Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***Crowd sensing platform***
Deliverable No.	***D8.2***
Delivery Date	***31 August 2016***
Authors	***Stefano Cresci (CNR), Ian Roberts (USFD), Kalina Bontcheva (USFD)***

SoBigDatareceives funding from the European Union's Horizon 2020 research and innovationprogramme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D8.2
Deliverable Title	Crowd sensing platform
Contractual Delivery Date	31 August 2016
Actual Delivery Date	14 September 2016
Author(s)	Stefano Cresci (CNR), Ian Roberts (USFD), Kalina Bontcheva (USFD)
Editor(s)	Stefano Cresci (CNR)
Reviewer(s)	Dominic Rout (USFD), Valerio Grossi (CNR), Stefano Cresci (CNR)
Contributor(s)	Dominic Rout (USFD)
Work Package No.	WP8
Work Package Title	JRA1_Big Data Ecosystem
Work Package Leader	CNR
Work Package Participants	CNR, USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETHZ, TUDelft
Dissemination	Public
Nature	Report + Other
Version / Revision	V1.0
Draft / Final	Final
Total No. Pages (including cover)	22
Keywords	crowd sensing, opportunistic sensing, participatory sensing, social media

DISCLAIMER

SoBigData(654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

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

The document has been produced with the funding of the European Commission. 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.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
RI	Research Infrastructure
VRE	Virtual Research Environment
API	Application Programming Interface
Crowdsensing	Also used in the form: *crowdsensing*. Crowdsensing is a data collection paradigm according to which relevant data about a phenomenon of interest is gathered by combining readings from numerous devices carried by a large number of individuals. In this document the term *crowdsensing* will be used interchangeably with the terms: *citizen sensing*, *social sensing*, and *humans as sensors*.
AJAX	Asynchronous Javascript and XML. A set of Web development techniques using many Web technologies on the client-side to create asynchronous and interactive Web applications.
JSON	Javascript Object Notation. A data exchange format used in the majority of Web applications and services.

TABLE OF CONTENT

DELIVERABLE SUMMARY

This deliverable thoroughly describes the design, development, andintegration activities of the crowd sensing platform that have been carried outas part of the task T8.2 “Participatory & Opportunistic Crowd Sensing”. It also draws a plan of workfor future activities that will be part of T8.2 for the upcoming years.

EXECUTIVE SUMMARY

The deliverable is organized in different sections in order to thoroughly describe the methodologies and the techniques developed in T8.2 for the SoBigData crowdsensing platform. Specifically, it describes the sensing paradigms of participatory and opportunistic crowdsensing, highlighting the strenghts and drawbacks of the two approaches. Then, it describes the Twitter Monitor sensing platform, designed and developed to enable crowdsensing activities in the SoBigData infrastructure. Integration details and activities are also described. Finally, this document also draws a plan of work for the development and integration of additional features in the sensing platform. Such features will be gradually added to the platform as part of the future activities of T8.2.

1 RELEVANCE TO SOBIGDATA

Among the aims of SoBigData is the capability to provide a set of readily available datasets and methods to scientific communities. Typically, users of the SoBigData eInfrastructure can leverage any of the datasets released within the SoBigData project itself, or upload and share their owns. Another appealing possibility is to provide end-users and stakeholders with a tool that allows them to build and share new datasets. This deliverable specifically focuses on the methods, techniques, and tools used in SoBigData to allow users to collect data and build new datasets.

1.1 PURPOSE OF THIS DOCUMENT

The purpose of this document is to describe the methods and the techniques used within the SoBigData eIfrastructure to enable the creation of new datasets by the end-users of the platform. The tool that have been developed and integrated for this purpose is specifically focused on the Twitter social network. More specifically, this document aims to describe:

- the crowdsensing methods (i.e., opportunistic and participatory) that can be leveraged to automatically collect data from Twitter;
- the Twitter Monitor tool, a Web-based crowdsensing tool that implements both the participatory and the opportunistic approaches in order to allow users to easily collect and organize a Twitter dataset;
- the activities carried out to integrate the Twitter Monitor tool into the SoBigData eInfrastructure;
- a plan of work for the development and integration of additional features in the Twitter Monitor tool.

