IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 December 2022, accepted 27 January 2023, date of publication 9 February 2023, date of current version 17 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243791

RESEARCH ARTICLE

A Citizen Science Approach for Analyzing Social Media With Crowdsourcing

CARLO BONO[1], (Graduate Student Member, IEEE), MEHMET OĞUZ MÜLÂYİM[2], CINZIA CAPPIELLO[1], MARK JAMES CARMAN[1], JESUS CERQUIDES[2], JOSE LUIS FERNANDEZ-MARQUEZ[3], MARIA ROSA MONDARDINI[4], EDOARDO RAMALLI[1], (Graduate Student Member, IEEE), AND BARBARA PERNICI[1], (Senior Member, IEEE)

[1]DEIB, Politecnico di Milano, 20133 Milan, Italy
[2]Artificial Intelligence Research Institute (IIIA-CSIC), 08193 Cerdanyola del Vallès, Spain
[3]Centre Universitaire d'Informatique, University of Geneva, 1211 Geneva, Switzerland
[4]Citizen Science Center Zurich (UZH and ETHZ), 8006 Zürich, Switzerland

Corresponding author: Barbara Pernici (barbara.pernici@polimi.it)

This work was supported by the European Commission Horizon 2020 (H2020) Project Crowd4SDG "Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience" under Grant 872944.

ABSTRACT Social media have the potential to provide timely information about emergency situations and sudden events. However, finding relevant information among the millions of posts being added every day can be difficult, and in current approaches developing an automatic data analysis project requires time and technical skills. This work presents a new approach for the analysis of social media posts, based on configurable automatic classification combined with Citizen Science methodologies. The process is facilitated by a set of flexible, automatic and open-source data processing tools called the Citizen Science Solution Kit. The kit provides a comprehensive set of tools that can be used and personalized in different situations, particularly during natural emergencies, starting from images and text contained in the posts. The tools can be employed by citizen scientists for filtering, classifying, and geolocating the content with a human-in-the-loop approach to support the data analyst, including feedback and suggestions on how to configure the automated tools, and techniques to gather inputs from citizens. Using flooding scenario as a guiding example, this paper illustrates the structure and functioning of the different tools proposed to support citizens scientists in their projects, and a methodological approach to their use. The process is then validated by discussing three case studies based on the Albania earthquake of 2019, the Covid-19 pandemic, and the Thailand floods of 2021. The results suggest that a flexible approach to tools composition and configuration can support a timely setup of an analysis project by citizen scientists, especially in case of emergencies in unexpected locations.

INDEX TERMS Citizen science, crowdsourcing, data analysis pipelines, emergencies, image filtering, social media analysis.

I. INTRODUCTION

The extraction of information from social media is challenging. It often requires the analyses of large volumes of data with a wide variety of information, made of short and informal text with no structure and possibly containing noise, i.e., information that is unrelated or not useful. However, social media posts have huge potential since they often document ongoing events and can provide timely information on developing situations, including images that can support further knowledge extraction.

Citizens already play an important role in this process by posting their contributions on social media platforms, from where these contributions can be retrieved using different

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

information discovery techniques. The role of citizens can then be further expanded with Citizen Science (CS) methodologies and tools. CS is defined in the Oxford dictionary as ''scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions''.[1] Several methods for contributing in CS have been studied [1], ranging from performing data collection and data analysis tasks, to actively providing ideas and evidence to decision makers.

This paper proposes an approach based on data processing pipelines, in which AI-based data analysis tools are combined with crowdsourcing techniques. It describes the results of using this approach to develop different types of projects. Along the pipelines, CS project organizers and data analysts are in control of all phases, including selecting the information sources and how to crawl them, providing input in the configuration of automated analysis tools applied to the data, and allowing citizens to contribute ideas and annotations.

As with many other kinds of CS, social media analysis efforts suffer from several difficulties related to engaging and retaining participants [2]. These include finding participants willing to contribute on a voluntary (unpaid) basis, and giving them the ''right tasks'' to perform, i.e., appropriate to their knowledge and skill level. Accordingly, the research illustrated in this manuscript focuses on the need to be effective when assigning tasks to citizens, providing them only relevant posts for analysis, and, most of all, involving citizens scientists in decisions related to how best to perform these task. Another research focus of the approach is to allow citizen scientists, as project organizers, to choose between different forms of consensus assessment for participants performing assigned tasks, including advanced methods for consensus analysis.

One of the main results of the efforts to address these challenges is a set of dedicated tools called the *Citizen Science Solution Kit* (CSSK). The kit includes several tools, including a tool for social media data filtering (VisualCit), a tool for crowdsourcing social media data analysis (the CS Project Builder, or CSPB), and a tool for crowd analysis (Crowdnalysis). These components can be used independently or in combination, and they can be easily configured by citizen science project organizers for the needs of specific projects. The kit provides CS project organizers with flexible ways to maintain full control over the project without the need for technical interventions by tool providers during its realization, and allows them to combine the tools in various ways depending on their analysis goals. The CSSK has been developed primarily to explore the impact of leveraging CS to achieve Sustainable Development Goals (SDGs), as will be detailed later.

The remainder of this paper is organized as follows. Section II presents both the state of the art and open research issues. In Section III, based on the steps of a plausible flooding scenario, we describe the overall approach of incorporating CS in the development of social media analysis pipelines. In Section IV, we present the different components of the kit, including the architecture of tool composition, a detailed illustration of each tool, and a methodological approach to their use. Section V focuses on validation for CS projects realized with the CSSK and compares the results obtained in three different case studies. Finally, conclusions and future research directions are discussed in Section VI.

II. RELATED WORK

CS has been proposed as a way to support the collection of data for reporting on the SDG indicators. In the paper ''Mapping Citizen Science contributions to the UN sustainable development goals'' [1], the authors perform a detailed analysis of existing projects already covering some of the SDG indicators and suggest potential areas for further development of new initiatives, including climate change. These are areas where collection of information is difficult with more traditional methods, such as household surveys and censuses. The United Nations Statistics Division[2] also mentions this, which highlights how achieving and monitoring the SDGs require innovative ways to produce and apply data and statistics in addressing the multifaceted challenges of sustainable development.[3]

In the literature, social media is considered a form of crowdsourcing and not necessarily an activity of CS [3]. However, the effect of combining the two approaches in a multi-dimensional framework has been studied in different research communities [4]. Also, the combination of automatic analysis of images contained in posts and the collaboration of human-based computing has been discussed by several authors in the literature [5], [6], [7]. For automatic analysis of images, many recent approaches are based on AI and, in particular, neural networks [8], [9]. Hybrid deep-learning and crowdsourcing approaches were analyzed as well [10].

Social media presents a huge potential to provide key information for monitoring SDGs and supporting decision makers. Thanks to their ability to provide timely information about ongoing events [9], [11], [12], their role has been studied in several contexts, e.g., earthquakes [13], demonstrating how this timeliness can help prevent further losses. The EU Crowd4SDG project has explored the potential of social media to provide key statistics and local indicators [14] based on lessons learned from case studies for SDG monitoring.

Many research papers have been published on social media for emergencies, including a special track at the International Conference on Information Systems for Crisis Response and Management [15]. An analysis of the opportunities and challenges of such applications [16] revealed how several challenges are related to the need of geolocating posts to leverage their contents, and the issues of analyzing posts in multilingual contexts [17], [18]. Several approaches for

[1] https://povesham.wordpress.com/2014/09/10/citizen-science-in-oxford-english-dictionary/

[2] https://unstats.un.org/UNSDWebsite/

[3] https://unstats.un.org/sdgs/report/2017/harnessing

TABLE 1. Comparison of approaches for crowd-sourced social media analysis. For the management of the tools/platforms, the roles of Tool Owners (TO) and Project Organizers (PO) are highlighted.

Approach	Description	Autonomous pipeline preparation			Data Collection from Social Media			Crowdsourcing	Flexibility and scalability
			TO	PO		TO	PO		
Automatic Tweet Analysis Platforms	Predefined services and tools for specific purposes, usually emergencies, e.g., [12], [17]	Pipelines are predefined	✓		Crawling with keywords with either text or images considered	✓	✓	Sometimes used for quality evaluation	Depends on infrastructure
		Focus on pre-trained ML classifiers, focus on images and geolocation	✓						
Combining human and machine computing	Automatic classifiers incrementally trained for a specific context, e.g., in AIDR [7]–[9]	Pipelines are predefined	✓		Crawling with keywords with focus on images	✓	✓	In some platforms used for classification evaluation	Depends on infrastructure
		Leverage on human feedback for improving ML classifiers	✓						
E2mC	Automatic collection of images from social media combined with crowdsourcing for emergencies [5], [6]	Pipelines are predefined	✓		Crawling with keywords with focus on images	✓		Used for classification by citizen scientists	Depends on service provider
		Focus on images and geolocation	✓						
Community platforms	Services provided for collecting information from involved citizens, e.g., Ushahidi [19], Facebook Crisis Response [20]	Predefined services provided for general use	✓		Direct collection of posts within the platform	✓		Used for information collection	Depends on service provider
Generic Citizen Science and crowdsourcing platforms	CS platforms for involving citizens in problem solving tasks, e.g., Zooniverse [21], AWS MTurk [22]	Pipelines are configurable		✓	Can be uploaded as dataset	✓	✓	Used for classification by citizen scientists	Depends on service provider
		Single service provided for general use	✓	✓					
Crowd4SDG Citizen Science Solution Kit	Services provided to enable citizen scientists configure their own project and collect data from social media	Pipelines are configurable	✓	✓	Crawling with keywords focus on text, images and geolocation	✓	✓	Powered by CS with advanced and tailorable concensus mechanism	Extensible, scalable, and reproducible (Open Source)
		Multiple services provided for general use	✓	✓					

geolocation and geocoding have been proposed in the literature [19]. Precision in the location is important, but difficult to achieve as one can not rely on native locations of, for example, tweets. This is due both to their scarcity (most users keep location information disabled, as per the default setting), and because the location of the posting is often different from the location of the event occurring during an emergency. Many of these challenges have been addressed by the combination of automatic analysis and crowdsourcing techniques. Automatic analysis enables the processing of millions of tweets in an almost real-time fashion, while crowdsourcing is used to curate the final dataset by validating and complementing the information, overcoming the limitations of automatic techniques (see for example [6]).

For emergencies, some community platforms have been developed over the years. The first and most known example is Ushahidi [20], which allows crowdsourcing information about an ongoing crisis and displaying the information on maps. Other social media, e.g., Facebook with the Crisis Response Service [21], provide an open space to collect awareness information about emergency situations and notification services.

Generic Citizen Science and Crowdsourcing tools such as Zooniverse [22] and Amazon MTurk [23] allow the design and implementation of interfaces to enable the crowd to perform social media data analysis. In these cases, the social media data needs to be updated and can not be mined from those platforms.

In this study, we focus on combining automatic analysis of images present in socical media posts, with geolocation of the posts and crowdsourcing, inserting a human-in-the-loop approach in all phases. The goal is to refine the automatically extracted information and improve the selection of relevant posts. The CSSK presented in this study enables the implementation of such ‘‘analysis pipelines’’ and the easy combination and configuration of tools.

