PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student:

Cycle:

Training and Research Activities Report

Year: First

Marco Varlese

Tutor: Prof. Simon Pietro Romano

Co-Tutor:

Date: October 31, 2024

Training and Research Activities Report

PhD in Information Technology and Electrical Engineering

Cycle: **Author:**

1. Information:

- **PhD student:** Mario Varlese
- **DR number:** DR997197
- **Date of birth:** 13.03.1996
- **Master Science degree:** Computer Engineering **University:** Federico II
- **Doctoral Cycle:** XXXIX
- **Scholarship type:** PNRR - DM 118/2023 Dottorati Pubblica Amministrazione (CUP: E66E23001050002)
- **Tutor:** Prof. Simon Pietro Romano
- **Co-tutor: //**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Virtualization Technologies and their applications	Course	23	5	08/01/2024 - 26/02/2024	Prof.. Luigi De Simone	Y
Using Deep Learning properly	Course	12	4	23/01/2024 – 08/02/2024	Dr. Andrea Apicella	Y
Strategic Orientation For STEM Research & Writing	Course	24	5	07/12/2023 - 23/02/2024	Dr Chie Shin Fraser	Y
Big Data Architecture and Analytics	Course	21	5	06/05/2024 – 31/05/2024	Prof. Giancarlo Sperlì	Y
Hands-on Network Intrusion Detection via Machine and Deep Learning	Course	11	4	09/01/2024 - 18/01/2024	Dr. Antonio Montieri	Y
Computer Forensics	Course	48	6	01/03/2024 - 18/06/2024	Prof. Lourenzo Laurati	Y
Software Security	Course	48	6	01/03/2024 - 18/06/2024	Prof. Roberto Natella	Y
Statistical Data Analysis for Science and Engineering research	Course	12	4	15/02/2024 – 29/02/2024	Prof. Roberto Pietrantuono	Y

Open and programmable 6G networks in the cloud/edge continuum: research challenges and experimentation tools in SLICES Research Infrastructures	Doctoral School	40	6	07/07/2024 – 13/07/2024	Sergio Palazzo @ https://netprog24.lipparischool.it/	Y
Introduction to Multi-Agent Reinforcement Learning	Seminar	2	0.4	21/11/2023	Prof. Beniamino Di Martino	Y
Generative Power of Deep Learning Variational Auto-Encoders and Generative Adversarial Networks	Seminar	2.5	0.5	14/12/2023	Prof. Beniamino Di Martino	N
Blockchain & IoT	Seminar	3	0.6	15/12/2023	Meditech - segreteria@meditech4.com	N
Media Forensics in the era of Generative AI	Seminar	2	0.4	11/12/2023	Prof. Luisa Verdoliva	Y
Analytic center selection of optimization-based controllers for robot ecology	Seminar	1	0.2	09/04/2024	Prof. Bruno Siciliano	Y
Complexity of Quantum Computing vs. Complexity of Classical Computing	Seminar	2	0.4	13/03/2024	Prof. Beniamino Di Martino	Y
Current Limitations of Quantum Computing	Seminar	2	0.4	27/03/2024	Prof. Beniamino Di Martino	Y
Exploring the Frontiers of Modern Cryptography	Seminar	1.5	0.3	12/04/2024	Prof. Simon Pietro Romano	Y
Modellazione di sistemi con linguaggio UML	Seminar	2	0.4	16/04/2024	Prof. Beniamino Di Martino	Y
Motivation and Basic Ideas of Quantum Computing	Seminar	2	0.4	13/03/2024	Prof. Beniamino Di Martino	Y
Intelligenza Artificiale e Regole Del Mercato	Seminar	2	0.4	14/05/2024	Prof Maria Antonia Tulino	Y

