REACH

NEXT GENERATION DATA INCUBATOR

EXPLORE PHASE TECHNICAL SPECIFICATIONS

11/05/2023

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS

1 ANNEX I. Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** The mock-up solution is suitable and correctly addresses the challenge/theme selected over the REACH dataset/s. The Big Data solution architecture proposed is adequate to tackle the data management issues associated to the solution in mind. "To what extent does the applications handle the data provided?"

The goal of technical scope refinement is to oversee the transfer of mature Big Data technologies into the Platform use cases (hereafter "pilots"). The transfer will happen in three cycles: (1) initial prototypes, based on pilot requirements; (2) updated prototypes, based on pilot internal validation; (3) final implementations, based on pilot external validation. Each cycle will be described within the work packages deliverables. The final outcome of technical specifications is building the technical Blueprint, it will also allow to do a mapping between the requirements and the technical components. The mapping matrix will be multidimensional, taking into account aspects of technologies, of pilots/businesses, and of communities, as well as aspects of specific data sources that have different velocities, e.g. social media data vs. web site data. We have chosen the Free-choice track 3 and working with **mention** a social media data provider. We will focus initially at first on the use case mock up: social media and Web Monitoring of Health disinformation[1] like Promotion of a selection of harmful products which have a huge impact on European citizens[2] health and targeting elderly people. The data domains are composed of the following sources (Health Data, Social Network captured with the Data provider **Mention**[3] using its APIs**,** Visual Media, Health Misinformation, other sources of Misinformation (Web sites, Open data, blogs…), large scale science to provide dataspaces to the scientific community). The security: Data Privacy, Integrity and GDPR is an essential part of this project).

2. **SELECTION OF ALGORITHMS AND TOOLS:** The indicated Data Science approach, i.e. algorithms chosen, and Big Data architecture approach, i.e. tools chosen may successfully accomplish the required data governance, processing and analysis. A clear understanding of the used REACH dataset/s is demonstrated.

This misinformation detection will be based on a data model collection data related to our use case scenarios Health disinformation on harmful products from multiple sources. Our platform will detect misinformation using two strategies: Big Data abnormal content spread detection and NLP/ML detection and spread detection (social media, Facebook, Instagram, twitter, Tumblr, Reddit, blogs). Based on our research, we identified that misinformation spread pattern is very specific, a small number of users heavily sharing the content[4].Our platform will monitor, detect, and flag such content spread behaviour. This approach allows to rapidly detect new misinformation trends that are not related to previous predictable topics detected by machine learning. We will use multi-type of Big Data ingestion system provides data analytics using AI and cognitive computing technologies including analysis of text natural texting processing (NLP) & semantical technology, image recognition, social interactions and other relevant metadata. An AI/ML assisted algorithm will flag the misinformation in text, and in image recognition using the four dimensions of analysis to detect misinformation developed by the researchers and outlined in the following four main questions: consistency of message, coherency of message, credibility of source and general acceptability of message. The selected data will be cross-sector and multi-stakeholder as it will cover Citizens, Hospital, Institutions like the World Health Organisation and multiple countries / Languages (Spanish, French, English). We will use multilingual keywords to ingest and refine the monitoring and listening

1 Examples: https://www.weforum.org/agenda/2022/01/covid-misinformation-omicron-and-how-to-combat-it/ and https://www.frontiersin.org/articles/10.3389/fpsyg.2022.783909/full

2 Examples: https://genevasolutions.news/climate/un-expert-slams-chemical-industries-for-spreading-fake-news-about-risks and https://extension.colostate.edu/topic-areas/nutrition-food-safety-health/nutrition-misinformation-how-to-identify-fraud-and-misleading-claims-9-350/

3 https://dev.mention.com/current/src/index.html

4 Kumar and Geethakumari Detecting misinformation in online social networks using cognitive psychology, Human-centric Computing and Information Sciences, 2014, p.13

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

3. **TECHNICAL SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** The solution can truly cope with humongous and increasing datasets, potentially from diverse data providers, and is flexible it to adapt to other related domains.