To illustrate the main differences and characteristics of the above mentioned approaches, Table 1 summarizes and compares some of them, representing the state of the art in the field. They include tweet analysis, combining human and machine computing, emergency image extraction from Twitter (E2mC), community platforms, and generic CS and crowdsourcing platforms. The comparison is done by analyzing each approach along four dimensions:

1) *Autonomous pipeline preparation*: as CS initiatives are often created bottom-up, organizers may want to develop their data collection and analysis autonomously. This dimension looks at the availability of configurable functionalities and services (e.g., filters) that allow the project organizers – ideally without the need for coding – to collect, filter, clean and analyze social media data.
2) *Data collection from social media*: this dimension looks at the ability of the system to dynamically crawl social media with keywords selected by the user, automatically handling multimedia content, and visualizing the collected data, including spatial data.
3) *Crowdsourcing*: examines the availability of crowdsourcing tools and methods to allow both the collection and evaluation of human contributions and feedback.
4) *Flexibility and scalability*: refers to the ability of the platform to easily adapt to specific – and sometimes quickly changing - project requirements, including the ability to scale up to a large number of users. In the table, we focus on how flexible the service provisioning is with respect to its deployment on the computing infrastructure, provisioning as a service by the tool owner or a third party, and availability of the code as open source.

Also highlighted in the table is the possibility by Project Organizers (PO) to configure and execute their own project by acting directly on the available features and services, versus what only Tool Owners (TO), defined as producers or managers of tools and platforms, can do.

The conclusion from this overview of related work is that the CSSK approach brings the solution closer to citizens and communities, and provides a truly unique combination of automatic and crowdsourced data analysis that allows for easy creation of powerful projects, without the need for programming or technical expertise.

III. SOCIAL MEDIA ANALYSIS PIPELINE

Social media have been investigated in many crisis and emergency scenarios, for two main reasons: i) they provide a viable source of information covering areas and situations in which it is difficult to set up systematic data collection, and ii) they potentially allow the extraction of timely information concerning the affected areas. One of the main goals of the research in this area is to automatically extract information to characterize, highlight and react to emergencies. This domain particularly seeks visual information because it provides substantial evidence of ongoing events [5], [24].

Extracting relevant information from tweets presents several challenges. The contents of the tweets must be relevant to the event and to the type of information being searched. The definition of ''relevance'' is specific to each project as it depends on the use of the information and the actors involved, as discussed later.

The extraction process requires several phases of data preparation and analysis. Raw datasets extracted from social media usually contain a very low percentage of useful information, estimated to be between 0.5% and 3% in the literature [17], [25]. However, even a very small number of relevant posts can be useful in providing a significant overview of the areas in which the event occurs [26], [27]. Given the massive number of posts to be examined, a purely manual analysis is not viable.

Figure 1 illustrates the social media analysis approach proposed in this study, interleaving fully automatic and human-supported phases. The information extraction process can be split into two phases: in the first one, the *Data Preparation* phase, data are automatically prepared for the analysis, while in the second phase, *Data Analysis*, further information is provided by citizen scientists and data are processed by data analysts.

Let us take as example the analysis of a flooding event. The first step of the Data Preparation phase is to retrieve potentially ''relevant'' posts from social media (*Social media crawling*), selecting the appropriate query for the search. The selection of keywords is critical, as it needs to maximize the amount of useful information while minimizing the collection of irrelevant content contributed by generic keywords. The native geolocation of posts can be also used for filtering the data. However, as the proportion of natively geolocated tweets is very low (estimates are around 0.3% to 3%), there is a high risk of excluding useful information if only the geolocated tweets are considered.

The second step, *Cleaning and filtering*, aims at filtering posts to retain only those that are possibly relevant to the event, as quality issues are often present when extracting information from social media. Not only could texts and images be not relevant, but they may also contain noise in the form of fake news or biases. Moreover, posts frequently contain generic texts, images with memes, or information only marginally related to the event. Filtering mechanisms must be put in place to clean the data and prepare it for analysis. This includes the removal of duplicates (both exact duplicates and near duplicates) and basic cleaning actions on the raw data, e.g., the removal of non-informative text. Several models for automatic image and text classification in social media have been developed, mainly based on deep neural networks (e.g., [24]). One of the open problems, beyond the development of the classifiers themselves, is the selection and configuration of classifiers tailored to a given situation, as events can present distinctive characteristics in different contexts and locations around the world. As a result, the *Cleaning and filtering phase* is usually a rather complicated one, which is often developed specifically for each data analysis project. In the proposed approach, this phase is made easily configurable by CS without technical assistance, based on a set of available cleaning and filtering components.

The third step is an *Automatic annotation* to enrich posts with additional information derived from the automatic analysis of their content. It includes labeling through automatic classification, for example, classification of the impact level of a disaster based on visible damages in an image. It may also

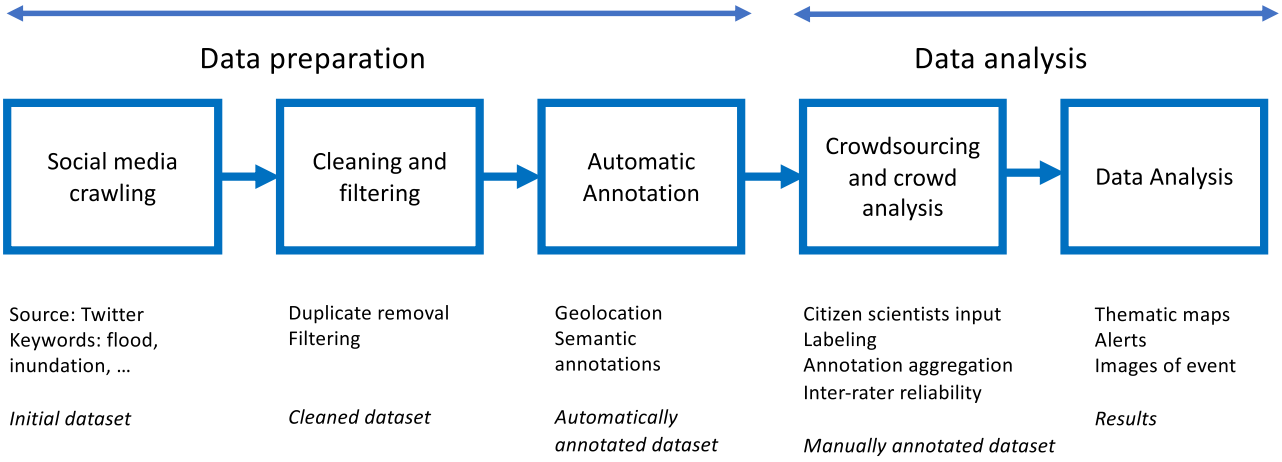


FIGURE 1. Process of social media analysis using the combined "human-computing" approach discussed in this study.

include automatic geolocation of posts that are not natively geolocated. This information is critical for identifying the posts which are actually relevant for the events to be studied.

The curated dataset is now ready for the *Data Analysis* phase, in which data analysts can extract information and explore hypotheses, for example determine whether a given area was affected by an event or not. To support this analysis, CS can be leveraged to provide additional contributions from citizen scientists. For example, they can check and refine the output of the automatic annotation and cleaning performed in the previous steps. CS activities generally consist in labeling and/or annotating posts, but may include tasks that are inherently difficult, e.g., identifying if there are persons in a blurred image from a distance. As citizen scientists are often not a specialized crowd, disagreements may arise – indeed even experts sometimes disagree with one another. A systematic analysis of the crowdsourcing-derived information must then be performed to assess its validity, aggregating crowd answers and assessing inter-rater reliability (as a whole, *Crowdsourcing and crowd analysis* step).

Only after this step, the actual *Data Analysis* can be performed to derive valuable knowledge. Different types of analysis can be envisioned, and the following are considered in this study:

- *Images of events*: aimed at retrieving images from a defined location, usually visualizing them on interactive maps or integrating them in a GIS (Geographical Information System).
- *Thematic mapping*: evidence, such as the intensity of a flood event emerging from the prepared posts, is aggregated in given geographical areas, normalized and often weighted, e.g., based on population information.
- *Alerts*: timeliness is the primary goal of this analysis, that aims at generating early alerts for a starting event. In this case, the balance between the manual and automatic activities to be performed must be assessed, as the results are useful if they are delivered in the first hours after the event onset [10].

A more advanced type of analysis would be to solicit and then tackle problems proposed by the citizens themselves via crowdsourcing. This requires the ability to organize and analyze proposals, that can be provided by decision support tools, such as for instance Decidim4CS.[4] We do not discuss these tools further in the present study, which mainly focuses on social media analysis.

When applying the proposed pipeline to the different analysis envisioned, some critical issues emerge. They come from the experience of projects such as Crowd4SDG[5] and E2mC [5]), as well as from the literature (see Section II):

- Depending on the type of task being performed, *time constraints* can be more or less stringent. In particular, there is a significant difference between alerts, which must be almost immediate to be useful, and thematic mappings, which are usually expected some time after the event. In some cases, there can be little or no time for human interventions in processing posts. Timeliness is a key parameter to be considered when designing a system in this context.
- While the number of retrieved posts can be high, relevant and useful posts may be few due to local circumstances or situations that may hinder the use of social media. A sufficient level of recall is required during the filtering activities. Moreover, the recall of relevant posts and the specificity of the collected data must usually be balanced for practical purposes.
- *Usability* of the tools is also an important aspect. Tools should be designed keeping the non-IT final users in mind. They should be flexible in their configuration and provide user-friendly interfaces and interchangeable data formats.

[4]https://decidim4cs.ml/
[5]https://crowd4sdg.eu/

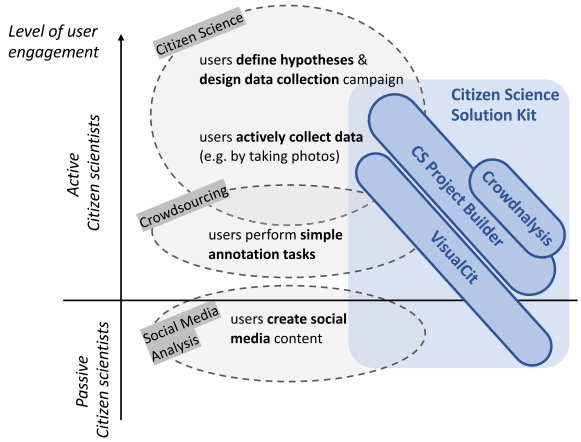


FIGURE 2. Citizen Science Solution Kit levels of engagement.

IV. CITIZEN SCIENCE SOLUTION KIT

In this section we present the *Crowd4SDG Citizen Science Solution Kit*[6] (CSSK). The aim of the Crowd4SDG (Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience) H2020 European project is mining the contribution of citizen scientists to tackle critical issues related to the UN Sustainable Development Goals (SDG).

Crowd4SDG focuses on SDG 13 - Climate action - and on ways to reduce the impact of natural disasters and emergency events on the population. The project uses SDG-related challenges to explore different modalities for citizens' involvement.[7] Citizen scientists are invited to propose ideas for projects to tackle the challenges. Selected projects/proponents join a multi-phase coaching program that provides them with crowdsourcing methodologies and tools. The process is used both to refine project ideas and to evaluate the tools offered for their implementation. A particular interest of Crowd4SDG is to explore innovative approaches and tools that support CS projects for the analysis of content posted in social media. The approaches advocate for an active role of citizen scientists in all phases of social media analysis.