1.2 RELEVANCE TO PROJECT OBJECTIVES

The focus of SoBigData is the development of a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”. As such, providing platform users with the possibility to collect, build, and share new datasets is of the highest importance and can significantly increase the usefulness of the RI and, ultimately, user engagement with the platform.

1.3 STRUCTURE OF THE DOCUMENT

Section 2 provides a description of the main crowdsensing methods that can be used to fecth data massively produced by Web users.

Section 3 describes how these crowdsensing methods are employed in the Twitter Monitor tool. This section also describes the integration activities carried out.

Section 4 describes the future activities to be undertaken in order to futher expand the set of functionalities offered by the Twitter Monitor.

Section 5 concludes this document.

2 CROWDSENSING METHODS

The rapid growth of social networking platforms and the ubiquitous proliferation of mobile devices produced a great interest in assessing how massive real-time social data can be used as a mine of information in numerous domains. With respect to this field, a sensor is not only a physical device but also a logical or social metaphor, implemented by the "human as a sensor" paradigm. This paradigm is also referred to as "social sensing", "citizen sensing", or "crowdsensing", giving focus to the involvement of a number of people. Because of their massive number of users, their real-time features, and their ease-of-use, social networking platforms such as Twitter and Facebook have been the source of information for many crowdsensing systems.

2.1 PARTICIPATORY CROWDSENSING

Depending on their awareness and their involvement in the system, users are confronted with either an opportunistic or a participatory sensing approach. When users consciously opt to meet an application request out of their own will, this is called **participatory crowdsensing**. More specifically, in participatory crowdsensing the user is aware of the sensing action, e.g. by photographing locations or discussing events and by intentionally sending such information to the sensing system. Systems exploiting participatory crowdsensing require intentional participation and must therefore provide incentives to the users to perform the sensing action. Thus a key challenge in participatory crowdsensing is the attraction of a significant user base. This may pose serious limitations to newly deployed systems and may ultimately lead to unsatisfactory results due to the lack of available data.

2.2 OPPORTUNISTIC CROWDSENSING

In contrast with participatory crowdsensing, in **opportunistic crowdsensing** the user spontaneously collects and shares data as he goes for his daily life. Relevant data is then transparently intercepted by a situation-aware system. Opportunistic crowdsensing platforms do not require a specific user base since they rely on already publicly available data. Here the challenge is posed by the acquisition, preprocessing, and analysis of unstructured and heterogeneous data that is not specifically targeted to the sensing system.

3 CROWDSENSING TECHNOLOGIES

Arguably, social media platforms are the most effective, sophisticated and powerful way to gather preferences, tastes, and activities of groups of users in the context of Web 2.0. In turn, this large amount of information may generate in-depth knowledge abouttopics of interest. As such, because of their massive number of users, their real-time features, and their ease-of-use, social media platforms such as Twitter and Facebook have become the main source of information for the majority of crowdsensing systems.

For all these reasons, the crowdsensing platform employed within SoBigData and described in this section, so-called **Twitter Monitor**, is based on the Twitter social network[1].

3.1 TWITTER MONITOR

3.1.1 FEATURES AND USAGE

The Twitter Monitor features an interactive Web application designed to access the Twitter stream by exploiting the public TwitterStreaming APIs[2].The application is able to manage concurrent monitors: it is possible to launch parallel listening sessions (i.e., more than one Twitter crawler at the same time) using different parameters and collecting different sets of data. In addition to offering aninteractive Web interface in order to ease all the operations related to Twitter crawling, the Twitter Monitor also offers a set of funcionalities aimed at minimizing the loss of data due to network or local machine problems. The Twitter Monitor is automatically capable of detecting and recovering from simple error situations, such as a closed or disconnected Twitter stream. It is also capable of detecting more serious issues, such as Twitter refusing to open new streaming connections, and automatically sends targeted alerts to system administrators.

