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

PhD Student: Valerio La Gatta

Cycle: XXXVI

Training and Research Activities Report

Academic year: 2021-22 - PhD Year: Second

Tutor: prof. Vincenzo Moscato

Co-Tutor:

Date: October 26, 2022

Training and Research Activities Report

PhD in Information Technology and Electrical Engineering

Cycle: XXXVI **Author: La Gatta Valerio**

1. Information:

- **PhD student: Valerio La Gatta** **PhD Cycle: XXXVI**
- **DR number: DR995141**
- **Date of birth: 15/01/1996**
- **Master Science degree: Computer Engineering** **University: University of Naples Federico II**
- **Scholarship type: UNINA**
- **Tutor: prof. Vincenzo Moscato**
- **Co-tutor:**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Web and Real Time Communication Systems, prof. Simon Pietro Romano, Corso di Laurea Magistrale in Ingegneria Informatica	Course	48	6	20/09/2021 – 17/12/2021	Prof. Simon Pietro Romano	N
Cyber security in Akka Technologies	Seminar	2	0.4	03/11/2021	Proff. Domenico Cotroneo, Roberto Natella, Simon Pietro Romano	Y
Possible Quantum Machine Learning Approaches in HEP	Seminar	2	0.4	12/11/2021	Prof.ssa Angela Sara Cacciapuoti	Y
Single cell omics leverage Machine Learning to dissect tumor microenvironment and cancer immuno editing,	Seminar	2	0.4	02/12/2021	Prof.ssa Anna Corazza	Y
The learning landscape in deep neural networks	Seminar	1	0.2	21/01/2022	Prof. Michele	Y

and its exploitation by learning algorithms					Ceccarelli	
The quest of quantum advantage with a photonics platform	Seminar	1	0.2	03/02/2022	PHD programs in Advanced Mathematics and Physical Sciences for Advanced Materials and Technologies	Y
Project Vac: Can a Text-to-Speech Engine Generate Human Sentiments?, 28/02/2022	Seminar	1	0.2	28/02/2022	Picariello Lectures on Data Science	Y
From basic principles in spintronics to some recent developments toward spin-orbitronics	Seminar	1	0.2	31/03/2022	Scuola Superiore Meridionale	Y
Towards a Political Philosophy of AI	Seminar	1	0.2	11/04/2022	Picariello Lectures on Data Science	Y
Big Data Architecture and Analytics	Ad hoc Course	18	5	06/04/2022 – 11/05/2022	Prof. Giancarlo Sperlì	Y
5G Networks in Action – The Private Mobile Era	Seminar	1	0.2	11/05/2022	5G Academy’s Seminar Series	Y
Teaching activities regarding practical lectures/seminars during the courses of “Big Data Engineering” and “Machine Learning and Big Data per la salute”	Tutorship	40	1.6	01/05/2022 - 30/06/2022	Prof. Vincenzo Moscato, Prof. Giancarlo Sperlì	N

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	**6**	**1.2**	**4.8**	**0**	**12**
Bimonth 2	**0**	**0.6**	**11.4**	**0**	**12**
Bimonth 3	**0**	**0.4**	**11.6**	**0**	**12**
Bimonth 4	**5**	**0.2**	**11.8**	**1.6**	**18.6**
Bimonth 5	**0**	**0**	**15**	**0**	**15**
Bimonth 6	**0**	**0**	**5.4**	**0**	**5.4**
Total	**11**	**2.4**	**60**	**1.6**	**75**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

During my second year of PhD course, I carried out two research activities within my research field, namely “Multimodal detection of previously fact-checked information” and “Exploiting external knowledge to improve the detection of hateful Internet memes”.

- **Multimodal detection of previously fact-checked information**

The rise of social media has revolutionized human communications providing individuals with new and faster solutions to share their opinions and emotions, increasingly through multimodal content (e.g. memes, short animated frames) which has been proven much more attractive and credible [1]. However, the misuse of this freedom of expression, often referred to as disinformation, has supported fake news dissemination to mislead people decisions and has encouraged hostility behaviors in the form of hate speech and cyber-bulling [2].

