Understanding Discussions of Citizen Science Around Sustainable Development Goals in Twitter

DAVID ROLDÁN-ÁLVAREZ, FERNANDO MARTÍNEZ-MARTÍNEZ, ESTEFANÍA MARTÍN, AND PABLO A. HAYA

1Escuela Técnica Superior de Ingeniería de Telecomunicación, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
2Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
3Instituto de Ingeniería del Conocimiento, 28049 Madrid, Spain

Corresponding author: David Roldán-Álvarez (david.roldan@urjc.es)

This work was supported in part by the European Community’s call H2020-SwafS-2019-1 (H2020-EU.5.c) under Agreement 872522, in part by the Madrid Regional Government through the Project eMadrid-CM under Grant S2018/TCS-4307, and in part by the Structural Funds [Fondo Social Europeo (FSE) and Fondo Europeo de Desarrollo Regional (FEDER)].

ABSTRACT Citizen science (CS) involves volunteers who participate in scientific research by collecting data or by addressing the needs of the project they are involved in. In the last years, there has been an increasing interest in how CS can contribute to the achievement of the UN Sustainable Development Goals (SDGs), that aim to reach a sustainable future. Research about data quality has taught us that through using an appropriate methodology, CS can foster scientific knowledge and promote specific actions to achieve accomplish goals. However, there is not much information about the SDGs that CS is more interested in. This paper presents a long-term study on how CS discuss about SDGs in Twitter, aiming to classify the discussion around the SDGs. The paper reports on a variety of topics such as open science, innovation and biodiversity, among others, but the results show that the most addressed topic in CS discussions in Twitter is about climate change, with corresponds to the SDG 13. Based on these findings, it is possible to affirm that climate change is a hot topic in CS in Twitter. However, there are also other SDGs that although underrepresented, are also discussed in CS.

INDEX TERMS Citizen science, social networks, Twitter, sustainable development goals, network analysis, SDGs.

I. INTRODUCTION

There are several definitions, but CS is often considered as the participation of non-scientist people in scientific processes [1]. CS gathers people with different profiles in order to collect, comment, transcribe and analyze data, methodology that some authors consider beneficial when it comes to transitioning from research and policy to sustainability [2]. Although public engagement in scientific research is not new (records dating back to 1980 [3]), there has been a significant increase in the number of academic researchers participating in citizen science(CS) initiatives in the recent decade [4].

A citizen scientist, without necessarily a scientific background, volunteers to collect or process data for scientific research [5]. Thanks to web-based technology, CS has included new ways of interaction apart from the traditional and physical ones and with them, new ways of learning from CS projects have emerged in the past few years [6]. Due to an increased accessibility and tolerance of Web technologies, CS has changed the professional-amateur relationships through permitting collective participation, less-mediated sharing of results and potential co-production of knowledge [7].

The use of existing technologies and the adoption of emerging technologies will enhance the ability of scientists and practitioners to centrally consolidate scientific information across projects, promote collaborative writing, and create virtual forums and communities [8]. Models have shown that collective actions can be sustained by social interactions and in this scenario, social networks emerge as useful tools to carry out these interactions [9]. Social media, defined as a set of online tools designed to enable and promote social interactions enable citizen scientist to have unique identifiable profiles, who consume, interact with and contribute to streams of user generated content and publicly articulate connections they create [10]. Although social media have

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
been used in a variety of fields for gathering real-time understanding of real-time phenomena, there is little understanding of how social media contributes to citizen science.

Monitoring and identifying current trends is crucial for government policy, research and development, strategic planning, social investment and enterprise practices [11]. In social networks, citizen scientists are no longer passive message recipients, but actively co-create communication channels [12]. The activities they conduct vary from expressing their opinions to broadcasting and planning activities and sharing the results of their research [13]. This is the main reason to pay attention to an analysis of the communicated content. Content analysis allows us to understand key aspects of a discussion, identify key topics and potentially engage interest groups [14].

Among all the existing networks, Twitter, with its micro-blogging system, is being used to share results and to create communication channels in CS [15]. Twitter is no stranger to new trends as they emerge over time and during the last years, there has been a common topic around the world that generates a great deal of conversation around it, the 2030 Agenda. In 2015, the United Nations’ 2030 Agenda for Sustainable Development set 17 interconnected Sustainable Development Goals (SDGs) and 169 associated targets that aimed to shape global development policies and actions [16]. Several features of the 2030 Agenda can build on CS, such as encouraging participation, partnerships and collaborations, education, sustainable living and global citizenship [17]. Studying the conversation around CS and the 2030 Agenda in social media can help to understand how goals evolve, the relationship between them and users and the trends at particular times [18]. Therefore, understanding how CS combines with SDGs in Twitter may help to develop strategies to further exploit communication tasks.