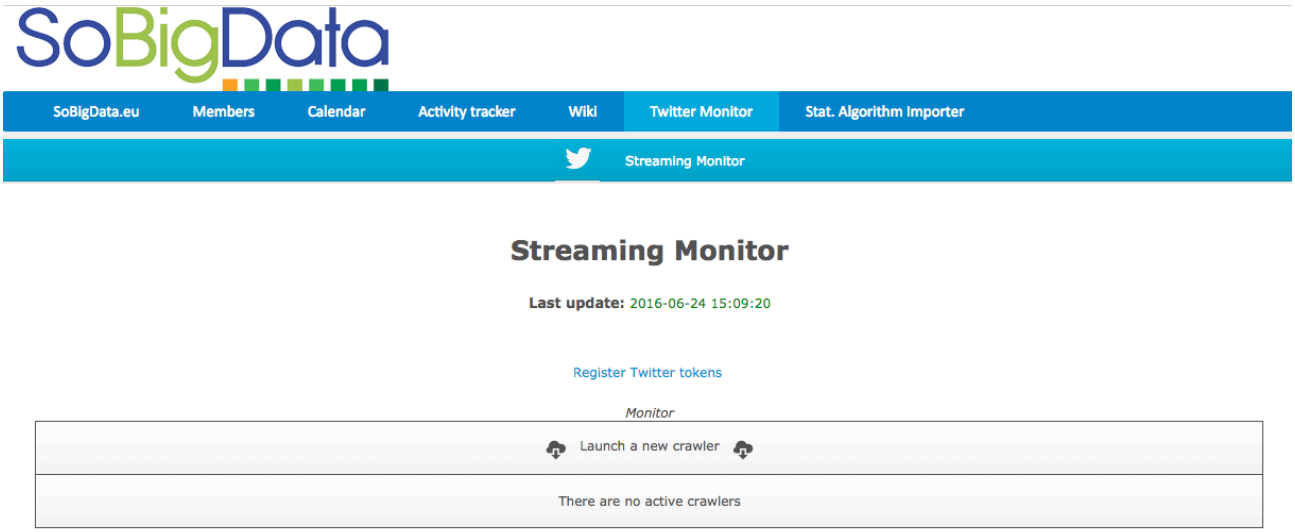

Figure 1: Landing page of the Twitter Monitor Web Application.

The Twitter monitor can be accessed via a dedicated tab within the SoBigData VRE. Once accessed, the main page of the Twitter Monitor appearsas shown in Fig. 1. In this page are reported:

- general information about the status of the Web application;
- information about active crawlers and their statuses (whether *running*, *paused*, or with *errors*).

The page automatically updates thanks to an AJAX Web interface and ad-hoc PHP APIs,hence always reflecting the latest status of the system. This page also containes the button that can be used to *launch a new crawler*. When clicked, a popup containing a Web form will appear (Fig. 2). The form can be used to set general parameters of the crawler, which are useful to label and easily retrieve information collected by the crawler, and the specific crawling settings that define the characteristics of data that the crawler will collect.

