DATA ANALYTICS IN HEALTH 4.0: EXTRACTING KNOWLEDGE FROM BIG DATA IN PANDEMIC TIMES

Daniela Alderuccio[1*], Rossana Cotroneo[2]

[1]*ENEA – TERIN-ICT-HPC – Rome, Italy*[1]
[2]*ENEA – ISV –DST-PM Rome, Italy*

ABSTRACT. Knowledge extraction from Big Data in Digital Health sector is relevant in the analysis both of bidirectional interactions between healthcare providers and patients, and of exchange of information among patients. Extracting re-usable knowledge activates a process of creating value, able to support prediction and prescriptive models in the healthcare sector. In this paper, we present an exploratory study on Digital Health Data in pandemic times. In Health 4.0, Knowledge is organized in: Providers' Knowledge, Patient's Knowledge, both accessing Organization knowledge (data/information available for providers and patients). Health Knowledge is used for disease prediction and prevention, personalized medicine, enhanced patient-centred care, and proactive treatment. In this on-going research we process e-Health multilingual data collection, and extract specific Knowledge and keywords, in order to adopt ad hoc strategies for crawling and analysing data from open sources (online news, websites, social media platforms, etc.). We design a pipeline of sequential steps to be performed, for processing and analysing data in digital Health domain. This pipeline could be standardized and used as a model to be applied to other domains of interest. First results show that - before crawling, extracting, processing, and analysing data using Text Mining methodologies – the adoption of a cross-linguistic semantic perspective to multilingual data is a propaedeutic step in the process and refines crawling and analysis strategies. Pandemic has created high level importance to the digital health and propelled to the ongoing developments. Data acquisition, processing, analysing, visualization, sharing, storage, lay the groundwork for a healthier society, combining ICT with medical and behavioural knowledge.
Keywords: Big Data - Social Media Analytics - Knowledge Extraction - Digital Health – Health 4.0 – Text Mining – Digital Epidemiology

1. Introduction

Big Data, Cloud Computing, and the Internet of Things are revolutionizing the whole eHealth ecosystem. In 2012 Digital Healthcare data was estimated to be equal to 500 petabytes and in 2020 it was expected to be 25,000 petabytes [1]. In 2022 worldwide digital healthcare data is estimated to currently equal between 25 exa-bytes and 35 zeta-bytes, with an annual increase of between 1.2 and 2.4 exa-bytes per year[2]. Such a huge amount of health data is generated by a variety of lab systems and health information systems (e.g., EHRs)[3][2]. Epidemics and pandemics have created high level importance to the digital health and propelled to the ongoing developments with the transformation in digital healthcare through online symptom checkers, remote patient monitoring tools, telehealth, patient portals, etc.

[1] Corresponding author*. E-mail: daniela.alderuccio@enea.it
[2] Elaboration by Sin [2] from IEEE Big Data source (https://bigdata.ieee.org/) , on 1 February 2022
[3] An Electronic Health Records (EHR) is an electronic version of a patient medical history (supported by the provider over time) with key heath information and administrative clinical data.

2. Health Knowledge

In the Health sector, Knowledge is defined as: *(i) Provider knowledge:* it typically contains both explicit (standard medical practice for a particular condition) and tacit[4] knowledge (practice and experience is internal knowledge and complement standard treatment); *(ii) Patient knowledge:* it is generally tacit (it is a patient's self-knowledge of current and past medical conditions); providers use this "health status" to diagnose, prescribe for and treat diseases; *(iii) Organization knowledge:* data (structured, semi-structured and unstructured data) and information available for providers and patients access. It often contains information collected from text-based materials, diagnostic systems, and other medical providers. Government and healthcare institutions are providers. They guarantee inclusive and easy access to healthcare services, for a healthier Society. Each government or healthcare institution has its own warehouse (silo) of public health information as well as confidential data. All these data are collected according to security and privacy issues.
Stakeholders in the Healthcare sector are patients, medical practitioners, hospital operators, pharma and clinical researcher, healthcare insurers [3]. Any knowledge created by one sector is important to all others. So, knowledge management in the healthcare sector must find a way to manage the creation, storage, sharing and use/reuse of this valuable information. COVID-19 Pandemic has dramatically accelerated the process and forced governments, public healthcare stakeholders and research institutions to collaborate and make available decades of stored usable, interoperable, searchable, and sharable according to the *once-only principle*. New sources (such as social media platforms, wearable devices, etc.,) have been used by Providers in addition to traditional sources (i.e., patient medical history, diagnostic/clinical trials data and drug information) with the aim to find relationships and pattern of interest among these heterogeneous sources. In this scenario, ENEA-CRESCO infrastructure faced the outbreak of COVID-19 improving the capability to analyse and to predict from huge amount of data, offering High Performance Data Analysis of Big Data (HPC/ HPDA) [4]. Social Media Analytics in ENEA focus on Knowledge Extraction in real-case scenario [5]. The challenge is to extract actionable knowledge from data, in order to perform predictive and prescriptive analysis, to support activities. The goal of knowledge management in healthcare is to provide decision-makers with the tools to turn information into a knowledge asset, transforming the organization: (i) into a place where new knowledge is generated and shared, (ii) into a learning organization, where information flows from ever-increasing sources and in great volume.

