frequent Terminology
05
Context analysis

Frequent Terminology

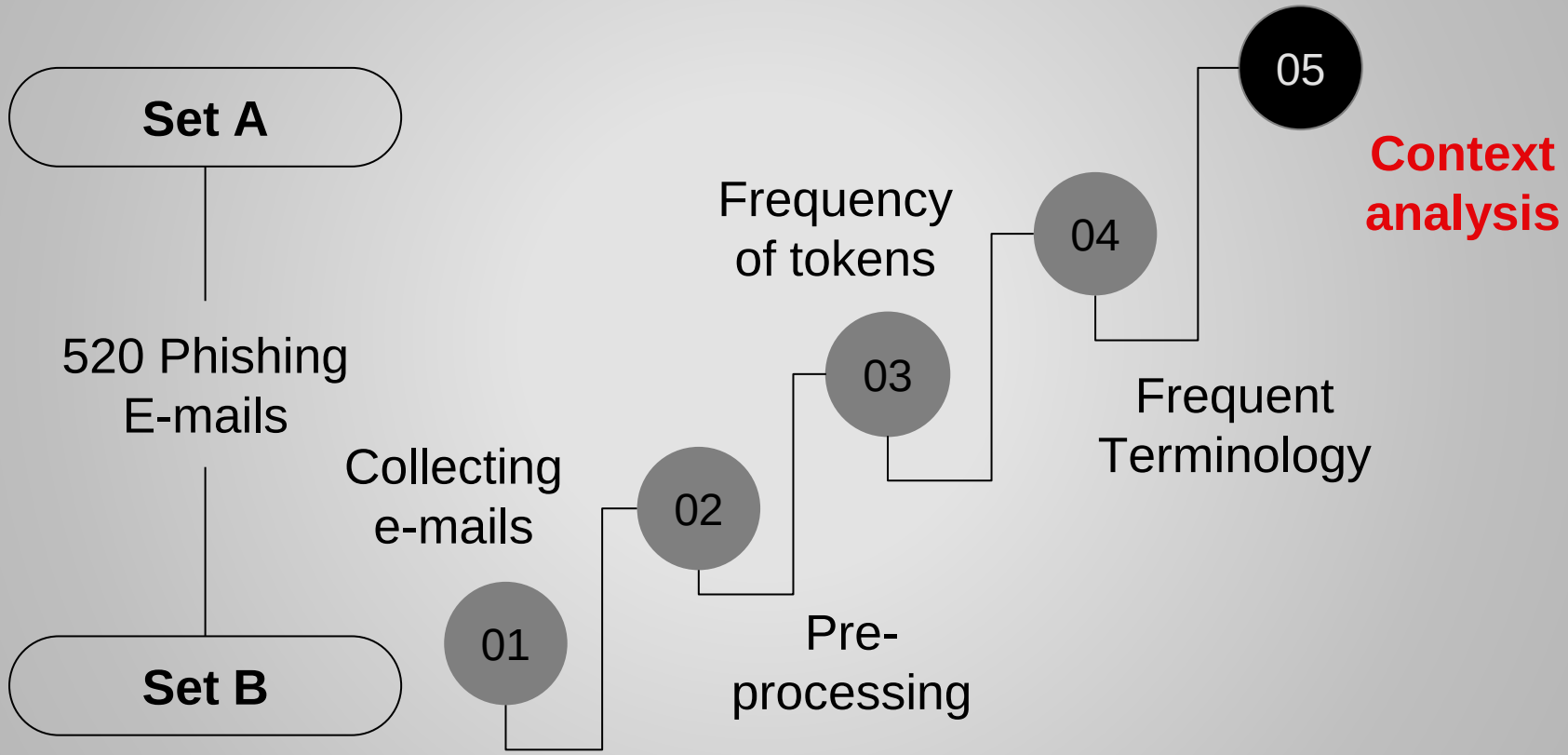
Set A

contains mostly nouns and vocabulary related to **financial fraud** (*banka – bank*), **valuable data** (račun – account), verbs reflecting *communication of receiving and giving, sending*, helping, replying, and mentions.