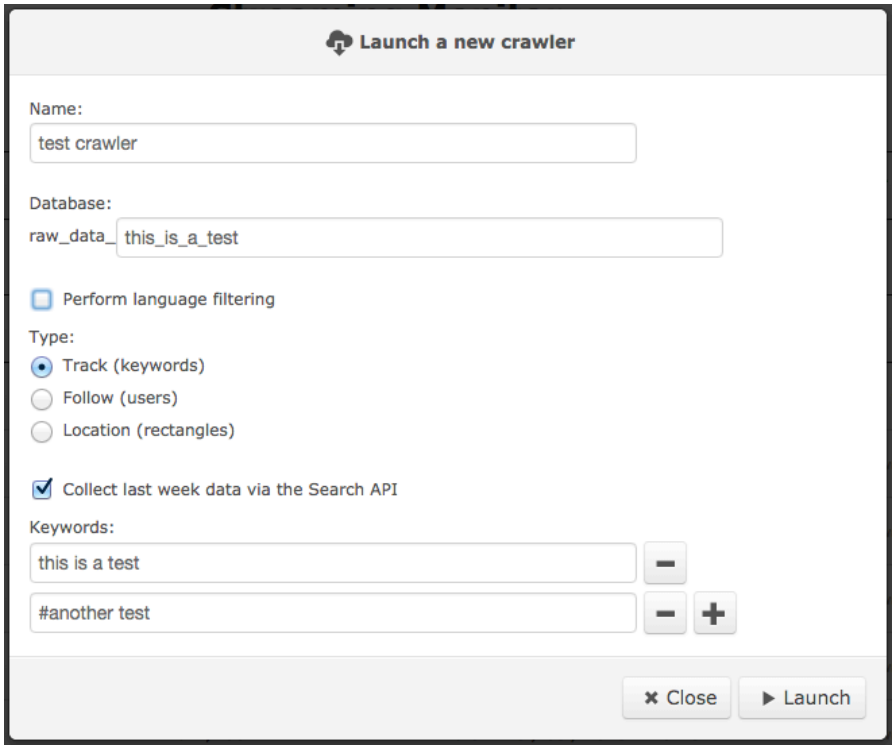

Figure 2: Popup and Web form to launch a new crawler.

Specifically, the Twitter Monitor works with three different types of searching parameters:

- **Keywords - Track mode**: in this way tweets containing specific keywords will be collected. It is possible to indicate both hashtags or simple keywords. In this case it is also possible to retrieve data published during the last week by accessing the Twitter Search API [3]. A maximum number of 400 keywords per crawler can be specified.
- **Users - Follow mode**: by using this mode it is possible to collect tweets published or mentioning a specific set of Twitter users. In contrast with the **Keywords - Track mode**, here the focus is on users instead of the content of tweets. In this mode of operation, a maximum number of 5000users per crawler can be specified.

- **Rectangles - Location mode**: this functionality allows to collect tweets published within a specific geographical area, defined by the lon/lat coordinates of the corners of a bounding box. The bounding box is represented as a geographical rectangle. The coordinates needed to define a bounding box are those of its south-west and north-east corners. In this mode of operation, a maximum number of 25bounding boxes per crawler can be specified.

From the main page of the Twitter Monitor it is also possible to check and interact with crawlers that have been launched previously. In particular it is possible to:

- check the status of crawlers (whether *running*, *paused*, or with *errors*);
- pause or permanentlyremove crawlers;
- add/remove keywords/users/bounding boxes from a running or paused crawler;
- check the amount of data already collected.

Figure 3: Twitter Monitor dashboard showing the details of every crawler.

Fig. 3 shows the Twitter Monitor dashboard. The dashboard is a table listing a crawler for each row and detailed information in each column. For each crawler, the following information are provided:

- status of the crawler (whether *running*, *paused*, or with *errors*);
- name of the crawler;
- number of specified crawling parameters(whether *keywords*, *users*, or *bounding boxes*, depending of the type of the crawler);
- date and time when the crawler was launched;
- number of tweets collected so far;
- a list of possible operations that can be performed on a crawler:
    - add a new searching parameter;
    - pause the crawler: the Twitter Monitor stops crawling data related to the searching parameters. The crawler and all its related parameters will still be displayed in thedashboard. Furthermore, it will be possibile to resume a paused crawler without

needing to insert the configuration parameters once more. Data already collected will remain in the database;
- resume a paused crawler;
- permanently remove a crawler: the crawler will be stopped and its configuration parameters will be permanently deleted. The crawler will not be listed in the dashboard anymore. Data already collected will remain in the database;
- retrieve last week data: launches a new historical crawler exploiting Twitter Search API[3] for retreiving last week's data. This functionality is only available for **Keywords - Track mode** crawlers.

3.1.2 TECHNICAL DESCRIPTION

This section lists the main implementation details of the Twitter Monitor, as well as the main technologies it is built upon.