Verso una gestione intelligente delle risorse idriche con il supporto dell'innovazione digitale	Seminar	1	0.2	14/05/2024	Prof Maria Antonia Tulino	Y
Resource management and orchestration for mixed-criticality cloud/distributed systems	Seminar	1	0.2	27/06/2024	Prof. Marcello Cinque	Y
The maximal covering location problem with edge downgrades	Seminar	1	0.2	28/06/2024	Proff. Claudio Sterle, Maurizio Boccia, Adriano Masone	Y
Applicazioni dell'intelligenza artificiale ai sistemi informativi del ministero di giustizia	Seminar	1	0.2	09/05/2024	Prof. Beniamino Di Martino	N
Nato Cyber Coalition	Research	40	3	27/11/2023 - 01/12/2023	COR (Comando per le Operazioni in Rete)	Y

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**0**	**1.9**	**8.1**	**0**	**10**
Bimonth 2	**14**	**0**	**2.5**	**0**	**16.5**
Bimonth 3	**8**	**2.1**	**2.5**	**0**	**12.6**
Bimonth 4	**17**	**1.2**	**1.5**	**0**	**19.7**
Bimonth 5	**6**	**0**	**2**	**0**	**8**
Bimonth 6	**0**	**0**	**1.8**	**0**	**1.8**
Total	**45**	**5.2**	**18.4**	**0**	**68.6**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

During my first year of PhD, I researched by collaborating with ***Direzione Nazionale Antimafia and Antiterrorismo (DNAA)****. The outcome of this collaboration has been finalized in the implementation of several Proof Of Concept (PoC), listed below:*

Named Entity Recognition from unstructured sensitive documents *(PoC n. 1)* and Design of a retrieval-efficient document repository *(PoC n. 2)*

The process of extracting and managing information in the fourth industrial revolution era is far from a trivial task. Complicating matters further is the substantial amount of data available to work with. The problem is clear and well-recognized within the context of the Direzione Nazionale Antimafia e Antiterrorismo (DNA) as well. Indeed, the huge amount of data extracted from documents needs to be quickly and accurately analyzed to support the investigative process. Revisiting and improving existing systems and processes involved in day-to-day operations can significantly enhance efficiency and productivity while reducing human errors caused by repetitive activities. Machine Learning architectures are perfectly aligned with this context mainly for their capabilities to process and extract knowledge from data, easing the workload of human operators; and allowing them to focus on the investigation activities. With this purpose in mind, and to demonstrate how human operators can exploit artificial intelligence capabilities, more specifically ChatGpt, in their day-to-day work activities we have performed several experiments on manual information extraction. Such experiments have highlighted how a “scribe” intending to leverage Large Language Models (LLMs) to automate part of their work.

1. Extract and copy the text from the document to be submitted to the artificial intelligence model;
2. Define a Prompt, which is the ”command ” to be submitted to the artificial intelligence model. It can be expressed in the form of a question, providing a structure of the response and any examples to support the generation of a solution;
3. Extract the model’s response and verify that it is correct.

Although the described process is feasible, it is indeed inapplicable concerning DNA. The main motivation that forbids the application of this kind of approach is related to the sensitivity of information. Thus, the information contained in investigation documents is often kept confidential during the investigation phase and cannot be stored or processed on external platforms, not previously approved by DNA. 2. To face the problems described, PoC n.1 and n. 2 suggest:

- The creation of a platform enabling the extraction of entities of interest and the Human In The Loop to evaluate and retrain the model;
- The utilization of locally installed open-source LLM models to ensure data handling in secure environments, while simultaneously minimizing the risk of information leakage;
- The development of an intuitive dashboard allowing the use of the proposed platform even by non-expert users.

Entity linking / relationships extraction from unstructured text *(PoC n. 3)*

The "National Anti-Mafia and Anti-Terrorism Directorate (DNAA)" plays a fundamental role in combating organized crime and terrorism in Italy, coordinating investigative activities on a national scale. This is not limited to the operational management of investigations but also includes activities aimed at countering organized crime, such as the mafia, drug trafficking, money laundering, and human trafficking. National Deputy Prosecutors, in addition to coordinating investigations within various "DDA" (District Anti-Mafia Directorates), are tasked with following these matters, collecting and analyzing information useful for developing new investigative strategies. A supporting tool for investigative activities is the "National Database," ("*Banca Danti Nazionale" - BDA)* fed by the procedural documents of the DDAs.