2.1 Big Data in the Health domain

Jee & Kim [6] redefined the characteristics of healthcare big data into three features: *Silo*, *Security* and *Variety*[5], instead of Volume, Velocity and Variety. Palanisamy and Jee & Kim [3, 6] stated that Big Data in the business sector differs from Big Data in Health domain[6]. Big Data in e-Health sector support comprehensive healthcare solutions such as clinical decision support, disease surveillance and public health monitoring and management. The traditional *disease-centric model* is now shifting towards a *patient-centric model*, as requested by National Recovery and Resilience Plan (PNRR) too, where an active participation of patients is needed. The challenges for healthcare involve sharing

[4] Explicit knowledge: It is recorded and communicated through different mediums (print, audio, etc.). It can be transmitted quickly and easily from one individual to another and is organized systematically. Tacit knowledge: It involves a person's ability (or specialized inherent knowledge). Tacit knowledge can be achieved only through experience.
[5] *Silo* is databases having healthcare information maintained by stakeholders (i.e., hospitals). *Security* feature is related to the care needed to keep healthcare data. *Variety* shows the forms of healthcare data: structured, unstructured and semi-structured.
[6] Big Data in business sector has been used to identify consumers' behavioral patterns, to develop business services and solutions. [3]

healthcare information[7] hosted in silos (warehouse) in order to provide data integration, to maintain the control over the data, to implement regulation on security and compliance and to transform the huge amount of e-health data into actionable knowledge. We will show a pipeline of sequential steps to be performed (Fig. **1**), in order to process and analyse health data. This pipeline is designed on health data but could be standardized and then used as a model to be applied to other domains. As a first step, we extract specific knowledge for keyword identification and selection before crawling from open data coming from different sources (online news, websites, social media platforms, etc.). The second step is Data Crawling. Further steps are Data Stream including Text Analysis, validation, and Sentiment Analysis.

Fig. 1: *Pipeline*

3. Results

As a first step, we explore Twitter Trends (2020-2022), in order to select keywords, to integrate the semantic field of hashtags to be used for crawling and analysis tasks. In health domain, Twitter data might offer knowledge about patients' symptoms, opinions, etc. [7]. In our analysis we proceed exploring the semantic field of topics for keyword identification and selection, taking into account multi-language and space-time distribution. The first idea was to use the definition of the disease (SARS-cov 2- COVID-19) or virus denomination (coronavirus, rna-virus), but investigating deeply we found additional keywords. Then we use these keywords to start crawling task and then analyse results. In particular, in Twitter Trends we searched for the keyword to be used to select tweets in Italy from 16th to 18th March 2020. We found following trends related to pandemic, and we extracted following hashtags containing: virus denomination and sometime geo-localization: *(#coronavirusitalia, #coronavirusitaly, #COVID19italia, #covid19uk, #bergamo, #coronavirus-updates, #CoronaVirusUpdates,);* keyword linked to: the pandemic emergency (*#emergenzacovid19, #EmergenzaCovid19*); to social activities/actions supporting people (*#CoronaVirusChallenge, #flashmob, #flashmobitalia, #flashmobsonoro, #balconi, #PaesaggiDaCartolina, #stayhome-challenge);* to emotion to support (*#orgoglioitaliano, #DomaniUsciraIlSole, #CuraItalia, #OrgoglioItaliano, #celafaremo, #scrivoquelchesento, #DomaniUsciraIlSole);* to negative emotion (*#coronapocalypse);* to the disease, protection devices and recommended behaviours: (*#quarantinelife, #Mascherine, #mascherine);* to governmental rules: (*#decretocuraitalia, #CuraItalia);* to politicians: *(#GovernoConte, #governoconte, #IoStoConConte, #contedimettiti, #De Luca, #boris Johnson);* and other hashtags without any semantic correlations, that apparently did not have nothing in commons with the pandemic (*#leCose, #iorestoacasa, #unitaditalia).* Preliminary results show the trends of emotion (expectation, anxiety, fear, reaction, sadness, and joy public acceptance/not acceptance) and the narratives underlying those emotions during the different phases