The Twitter Monitor has been developed using PHP 5.5.9-1ubuntu4.4. At the time of writing, the backend of the application is hosted on a virtual machine running an Ubuntu 14.04 LTS (GNU/Linux 3.13.0-24-generic x86_64) operative system. As already detailed in Section 3.1.1, the application is accessible via an interactive and responsive Web interface. The interface has been developed using AJAX, HTML5, CSS3, Javascript, and the Bootstrap[4] Javascript library.Users can perform a number of complex operation simply by interacting with the Web interface of the Twitter Monitor. Actions performed by users on the Web interface are translated by our system's PHP APIs into specific API requests issued to Twitter.

Regarding data capturing, the Twitter Monitor fetches data from Twitter streams in a real-time fashion. Data is acquired in JSON format and stored in a number of MySQL databases. The Twitter Monitor uses MySQL version 5.6.17-0ubuntu0.14.04.1. Once launched, a streaming crawler remains active until the user explicitly issues a command to *pause it* or to *permanently remove it*. Every crawler keeps track of the operations it performs, of collected data, and of possible errors in a separate log file. Log files are scanned and analyzed periodically by the Twitter Monitor in order to guarantee that crawlers are working without problems and to minimize possible loss of data.The checks on system's and cralwers' logs, as well as the scheduling of tasks, are performed by a script installed in the Unix crontab of the machine hosting the application.

3.2 INTEGRATION INTO THE SOBIGDATA E-INFRASTRUCTURE

As shown in Fig. 1, the Twitter Monitor is already loosely integrated within the SoBigData RI. It has been integrated as a specific tab from within the SoBigData VRE. As such, it is easily reachable and accessible by every user who is registered to the SoBigData VRE. The service running the Twitter Monitor application is hosted on a sigle, centralized, virtual machine of IIT-CNR, in Pisa, Italy.

In the long run, this integration choice might represent a limitation to the scalability of the service, since a single virtual machine will hardly serve all the requests coming from the SoBigData RI users. However, this choice represents a quick way to integrate the tool, thus allowing us to provide a fully functional version of

the Twitter Monitor tool also in the first months of the project. In the upcoming Section 4 of this document are described the plans related to the implementation of a new integration strategy for the Twitter Monitor. This plan will result in the Twitter Monitor being tightly integrated within the SoBigData RI, thus providing searchability, extendibility, and scalability of its services by means of parallelization on multiple nodes of the RI.

4 CROWDSOURCING TECHNOLOGY

Crowdsourcing is complementary to crowdsensing, in that it engages human volunteers or paid-for contributors to provide judgements and inputs on useful computational tasks, such as image tagging. As such, crowdsourcing is a method for data collection, which similar to crowdsensing, relies on the wisdom of the crowd to collect valuable data.

The ability to hardness crowdsourcing effectively is particularly important for SoBigData, as many existing gold-standard datasets and evaluation corpora were created before social media became ubiquitous and therefore, they do not reflect well the kinds of language and culture encountered in social media. Reliable natural language processing of social media, in particular, has been shown to be still an open challenge, due to lack of sufficient human-annotated training data. An increasing number of researchers have turned recently to crowdsourcing, as affordable and scalable means for the annotating such much needed data.

The open-source GATE text analytics infrastructure (which forms the key pillar for text analytics tools in SoBigData) comes already with some support for crowdsourcing annotations on text classification (e.g. sentiment annotation, part-of-speech tagging) and sequence annotation (e.g. named entity recognition, noun phrase chunking) tasks. It harnesses CrowdFlower as the underlying crowdsourcing platform.

Figure 4 shows the GATE user interface for automatic creation of classification jobs on CrowdFlower, based on linguistic annotations and document text. In this example, the aim is to create tasks asking the crowd to disambiguate mentions of person names to the correct Wikipedia entry or indicate that no suitable entity is shown. As shown in Figure 5, the automatically generated CrowdFlower tasks contain in this case tweet text, any clickable URLs, and disambiguation options. All linguistic information, including the Wikipedia entries, is added automatically to the CrowdFlower jobs when OK is selected in the GATE Crowdsourcing GUI.

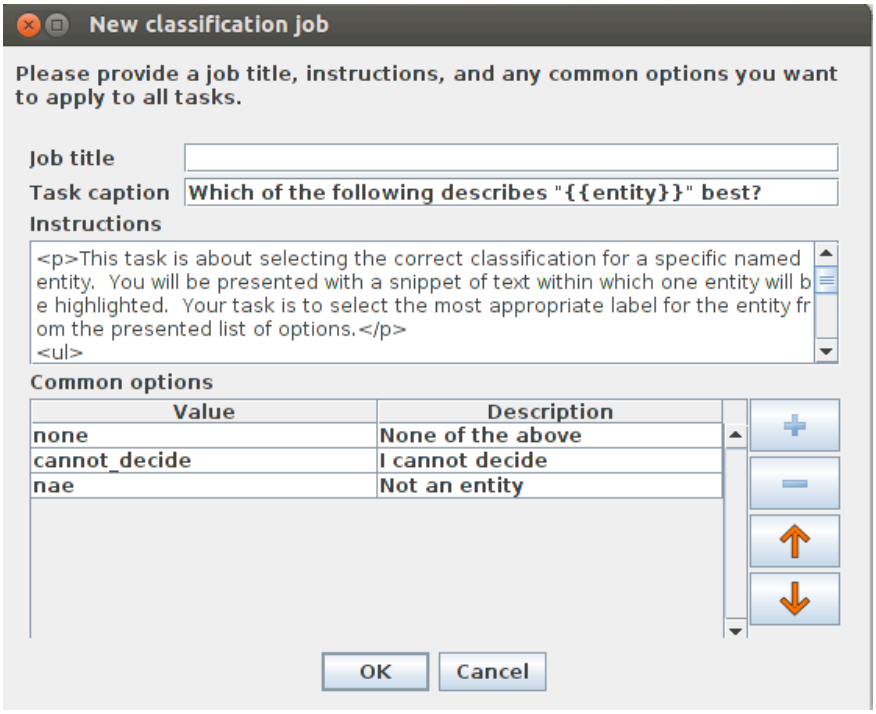

Figure 4 GATE Crowdsourcing Interface for Creating CrowdFlower Classification Tasks Automatically

Figure 5 An Example Entity Disambiguation Task in CrowdFlower

As part of our work on SoBigData, we extended the GATE Crowdsourcing plugin to be able to crowdsource input not just from tweet/document text, but instead from relevant metadata, such as Twitter user profile text. This was required in order to collect a dataset for training and evaluation of methods for entity classification of Twitter user names (persons, locations, organisations, products, other).

Another limitation of the previous version of the GATE Crowdsourcing plugin was that it did not allow the crowd to correct machine-made judgement, but instead assumed all crowd input is provided from scratch. This is limiting also in circumstances where we want to show the crowd annotations already made by other human volunteers/crowd workers.

Therefore, as part of our work here, we updated the plugin and the automatically generated CrowdFlower job interfaces to support cases where some selections are made already, e.g. one of the radio buttons in Figure 5 is already selected, based on the output of an entity disambiguation algorithm in GATE.

The updated code is available as open-source, under an LGPL license, with details available here:

https://gate.ac.uk/wiki/crowdsourcing.html

Integration in the SoBigData VRE is forthcoming, as part of future work.

5 FUTURE ACTIVITIES

This section describes the future activities of T8.2 with regards to both the development of new functionalities and the plans for a tighter integration of crowdsensing tools within the SoBigData RI.