The scalability and flexibility of the solution is demonstrated through 1**) Knowledge on misinformation** detection and analysis, we have won a visible project at the World Health Organisation (WHO) related to misinformation and the promotion of harmful products. The scope of analysis was the breach of the international code of Marketing in promoting misinformation of breastmilk substitutes to the mothers[5], elaborating a Marketing segmentation methodology based on scientific finding to ingest Big Data using AI and Machine Leaning to capture text using semantic technologies and NLP, Image recognition and sentiment analysis on social media and Web sites. Our solution consumes our own algorithms ingest data and technical APIs from social media third parties' partner. We will apply a similar segmentation methodology by collecting data from multiple datasets. The result of our platform and study has been published end of April 2022[6]. 2) **The functional pilot** for the WHO has demonstrated that we can cope with humongous and increasing datasets which are particular to the Big Data and AI projects and considering we have to tackle the **5V principles** volume, value, variety, velocity, and veracity.3) **The usage of keywords** to "listen" and monitor social media need to take into account of the **variability** (data in change) of the data and the changing nature of the data which will be captured, managed and analysed using different techniques- e.g., in sentiment or text analytics, changes in the meaning of key words or phrases given the initial 3 languages of the proposed experiment. 4) **The scalability of the platform** will allow to deploy multiple data spaces and integrate into Data Value Chains. These data spaces will be governed by different stakeholders like end users, scientific, Health organisation and other interested constituents. The governance building blocks will be the artefacts that regulate the business relationships between the groups of stakeholders that can be identified in data-driven business ecosystems: Data owners, Data provider, Data processor and data marketplace operator.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Data sharing challenges, data governance and legal compliance, must be observed. The proposed solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

We are committed to be responsible for managing of the data generated and collected during the project. Yet, the Exploitation and Dissemination which will include partners like academic and researchers as well as staff from the Project. This will be the govern body to specifically indicate how the data will be organized, identified, exploited and made accessible for verification and re-use. The Data Management Plan (DMP) will take into consideration the following issues, which are expected to be enhanced and updated throughout the project: 1) **Data management**: the structure of the data packages, and the links between different datasets. 2) **Data description**: the project team will define the metadata required for their data, ensuring, when needed, standard description terminologies.3) **Data distribution**: the team will define the rules for data use and for shared generation of datasets. 4) **Data Protection**: Data protection including ensuring the compliance with laws and regulation as well as the deployment of leading-edge state-of-the-art security technologies in protecting data and controlling data access.5) **Data Privacy**: Data privacy and anonymisation with handling and deletion of personally identifiable information (PII) in compliancy with laws and regulations such as the EU GDPR (General Data Protection Regulation) as well as the deployment of anonymisation technologies.6) **Data Governance**: taking into account Privacy and Protection define the rules to access and share data. This includes standardization of sharing metadata, technologies such as encryption as well as the necessary solutions to orchestrate the agreed governance.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Feasible and credible quality process followed for the final product generation. The potential risks in all the phases of the project (design of the solution, development, testing, deployment…) are identified and convincing mitigation plans put in place.

QA will be carried out periodically to verify the progress status and quality of the deliverables. Risk management is foreseen to timely address adversities during the project implementation. This will be carried out by the project manager, we will continuously revise, update, monitor the risks and the mitigation strategies to solve the identified risks and trigger contingency plans. In terms of Big Data and AI Risk management Framework, we use an AI Risk Management Framework which covers all the AI Lifecycle Risks Managements and mitigation actions based on the EU AI risk-based approach[7] which uses used by our team to classify the level of exposure and type of risks defines 4 levels of risk in AI (Unacceptable risk, High risk, Limited risk, Minimal or no risk)

[5] Marketing of breast-milk substitutes: national implementation of the international code, status report 2020 https://apps.who.int/iris/handle/10665/332183

[6] https://www.who.int/news/item/28-04-2022-who-reveals-shocking-extent-of-exploitative-formula-milk-marketing

[7] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS