PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Luigi Gallo

XXXIV Cycle

Training and Research Activities Report – Third Year

Tutor: Prof. Alessio Botta

Template

1. Information
    a. Name Surname, MS title – University
    b. XXIX Cycle- ITEE – Università di Napoli Federico II
    c. Fellowship type
    d. Tutor
2. Study and Training activities
    a. Courses
    b. Seminars
    c. External courses
3. Research activity
    a. Title
    b. Study
    c. Research description
    d. Collaborations
4. Products
    a. Publications
        i. Books, Book Chapters, Journal papers, Conference papers (mark international products)
        ii. List those in preparation
    b. Patents
5. Conferences and Seminars
    a. Details (Conference name, place, dates, number of papers)
    b. Presentations made
6. Activity abroad
    a. Details (Place, dates, number of papers, contact persons)
7. Tutorship
    a. Type, subjects, hours

1 – INFORMATION

Luigi Gallo

Computer Engineering

Università di Napoli Federico II

PhD in Information Technology and Electrical Engineering

Università di Napoli Federico II

No Fellowship

Tutor: Prof Alessio Botta

2 – STUDY AND TRAINING ACTIVITIES

The study and training part (of the third year) includes the following short courses and seminars.

Year	Seminar Title	Credits	Lecturer	Lecturer affiliation	Organization
3	AI4NETS–AI/ML for datacommunication Networks | Tutorial	0.6	Dr. Pedro Casas	Austria Institute of Technology	Politecnico di Torino
3	"Images, Texts, Emojis and Geodata in Sentiment Analysis Pipeline"	0.3	Dr. Serena Pelosi	University of Salerno	University of Napoli Federico II
3	"Data Driven Transformation in WINDTRE through Managers voice	0.4	Marcello Savarese	WindTRE	University of Napoli Federico II
3	AI LEGAL: Artificial Intelligence for notary's sector – a case study	0.2	Salvatore Palange		University of Napoli Federico II
3	"Approaches to Graph Machine Learning"	0.2	Miroslav Cepek		University of Napoli Federico II

3	"The coming revolution of Data driven Discovery (a fourth Methodological Paradigm of Science)"	0.3	Prof. Giuseppe Longo	University of Napoli Federico II	University of Napoli Federico II
3	"Distributional Semantics Methods: How Linguistic features can improve the semantic representation"	0.4	Alessandro Maisto	University of Salerno	University of Napoli Federico II
3	"5G: l'architettura, le applicazioni la rete di accesso radio"	0.4	Ing. Mollica Francesco	Vodafone	University of Napoli Federico II

Moreover, I attended the following internal courses and doctoral schools.

Year	Module Title	Type	Credits	Lecturer	Organization
3	"HOW TO BOOST YOUR PHD"	External module	3	Dr. Antigone Marino	University of Napoli Federico II
3	IMPRENDITORIALITA' ACCADEMICA	Ad hoc module	4	Prof. Pierluigi Ripa	University of Napoli Federico II

	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	20	1,6	0	3	0	6	5	15,6	14	0	3	0	9	2,8	0	14,8	10	0	0	3	4	0	0	7	37,4	30-70
Seminars	5	0,4	0	0,4	1,5	0	0	2,3	6	0,4	0	1,6	2,7	4	1,4	10,1	7	1,3	0,4	0,7	0,4	0	0	2,8	15,2	10-30
Research	35	8	10	6,6	8,5	4	5	42,1	40	9,6	7	8,4	3,3	0,7	6,1	35,1	43	8,7	9,6	6,3	5,6	10	10	50,2	127,4	80-140
	60	10	10	10	10	10	10	60	60	10	10	10	15	7,5	7,5	60	60	10	10	10	10	10	10	60	180	180

3 – RESEARCH ACTIVITY

Analysis and identification of cyber threats and frauds in the mailboxes of large companies

My research activity concerns the study and experimentation aimed at the design, development and testing, on real contexts, of the defense systems against current and future cyber attacks. The "real context" is made available by the collaboration

with the Cyber Security Lab of TIM S.p.A (Telecom Italia Lab, Turin), which provides real data and environments (in compliance with current regulations on privacy).

During the first year, the first step was to study the basic knowledges, open issues, and the challenges to be faced by this type of research work. During this phase I have identified a major point to focus on, in order to reach the origin of a large number of cyber attacks hurting people and companies: the identification of cyber threats in the mailboxes. In the second year this branch of research was strongly developed producing interesting results, which have been refined, validated, and expanded in this third year. To support this work I had to further deepen the studies already started in the first and second year, about Machine learning, Big Data Analysis and Cyber Security.

The context is the following: the email threat landscape is constantly evolving, making current countermeasures ineffective in protecting companies, especially because actually dangerous emails are able to evade carrier-grade spam filters and also deceive users. For this reason, Email is still one of the most used channels for making cyber attacks. Several law enforcement bodies (e.g. FBI, EUROPOL) and data protection agencies are constantly raising alarms in this regard, as more than 80% of financial fraud is executed by email causing huge monetary losses. The outcome is that companies typically rely on teams of security analysts to perform manual inspection on such emails. However, spam emails that pass the spam filter check, especially in the case of large companies, are too many for such analysis to be effective. This research project aimed at providing a contribution to this important problem and leverages the collaboration with TIM S.p.A.

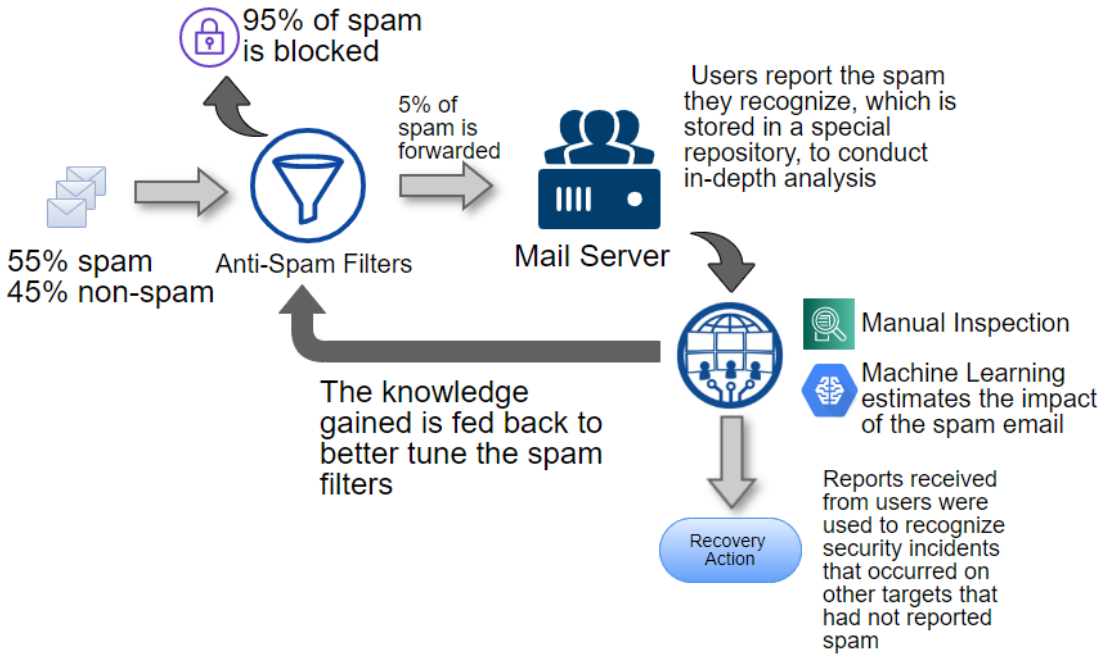

Figure 1 Collaborative Ecosystem

I designed a collaborative framework (Figure 1) supporting security analysts of the company in collaboration, to analyze the several malicious emails that evade the spam filter and cause security incidents. Thanks to this framework we collected a large labeled dataset, composed of real spam emails received in the company, each classified as critical or not relevant. Using this labeled dataset I have shown that some machine learning algorithms, with a properly designed feature set, can well identify the emails actually threating the security of the company. The complete set of features is shown in Table 1. I have also identified the main features that make a spam email dangerous and the best techniques and technologies to rely on for the defense. I have used both legacy and novel features and evaluated their relevance and correlation with the target. Using the best feature set maximizing the f1-score performance, the supervised approaches reaches 95.2% of precision and 91.6% of recall. I have also identified a reduced feature set that greatly reduce costs with a small impact on the performance.

