Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D10.5

SoBigData Interest groups report 1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (60 months)
ENDING DATE	31/12/2024
PROJECT WEBSITE	http://www.sobigdata.eu
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042

DELIVERABLE INFORMATION	
WORK PACKAGE	WP10 JRA3 - Exploratories
WORK PACKAGE LEADER	KTH
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETH Zürich, PSE, UNIROMA1, CNRS, CEU, URV, CSD, BSC, UPF, Eli, CRA, UvA
DELIVERABLE NUMBER	D10.5
DELIVERABLE TITLE	SoBigData Interest groups report 1
AUTHOR(S)	Luca Pappalardo (CNR), Roberto Pellungrini (UNIPI), Aris Gionis (KTH)
CONTRIBUTOR(S)	Michele Gentili (UNIROMA1), Aris Anagnostopoulos (UNIROMA1), Luca Pappalardo (CNR), Roberto Pellungrini (Unipi)
EDITOR(S)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
REVIEWER(S)	Michela Natilli (CNR), Vaiva Vasiliauskaite (ETHZ)
CONTRACTUAL DELIVERY DATE	30/06/2021
ACTUAL DELIVERY DATE	02/07/2021
VERSION	V1.2
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	12
KEYWORDS	Interest groups, resources available

EXECUTIVE SUMMARY

The deliverable contains the activities in the interest groups, reporting the creation of new ones and the status of the resources available.

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

This document contains information on SoBigData++ core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData++ Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The content of this publication is the sole responsibility of the SoBigData++ Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

GLOSSARY

EU	European Union
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
MP	Micro-projects
AI	Artificial Intelligence

TABLE OF CONTENTS

1 Relevance to SoBigData++

This document describes: (i) the activity carried out within the SoBigData++ interest groups since the beginning of the project; and (ii) the report about the creation of a new exploratory, the "Network Medicine" exploratory. For each interest group, we report the results achieved and the activities carried out in terms of conferences/workshops, hackathons, data collection, and software development. The topics and activities described in this document are relevant to milestone MS3 "new exploratories coming from Interest groups become operative".

Since in the document we also describe some activities made or planned for the next period, this deliverable is also related to work packages WP3 - Dissemination, Impact, and Sustainability (because of workshops and conferences have been made or planned), WP4 - Training (because hackathons have been made or planned), and WP7 - Virtual Access (because data sets and software have been made available on the infrastructure or planned).

1.1 Structure of the document

In Section 2, for each interested group, we report the results achieved and the activities carried out since the beginning of the project, as well as the topics and activities planned for the next period. In Section 3, we report about the creation of a new stable exploratory in SoBigData++.

2 Activities in the interest groups

Interest groups are possible future exploratories which will be investigated by the consortium to understand if there are interests and experiences which may be transformed in services. Those interest groups will organize meetings with experts in the field, researchers and industries to eventually become exploratories in SoBigData++.

Since the beginning of the project, we have investigated two interest groups: one in Network Medicine and the other in Computational Epidemiology.

2.1 Interest group on Network Medicine

During last year, the interest group on Network Medicine investigated several research topics, activating some micro-projects (see Section 2 of deliverable D10.2), producing stories and organizing seminars.

2.1.1 BIOLOGICAL RANDOM WALK

Partners involved: UNIROMA1

We carried out research and experiments on the development of this new Disease Gene Prioritization Algorithms. We obtained interesting results both in comparisons with other state-of-the-art algorithms and in the biological quality of the outputted ranking. We also wrote a blog post on the SoBigData++ website in October 2020: Network Medicine: Disease Genes Prioritization Problem. Currently, we are writing a draft and we expect to submit our work to a journal of the area next year.

2.1.2 PH-LIKE LEUKEMIA PATIENTS

Partners involved: UNIROMA1

We carried out research and experiments on a genetic study of Philadelphia Like Leukemia Patients. Unfortunately, the result we obtained did not confirm what we expected: we could not find any new statistically associated gene with the different phenotypes of this disease. Most of the results were already present in literature. So, we decided to put a hold on this project at least for the next year.

2.1.3 CROHN DISEASE

Partners involved: UNIROMA1

We developed a new Local Community Detection algorithm and we started the analysis of Inflammatory Bowel Disease. We published a blog post on SoBigData++ blog on November 2020: Network Medicine: Local Community Detection on Inflammatory Bowel Disease. However final results on this disease did not bring good enough results. In the next year, we will expand the analysis on a pulmonary disease, in particular, COPD (see below).

2.1.4 RECURRENT RISK OF THYROID CANCER

Partners involved: UNIROMA1

We applied machine learning techniques for recurrent risk prediction of thyroid cancer. In particular, after using advanced supervised machine learning algorithms, such as extra tree, xgboost and random forest, we preferred to use a single decision tree to have a better interpretation of the results. In the end, we were able to provide a better risk assessment with respect to the current one used by clinicians, called the, ATA (American thyroid association) risk. The plan of the exploratory for the next year is the submission of an article to a journal of the area.

2.1.5 PARTIAL CORRELATION FOR FUNCTIONAL COPD SUBNETWORK GENES DISCOVERY

Partners involved: UNIROMA1

We developed a new software, leveraging Partial Correlations and Protein-Protein interactions network to create a subnetwork of related genes in the chromosome 4 region in lung tissue. In particular, we found statically associated relations, $p<0.01$ and n~15k edges, that led to the discovery of new key genes involved in the COPD phenotype. We are currently writing the draft of the work done and for the next year we plan to submit the work to a journal and write a blog post on the SoBigData++ blog.