The work reported in this paper aims to gather understanding about the conversations around SDGs that happen in CS on Twitter, taking into account the networks that are formed as a result of the interaction on this platform. Studying the information available can help to understand the conversations around CS and SDGs, including how users interact with each other, who are the users that initiate discussions, finding trending topics in general and in specific time periods and how topics are related with each other. With this analysis, we aim to answer the following questions:

RQ1: Are CS and SDGs related in Twitter?
RQ2: Which are the most discussed SDGs on Twitter? Which are the most used words and hashtags?
RQ3: Are there SDGs related with each other?
RQ4: Who are the most important users talking about CS and SGS on Twitter?

II. CITIZEN SCIENCE AND 2030 AGENDA

CS related activities can address sustainability challenges and contribute to the implementation of SDGs. CS encourages social cohesion, a crucial element for the moral dimension of the 2030 Agenda, which aspires to benefit all people and leave no one behind through global citizenship and shared responsibility [16]. In order to work towards SDGs, a problem-solving attitude is needed, where all stakeholders are engaged [19]. As a consequence, CS actions performed by individuals, teams or networks of volunteers with a significant contribution to the societal changes cannot be neglected [17]. It is not only reasonable to mention that CS can advance scientific knowledge [20] but it can also foster societal transformation and the 2030 Agenda goals [21].

In the literature, we can find some research about the relationship between CS and the SDGs of the 2030 Agenda. For instance, in Shulla’s et al. work [17] it is stated that CS activities contributed mainly with SDG4 Quality Education, SDG11 Sustainable Cities and Communities, SDG 13 Climate Action and SDG15 Life on Land. However, these results need to be treated with caution, since they come from an international survey that received 84 answers, which seems a low number in order to get a good representation about how CS impacts SDGs.

We can find more research where the authors interviews CS project coordinators in order to gather more understanding about how CS can contribute to SDGs [23]. In this work, they considered the importance of finding ways to measure the impact of CS projects. In this scenario, it seems that specific frameworks that enable measuring the impact of CS in SDGs need to be developed, being built upon existing assessment techniques. Nowadays, it seems that there is a lack of standardised approach for assessing impact across different domains [24] such as measuring impact in social networks or studying the relevance of the gathered data, which derived on a call for the CS impact assessment field to reflect on its own approaches [25]. Although only 11 coordinators participated in this research, the literature show that measuring the impact was already a concern years prior to this study [26]–[28].

III. CITIZEN SCIENCE AND SOCIAL NETWORKS

The way of gathering data is something that has been relevant in the literature. For instance in Fritz et al. work [22] it is pointed out that traditional data source are not sufficient for measuring SDGs. Social media has revolutionized communication among peers and sharing data, enabling citizens to give their opinions and interact with other researchers easily [15]. In addition, communities may be set up in social networks in order to enhance data gathering and promote the aforementioned interaction among citizen scientists [29], with many authors pointing out that social networks promote greater civic participation [30], [31]. In this scenario, a social network as Twitter allows the users to interact with each other and to share information easily.

Twitter brings together millions of users around its minimalist concept of micro-blogging. Recently, the number of characters allowed in each message has increased to 280 which means that the amount of information present in a single message can double compared to the past. The API Twitter provides, make it an ideal medium for the study of online data. Across the years, researchers have used Twitter to
analyse networks of tweets, mentions and retweets out, along with user’s profiles and activity [32]. These analysis have led the research community towards solving questions about influence measuring, scientometry or geolocation [33]–[35].

CS projects tend to establish an online presence in platforms such as Twitter, aiming to develop and create a strong community and a dissemination medium [36]. Although it is important to understand the content shared in Twitter, it is also important to study the structure and nature of communities on social networks themselves in order to learn how communities and individuals behave. While there are many studies aiming to provide some understanding in a variety of contexts [37], such as biodiversity and methodologies, we could not find much research about how CS, 2030 Agenda and Twitter work together. Data gathered for CS research is usually downloaded from other sources [22] that do not allow the analysis of the aforementioned communities. Across the literature, we can find that in Mazumdar and Thakker work [15] they studied the behaviour of CS communities in Twitter trying to propose tools to increase the presence of CS in Twitter. They discovered that most of the discussions evolved around retweeting and posting additional comments, replying or adding supporting statements to the original tweets.