A. CONCEPTUAL FRAMEWORK

The CSSK has been developed as a set of flexible tools able to support the collection, curation, and analysis of data for different goals and applications. The kit supports both automatic computation and citizen participation in the analysis process.

In particular, Crowd4SDG has been using the CSSK as an integrated framework for social media analysis based on three of its components: VisualCit, CS Project Builder, and Crowdnalysis. The framework supports analysis pipelines that progressively transform a raw dataset of information collected on social media into a curated — i.e., cleaned, processed and annotated — set ready for data analysis, following the approach illustrated in the previous section. In the curated dataset, posts have been cleaned from irrelevant information, marked with automatically extracted geographical information, and further localized and annotated by citizen scientists via crowdsourcing.

The three components of the CSSK have different roles and functionalities. The goal of VisualCit (short for *Visual Cit*izen Scientist) is to enable citizen scientists and data analysts to query social media in order to obtain data and to interactively define automated filters and annotations to clean and refine the selected data. This allows reducing the number of irrelevant posts that will then feed into the manual analysis. The Citizen Science Project Builder (CSPB or CS Project Builder) allows the creation of crowdsourced data analysis projects where participants perform various analysis tasks on existing digital data (text, images, etc.). This data may originate from citizen-generated content on social media. Tasks performed can include confirming the relevance of selected images/texts, identifying information in images/texts and contributing new information by assigning predefined labels or annotating them with free text, and so on. As crowdsourced information is known to be subject to possible biases or errors by the contributors, an overall analysis of the crowd behavior is necessary to identify possible problems and to aggregate and evaluate citizens' contributions. The analysis of crowdsourced information is performed by the Crowdnalysis tool. Each of the CSSK components is described in detail in the following sections.

One of the key elements to consider when designing tools based on citizen participation is their desired "level of engagement", i.e., how actively citizens are supposed to provide support, process information, and suggest solutions. For geographical applications, Craglia and Shanley propose an analysis framework [28] in which candidate tools and initiatives are classified according to the level of required engagement of citizens and the implicit/explicit nature of the geographical information. Fig. 2 uses the level of engagement, as defined in [28], to analyze the CSSK components and their associated activities. In the present study, we consider both active and passive citizen participation. Purely passive participation consists in providing information about ongoing events by posting on social media without being aware or interested in the possible use of the information. In this case, the CSSK can process the information without further citizen involvement. In other scenarios, we consider a more active participation, where citizen scientists can contribute by analyzing, validating, and annotating information, or contributing new data. In VisualCit for example, the focus is on collecting information from passive, citizen-generated content. However, citizens can also actively participate in defining the type of information to be collected. VisualCit can also be used as an active collection tool, in cases where specific hashtags focusing on a given event or situation are used [29].

VisualCit, CSPB, and Crowdnalysis are flexible tools and can be combined in various ways. Fig. 3 illustrates several possible uses and combinations, which have been

[6] https://crowd4sdg.github.io/

[7] Some Crowd4SDG challenges are presented in https://crowd4sdg.eu/gear-cycle-01/ and https://crowd4sdg.eu/gear-cycle-02/

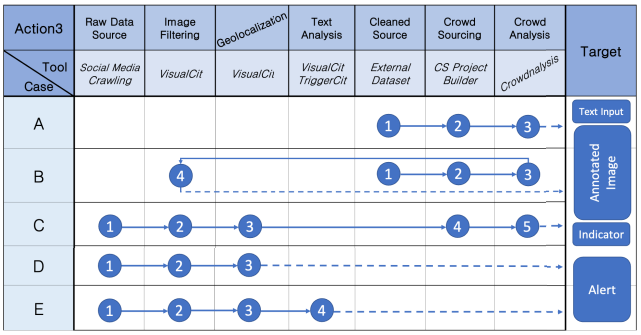


FIGURE 3. Use cases for the Citizen Science Solution Kit.

experimented for scenarios similar to the one discussed in Section III. Use case (A) starts with an existing set of data (e.g., images) that need to be analysed with the help of the crowd (CSPB) and the results are verified (Crowdnalysis). This can be used both for collecting different kinds of annotations and text input (e.g., descriptions or translations) or for keywords-based tagging. In use case (B), crowdsourced information is filtered by VisualCit after crowdsourcing and crowd analysis, to further reduce the percentage of non-relevant data. Case (C) illustrates a full-fledged data preparation and analysis pipeline where all components are used for searching, filtering, annotating and geolocating images, for example to derive indicators or for building a thematic map. In all cases in which CSPB (crowdsourcing) is used, the Crowdnalysis tool enhances the analysis and results of crowdsourced information. The use cases in which the analysis does not include crowdsourcing (D and E) are cases where timeliness is critical (producing alerts), so human activities are minimized.

The above typologies (A-E) and their associated methodology have been extensively validated by Crowd4SDG using large case studies, as discussed in Section V-B.

B. CSSK USER ROLES

To clarify the possible ways to interact with the tools during the pipeline process, we distinguish two main roles in the use of the CSSK: CS project manager and ''citizen scientist''.

CS Project Managers could be single individuals (e.g., researcher or citizen scientist) or teams putting forward an idea and implementing its corresponding data analysis protocol. The CS Project Manager is responsible for defining and refining the goals of the project, the type of analysis to be performed and the desired outcome. CS Project Managers can play three different specific roles:

- *Organizer*: When CS Project Managers have set their goals, they must configure the tools and data preparation pipeline, select the analysis scenario, the volume of data to be analyzed and the frequency of the analysis. If crowdsourcing is part of the scenario, the manager must identify citizens' specific tasks.
- *Executor*: Once the data preparation pipeline has been configured, the CS Project Manager can execute it starting from an initial set of (raw) data. This is an iterative process: the assessment of intermediate results might initiate a reconfiguration of some of the tools to improve quality, or even a revision of the initial goal.
- *Analyst*: In case of crowdsourced tasks, the CS Project Managers should analyze the results of the crowd. This includes the identification of the most appropriate model for computing consensus, and the analysis of the behavior of the crowd, looking for anomalous or opportunistic behavior that might invalidate the results.

With the term ''citizen scientist'', we indicate anybody contributing to the preparation or analysis phases, i.e., anybody involved in performing tasks or providing information useful for the analysis goals.

A visual interface is provided to support configuration operations and quality-improving actions on the VisualCit and CSPB tools. In VisualCit, execution can be performed via an interactive interface for the initial samples, whereas execution on large volumes of data is performed using a service-based architecture. In CSPB, an interactive interface is provided both for the setup and for the execution of tasks by the crowd, and to extract data for analysis by the Project Manager. Both components provide a set of web services that can be invoked via APIs from both interactive and programmatic interfaces. Once the data are prepared with VisualCit, crowdsourcing can be performed. A common data exchange format is defined for the three tools based on CSV files.

C. SELECTING AND FILTERING DATA: VISUALCIT

1) FILTERING

VisualCit is a social media processing tool aimed at exploring and building data processing pipelines. It is meant to be general purpose, pluggable across different data sources, and easily extensible.

VisualCit functionalities are available in two ways. The first is an interactive web interface (see Fig. 4), which is useful for exploring data and checking the results of applying actions to the data. This tool can be used to quickly sketch data processing pipelines and adjust parameters with immediate visual feedback.[8] The second mode is an HTTP web service that exposes the GET and POST methods. The web service accepts single processing requests or composite pipeline configurations, executes them, and sends the results back. A high-level overview of the architecture is presented in Fig. 5.

At its core, VisualCit exposes a set of configurable data processing actions. The action library is mostly aimed at actions that can be applied to media. Supported media are currently images. The classes of data processing actions currently supported by VisualCit are i) *ingestion*, which collects data from social media with configurable criteria; ii) *cleaning*, mainly to reduce the number of duplicate posts and images; iii) *selection*, to select posts that are relevant to the goals of the social media data analysis to be performed

[8] A demo of the interface is available at http://visualcit.polimi.it:20003/

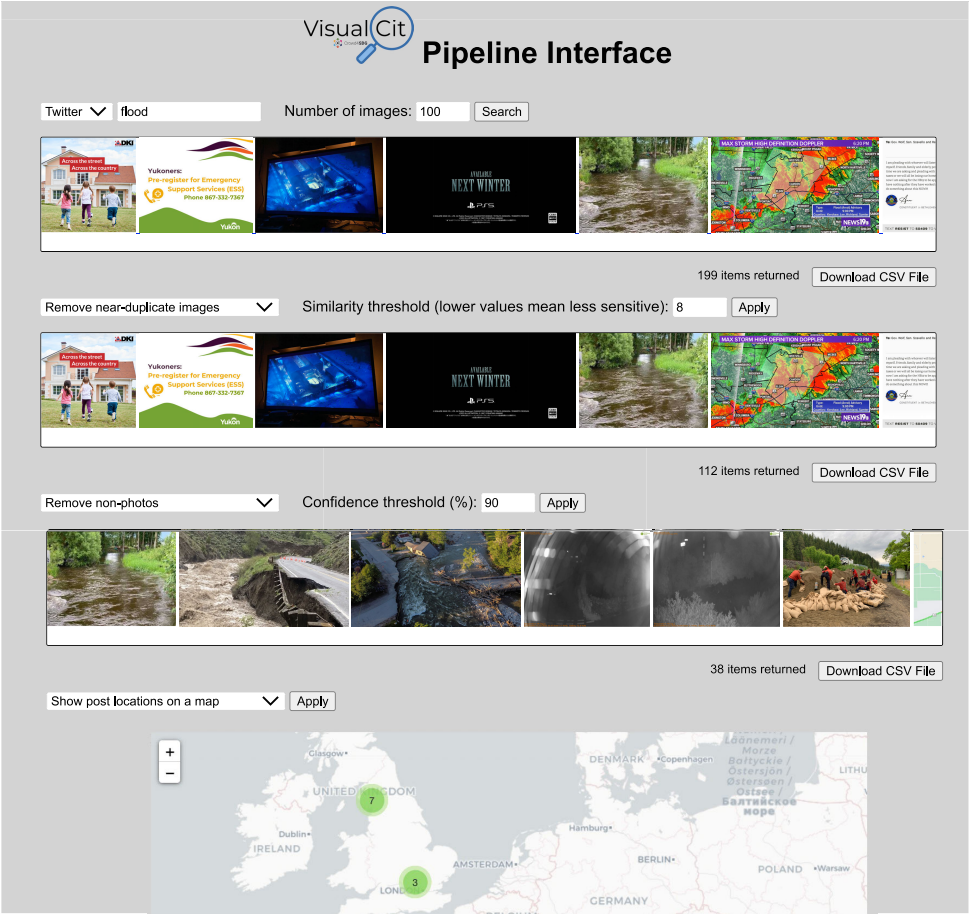


FIGURE 4. VisualCit interactive interface.

(e.g. selecting all images containing a particular type of object); and iv) *augmentation*, to automatically derive further information (such as the location) for the posts being analyzed.

Actions may have one or more configurable parameters, such as confidence thresholds required by Machine Learning based image classifiers. These parameters can be provided from either of the two interfaces. If they are not provided, default configurations are used. In the web interface, parameters can be configured via form fields, whereas in the web service they can be provided as key/value pairs, where keys correspond to the parameter names and values to the parameter values. Key/value pairs can be provided as HTTP request parameters, or alternatively as a JSON payload. The web service also accepts multiple actions to be executed together as a pipeline. This is a natural extension that enables the execution of an ordered list of actions, which can be regarded as a composite transformation or a data preparation pipeline. Currently, in the front-end interface, there is a prototype for exporting a processing configuration that is built interactively. This configuration can be sorted, edited and sent directly to the web service backend, for large-scale and batch processing.

The data interchange format, for both input and output, is based on the CSV format. The minimal set of attributes to be provided consists only of a media URL attribute, referencing the media to be analyzed.

Data can be provided to VisualCit explicitly as records in a CSV file. The data to be processed can also be ingested directly from social media. An adapter for extracting social media posts should be implemented for each compatible platform. Crawling is then configured through a query string, accounting for keywords, key phrases and possibly advanced search operators. Options such as date and location filters are also possible depending on their availability on the selected social media platform. The adapter accounts for paginating the results and building a unique dataset, which can then be processed. Crawling can be invoked as any other action and is executed before all the actions. In this way, querying the data source and processing the results with a pipeline can be performed by providing a single configuration. If no crawling action is provided as a starting point, the data to be elaborated must be sent as a POST payload.

For efficiency reasons, the media content is cached when the first action is executed on an item. Moreover, data can be processed using a configurable level of parallelism. This is

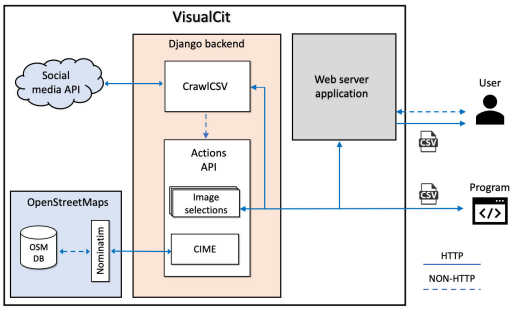


FIGURE 5. VisualCit architecture for selecting filters and annotating geolocations.

done to take advantage of the possible underlying hardware architectures. For CPU computing, parallelism was obtained through an asynchronous thread pool of workers. For GPU computing, if enabled in configuration, the parallelism level represents the number of items concurrently loaded in the GPU memory. This is especially convenient when actions involve the use of certain types of deep neural networks, such as convolutional neural networks (CNNs), which can take advantage of such a setup.

Among cleaning actions, VisualCit implements deduplication functionalities. Deduplication can be performed against 1) identical URLs, to avoid analyzing duplicated links, 2) exactly matching images, through standard hashing techniques, and 3) almost identical images, using a perceptual hashing algorithm, with a configurable similarity threshold.

Selection actions generally leverage CNNs to enable queries on the contents of image data, in order to perform filtering at a semantic level. The application of this kind of deep learning based image-filtering approach have been the subject of previous studies in the field [25] and [30]. Some of the provided classifiers are off-the-shelf, while others are custom. In VisualCit, state-of-the-art object classifiers are available[9] so that queries can be made about the presence of an object (e.g. a person, automobile, etc.) or a number of these objects in the image. Depending on the operational requirements, a confidence level can be provided to accept or reject the presence of an object with customized sensitivity. A scene classifier is also available,[10] enabling queries about specific kinds of scenes or aggregated scenes such as indoor/outdoor spaces or private/public spaces. An action for removing NSFW content is provided.[11] An action for detecting and filtering non-photographic content is also provided in order to remove images corresponding to drawings, screenshots, memes and other types of non-photographic content. This is usually helpful when looking for visual evidence during emergency events. Finally, a custom flood classifier can be used as an example of an event-specific classifier. This classifier attempts to estimate whether the images depict a flood event.

[9]Currently, DETR and YOLOv5.

[10]PlacesCNN, trained on the Places365 database.

[11]NSFW means "not suitable for viewing at most places of employment" according to the Merriam-Webster online dictionary.

VisualCit is not limited to data filtering or selection. Augmentation actions are also possible, for example, by enriching post data with external data sources. A notable example is geolocation, through which geographical positions are automatically associated with social media posts. This is usually critical for obtaining a spatial representation of an event. Geolocalization in VisualCit is performed using the CIME algorithm, which is detailed in the next subsection.

2) GEOTAGGING

To provide evidence that can be used in an emergency response, the selected content is often not useful unless associated with a position. For example, if the specific location depicted in a image is known it may be associated with the image, if not, a wider geographic area (such as a city) might be associated with it. There is need to assign this information to the original data along with some confidence for the annotation. As noted, the number of social media posts that are natively associated with a geographical location is usually small. Posts may also contain other geographical references in textual form that can be exploited to associate possible locations with a post. Candidate locations are extracted from the text using a multilingual named entity recognition (NER) library.[12] The NER functionality extracts strings that could refer to some kind of entity, in our case a geographic location. The language used for the analysis of the posts can be provided externally using a language detection algorithm, possibly propagating known information about the language (e.g., coming from the social media platform), or automatically inferred by CIME. Disambiguation of the candidates must then be performed to isolate meaningful candidates. This is usually a challenging task because the same entity could be referred to with many names, and different places could share a name, or parts of it. To link candidate strings to location entities, Nominatim[13] is used as a gazetteer. Nominatim leverages OpenStreetMap data to link geographical entities to the detected candidates. After disambiguation, the set of most promising results is returned. Further details on the operation of CIME can be found in [17]. Further priors on the location of the post could also be leveraged to enhance the results. For example, information about the language of the post can be used together with associated administrative boundaries to filter out some of the candidate locations.

3) KEYWORD ENRICHMENT MECHANISMS

Using VisualCit, we focus on visual evidence that can support an emergency response. This is somewhat complementary to text-based approaches that already have wide applications in the field. In some tasks, the textual dimension is indispensable. This is the case, for example, in building queries that are used to ingest data from social media platforms. In [31], we focused on light dictionaries to recognize the onset of natural disaster events. An appropriate dictionary

[12]Polyglot https://polyglot.readthedocs.io/en/latest/

[13]https://nominatim.org/

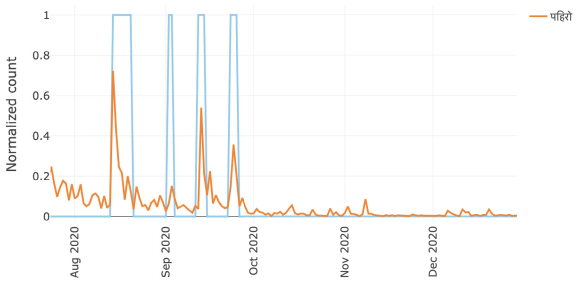


FIGURE 6. Occurrence of the term "landslide" compared with flood events reported by GDACS in Nepal.

can enhance both the recall and precision of the collected data, as discussed in Section III. We then studied how to automatically extract a relevant dictionary when a ground truth about the events of interest was available. An intuition of the approach is illustrated in Fig. 6.

By sampling posts from days in which the events occur, together with days in which they are not, a measure of the correlation between keywords and events can be established. We used time-lagged cross-correlation between the word counts and event onsets to measure the intensity of their relation. A one-sided variant was used to avoid to measure inverse causality. The correlation measures are usually weak but significant in terms of recall augmentation. A manual filtering step, coupled with automatic translation, was also involved. Filtering was implemented using checkboxes. This method enables an effortless construction of recall-oriented, language-specific dictionaries.

At the intersection of image-based and text-based approaches, the available post data with filtered images can be used to augment the results by looking at posts with no attached images. A straightforward approach is to extract tweets that are similar to those selected by VisualCit, which are generally associated with a better relevance to the goal. Textual frequency analysis can then be used to identify keywords that were not initially considered. These new keywords can be used to select other posts, thus obtaining a wider result set (e.g., posts that were not fed to VisualCit because they had no media attached), and extract a new dataset from social media. This approach can be used to increase the number of locations extracted from a dataset. Textual similarity scores were used to rank the correspondence of the posts to those filtered by VisualCit based on media content. Finally, the most promising posts were fed to CIME for geolocation. This can enhance the geographical description of the event, since more data points become available. More details on how this approach can be used are provided in [31] where TriggerCit, a VisualCit extension aimed at generating early alerts, was presented.

D. BUILDING CROWDSOURCING PROJECTS WITH CITIZEN SCIENCE PROJECT BUILDER

Depending on the domain and scientific discipline of application, Citizen Science has multiple practices associated with it. Recent literature [32] proposes a general classification based on the level of engagement and type of Citizen Science activity, namely:

- *Volunteer sensing* - where participants use available sensors (e.g., in smartphones) to collect data that are then used by scientists for analysis.
- *Volunteer computing* - a method in which participants share their unused computing resources on their personal computers, tablets, or smartphones and allows scientists to run complex computer models.
- *Volunteer thinking* (or analyzing) - where participants contribute their ability to recognize patterns or analyze information that will then be used in a scientific project.
- *Self-reporting* - where participants share and compare medical information as both qualitative (self-reported symptoms and illness-narratives) and quantitative data (patient records, genomic and other laboratory test results, and self-tracking health data).
- *Making* - these practices are based on "making" things, often low-cost DIY sensors, and use them to collectively produce knowledge.

The first three categories can also be grouped under the heading "citizen cyberscience", a term created to describe activities that rely on the use of computers and the Internet [33]. The *Citizen Science Project Builder* (CSPB) is a web-based tool that allows researchers, students, and all members of the public to create and run, and also participate in, volunteer thinking projects. Specifically, projects where volunteer contributors are asked to perform complex data classification tasks that are still best performed by human minds and skills, such as classify, tag, describe, or geolocalize existing digital data. Examples include classifying images of snakes, transcribing handwritten German dialect, or describing the content of video clips. In particular, the CSPB supports projects based on digital data in the form of images, text, PDF documents, social media posts (e.g., tweets), sounds and video clips.

In practice, the CSPB is a web interface that provides access to a dashboard where projects can be set up following a simple step-by-step process. Each step is clearly described in a text/image-based tutorial available on the platform, and the process can be summarized as follows (see Fig. 7):

- *Step 1: Project description* - Provide a nice and catchy title for the project, an image to represent it, and some additional text to explain in simple terms the purpose of the project, why it is important, how volunteers can contribute, how their contributions will be used, who is behind the project, and links to additional information available online.
- *Step 2: Data source and type of task* - Select the type of source files to work with. The options include: images (.jpg and .png), videos (.mp4), audios (.mp3), tweets, PDF, and maps. Then, select the task the contributors will need to do, including "survey" format for tasks such as classification, description, and counting, and

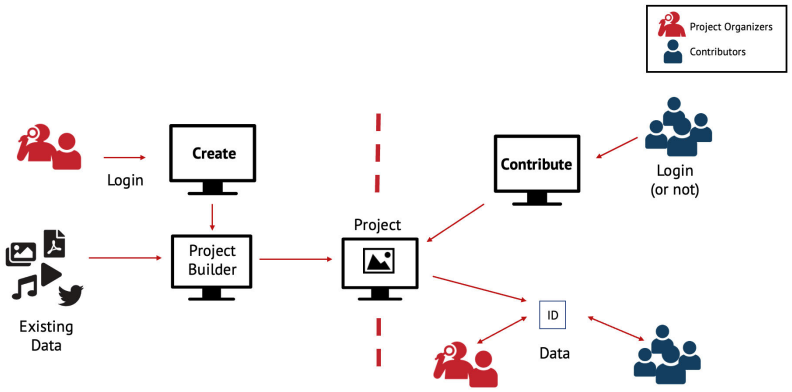


FIGURE 7. CS Project Builder overview.

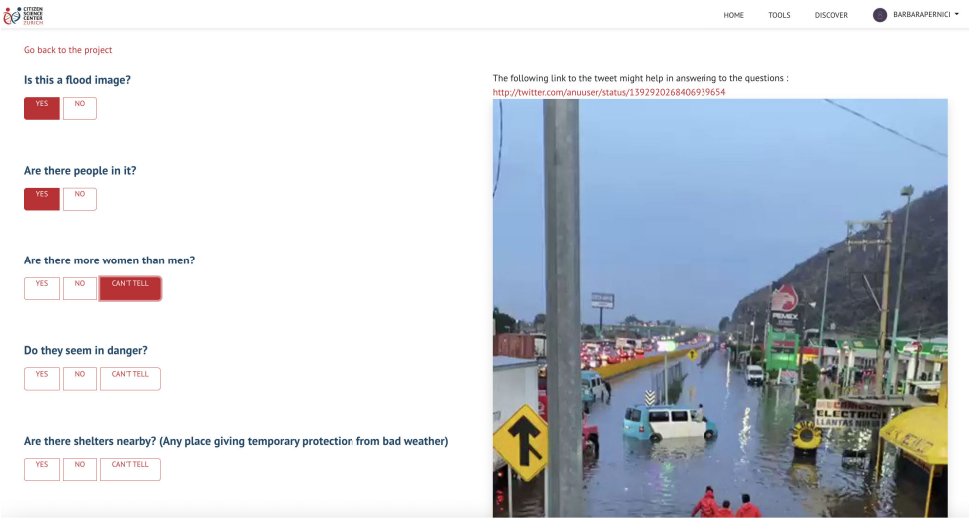


FIGURE 8. Example of CS Project Builder crowdsourcing interface.

"survey plus geolocation" if the task also includes geolocation on a map (e.g., identifying a landmark in an image and locating it in a city).

- *Step 3: Design your task protocol* - Design the survey protocol, including many options of questions and answers, or text fields, just as easy as a typical survey-building interface.
- *Step 4: Select the location of source files* - Import the source files (images, videos, sounds, etc.). The possible data sources include: Dropbox, Flickr, Amazon S3 buckets, and CSV files. There are particular cases for Twitter, including files built with VisualCit, or for data collected with the CS Logger (another tool being developed as part of the CSSK).
- *Step 5: Test and Publish* - From the project dashboard page, the draft can be modified and shared with collaborators for feedback and iterations. Several parameters of the project can be monitored and set from this dashboard, including statistics and tasks and data management. Once ready to go public, the project can be submitted for publication. The Administrators team will ensure that the project adheres to legal/ethical criteria before making it available for public contributions. Once the project has been published, anyone can participate in it.

The entire process requires limited technical knowledge and little or no coding skills. However, the project dashboard also provides access to a coding interface where the project protocol code is accessible and modifiable. Users with Vue.js coding skills can act directly on the code to make any desired modifications to the interface.

Fig. 8 shows an example of the typical interface generated by the CSPB: the right side of the screen displays the digital data, in this case a tweet including both text and an image. The left part features the survey protocol, in this case a series of conditional questions about the content of the tweet and related images.

The CSPB implementation is based on the open-source crowdsourcing framework PyBossa and its code is publicly available under the *CitizenScienceCenter* organization on

GitHub. PyBossa[14] is an ongoing open-source development project. It is an extremely flexible and versatile technology used for the development of platforms and for data collection within collaborative environments, analysis and data enrichment. PyBossa is implemented with a RESTful API[15] to easily distribute tasks and collect, maintain, and process data from volunteer participants. In addition, the PyBossa RESTful API allows creating/removing projects, as well as the implementation of graphical interfaces.

The PyBossa software architecture is composed of three major components:

- The *Task Importer* allows the creation of a new PyBossa project and tasks. This can be called from both the web interface and command line using the PyBossa API.
- The *Task Presenter* eases the creation of the front end of the PyBossa projects. The task presenter accepts Vue.js code, which provides total flexibility to develop complex data analysis interfaces for audio, video, photos, or PDF documents.
- The *PyBossa Core* builds on a PostgreSQL database to store all information related to users and projects. The PyBossa Core handles the scheduler to allocate tasks to users, redundancy, and export options. It also handles the execution of webhooks to manage the execution of cron processes.

The CSPB as well as PyBossa allows three different user roles:

- The *Project Organizer* is assigned as soon as the user creates a new project, and is related and limited to the projects that the user owns. The project Organizer has the right to edit the content of the project, update task visualization, add/delete tasks, export data and results.
- The *Contributor* is any user who participates in a CSPB project by analyzing data and submitting their results. Contributions can be assigned (after login) or anonymous.
- The *Admin* role is reserved for the administration of the platform, and it gives similar privileges to project organizers over all projects.

The CSPB builds on the PyBossa RESTful API, providing a graphical interface that exploits the potential of PyBossa without the need for interacting using bash or any programming language.

E. SOCIAL MEDIA CONSENSUS ANALYSIS WITH CROWDNALYSIS

When the citizen scientists contributing to a Citizen Science project in CSPB complete the classification tasks (e.g., labelling the severity of damage in photos) on the platform, the individual annotations of the citizen scientists need to be aggregated to achieve a consensus on each task. Majority Vote is the most common aggregation method where the consensus on a task is the most voted class, and each annotator's vote has an equal contribution. This method counts on the wisdom of the crowd, but it cannot tackle situations such as when i) we are unsure of how many annotators are required to obtain reliable results and ii) some annotators perform better on some tasks than others.

[14]https://pybossa.com/
[15]https://docs.pybossa.com/api/intro/

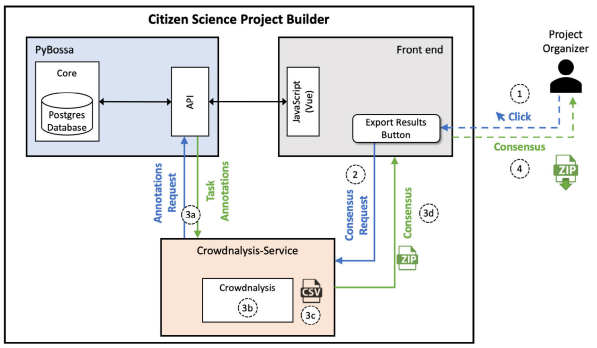


FIGURE 9. The workflow of the automatic computation of consensus in CSPB via Crowdnalysis. Upon a single click, the user downloads the consensus on project tasks without leaving the front end.

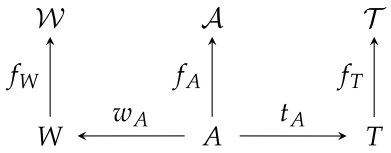


FIGURE 10. Spaces and mappings in Crowdnalysis' mathematical model.

We developed the Crowdnalysis [34] software library to address the critical needs that arise in the planning of a crowdsourcing project and the analysis of crowd-sourced data in CSPB. Crowdnalysis incorporates advanced probabilistic consensus models that estimate individual annotator error rates for a given set of tasks even when the ground truth is unavailable. These models enable Crowdnalysis to weigh the contribution of each annotator and thus, yield a more reliable consensus.

Crowdnalysis is integrated into CSPB as an on-demand service. Fig. 9 illustrates the use case: 1) The project organizer clicks the "Export Results" (of annotations) button on the front end; 2) The request is forwarded to the Crowdnalysis-Service; 3) The service: a) Calls the PyBossa API to extract task and annotation data; b) Computes the consensus on tasks for each question asked to the crowd by using Crowdnalysis with the given consensus model (e.g., Dawid-Skene [35]); c) Creates a Comma-Separated Values (CSV) file for each consensus; d) Sends the consensus files to the front end in a ZIP archive file; 4) The project organizer downloads the ZIP file without leaving the front end in any of the above steps.

We introduced the conceptual and mathematical framework of Crowdnalysis in [36]. The main concepts (inspired by [37]) in our mathematical model are as follows:

Worker is any of the participants in the annotation process.
Task is the minimal piece of work that can be assigned to a worker.

Annotation is the result of the processing of a task by a worker.

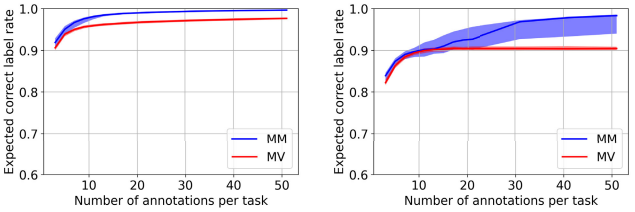
Spaces and inter-mappings regarding these basic concepts are shown in Fig. 10. Workers, tasks and annotations are given by the finite sets of W, T and A, respectively. Assuming that each worker, task and annotation can have distinctive features, we define the *feature spaces* $\mathcal{W}$, $\mathcal{T}$ and $\mathcal{A}$, and *feature mappings* $f_W : W \to \mathcal{W}$, $f_A : A \to \mathcal{A}$ and $f_T : T \to \mathcal{T}$ that altogether describe the individual characteristics of each member of the corresponding concepts. Finally, the functions $w_A : A \to W$ and $t_A : A \to T$ map each annotation to its worker and task, respectively.

From an epistemological point of view, we refer to crowdsourcing when we need to determine a specific characteristic(s)—which is(are) unknown to us—of each item in a set of tasks with the aid of workers annotating them for us. Accordingly, we assume that we can factor the task feature space as $\mathcal{T} = \mathcal{T}_O \times \mathcal{T}_C \times \mathcal{T}_H$, where $\mathcal{T}_O$ contains the observable task characteristics, $\mathcal{T}_C$ contains the unobservable (i.e., latent) characteristics in which we are interested in for the consensus and $\mathcal{T}_H$ contains the characteristics that are unobservable and we are not interested in. We model the lack of knowledge by means of a probability distribution, and subsequently aim to determine a probability distribution over $\mathcal{T}_C \times .^{\mathtt{t}}. \times \mathcal{T}_C$ where $\mathtt{t}$ denotes the number of tasks. We call this distribution a *joint consensus*.

With the above conceptual framework, in [36], we defined the consensus problem, introduced the abstract consensus model as a probabilistic graphical model, and demonstrated how we query this probabilistic distribution to obtain the marginal distribution of the joint consensus for each task. Moreover, we showed how discrete annotation models can be accommodated within our framework, giving pooled Multinomial [38] and Dawid-Skene [35] models as examples.

Once it fits a probabilistic model to a given annotation data, Crowdnalysis can also leverage this model to conduct a *prospective data quality analysis* for the crowd of interest. Specifically, Crowdnalysis uses the estimated parameters for tasks and annotators (i.e., marginal probabilities of classes and annotator error rates, respectively) to generate *synthetic* annotation data of different sizes, mimicking crowd behavior. Then, performance analysis on this synthetic data provides the expected values of a given metric – say, accuracy – over the "number of annotations per task". This analysis allows the determination of the redundancy in annotations needed for the crowd to reach the desired accuracy. Consequently, given the performances of different communities on the same tasks, prospective analysis also helps choose the community to work with in a future project with similar task characteristics. This feature is particularly beneficial when some communities are not free of charge (paid workers) or hard to assemble (experts).

In [36], we present a case study of the 2019 Albania earthquake[16] analyzing the annotation data from three different communities, namely, volunteers, paid workers and experts, and comparing their performances by Crowdnalysis. All three communities were provided with the same set of social media images and asked to annotate the severity of damages in the photos with one of the labels in $\mathcal{A} = \{$irrelevant, no-damage, minimal, moderate, severe$\}$. We provide a concise description of the communities and crowdsourcing settings in Section V-B. In Fig. 11, we present part of the prospective analysis of this case study. We performed this analysis in three steps: 1) In the first step, we fit the Multinomial Model (MM) with annotation data from experts and calculated the consensus. Then, we fit another MM to the volunteer data, using the experts' consensus as the ground truth. 2) In the second step, we used the prior probability distribution for the above labels, estimated by the experts' MM, to generate a dataset of 10,000 synthetic tasks. Then, we generated synthetic annotation datasets for both experts and volunteers with different redundancies per task using their estimated error rates computed in the first step. 3) In the last step, we computed the consensuses of synthetic expert and volunteer annotations with MM and Majority Vote (MV). Finally, we compared their expected accuracies over the number of annotations per task, using the labels of the synthetic tasks generated in the second step as the ground truth.



(a) Experts (b) Volunteers

FIGURE 11. Prospective analysis for correct label rates for experts and volunteers with the Multinomial Model (MM) and Majority Vote (MV) in the 2019 Albania earthquake case study.

In Figures 11a and 11b, we observe that our proposed probabilistic model outperforms the MV. The performance is comparable only for the lowest redundancy, that is, three annotations per task. The figures also show that when we use the probabilistic model for consensus, the accuracy increases with redundancy in the annotations. This observation is also valid for MV up to a certain redundancy value, after which we achieve no, or very minuscule, increase in accuracy.

Owing to Crowdnalysis' probabilistic modeling for consensus and prospective analysis features, CS project organizers can make more informed decisions on the crowd to work with and the size of that crowd. For example, to obtain a 0.95 consensus accuracy in a scenario similar to an earthquake, Fig. 11 indicates that they would need 6 expert annotations per task, whereas it would take more than 20 volunteer annotations for the same accuracy (assuming these two types of annotators exhibit the same or fairly similar annotation skills as the members of the corresponding community that was previously worked with).

[16] https://en.wikipedia.org/wiki/2019_Albania_earthquake

Crowdnalysis is implemented with the Python programming language and distributed via the Python Package Index[17] (PyPI) software repository. Therefore, it can be easily imported and used in any Python script. Its source code is publicly available at GitHub[18] and bears an exhaustive test suite. Crowdnalysis also comprises consensus model implementations in the Stan probabilistic programming language [39] thanks to the CmdStanPy[19] command-line interface for Python. Furthermore, Crowdnalysis has the following additional features:

- Setting inter-dependencies between questions to filter out irrelevant annotations (e.g., ignoring the following answer for a specific task if the previous answer is ''Not relevant'').
- Distinguishing real classes for answers from reported labels (e.g., ''Not answered'').
- Calculating inter-rater reliability using different measures (e.g., Fleiss' kappa).
- Visualization of annotator error rates and consensus.

V. VALIDATION

A. VALIDATION METHODS

The quality of the results obtained by crowd-based data analysis depends on different factors: the quality of the dataset obtained from the previous preparation and analysis phases, how the task has been managed and described, and the skills and expertise of the citizen scientists. In our case studies, we based our validation on the quality of the resulting data. Adopting this end-to-end approach, we can evaluate the quality of the result by considering the effect of other components: the filtering components in VisualCit, which can also present quality issues, as they are based on models trained with machine learning methods, and for which configuration parameters can have an impact on the final results, as well as; the quality of the crowdsourcing tasks assigned to the crowd and the quality of the results of crowdsourcing activities, as described in Section IV-E.

In order to assess this quality several methods can be applied:

- *Individual*: The role of humans in crowdsourcing can be extended to the quality assessment phase. The accuracy of a given output is evaluated by individuals, e.g., by citizen scientists or external experts.
- *Group*: The assessment is performed by a group of people (typically citizen scientists), e.g., through voting.
- *Computation-based*: This includes assessment methods that can be performed by a machine without the involvement of humans.
- *Validation datasets*: Validation can be performed by comparing the obtained results with external sources focusing on the same problem, e.g., surveys in the field.

The computation-based methods can support the evaluation of the quality of results from different perspectives and consider different quality dimensions:

- *Accuracy*: This can be computed considering a given ground truth and suitable comparison operators to evaluate the similarity between the obtained results and the desired values. The assessment of accuracy can be precise depending on the reference source considered.
- *Timeliness*: Some data analyses have stringent time constraints. Timeliness measures the temporal validity of the input data; only not outdated values should be considered. This means that the results should be produced within a given time after the event occurs.
- *Provenance*: The reliability of the input data increases if information about acquisition methods, processing steps, and the way in which crowdsourcing has been performed (e.g., crowd characteristics, redundancy, and aggregation methods) is available. It is important in the evaluation of the results to obtain information on the data provenance: the more input data are reliable, the more trustworthy the analysis results are.

A quality control methodology for datasets derived from social media has been studied in Crowd4SDG to assess whether extracted data can be used as non-traditional data sources by national statistical offices (NSOs) [40]. A scoring approach was proposed to evaluate the quality of the dataset. In addition to the validation criteria mentioned above, a few new criteria were added for validation focusing on the data production process, confidentiality, and impartiality.

The results of the validation for the three case studies are discussed in the next section and listed in Table 2. In the table, the main validation criteria are discussed for the results of each of the CSSK components used in the case studies. The results derived from the quality control methodology are reported in the Data quality assessment column. Finally, a general end-to-end evaluation is given under the General evaluation column and the main issues mentioned under Problems.

B. CASE STUDIES

In the following, some different case studies, covering different typologies of usage scenarios of the CSSK, are presented to illustrate the different contexts of data analysis and their challenges and results.

- *2019 Albania earthquake*: In this case study, we aimed to provide a working example of how we fit the conceptual and mathematical framework behind Crowdnalysis to real-world data, and demonstrate how we could benefit from our software library in an emergency scenario, where we refer to crowdsourcing. The framework and case study results are published in [36].
  The imagery dataset we worked with was the courtesy of the Qatar Computing Research Institute, which contains social media images posted for the 2019 Albanian earthquake filtered by their Artificial Intelligence for

[17] https://pypi.org/project/crowdnalysis
[18] https://github.com/Crowd4SDG/crowdnalysis
[19] https://mc-stan.org/cmdstanpy

TABLE 2. Case studies.

Case study	Goal general/specific	Source	VisualCit task	VisualCit configuration and validation dimension	Crowdsourcing tasks	Crowdsourcing configuration criteria	Crowd analysis	Crowd validation	Data quality assessment of resulting dataset	General evaluation	Problems
Albania earthquake [36]	general goal: collect images with damage assessment of buildings use cases: A, B specific goal: relevant images	cleaned images dataset from Twitter 907 images	eliminate non relevant images	64% non relevant discarded, 83% precision, 90% recall and 82% overall accuracy	grading tasks tool: Amazon MTurk redundancy: 10	time to complete a task difficulty of task	applied (all methods)	TPR: 98%, TNR: 52%	metadata are needed	annotator error rates remained fairly the same; suggesting we can trade crowd time for higher recall	a qualified crowd would be preferrable
COVID-10 social distancing [25]	general goal: derive indicators about social behaviors in COVID-19 by country use case: C specific goal: annotated geolocalized images	Twitter posts with images 500K/week (three weeks)	eliminate non relevant images geolocate	0.5% classified as relevant and geolocated accuracy of geolocation: 84% accuracy of filtering: 92-99% (depending on applied filter)	question answering tool: Project Builder redundancy: 3	high volume: sampling needed non expert crowd sufficient difficult of finding motivated crowd	Majority Vote; Dawid-Skene	crowd accuracy for social distancing: 71% (w.r.t. crowd consensus by Dawid-Skene method without ground truth)	comparison with independent survey proved satisfactory dataset evaluation positive, more multilingual approach requested	sampling reduced significantly the amount of analyzed data	volume crowd recruitment
Thailand floods [31]	general goal: timely alerts (<24h) of floods monitoring social media use case: E specific goal: spatial representation of the event	Twitter posts, 4 million in two days	extract posts with images, filter out non relevant images, geolocate	completeness of the results improved with text analysis (45% increase)	not applied due to required time constraints	-	-	-	comparison with independent survey proved satisfactory dataset evaluation positive, general multilingual approach required	coverage of regions not uniform compared with Tweet native gelocation only, accuracy improved and less bias on capital city	volume, timeliness, relevance

Disaster Response (AIDR) platform [41]. The AIDR collected Twitter posts for four days after November 26th, 2019, when the earthquake had occurred, which was the strongest for the country in the last 40 years. The AIDR automatically classified 907 images as relevant out of 9, 241 collected.

As mentioned in Section IV-E, crowdsourcing was performed using three different groups of citizens to annotate the relevant images and the severity of damage seen in the photos by humans compared to the AIDR. Specifically, to obtain a ground truth for the labels, we first worked with ten disaster experts and configured a redundancy of 3 for annotations on the Crowd4EMS platform [6], a precursor of CSPB. We then referred to a group of 50 volunteers on Crowd4EMS with a redundancy of 3. Finally, 171 paid workers annotated the same dataset on the Amazon Mechanical Turk (MTurk) platform[20] with a redundancy of 10. Using Crowdnalysis, we first computed the experts' consensus as the ground truth and, successively, calculated the error rates—in labeling—of the three crowds. We also carried out a prospective analysis of the communities detailed in Section IV-E of this article. This case study is related to the use case A in Fig. 3.

Subsequently, we extended our analysis with this dataset incorporating VisualCit for further filtering the AIDR dataset for relevant images; hence, we adopted the use case B in the same figure. We provide the confusion matrix for VisualCit in Table 3, which shows a 0.83 precision, a 0.90 recall and a 0.82 overall accuracy of VisualCit, with respect to the experts' consensus on relevance.

[20]https://www.mturk.com

TABLE 3. Number of relevant and irrelevant images labeled by VisualCit w.r.t. the experts' consensus.

		VisualCit	
	Relevant	Yes	No
Experts	Yes	546	59
	No	108	194

Moreover, we conducted a hypothetical experiment to determine whether the results would be accurate if the experts and the MTurk crowd annotated the subset of the AIDR dataset filtered by VisualCit. Specifically, after removing the annotation data for the 253 tasks excluded by VisualCit, we calculated the new experts' consensus and error rates of the experts and MTurk crowd by Crowdnalysis. We observed that the error rates for both communities remained almost the same (e.g., +0.01 in experts' recall, −0.04 in MTurk crowd's specificity).

Our extended analysis suggests that when we can trade crowdsourcing efforts for higher recall, we can safely bring VisualCit to automatically filter relevant images, thus reducing the number of irrelevant tasks to be processed by citizen scientists.