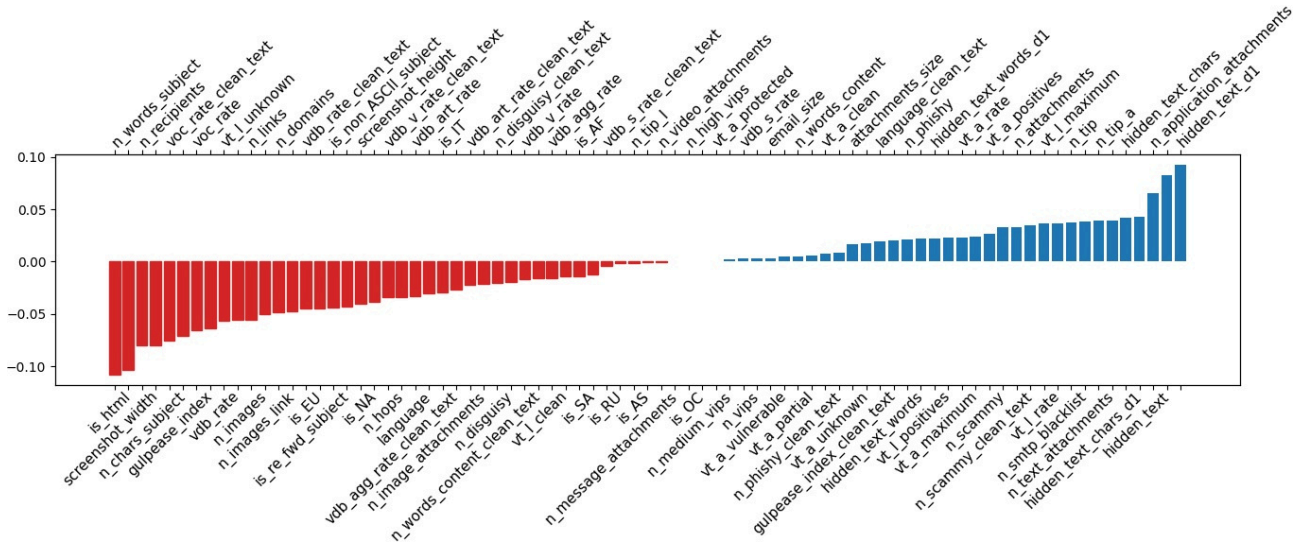

Figure 2 Feature importance with wrapper method (SVM classifier as Wrapper)

The feature ranking work (Figure 2 shown the results) also provides an important explanation on how critical emails are built and can be detected. This knowledge led to the design of a week-long awareness campaign, which involved all 40,000+ employees of the partner company, including top managers and executives. This large social experiment confirmed that the developed system correctly models the phishing phenomenon and, together with well-trained people, represents a global defence ecosytem robust to the majority of email attacks.

In the last few months of this third year, I have been further investigating the cognitive aspects of phishing attacks, which vary from person to person. Therefore, I have begun collecting new specific data in order to subsequently infer about the correlation between certain personality traits with certain vulnerabilities to characteristics that fraudulent emails may have. As future work, I plan to release this new dataset to the scientific community.