Whilst representatives of online platforms, leading social networks and advertising industry, also in accordance with new governmental policies, are increasingly adapting to mitigate this problems, fact-checking still represents the leading strategy to debunk false information though domain experts' analyses and semi-automatic systems assessing news truthfulness [3]. Concretely, the fact-checking process comprises a four-stages pipeline [4]: (i) spotting check-worthy claims, i.e. selecting and prioritizing a set of claims according to their importance; (ii) verified claim retrieval, i.e. detecting previously fact-checked information; (iii) evidence retrieval, i.e. finding the evidences which support or refute a claim; (iv) claim verification, i.e. assessing claim's veracity (even partially) based on the retrieved evidences.

Considering the increased number of fact-checked news and that the same viral claim is often re-posted by thousands of people, the verified claim retrieval task represents a promising approach to improve the fact-checking process. Thus, detecting previously fact-checked information can ease the manual fact-checkers' effort by filtering out already verified information and providing relevant and reliable information which could increase their productivity and thus their effectiveness.

The verified claim retrieval task can be phrased as an ad-hoc information retrieval (IR) problem whose aim is to retrieve a list of verified documents according to their relevance with an input claim. During the second year of PhD, I focused on the multimodal settings of the problem: we exploit the texts and the images of the input claims and of the verified documents and leverage the most advanced fusion techniques to extract powerful representations which capture the complex relationships between these modalities. Specifically, we propose a simple, yet effective, multimodal IR system which could adopted for both retrieval settings, i.e. retrievers select the potential set of (verified) documents relevant to the input claim, and re-ranking settings, i.e. re-rankers reorder that set of candidates with more powerful and fine-grained techniques. While the latter settings have been recently studied [5] [6], we first instigate neural network-based ranking systems as potential retrievers. In addition, even if [7] shares our multimodal problem formulation, we first evaluate the contribution of each modality to the final performance.

Our results show the superiority of the proposed system: (i) it achieves the best performance in the re-ranking scenario, improving the state-of-the-art performance up 15 NDCG points; (ii) when compared with other retrievers, our model is the only one which reaches and (slightly) overcomes a standard IR baseline, i.e. BM25 [7] algorithm.

- **Exploiting external knowledge to improve the detection of hateful Internet memes**

Nowadays, the misuse of social media platforms has supported fake news dissemination to mislead people decisions and has encouraged hostility behaviors in the form of hate speech and cyber-bulling [2]. In addition, social media contents are more and more multimodal, i.e. the text is combined with other modalities (e.g. images, videos) in order to make the content much more attractive and impactful (e.g. memes, short animated frames) [9]. From the disinformation perspective, it has been proven that multimodal posts are more credible than simple text messages as well as the detection of harmful content cannot be pursued unimodally because that content becomes offensive only when two or more modalities are combined [10] (e.g. meme texts are often innocuous but turn into weaponized posts when combined with specific images).

During the second year of PhD, I focused on studying harmful Internet memes, in the context of racist, sexist and misogynistic content. Recently, several detection systems have been proposed during the "Hateful Memes Challenge" [10] and "Multimedia Automatic Misogyny Identification Challenge" [11], showing that the ensemble of different state-of-the-art vision-language pre-trained models (VL-PTM) [12], can reach good classification performance. However, an Internet meme is a complex piece of information whose understanding cannot only rely on the knowledge acquired during any model's (pre-)training process, even if it is on large-scale datasets. On the contrary, we conjecture that the key to understand a meme is not within the content but requires some external background knowledge. In other words, understanding whether a meme is harmful requires being familiar with the entities it represents and the context in which it is used. Inspired by this intuition, we have designed KERMIT (Knowledge-EmpoweRed systeM In hateful meme deTection), a novel framework which firstly interconnects the entities within a meme with external factual knowledge, and then performs a reasoning step to perform the hateful classification. Concretely, our framework is based on two sequential steps: (i) we first build a knowledge graph combining the entities in the text and the image of the meme with related external knowledge, retrieved from Wikidata or ConceptNet; (ii) we train a reasoning model based on attention mechanism to extract the subset of the above-mentioned knowledge graph that is most useful to assess whether the meme is hateful or not.

Our preliminary experiments on two datasets, i.e. Hateful Meme Challenge and SemEval Task 5, show promising results: (i) the knowledge graph extracted from both Wikidata and ConceptNet always improves the classification performance; (ii) the reasoning step further improves the performance even if it requires a detailed hyper-parameters tuning.