- *COVID-19 social distancing case study*: This case study, developed in 2020 during the initial phases of the pandemic, aimed to derive indicators of social distancing behavior and face mask usage in different countries [25]. This case study is related to use case C shown in Fig. 3, with a complete data preparation pipeline. Selected raw data were captured from social media from May to August 2020, crawling Twitter with generic COVID-related keywords, and then filtering posts with VisualCit, selecting only images that are classified as photos (excluding memes, maps, drawings, etc.), in outdoor spaces, and containing at least two persons. When a native geolocation was not present in the tweet, the

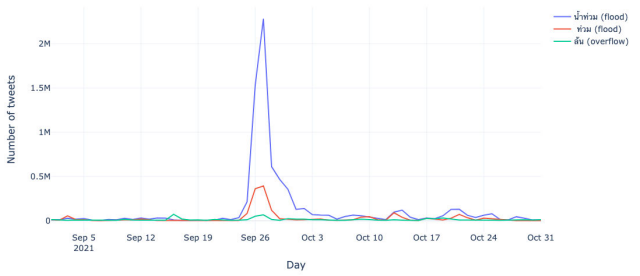


FIGURE 12. Daily tweet counts for Thai dictionary entries.

posts were geolocated using the CIME algorithm [17]. Crowdsourcing performed with the CS Project Builder was used to confirm the relevance according to the above-mentioned criteria and the automatic geolocation of the images, and to assess whether face masks were correctly used and social distancing rules followed. The majority vote was used to compute the consensus on the crowd annotations. The resulting images were aggregated to build behavioral indicators on a national basis.

- *Thailand floods alerting case study*: In this case study, a large-scale flood event in Thailand was evaluated. Tropical storm Dianmu[21] hit Thailand in late September and October 2021. According to the UNOSAT Thailand flood monitoring dashboard,[22] as a preliminary assessment around 1.4 million people were affected by flooding. The goal of this case study was to assess the ability to capture the onset of large scale emergencies, while enabling a spatial description of the event. This case study refers to the use of Case E in Fig. 3. An overall description of this case study can be found in [31]. Tweets corresponding to event onset were crawled using a small set of flood-related Thai keywords, provided by the UNOSAT office in Bangkok. The time series corresponding to keywords' usage during event onset is shown in Fig. 12. After detecting the onset trend, a VisualCit pipeline was executed to remove near-duplicates, non-photos, and NSFW images, to exclude irrelevant content. The filtered posts were then geolocated with CIME, focusing on locations within Thailand's national borders. The results were then aggregated by administrative region and normalized using the population count.

The aggregate descriptions of the results and insights from the case studies are represented in Table 2.

We present here the main results and outcomes of the three case studies presented above and discuss their validation procedure. The case studies were aimed at a range of goals. In the Albania case study, the focus was on analyzing images from posts assessing damage to buildings, while in the COVID and Thailand case studies, the goal was to generate thematic maps providing evidence about ongoing situations.

The case studies were characterized using different types of source data. In most cases, social media data are obtained through the crawling of Twitter, while in one case (the Albania case study), the initial dataset is available as the result of previous analyses. Consequently, the volume of posts to be analyzed was significantly different. In addition to the volume aspect, the three case studies were characterized by different velocity characteristics, and timeliness requirements. In particular, the Thailand flood case study has a large number of posts to be analyzed rapidly to provide a rapid assessment of the situation, while in the COVID case study, the volume is high, while timeliness is less stringent, as the analysis is performed on a weekly basis. The VisualCit results are validated in different ways, according to the tasks to be performed. In the Albania case study, VisualCit was used to further pinpoint relevant images from the initial dataset, and a significant improvement in accuracy was obtained. In the COVID case study, filtering resulted in only 0.5% of the images being considered relevant for the task to be performed. A detailed evaluation of precision and recall for different filters was published in [25], and the accuracy of geolocation measured on a country basis was 84%. In the Thailand case study, low recall was compensated by analyzing the text of the tweets, thus improving completeness in the data for the considered regions. Crowdsourcing was only applied to the first two case studies. In the Albania case study, crowdsourcing was performed using three different communities: a group of experts, a group of volunteers, and a group of paid workers on MTurk, as detailed above. In the COVID case study, characterized by a high volume even notwithstanding the high VisualCit selectivity, the CS Project Builder was used with redundancy set to three and majority vote. The main problem encountered in this case was related to the size of the data and the availability of a crowd of sufficient size for performing the analysis, and the costs of the analysis when a paid crowd is employed. In the Albania case study, the performance of the crowd was analyzed using Crowdnalysis, based on comparisons with expert evaluations, as illustrated in Section IV-E. However, the size of the COVID case study did not allow extensive data validation of the results, so an indirect validation procedure was adopted, comparing the results with similar analyses conducted at the same time with surveys, highlighting a significant correlation of the results [25]. In addition, in the case of the Thailand floods case study, validation was performed by comparing the obtained data with data collected in the field, showing good qualitative results of the method [31].

The selected case studies represent various goals and operating conditions. In general, the volume of data is high, and timeliness may be a constraint, whereas the number of relevant tweets for a given goal is usually small. Fast automatic filtering tools such as VisualCit can be successfully employed to speed up the analysis and reduce the number of irrelevant elements to be examined. When timeliness poses constraints

[21] https://floodlist.com/asia/thailand-tropical-storm-dianmu-floods-september-2021

[22] https://unosat-geodrr.cern.ch/portal/apps/opsdashboard/index.html#/4f878691713a40f3b8ef3140e63c9f6d

of less than 24h to analyze a situation, acceptable results can be obtained with completely automated tools, while better accuracy can be obtained by adding crowdsourcing if time is available for the analysis. In general, while citizen scientists for crowdsourcing activities are difficult or costly to recruit, better results can be obtained with a selected crowd or experts, thus showing the importance of crowd analysis.

VI. CONCLUDING REMARKS

In this work, we presented an innovative approach to support CS project managers and analysts in the analysis of social media data using a combination of automatic classification, filtering, and geolocation tools and a human-in-the-loop Citizen Science approach. The proposed configurable ''data analysis pipeline'' retrieves information from large sets of posts—many of which may be irrelevant—using automated AI-based filtering tools, feeding crowdsourcing projects to gather qualified inputs from citizens. We also illustrate the structure and use of dedicated tools to support the CS project implementation in each step of the process. These tools were developed within the Crowd4SDG European project and are included in the ''Citizen Science Solution Kit'' (CSSK). A systematic validation of the approach with three case studies is presented, and the validation criteria are discussed. Further details on the tools can be found on the project web site.[23]

The research community is currently very active in developing tools for automatically analyzing social media. However, collecting such information and images may introduce several types of bias that should still be investigated. These include biases in the analysis due to variations in the profiles and densities of contributors in different areas, that is, there is a risk of under-representing groups and focusing only on a subset of events related to the profiles of the contributors. In addition, analyzing social media in isolation presents all the limitations of a single source, and efforts should be made to integrate other sources of information. Systematic sampling, data augmentation, and bias reduction techniques should be investigated. Near real-time integration with other sources of information is also to be exploited, such as different social media, Voluntary Geographical Information such as OpenStreetMap [42], Earth observations such as Google Earth or sensor networks, official reporting agencies, and statistical information. The challenges of this integration on a global scale are high, because of the variety of available data; however, tools such as OpenStreetMap, which provide a rich source of crowdsourced information beyond geographical locations, if integrated into social media analysis, would greatly increase the understanding of events. More research is needed to integrate the solution kit with techniques that provide a fast assessment of the quality and reliability of the obtained results when high volumes of data are available. This is especially true in multilanguage setups, such as the ones enabled in [43], where the volume and variability of data are further amplified. Future work also includes incorporating hierarchical consensus models into Crowdnalysis and leveraging annotator models to efficiently govern task-annotator assignments within CSPB to make the most out of citizen contributions. Given that local knowledge can be favorable in crowdsourcing and emergency responses, we also plan to leverage the geolocation information to assign tasks to best-performing annotators in related regions, if available.

The CSSK responds to the needs of researchers approaching crowdsourcing methodology for the first time, providing them with easy-to-use and flexible tools for testing and implementing their projects. This is particularly important when aiming to support National Statistical Offices in the collection and analysis of data for official SDG reporting, as highlighted in the Crowd4SDG policy brief on using Citizen Science data to track SDG progress [44].

Several academic institutions and Non-Governmental Organizations have used the Citizen Science Project Builder tool in CSSK to setup projects in various scientific disciplines and social fields. Owing to the development and enhancements of the platform carried out in the context of the Crowd4SDG partnership, CSSK has been used specifically to test different approaches to crowdsourced disaster response, as described in the case studies and examined in the Crowd4SDG challenges. Multiple organizations have already profited from the more performing interface and the smoother user experience, including researchers at the Joint Research Centre of the European Commission. However, extensive outreach activities are needed to make the wider emergency community aware of the availability of the CSSK, and knowledgeable in the use of its different components.

Social media activities as a form of *passive* CS contribution—where useful information can be discovered through crawling and analysis—represent huge and untapped data. On the other hand, passive contributions complement – equally, if not more powerful – *active* ways for citizens to contribute to data collection and analysis. When active participation of citizens is required, for example by some of the CS tools, crowdsourcing poses several known challenges related to engaging and retaining participants. These include making sure that citizens and their communities benefit from participation, planning adequate rewards to recruit and keep the crowd active and interested, and the necessity to tailor the tasks to a variety of possible skills.

ACKNOWLEDGMENT

This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

The authors thank François Grey for his work coordinating the Crowd4SDG project, Miguel Armendáriz Jáuregui for his work on the VisualCit interface, and all the students who experimented with CSSK during its development.

They also thank all the Crowd4SDG partners for the discussions and common work during the project.

[23] https://crowd4sdg.eu/about-2/tools/

REFERENCES

[1] S. Fritz, L. See, T. Carlson, M. Haklay, J. L. Oliver, D. Fraisl, R. Mondardini, M. Brocklehurst, L. A. Shanley, S. Schade, and U. Wehn, "Citizen science and the united nations sustainable development goals," *Nature Sustainability*, vol. 2, no. 10, pp. 922–930, 2019.

[2] D. M. Wald, J. Longo, and A. R. Dobell, "Design principles for engaging and retaining virtual citizen scientists," *Conservation Biol.*, vol. 30, no. 3, pp. 562–570, Jun. 2016.

[3] L. A. Shanley, A. Parker, S. Schade, and A. Bonn, "Policy perspectives on citizen science and crowdsourcing," *Citizen Sci., Theory Pract.*, vol. 4, no. 1, p. 30, Dec. 2019.

[4] C. Franzoni, M. Poetz, and H. Sauermann, "Crowds, citizens, and science: A multi-dimensional framework and agenda for future research," *Ind. Innov.*, vol. 29, no. 2, pp. 251–284, Feb. 2022.

[5] C. Havas, B. Resch, C. Francalanci, B. Pernici, G. Scalia, J. L. Fernandez-Marquez, T. Van Achte, G. Zeug, M. R. Mondardini, D. Grandoni, and B. Kirsch, "E2mC: Improving emergency management service practice through social media and crowdsourcing analysis in near real time," *Sensors*, vol. 17, no. 12, p. 2766, 2017.

[6] A. R. Shankar, J. L. Fernandez-Marquez, B. Pernici, G. Scalia, M. R. Mondardini, and G. Serugendo, "Crowd4Ems: A crowdsourcing platform for gathering and geolocating social media content in disaster response," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 331–340, Aug. 2019.

[7] F. Alam, F. Ofli, and M. Imran, "Processing social media images by combining human and machine computing during crises," *Int. J. Hum.-Comput. Interact.*, vol. 34, no. 4, pp. 311–327, Apr. 2018.