In this context, PoC No. 3 fits as the natural continuation of the previous phases, to further extend the capabilities of the system through the integration of Large Language Models (LLMs). Using the entities extracted in PoC No. 1, the primary goal was to demonstrate the effectiveness of these models in automating the complex task of extracting interconnections of interest among the entities involved. More specifically, this work contributes across four areas:

- Evaluating the effectiveness of Large Language Models in the so-called Entity Linking / relationship extraction task;
- Analyzing the effectiveness of strategies like Chain of Thought and Rethinking in the context of extracting relationships from legal texts in Italian;
- Defining an effective processing flow (or pipeline) for extracting relationships of interest, which allows not only identifying such relationships but also understanding the reasons behind the model's responses (explainability of decisions);
- Identifying the appropriate technology to collect the extracted information to create a knowledge base used for detecting non-immediate connections among the entities under investigation.

Multimedia processing and automatic transcription *(PoC n. 4)*

Artificial Intelligence (AI) has revolutionized numerous sectors, offering advanced solutions for automation, data analysis, and enhancing operational efficiency. Automatic transcription technologies are gaining attention for their ability to convert speech into text with high precision. These technologies are being applied across various fields, such as healthcare, customer service, media, and the justice system, enhancing accessibility and enabling more effective management of audio data. The use of AI for automatic transcription not only reduces manual workload but also increases the quality and speed of the process, providing significant benefits in terms of productivity and accuracy. Despite these advantages, many organizations still rely on manual processes, which require considerable time and resources, and increase the likelihood of human error. Managing large volumes of audio data in this manner proves to be inefficient and costly. Within the DNAA, multimedia transcriptions are an essential component of investigations. Accurate transcription of conversations can provide crucial evidence and enhance the quality of investigations. However, the traditional method of manual transcription is labor-intensive and prone to errors, slowing down the investigative process and demanding substantial effort from staff. Moreover, the growing volume of audio data to be analyzed necessitates the adoption of more advanced and automated solutions. The DNA thus requires tools that can improve the efficiency and accuracy of transcriptions, allowing

investigators to focus their resources on more critical tasks. There is a clear need to automate this process to improve operational efficiency. The proposed solution is a Proof of Concept (PoC) that utilizes Whisper technology for the automatic transcription of audio files. This system is designed to be scalable, user-friendly, and efficient, significantly reducing manual workload and minimizing errors. The PoC demonstrates how AI can be integrated into the DNAA's investigative processes to enhance the productivity and accuracy of transcriptions.

4. Research products:

The above research activities are being adapted to be prepared and submitted to the scientific community.

5. Conferences and seminars attended

- ***HackInBo*** *Computer Security Conference (https://www.hackinbo.it/) – Bologna, 08 June 2024*
- ***RomeHack*** *Computer Security Conference (https://romhack.io/) – Roma, 28 September 2024*
- ***MOCA*** *Computer Security Camping (https://moca.camp/) – Teramo, 13-15 September 2024*
- ***NoHat*** *Computer Security Conference (https://www.nohat.it/) – Bergamo, 19 October 2024*

6. Activity abroad:

//

7. Activity in partner companies:

The three research activities described above have been carried out by collaborating with "*Direzione Nazionale Antimafia e Antiterrorismo" (DNA).* Several meetings and discussions were conducted to understand their actual systems in the context of the refactoring of their database, "*Banca Dati Nazionale*" (BDA). Our main objective was to demonstrate the applicability of cutting-edge technologies to their processes through a prototypal development approach. This has led to the realization of the three prototypes (*Proof Of Concepts* - PoC) described in section 5. During this year, in-person activities in their headquarters in Rome were needed to better understand their systems. Below are detailed the periods:

- May 2024 – June 2024
- October 2024 – November 2024

8. Tutorship

//