Set B

contains mostly **obscene vocabulary** (*djevojka – girl*), **mentions websites and e-mails** as *means of communication* (e-pošta – e-mail), **transaction-related data** (račun – account)

Set A		**Set B**	
Nouns	Verbs	Nouns	Verbs
banka	kontaktirati	djevojka	željeti
fond	moći	e-pošta	tražiti
e-pošta	poslati	stranica	razgovarati
ime	moliti	račun	podijeliti
račun	dobiti	pratnja	poslati
adresa	željeti	usluga	kontakirati
sredstvo	primiti	pozdrav	moliti
broj	odgovoriti	transakcija	pogledati


Set A
520 Phishing E-mails
Set B
01
Collecting e-mails
02
Pre-processing
03
Frequency of tokens
04
Frequent Terminology
05
Context analysis

Context analysis

- Concordance analysis - examining and comprehending context

Set A

-

often contain **urgency**, more focused on **financial fraud**, therefore using words such as *odmah – immediately* and *sada – now*

Set B

begging for help or asking for a conversation, hence the of use of verbs, such as *razgovarati – talk* and *podijeliti – share* (both used more than 40 times) accompanied by a **call to click on links t**hat were integrated into the text body of e-mails (95 times out of 260 e-mails).

Set A	Set B
odmah (44) – immediately	odmah (5) – immediately
savjetujemo da odmah pošaljete svoje ime, kontakt adresu i broj mobilnog telefona	razgovarati (44) – talk
ako budem plaćen, odmah ću uništiti video	pregledajte našu stranicu i **razgovarajte** -> https:
(MMF) Afričke regije odmah unutar sljedećih 168 sati.	privatne fotografije ***Razgovarajte*** sada (samo unesite svoju e-poštu)
Napomena: odmah po zaključku transakcije imate pravo na 45%	podijeliti (44) – share
kontaktirajte me odmah za daljnju komunikaciju	***Podijelit ću*** datoteku s vama
sada (60) – now	***Podijelit*** ću slike i više detalja o sebi čim mi se javite
Sada kontaktirajte Službu za korisnike UBA banke	sada (41) – now Možemo li ***sada*** uspostaviti video poziv

N-grams

- continuous sequences of words or symbols, or tokens

Set A

2-grams

poštovani korisniče – **dear user**, bankovni račun – **bank account,** banka africa – bank Africa, broj telefona – phone number, kartice adresu – **card address**, atm kartica – ATM card, pošaljite informacije – **send information**, milijun dolara – **million dollars**, pošaljite podatke – **send data**, visa kartice – Visa cards, prijenos sredstava – **money transfer**, najbliža rodbina – close relatives, odmah odgovorite – answer immediately etc.

3-grams

united bank africa – **united bank Africa**, atm kartice adresa – **ATM card address**, telefon godine spol – phone years gender, pošaljite podatke visa – **send Visa data**, iznos milijun tisuća – amount million thousands, poštovani korisniče računa/pošte – **dear account/mail user**, međunarodni monetarni fond – International Monetary Fund

Set B

2-grams

zgodna pratilja – **handsome companion**, povremeni spojevi **– occasional dates**, craigslist zakona – **craigslist law**, tajna zajednica – **secret community**, nabaviti zgodne – get a pretty, usluga pratnje – escort service

3-grams

noć povremene spojeve – night occasional dates, odjeljku craigslist zakona – craigslist law section, rastuća tajna zajednica – growing secret community, nabavite zgodne djevojke – find hot girls, djevojka na poziv – call girl, tajna zajednica nsa – secret NSA community, etc.

Typos on purpose

Spot the Difference?

maybank2u.com is not the same as
maybank2u.com

citibank.com is not the same as
citibank.com
(the first one is correct, the second one is from hackers)

The "a" in the later url is a cyrillic alphabet.

An average internet user can easily fall for this. Be careful for every mail requiring you to click on a link.