[8] M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," *Inf. Process. Manag.*, vol. 57, no. 5, Sep. 2020, Art. no. 102261.

[9] C. V. L. Pennington, R. Bossu, F. Ofli, M. Imran, U. Qazi, J. Roch, and V. J. Banks, "A near-real-time global landslide incident reporting tool demonstrator using social media and artificial intelligence," *Int. J. Disaster Risk Reduction*, vol. 77, Jul. 2022, Art. no. 103089.

[10] S. Anjum, A. Verma, B. Dang, and D. Gurari, "Exploring the use of deep learning with crowdsourcing to annotate images," *Hum. Comput.*, vol. 8, no. 2, pp. 76–106, Jul. 2021.

[11] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Comput. Surv.*, vol. 47, no. 4, p. 67, Jul. 2015.

[12] A. Adrot, S. Auclair, J. Coche, A. Fertier, C. Gracianne, and A. Montarnal, "Using social media data in emergency management: A proposal for a socio-technical framework and a systematic literature review," in *Proc. ISCRAM*, May 2022, p. 470.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931, Apr. 2012, doi: 10.1109/TKDE.2012.29.

[14] E. Proden. *Crowd4SDG Deliverable 5.1—Initial Report on Relevance and Quality-Related Considerations of Citizen-Science Generated Data*. Accessed: Sep. 13, 2022. [Online]. Available: https://crowd4sdg.eu/wp-content/uploads/2021/10/D5.1-Initial-report-on-relevance-and-quality.pdf

[15] H. Karray, A. D. Nicola, N. Matta, and H. Purohit, "Understanding reactions to misinformation—A COVID-19 perspective," in *Proc. 19th Int. Conf. Inf. Syst. Crisis Response Manag.*, 2022, pp. 1–15.

[16] M. Wiegmann, J. Kersten, H. Senaratne, M. Potthast, F. Klan, and B. Stein, "Opportunities and risks of disaster data from social media: A systematic review of incident information," *Natural Hazards Earth Syst. Sci.*, vol. 21, no. 5, pp. 1431–1444, May 2021.

[17] G. Scalia, C. Francalanci, and B. Pernici, "CIME: Context-aware geolocation of emergency-related posts," *GeoInformatica*, vol. 26, pp. 125–157, Jan. 2022.

[18] V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon, "Integrating social media into a pan-European flood awareness system: A multilingual approach," 2019, *arXiv:1904.10876*.

[19] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," *IEEE Intell. Syst.*, vol. 29, no. 2, pp. 9–17, Mar./Apr. 2014.

[20] O. Okolloh, "Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information," *Participatory Learn. Action*, vol. 59, no. 1, pp. 65–70, 2009.

[21] *Facebook Crisis Response*. Accessed: Mar. 12, 2022. [Online]. Available: https://www.facebook.com/about/crisisresponse/ Accessed:

[22] *Zooniverse*. Accessed: Mar. 12, 2022. [Online]. Available: https://www.zooniverse.org/

[23] *AWS MTurk*. Accessed: Sep. 12, 2022. [Online]. Available: https://www.mturk.com/

[24] A. Asif, S. Khatoon, M. M. Hasan, M. A. Alshamari, S. Abdou, K. M. Elsayed, and M. Rashwan, "Automatic analysis of social media images to identify disaster type and infer appropriate emergency response," *J. Big Data*, vol. 8, no. 1, pp. 1–28, Dec. 2021.

[25] V. Negri, D. Scuratti, S. Agresti, D. Rooein, G. Scalia, A. R. Shankar, J. L. F. Marquez, M. J. Carman, and B. Pernici, "Image-based social sensing: Combining AI and the crowd to mine policy-adherence indicators from Twitter," in *Proc. IEEE/ACM 43rd Int. Conf. Softw. Eng., Softw. Eng. Soc. (ICSE-SEIS)*, May 2021, pp. 92–101, doi: 10.1109/ICSE-SEIS52602.2021.00019.

[26] J. Fohringer, D. Dransch, H. Kreibich, and K. Schröter, "Social media as an information source for rapid flood inundation mapping," *Natural Hazards Earth Syst. Sci.*, vol. 15, no. 12, pp. 2725–2738, Dec. 2015.

[27] V. Scotti, M. Giannini, and F. Cioffi, "Enhanced flood mapping using synthetic aperture radar (SAR) images, hydraulic modelling, and social media: A case study of hurricane Harvey (Houston, TX)," *J. Flood Risk Manag.*, vol. 13, no. 4, p. e12647, Dec. 2020.

[28] M. Craglia and L. Shanley, "Data democracy—Increased supply of geospatial information and expanded participatory processes in the production of data," *Int. J. Digit. Earth*, vol. 8, no. 9, pp. 679–693, Sep. 2015.

[29] V. Grasso and A. Crisci, "Codified hashtags for weather warning on Twitter: An Italian case study," *PLoS Currents*, vol. 8, pp. 1–32, Jul. 2016, doi: 10.1371/currents.dis.967e71514ecb92402eca3bdc9b789529.

[30] D. T. Nguyen, F. Alam, F. Ofli, and M. Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises," in *Proc. 14th ISCRAM Conf.*, Albi, France, May 2017, pp. 1–13.

[31] C. Bono, B. Pernici, J. L. Fernandez-Marquez, A. R. Shankar, M. O. Mülâyim, and E. Nemni, "TriggerCit: Early flood alerting using Twitter and geolocation—A comparison with alternative sources," in *Proc. 19th Int. Conf. Inf. Syst. Crisis Response Manag.*, R. Grace and H. Baharmand, Eds. Tarbes, France, May 2022, pp. 674–686.

[32] B. Strasser, J. Baudry, D. Mahr, G. Sanchez, and E. Tancoigne, "'Citizen science'? Rethinking science and public participation," *Sci. Technol. Stud.*, vol. 32, no. 2, pp. 52–76, 2019.

[33] F. Grey. (2009). *The Age of Citizen Cyberscience*. Accessed: Jul. 2022. [Online]. Available: http://cerncourier.com/cws/article/cern/38718

[34] J. Cerquides and M. O. Mülâyim, "Crowdnalysis: A software library to help analyze crowdsourcing results," Zenodo, Jan. 2022, doi: 10.5281/zenodo.5898579.

[35] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Appl. Statist.*, vol. 28, no. 1, p. 20, 1979. [Online]. Available: https://www.jstor.org/stable/10.2307/2346806?origin=crossref

[36] J. Cerquides, M. O. Mülâyim, J. Hernández-González, A. R. Shankar, and J. L. Fernandez-Marquez, "A conceptual probabilistic framework for annotation aggregation of citizen science data," *Mathematics*, vol. 9, no. 8, p. 875, Apr. 2021. [Online]. Available: https://www.mdpi.com/2227-7390/9/8/875

[37] A. Dumitrache, O. Inel, L. Aroyo, B. Timmermans, and C. Welty, "CrowdTruth 2.0: Quality metrics for crowdsourcing with disagreement," 2018, *arXiv:1808.06080*.

[38] S. Paun, B. Carpenter, J. Chamberlain, D. Hovy, U. Kruschwitz, and M. Poesio, "Comparing Bayesian models of annotation," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 571–585, Dec. 2018. [Online]. Available: https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00040

[39] S. D. Team. (2022). *Stan Modeling Language Users Guide and Reference Manual*. [Online]. Available: https://mc-stan.org

[40] E. Proden. (2021). *Crowd4SDG Deliverable 5.2—Data Usability Assessment and Recommendations for SDGs GEAR Cycle 1*. [Online]. Available: https://crowd4sdg.eu/wp-content/uploads/2021/10/D5.2-Data-usability-assessment.pdf

[41] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 159–162.

[42] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

[43] C. A. Bono, M. O. Mülâyim, and B. Pernici, "Learning early detection of emergencies from word usage patterns on social media," in *Information Technology in Disaster Risk Reduction* (IFIP Advances in Information and Communication Technology). Springer, 2022, pp. 1–16.
[44] E. Proden, K. Bett, H. Chen, S. D. Valero, D. Fraisl, G. Gamez, S. MacFeely, R. Mondardini, L. See, and Y. Min. (2022). *Citizen Science Data to Track SDG Progress: Low-Hanging Fruit for Governments and National Statistical Offices*. [Online]. Available: https://tinyurl.com/unitar-policy-brief

CARLO BONO (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the Politecnico di Milano. He has a background in computer science and philosophy. His main research interests include data-driven applications, adaptive information systems, and automatic learning algorithms.

MEHMET OĞUZ MÜLÂYİM received the Ph.D. degree (cum laude) in artificial intelligence from the Universitat Autònoma de Barcelona, in 2020. He is currently a Postdoctoral Researcher at the Artificial Intelligence Research Institute (IIIA), Spanish National Research Council (CSIC). His research interests include machine learning applications to enhance citizen science and healthcare.

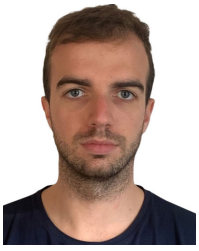
CINZIA CAPPIELLO received the Ph.D. degree in information technology from the Politecnico di Milano, Italy, in 2005. She is currently an Associate Professor of computer engineering at the Politecnico di Milano. Her research interests include data and information quality aspects in big data, service-based and web applications, and sensor data management.

MARK JAMES CARMAN received the Ph.D. degree from the University of Trento, in 2006. He has worked at the Università della Svizzera Italiana and Monash University. He is currently an Associate Professor of computer engineering at the Politecnico di Milano. His research interests include data science and deep learning techniques applied to information retrieval, natural language processing, and image analysis, as well as crowdsourcing and eXplainable AI (XAI).

JESUS CERQUIDES received the Ph.D. degree in artificial intelligence from BarcelonaTech, in 2003. He is currently a Scientific Researcher at the Artificial Intelligence Research Institute (IIIA), Spanish National Research Council (CSIC), where he is also heading the Learning Systems Department. He has coauthored more than 160 scientific articles. His research interests include probabilistic machine learning, explainability, causality, and their application on citizen science and health.

JOSE LUIS FERNANDEZ-MARQUEZ received the Ph.D. degree in artificial intelligence from the Universitat Autònoma de Barcelona in 2011. He is currently a Senior Lecturer at the University of Geneva and the Technical Coordinator of the Crowd4SDG EU Project. His current research interests include citizen science data quality analysis and methodologies to improve citizen science data quality, and make it suitable for decision policy makers.

MARIA ROSA (ROSY) MONDARDINI is currently the Managing Director of the Citizen Science Centre Zurich, a joint effort of the University of Zurich and ETH Zurich to promote and support the adoption of citizen science in the two institutions, Switzerland, and beyond.



EDOARDO RAMALLI (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the Politecnico di Milano. His main research interest includes knowledge extraction and management in data ecosystems, mainly for developing predictive models.

BARBARA PERNICI (Senior Member, IEEE) is currently a Full Professor of computer engineering at the Politecnico di Milano. She has published more than 70 articles in international journals and about 350 papers at international level. Her research interests include adaptive information systems, data quality, IS energy efficiency, and social media analysis. She is a member of the Editorial Board of the IEEE TRANSACTIONS ON SERVICES COMPUTING and *ACM Transactions on the Web*.

Open Access funding provided by 'Politecnico di Milano' within the CRUI CARE Agreement