- **Other research activities**

In the context of the collaboration with prof. Emilio Ferrara, at University of Southern California, Los Angeles, we are investigating whether and how different highly-moderated social media platforms can collaborate to improve their respective digital environments. Such collaboration does not necessarily imply the deployment of a unified moderation process and can be operationalized by sharing the decisions taken by the single moderation processes of the participants. For instance, YouTube deciding to remove some videos can be the trigger for Twitter to analyze its users who directly share or are indirectly exposed to them. Under such a YouTube-Twitter scenario, we investigate how eventually-moderated YouTube videos spread on Twitter to uncover peculiar interaction patterns which, in turn, might enable the early detection of harmful content diffusion. Concretely, we want to answer the following research questions (RQs):

1. Which characteristics exhibit Twitter users who share eventually-moderated YouTube videos? And are they different from those sharing non-moderated videos?
2. Do eventually-moderated and non-moderated YouTube videos spread differently on Twitter?
3. Can we predict whether a YouTube video will be moderated based on its diffusion on Twitter?

4. Research products:

- **V. La Gatta**, V. Moscato, M. Pennone, M. Postiglione, G. Sperli; “Music Recommendation via Hypergraph Embedding”; IEEE Transactions on Neural Networks and Learning Systems, IEEE TNNLS; published
- A. Ferraro, A. Galli, **V. La Gatta**, V. Moscato, M. Postiglione, G. Sperlì; “An epidemiological Neural Network model exploiting dynamic graph structured data”; IEEE World Congress on Computational Intelligence (IEEE WCCI2022); published
- A. Barducci, S. Iannaccone, **V. La Gatta**, V. Moscato, G. Sperlì, S. Zavota; “An end-to-end framework for information extraction from Italian resumes”; Expert Systems with Applications, ESWA; published

- A. Ferraro, A. Galli, **V. La Gatta**, M. Postiglione; "A Deep Learning pipeline for Network Anomaly Detection based on Autoencoders"; IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence, and Neural Engineering; IEEE MetroXRAINE 2022; published
- T. Chakraborty, **V. La Gatta**, V. Moscato, G. Sperli; "Information retrieval algorithms and neural ranking models to detect previously fact-checked information"; Neurocomputing; submitted
- R. Formisano, **V. La Gatta**, V. Moscato, G. Sperli; "A Novel Multimodal Retrieval System for Previously Fact-checked Information Detection"; ACM Transactions on Management Information Systems, ACM TMIS; submitted
- A. Galli, **V. La Gatta**, V. Moscato, M. Postiglione, G. Sperlì; "Interpretability in AI-based Behavioral Malware Detection Systems", IEEE Transactions on Dependable and Secure Computing, IEEE TDSC; submitted

5. Conferences and seminars attended

- 2nd Ital-IA Conference, organized by Consorzio Interuniversitario Nazionale per l'Informatica (CINI); presented the paper Trustworthy AI @ PICUS Lab; online attendance

6. Periods abroad and/or in international research institutions

I started an internship at University of Southern California, Los Angeles, under the supervision of Prof. Emilio Ferrara. The internship started on 5th June 2022 and thus the number of months spent abroad, during this second year, is four.

We are currently working on two research projects: (i) discover how false claims spread on Twitter during the first two months of Russian-Ukraine war; (ii) investigate whether and how different highly-moderated social media platforms, i.e. Twitter and YouTube, can collaborate to improve their respective digital environments.

7. Tutorship

- Co-supervisor of seven master theses in Computer Engineering
- Weekly two hours of teaching activities regarding practical lectures/seminars during the course "Big Data Engineering", Master Degree in Computer Engineering
- Weekly two hours of teaching activities regarding practical lectures/seminars during the course "Machine Learning e Big Data per la salute", Master Degree in Biomedical Engineering

8. Plan for year three

In the next year, I plan to:

- Extend KERMIT to other multimodal classification in the field of disinformation mining (e.g. fake news detection).
- Augment KERMIT interpretability using eXplainable Artificial Intelligence techniques.

- Spend two months at University of Southern California, Los Angeles, under the supervision of prof. Emilio Ferrara. We plan to complete the cross-platform study between Twitter and YouTube, to unveil how the two platforms can collaborate to improve their respective digital environments.
- Co-supervise master students and keep the tutorship for the course of Big Data Engineering.
- Write my thesis on the role of multimodality across several disinformation mining tasks