In our work we want to offer a panoramic view of the relationship between CS and Agenda 2030 on Twitter. Through different social network analytics and text analytics we will study which SDGs are the most addressed by the CS community, getting to know the most popular hashtags and the most important topics, and the networks that are formed as a result of the interactions on Twitter. Studying this information can help understand how SDGs are discussed in CS and what are the components (people, organizations, bots) that facilitate the dissemination of such topics. Gathering this information could be of interest to many stakeholders, such as project owners, practitioners and citizen scientists since it could provide unique insights into the structure of online communities.

IV. EXPLORATORY DATA ANALYSIS
The period of data collection was from the 30th of September 2020 until the 20th of June 2021. Our initial data set consisted of tweets with the words ‘citisci’ or ‘citizen science’ in them in different languages. The number of tweets collected in the period previously specified is 275,868 tweeted by 88,974 unique users. There are a total of 176,728 retweets and 136,898 unique retweets, meaning that there are 39,830 duplicated retweets (different users retweeting the same tweet). In addition, 4,441 tweets are replies to other tweets. It should be considered the fact that there may be people tweeting about this topic without using the words “citizen science” or “citisci”, which means that a very restrictive query was done from our side.

We reduced the number of tweets by applying a filter based on words related to the Sustainable Development Goals, such as: SDG, Clean water and sanitation, Affordable and clean energy… The resulting number of tweets was 19,543 tweeted by 10,186 unique users. Of those tweets, 5,960 are original (and unique) tweets and 13,583 are retweets. Of those retweets, 2,294 are retweets to the same tweet (Different users retweeting the same tweet). There are 760 tweets that are replies to other tweets.

The number of tweets obtained after filtering represent a 7.08% of the total number of tweets gathered in the aforementioned period, way below our expectations due to the importance given to this matter in a social network such as Twitter, and the level of social awareness related to the different SDGs. When analysing the tweets, we carried out preprocessing tasks in order to remove as much noise as possible. For instance, we removed punctuation marks, stopwords, urls and transformed the text to lower case. Further preprocessing tasks will be explained in the corresponding sections.

A. LANGUAGES OF THE DATASET
To identify the different languages present in our tweets, after the filtering process, we trained a NLP based model to recognize them. To do so, we used the Multinomial Naive Bayes classifier. To train this model we used a dataset containing 10337 different texts in 17 different languages including English, French, Spanish, and German, among others. Once the dataset was tested by the model all the languages present in the gathered tweets were unveiled. These languages were: English (78.2%), French (18.2%), Spanish (2%), German (1%), Portuguese (0.2%), Dutch (0.6%), Swedish (0.1%), Italian (0.4%) and Danish (0.5%).

In the original dataset, the languages found were: English (0.6%), Swedish (0.1%), Italian (0.4%) and Danish (0.5%). As we can see in these results, English is the predominant language, since it is the business, scientific and sometimes institutional language nowadays followed by French and then Spanish. The results for the rest of the languages were almost residual.

B. HASHTAG ANALYSIS
Hashtags are one of the main features in social media today. Usually, when a hashtag is highly used or retweeted it means that a lot of people is concerned about the subject the hashtag is about. In Figure. 1 the top 10 retweeted hashtags in these tweets are shown. When counting the number of hashtags retweeted, each hashtag is counted everytime it is retweeted by any user. For example, if one tweet with one hashtag is retweeted seven times, then the hashtag is counted seven times. Once we carried out the analysis, we saw that the #sdgs was the most mentioned but, since it is used in mostly all of the tweets related to SDGs, we took it out from the visualization so only descriptive hashtags remain.

The first hashtag is #sdgs, which stands for “Sustainable Development Goals”, with more than 8,000 retweets. The second is #climatechange, with around 1,200 retweets. The third one is #openscience, also near 1,200 retweets. The third hashtag is #fridaysforfuture with around 800 retweets,
showing the importance of citizen science in the goals to achieve with the agenda 2030. The rest of the hashtags are related to sustainability and climate change, another important part of the 2030 Agenda.