Please Stay Alert

Conclusion – part I

- **Set A** contains content related to finance, personal data, bank details, business offers, funding and payments
  - Mainly sent to undisclosed recipients
  - Messages often include e-mail addresses as a malicious means for connecting with victims
  - Content: finance, personal data, bank details, business offers, funding and payments

- **Set B** contains more mature and obscene content
  - Mainly sent to personal e-mail (addresses from web: research papers, speeches)
  - contain more **web links,** enticing potential victims to click on them and instructing victims to disclose private or sensitive information
  - Content: more mature and obscene content

Research - ML

Ensemble learning methods

Ensemble learning methods are used in machine learning to **combine predictions from multiple models** in order to improve predictive performance.

---

uses **different models** (e.g. Linear Regression for regression tasks, or Logistic Regression for binary classification) **on the same dataset** and **another model to learn how to make predictions**.

---

aim to **change the training data** in order to **focus on** examples that were **erroneously predicted** during previous fitting of models on the training dataset

---

III Ensemble learning methods

Bagging (Bootstrap Aggregation)

uses multiple *Decision Trees* on **different samples of the same dataset,** and then **averages predictions**. It is often used in statistics on **small datasets**, where many training datasets can be prepared to achieve an overall better estimation.

employs different members of an ensemble group; uses replacement, meaning that once the instance (row) is selected, it is returned, and can be selected again – used on small datasets

Random Forest – contains numerous Decision Trees on different subsets of a dataset - final decision based on the maximum votes of predictions (useful when individual trees are not correlated)

Ensemble learning methods

Stacking	uses different models (e.g. Linear Regression for regression tasks, or Logistic Regression for binary classification) **on the same dataset** and another model to learn how to make predictions.
Boosting	aims to change the training data in order to **focus on examples** that were **erroneously predicted** during previous fitting of models on the training dataset

Category	Description	Example
Finances	commercial phishing e-mails with the aim to scam users for monetary benefit	Obavijest: Vaša kompenzacijska bankovna kartica na bankomatu u vrijednosti od 1.500.000,00 dolara registrirana je kod voditelja kurira DHL-a g. Marka Adjovija. Radi trenutne isporuke kontaktirajte ga putem e-pošte (dhlcourierexpressbj1@outlook.com) za više informacija o tome kako ćete je zatražiti **Translation:** Your bank card valuable 1.500.000,00 is registered on DHL, by M.A. Please contact by mail (dhlcourierexpressbj1@outlook.com) how to ask your bank card:

Category	Description	Example
Health	offers and promises life changing products that do not exist, in order to acquire valuable personal information from victims	Naši vjerni kupci u posljednjih deset godina već su se uvjerili u učinkovitost prirodnih pripravaka Naturelle za poboljšanje zdravlja: Snaga daha štiti tijelo od svih vrsta sezonskih respiratornih alergija. Mountain Top uspješno povećava otpornost organizma na opasne viruse i bakterije. […] **Translation:** Our loyal customers in the last 10 years have been convinced that our products Naturell have successfully improved your health from respiratory alergies…..

Category	Description	Example
Adult content	e-mails containing erotic and raunchy content, or having allusions to pornographic topics, nudity, explicit sexual material or violence	Bok, ja sam lokalna djevojka iz vašeg područja. Možemo li se naći nasamo? Odgovori mi da ili ne na moju privatnu poštu >jadmesm256@gmail.com **Translation:** Hello, I am a local girl from the neigbourhood. Can we meet alone? Answer yes/ no on my mail
Short communication	short e-mails containing brief and generic greetings to encourage victims to continue a conversation in which they provide sensitive information to the attacker	Draga moja, jesi li primila poruku koju sam ti poslao? Pozdrav, Jerry Ngessan **Translation:** Hello dear, have you recived the message I have sent?

Dataset characteristics

A large amount of phishing e-mails focuses on the *financial aspect* of victims, asks for valuable data or directly for money, employs *blackmailing strategies*, presents *inappropriate services*, asks for *charity donations* or *promises lottery wins*.