5.1 INCLUDING SUPPORT FOR HYBRID CROWDSENSING METHODS

Section 2 highlighted fundamental differences, strenghts, and drawbacksof those systems based on a participatory sensing approach, with regards to those that are based on an opportunistic approach. Indeed, many participatory sensing systems are heavily focused on data collection problems and on proposing mechanisms to obtain and maintain users engagement with the system. In contrast, systems that are based on opportunistic sensing are much more focused on data analysis than data collection. This is because such systems intrinsically have access to a large user base, since they do not require any action from the users. Every user of a social networking platform or social media, is a potential contributor of an opportunistic sensing system, simply because she spontaneously shares content on those platforms. However, since data collected in this way is highly heterogeneous and diverse, opportunistic sensing systems have to invest much more on data preprocessing and analysis. In turn, this brief analysis on systems based on traditional sensing paradigms motivates the development of a hybrid approach, capable of combining the strengths of participatory and opportunistic sensing.

With the aim of combining the strengths of participatory and opportunistic approaches, and overcoming their limitations, within the SoBigData RI and its crowdsensing technologies we plan to provide support for a novel crowdsensing paradigm called *hybrid* crowdsensing. This new paradigm is based on the combination of two sensing phases. In the first phase, we can exploit advantages of the opportunistic paradigm with a twofold goal: (i) gather as much information as possible, and do so as fast as possible, by drawing upon data that is readily available, and (ii) discover possible participants for the subsequent phase. In the second phase, we can exploit a participatory approach by contacting a subset of users previously discovered and by directly asking them to provide more focused and more detailed information. As a result of the two sensing phases, a sensing system implementing the hybrid crowdsensing paradigm could be fully automated and capable of acquiring user contributions without the need of human intervention.

Therefore, among the future activities of T8.2 is the development of a tool able to leverage this hybrid crowdsensing paradigm. Such tool will be integrated within the Twitter Monitor that is already available to the users of the SoBigData RI, thus allowing such users to perform more complex and efficient crowdsensing tasks.

5.2 FURTHER INTEGRATION ACTIVITIES

Future activities of T8.2 will also focus on providing a better and tighter integration of the Twitter Monitor tool within the SoBigData RI. In particular, the code of the actual Twitter Monitor is already being refactored with the goal of splitting it in smaller and modular software components. New software components will include:

- a software module implementing the user interface (front-end);

- a software module representing the back-end of the Twitter Monitor and implementing the functionalities of the crawler scheduler and of the reliability and availability mechanisms;
- a set of software modules implementing the Twitter crawlers.

Such components will then be individually integrated and deployed in the SoBigData RI. As a result, software components, such as those implementing the Twitter crawlers, will be instantiated and deployed on-the-fly, should the need arise due to a peak of requests from the RI users. In turn, this will allow to provide extendibility and scalability of the Twitter Monitor services by means of parallelization on multiple nodes of the RI.

Other integration activities will be undertaken in order to allow the output of the Twitter Monitor (the data downloaded from Twitter) to be directly fed to other methods and algorithms integrated within the SoBigData RI.

5.3 ENRICHMENT OF COLLECTED TWITTER DATA

Additional components are under development to collect data on the social network connections of sets of Twitter users (which users they follow or which users follow them), and the content of URLs referenced within collected tweets, to allow for more detailed analysis than can be performed on the short tweets alone. These components can be seeded with data collected by the Twitter Monitor and/or by the results of other analytics components in the RI. They will be exposed via the SoBigData RI, as well as via the GATE Cloud infrastructure-as-a-service (https://cloud.gate.ac.uk) which, in turn, will be integratedwithin the OpenMinTeD infrastructure.

6 CONCLUSION

This document presented currently available crowdsensing methods and the crowdsensing tools developed and integrated within the SoBigData RI to allow its users to perform crowdsensing tasks. Current features and technicals details of the Twitter Monitor crowdsensing tools have been described, as well as the plans for providing further functionalities and for realising a tighter integration within the RI.

REFERENCES

[1] Twitter.com, (2016), *About Twitter, Inc.*, [Online] Available at: https://about.twitter.com/company

[2] Twitter.com, (2016), *The Streaming APIs*, [Online] Available at: https://dev.twitter.com/streaming/overview

[3] Twitter.com, (2016), *The Search API*, [Online] Available at: https://dev.twitter.com/rest/public/search

[4] Bootstrap, (2016), *About Bootstrap*, [Online] Available at: http://getbootstrap.com/about/