[7] The goals are: (i) providing better citizen's healthcare services, (ii) improving medical treatment, (iii) enhancing easy and equal access to health services, (iv) improving security, speed, interoperability, analytics capabilities.

of pandemic by North, Centre, and South of Italy. The outcomes show the relationships and pattern of interest among these heterogeneous sources of big data (providers and user-generated contents), and we formulate hypothesis on the dynamic of information spreading from hub to users, from the center to the periphery. These results are useful for management of health domain.

4. Conclusions

Digital Health Data provide an interesting scenario, in view of many applications such as personalized healthcare and public health, for a healthier Society[8]. The challenge to face is to transform the huge amount of health data into actionable knowledge, in order to perform predictive and prescriptive analysis, to support healthcare activities. Furthermore, real-time analysis of health data should be carried out, according to privacy and security laws [6]. In our research, we highlight the importance of selecting keywords before crawling tasks and of adopting a pipeline designed on the characteristics of the application domain. Before extracting, processing, and analysing health data using Text Mining methodologies, in this first exploratory study we adopted (i) a cross-linguistic semantic perspective exploring multilingual data, and enabling the selection of keywords and hashtags for crawling and analysis task; and (ii) we defined a pipeline to process Healthcare data. This pipeline could be standardized and used as a model to be applied to other domains. As a further step we will formulate hypothesis on the dynamics of information/knowledge spreading from institutional hub to users/patients, finding centre/periphery interactions, and among patients.

References

[1] J. Sun, C.K. Reddy. Big data analytics for healthcare. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013).

[2] P. Šín, A. Hokynková, M. Novakova; A. Pokorna, R. Krcˇ, J. Podroužek. Machine Learning-Based Pressure Ulcer Prediction in Modular Critical Care Data. *Diagnostics* 2022, 12, 850. Academic Editors: Keun Ho Ryu & Nipon Theera-Umpon - https://doi.org/10.3390/ diagnostics12040850 (2022) – cited IEEE Big Data. https://bigdata.ieee.org/ (accessed on Feb 2022)

[3] V. Palanisamy, R. Thirunavukarasu. Implications of big data analytics in developing healthcare frameworks - A review. *J. King Saud Univ. Comput. Inf. Sci.,* 31, 415-425.Science J. King Saud Univ. Comput. Inf. Sci. (2017)

[4] F. Iannone and HPC-CRESCO Team. ENEA HPC CRESCO in the time of Covid-19 and new Supercomputing Frontiers, *ENEA CRESCO in the fight against COVID-19* - ENEA - pp. 8-31, 2021 - ISBN:978-88-8286-415-6 (2021)

[5] D. Alderuccio, S. Migliori and ICT-HPC Team: Knowledge Extraction from Social Media Web Sources: Elements affecting Web Crawling and Data Analytics Tasks in ENEAGRID. *High Performance Computing on CRESCO Infrastructure: research activity and results 2019 ,* pp. 155-158 - ISBN: 978-88-8286-390-6 – ENEA *(*2020)

[6] K. Jee, GH. Kim. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthc Inform Res. 2013 Jun*;19(2):79-85. doi: 10.4258/hir.2013.19.2.79. Epub 2013 Jun 30. PMID: 23882412; PMCID: PMC3717441. (2013)

[7] B.A. Panuganti, A. Jafari, B. MacDonald, A.S. De Conde, Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. *Otolaryngology – Head and Neck Surgery 2020,* Vol. 163(3) 491–497 The Author(s) - DOI: 10.1177/0194599820932128 - http://otojournal.org (2020).

[8] COM(2018) 233 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0233&from=EN