	Set A		Set B	
	Training set	**Test set**	**Training set**	**Test set**
No. e-mails	**250**	25	**250**	25
Finances	89	11	90	10
Health	8	2	17	5
Adult content	94	8	103	9
Short communication	59	4	40	1

Prediction of categories

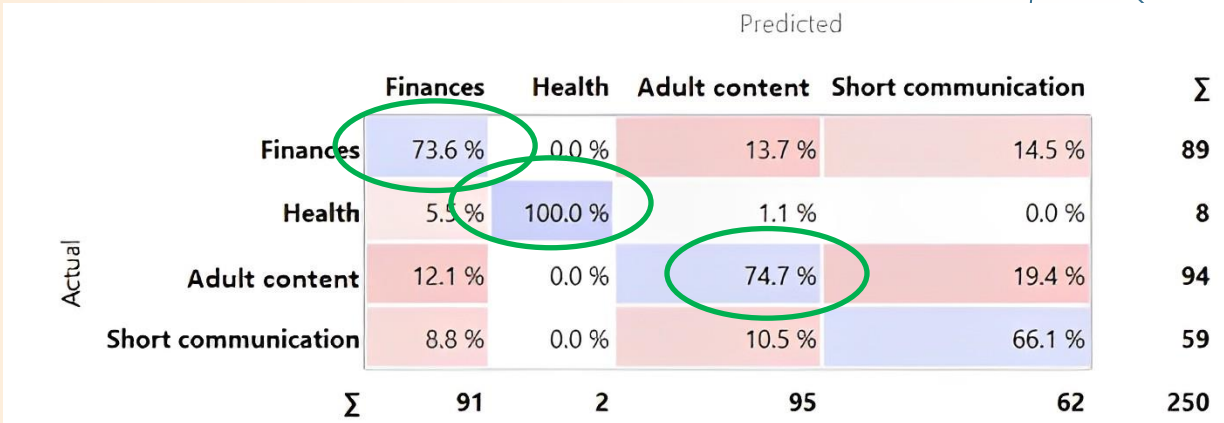
(Finances, Health, Adult content, Short communication)

When comparing prediction outcomes, accuracy and F1 measures achieved similar results. The highest predictions were achieved by the algorithms **Random Forest and AdaBoost**, ranging from ca. 71% up to ca. 77% for accuracy and F1 scores

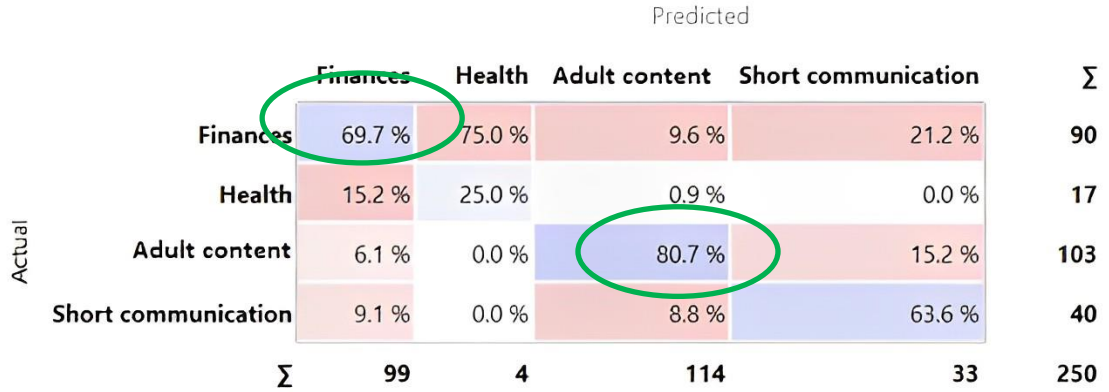
	Set A		Set B	
Algorithm	Accuracy	F1	Accuracy	F1
LR	0.684	0.686	0.680	0.675
RF	0.776	0.767	0.732	0.711
kNN	0.704	0.691	0.648	0.613
NB	0.460	0.454	0.448	0.453
AB	0.748	0.743	0.760	0.753

Confusion Matrix

The confusion matrix is a *predictive analytics tool* used widely in machine learning and in various classification tasks. It is a summarized NxN table that presents the **number of correct and incorrect predictions** in a classification task, and therefore **estimates the performance of a classification algorithm.**