Table 1 Features extracted from the raw data

Field	Feature	Description
General	is_html	if it is an html mail
	n_smtp_blackist	the number of smtp servers traversed in the blacklists
	email_size	the size of the email
	n_recipients	the number of recipients
	n_hops	the number of SMTP hops
	is_IT	if the email comes from Italy
	is_EU	if the email comes from Europe
	is_NA	if the email comes from North America
	is_SA	if the email comes from South America
	is_RU	if the email comes from Russia
	is_AS	if the email comes from Asia
	is_AF	if the email comes from Africa
	is_OC	if the email comes from Oceania
Content[3]	language[3]	the language of the mail
	voc_rate[3]	the rate of words of the content in the vocabulary
	vdb_rate[3]	the rate of words of the content within the basic vocabulary
	vdb_agg_rate[3]	the rate of adjectives within the content
	vdb_v_rate[3]	the rate of verbs within the content
	vdb_s_rate[3]	the rate of nouns within the content
	vdb_art_rate[3]	the rate of articles within the content
	gulpease_index[3]	readability index (Italian - Gulpease index [27], English - Flesch formula [15])
	n_words_content[3]	number of words in the content
	n_disguisy[3]	number of disguised words in the entire email (content, subject, address)
	n_phishy[3]	number of deceiving words, related to phishing, in the content and subject
	n_scammy[3]	number of deceiving words, related to scamming, in the content and subject
View	screenshot_width	the width of the email as it is displayed to the recipient
	screenshot_heigth	the heigth of the email as it is displayed to the recipient
	n_images	number of images
	n_images_links	number of images as links
	hidden_text[4]	percentage of text in the content not displayed to the recipient
	hidden_text_words[4]	number of words in the content not displayed to the recipient
	hidden_text_chars[4]	number of characters in the content not displayed to the recipient
Subject	n_words_subject	number of words in the subject
	n_char_subject	number of characters in the subject
	is_non_ASCII_subject	if the object contains non-ASCII characters
	is_re_fwd_subject	if the email is replied or forwarded
Links	n_links	number of links
	n_domains	number of link domains
	vt_l_rate	rate of links considered malicious by at least one engine of VirusTotal
	vt_l_maximum	maximum number of VirusTotal engines that consider a link as malicious
	vt_l_positives	number of links considered malicious by at least one engine of VirusTotal
	vt_l_clean	number of links not considered malicious by all engines VirusTotal
	vt_l_unknown	number of unknown links to VirusTotal
Attachments	n_attachments	number of attachments
	n_image_attachments	number of image type attachments
	n_application_attachments	number of application type attachments
	n_message_attachments	number of message type attachments
	n_text_attachments	number of text type attachments
	n_video_attachments	number of video type attachments
	attachments_size	average size of attachments
	vt_a_rate	rate of attachments considered malicious by at least one engine of VirusTotal
	vt_a_maximum	maximum number of VirusTotal engines that consider an attachment as malicious
	vt_a_positives	number of attachments considered malicious by at least one engine of VirusTotal
	vt_a_clean	number of attachments not considered malicious by all VirusTotal engines
	vt_a_vulnerable	number of attachments considered malicious by VirusTotal engines not including corporate antivirus
	vt_a_partial	number of attachments considered partially malicious by VirusTotal engines not including corporate antivirus
	vt_a_protected	number of attachments considered malicious by VirusTotal engines including corporate antivirus
	vt_a_unknown	number of unknown attachments to VirusTotal
Other	n_tip	number of entities in TIP
	n_tip_a	number of attachments in TIP
	n_tip_l	number of links in TIP
	n_vips	the number of vips among the recipients
	n_medium_vips	the number of managers among the recipients
	n_high_vips	the number of top managers among the recipients

Together with the main activity explained above, for a broader view of cyber security problems, I also conducted research activities on Malware Analysis, Anomaly detection in network traffic, security in 5G mobile networks and Cyber-Physical sistems, and Cloud Robotics.

4 – PRODUCTS

Products of the third year:

International journal papers

Luigi Gallo, Alessandro Maiello, Alessio Botta, Giorgio Ventre, 2 Years in the anti-phishing group of a large company, Computers & Security, Volume 105, 2021, 102259, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2021.102259.

International Conference and journal papers in preparation

"DewROS: a Platform for Informed Dew Robotics in ROS" (tentative title)

"Security testing methodologies for Network Traffic Analyzers" (tentative title)

"A game-based platform for phishing awareness testing" (tentative title)

5 – CONFERENCE AND SEMINARS

Attended MedComNet 2021 : 19th Mediterranean Communication and Computer Networking Conference

Attended Italian Networking Workshop (INW 22)

6 – ACTIVITY ABROAD

The pandemic did not allow for any periods spent abroad, nevertheless the research activities were carried out with multiple remote collaborations with foreign institutes

7 – TUTORSHIP

I have been involved as assistant to the exercises of the courses of "Fondamenti di Informatica" and "Computer Networks" (20 + 20 hours), for which I have also prepared some course materials.

In addition, I follow as co-advisor the preparation of the thesis by four MS students.