2.1.6 DUAL TRAUMA PROJECT

Partners involved: UAQ

The Dual Trauma research project was created to predict the possibility that a guy could harm himself, in order to intervene quickly in case of risk. The approach makes use of answers provided to a questionnaire internationally accepted in psychiatry.

The population adopted for the study is part of the Aquilano seismic crater and is made up of boys and girls who were 18 years old at the time of the test and attended high school. The dataset used for the predictions was created by giving a questionnaire of about 280 questions to 1010 students, therefore the resulting dataset is a 280 X 1010 matrix. Among the blocks of questions administered we report:

- Risk Family Questionnaire
- International Trauma Exposure Measure
- Cyber-bullying
- Dissociative Experiences Scale
- Resilience Scale for Adult
- Attachment style questionnaire
- Questions to measure emotional temperaments
- Self-harm inventory
- Evaluation of emotional regulation strategies
- International Trauma Questionnaire

- Assessment of anxious and depressive symptoms
- Assessment of pre-psychotic symptoms

In addition to the basic statistical analyses carried out by the doctors who gave the questionnaire, we applied machine learning techniques to predict all the seven variables of the self-harm inventory block. We used approaches and algorithms for the creation of decision trees, SVM and various configurations of neural networks to find the optimal prediction setting to maximize accuracy. High priority has also been given to minimize the number of false negatives, that is, those guys who are identified by the system as "safe" but who are actually "at risk".

We are working on the experiments that are reporting promising results. The work done will be published to two scientific international journals: one more centered on psychiatry topics and the other on IT and AI fields.

2.1.7 MICRO-PROJECTS

- Discovering side effects of drugs through denoised GCN
  - Status: active
  - Partners: UNIROMA1, UT
  - External partners: none
  - Expected output: Method, Experiments, Blog post, Preprint paper
- Predicting Thyroid Cancer Recurrence
  - Status: active
  - Partners: UNIROMA1
  - External partners: none
  - Expected output: Method, Experiments, Blog post, Preprint paper

2.2 Computational Epidemiology

The interest group on Computational Epidemiology aims at studying models to monitor and predict the diffusion of epidemic diseases using tools from complex systems and AI. Related topics investigated in this interest group are misinformation related to the COVID-19 pandemic, and the change in the behaviours of people given the pandemic.

2.2.1 COVID-19 MISINFORMATION ANALYSIS

See Section 3.1.1.1 in deliverable D10.2.

2.2.2 VACCINE MISINFORMATION ANALYSIS

See Section 3.1.1.2 in deliverable D10.2.

2.2.3 HUMAN MOBILITY AND COVID-19 PANDEMIC

See Section 3.3.1.6 in deliverable D10.2, points 2) and 3).

2.2.4 COVID AND CLIMATE CHANGE

See Section 3.3.1.1 in deliverable D10.2

2.2.5 IMPACT OF COVID-19 ON EMPLOYMENT RISK

See Section 3.3.1.5 in deliverable D10.2

2.2.6 MICRO-PROJECTS

- A dataset to assess mobility changes in Chile following local quarantines
  - Status: active
  - Partners: UNIPI, ISTI-CNR
  - External partners: Universidad del Desarrollo de Santiago de Chile (Chile), Telefónica Chile
  - Expected outputs: Paper, Dataset, Blog post
- Changes in visiting patterns to venues during the COVID-19 pandemic in the US
  - Status: active
  - Partners: ISTI-CNR
  - External partners: FBK
  - Expected output: Method, Experiments, preprint paper, blog post
- Vaccine disinformation micro-project
  - Status: active
  - Partners: USFD, WAFI - IIT, CSD
  - External partners: none
  - Expected output: Dataset, analysis services, Blog post

2.2.7 Pubblications

- V. Pieroni, A. Facchini, M. Riccaboni, COVID-19 and Unemployment Risk: Lessons for the Vaccination Campaign, https://arxiv.org/abs/2102.03619, 2021
- Cintia et al., The relationship between human mobility and viral transmissibility during the COVID-19 epidemics in Italy, https://arxiv.org/abs/2006.03141, 2021
- L. Pappalardo, G. Cornacchia, V. Navarro, L. Bravo, L. Ferres, A dataset to assess mobility changes in Chile following local quarantines, https://arxiv.org/abs/2011.12162
- Vasiliauskaite, Vaiva, Nino Antulov-Fantulin, and Dirk Helbing. "Some Challenges in Monitoring Epidemics." *arXiv preprint arXiv:2105.08384* (2021).

3 Report on next exploratories creation

Two interest groups have been created in the first year at half of the project, named “Network Medicine”, and “Computational Epidemiology”.

Network Medicine. The Network Medicine exploratory was coordinated by UNIROMA1 and mainly involved people from UNIROMA1 and UAQ. The interest group produced two active micro-projects (see Section 2.2.1.6) and several topics are currently investigated (see Section 2.1). Given the availability of UNIROMA1 to lead a new exploratory dedicated to network medicine, and given the interest from several other partners in collaborating on this topic, *we decided to create a new exploratory named “Network Medicine”, to be active since July 2021*. UNIROMA1 proposed a task leader and a user community activist for this new exploratory.

Computational Epidemiology. The interest group on Computational Epidemiology was mainly motivated by the striking impact of the COVID-19 pandemic in 2020, and intended as an inter-exploratory working group. Indeed, as shown in Section 2.2, most of the activities of this interest group have been developed within the context of other exploratory (e.g., Societal Debates and Misinformation Analysis, Sustainable Cities for Citizens). However, none of the partners proposed to lead a new exploratory in this topic so far. For this reason, *we decided not to create a new exploratory on this topic.*