Set A – Confusion Matrix

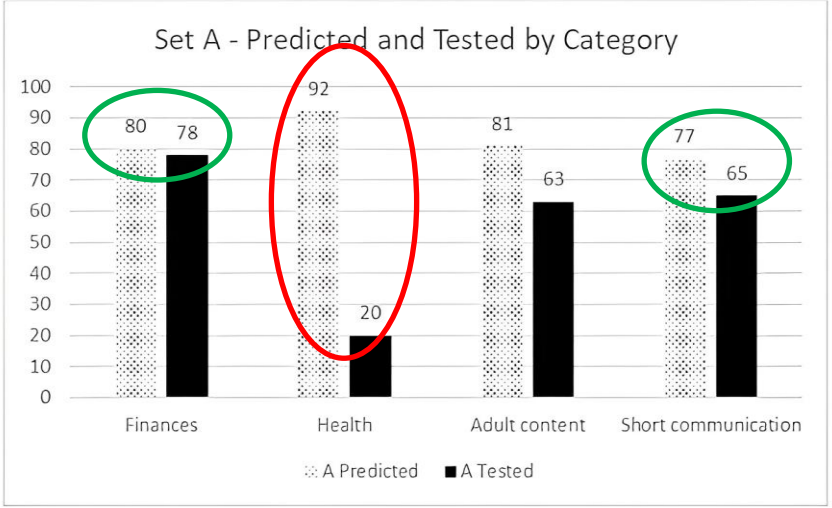


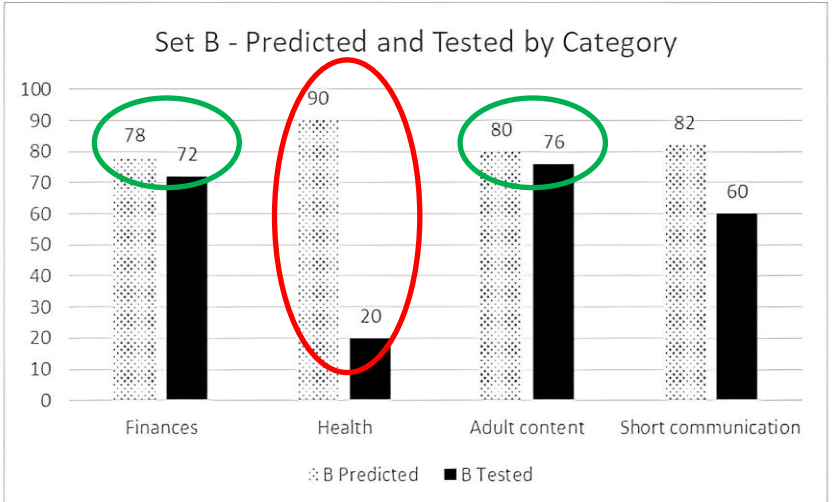
Set B – Confusion Matrix

Results of predicted and tested accuracies

- Table 1 - results of tested and predicted scores across all four categories for **Set A**
  - **accuracy** of tested scores is **lower** than accuracy of predicted scores, generally by 2-6%
- Table 2 - results of tested and predicted scores across all four categories for **Set B**
- Both - significant variations among categories

- **decline** of **tested accuracy** scores in all categories, but **smallest** in *Finances* (most represented)
- the most significant decline was in *Health* category (72% in Set A and 70% in Set B), with the smallest amount of training data





Conclusion – part II

- Research on two sets, differentiated by **language originality**
- Training and test sets were **imbalanced in terms of categories**
  - most represented: „Finances” and “*Adult content*”
  - least represented: “*Health*”
  - all of phishing e-mails deal more or less with financial issues, and therefore can belong to **multiple categories -** main limitation of this research
- Results in this paper confirm that **predicted accuracy increases with data quantity (best predictions were obtained by Random Forest and AdaBoost**.
- best predictions are obtained for categories that were most represented in datasets, whereas worst results were obtained for the least represented category “Health”

- In order to improve accuracy results -> increase sample size, equalize the number of phishing e-mails that are used for in training and test sets, and harmonize the length of phishing messages.