Analyzing the hashtags used outside the retweets, shown in Figure 2, a pretty similar situation can be observed. In this case retweets were deprecated (every tweet preceded by “RT:@” was removed) so the final sum shows us only the use of that hashtag. The hashtag #sdgs continues in first position. Followed by some of the same hashtags (#climatechange, #2degreesc, #openscience), although with lower numbers than before (users that are not retweeted). These results expose that SDGs are discussed in Twitter, being the topics related to climate and sustainability those with higher importance to the people.

C. TEMPORAL EVOLUTION OF THE USE OF HASHTAGS

To get a better understanding of the high number of retweets and use of the hashtag #sdgs, we performed a temporal analysis of the hashtags. In Figure 4 we can see the evolution of use of this hashtags from late September 2020 until June 2021. This linear graph shows higher values in those dates when the use of the hashtag increased. In this case we can see a high use of the hashtags #climatechange and #openscience in the retweets, with three big peaks in November 2020, December 2020 and February 2021. #climatechange and #savetheearth follow the previous ones in use. Around May 2021 the numbers of all the ten displayed hashtags reached a close equilibrium but quickly dropped again, leaving #openscience in first position once more.
becomes the most used again, followed by biodiversity, reaching this one the first position after March 2021.

Again, when combining both analysis the situation is similar. In Figure 6 the hashtag #cs_sdg2020 tops the use, but only until late October 2020 when it is overtaken by #climatechange, but only for a short period of time because #openscience quickly becomes the most used. After that, these two hashtags previously mentioned continue greatly close in use. Only #climatem strike and #climatechangeisreal get close to the last two from April 2021 to June 2021.

FIGURE 5. Temporal evolution of top 10 used hashtags outside retweets.

FIGURE 6. Temporal evolution of top 10 used hashtags.

D. MOST USED WORDS

Inside a tweet, normally three or four hashtags are used, but these are not the majority of words in the text, they are accompanied by more words. Despite the great importance of a hashtag, we decided to perform a count of the most used words in the tweets that we obtained once we carried the filtering process. With this purpose we selected all nouns, verbs and adjectives in the set of tweets. Numbers, hashtags and user names (ones preceded by # and the others by @) were eliminated. Then, we added the common stopwords (conjunctions, prepositions, etc.) in the languages of our dataset to remove any non-desired word that could have skipped the previous filter.

In the wordcloud shown in Figure 7 we can see those words that are most used, the bigger they are displayed the more times they appear in our set of tweets. Climate, citizen, science, help, join or change are words with a high usage, something that was expected, but having a closer look into the wordcloud we can find some other words, such as: environmental, projects, world, water, sdgs and planet. Again, the climate awareness is evident taking into account the words used in the tweets, but in any case we can see some words in relation to other topics, words like: inclusiveness, policy, health, research, social or work.

E. TF-IDF

Once we know the most used words in our set of tweets we decided to verify if these words were not only appearing a lot of times but also if they were relevant in the texts we had. The most common way is the calculation of TF-IDF, which stands for term frequency-inverse document frequency. This is a statistical computation of a high importance to determine the weight of the words appearing in a collection of documents in a corpus. In our database each document is a tweet, and the corpus is the collection of tweets. The estimation of the TF-IDF is done by calculating firstly the frequency of each word in the document and then multiplied by the inverse frequency, balanced by the number of documents. The result of this estimation leaves values of 1 (or near 1) to those terms of non-importance, and values close to 0 for those terms of higher importance or weight in the collection of documents.

In Figure 8 those terms of higher importance are displayed. In our case, we decided to select those terms with values of TF-IDF between 0.01 and 0.5 because of the elevated number of terms, statistical restrictiveness, and aesthetical purposes. As it can be observed, the bigger the word the more important it is, and the closer to zero its value. Words like social media, mediaquality, citizens appear in our word cloud in a big shape, understandable if we consider the first filter applied to obtain our database. Then we can see words like harmony, charity, world cities day, opportunity and some other related to climate, biodiversity, and development. This
obey to the second filter applied and shows us those words with higher importance in our set of tweets filtered by SDGs related words.

One of the main purposes of using the TF-IDF method is to discover words that could be used as keywords in a filtering process, also it could be used to find those highly repeated terms that may come from bots and should be removed. In our case, the words displayed in the word cloud should be taken into consideration when filtering in search of SDG conversation in Twitter.

V. UNDERSTANDING DISCUSSIONS

A. TOPIC MODELLING

Using keywords and hashtags to understand social media messages is helpful when having a dataset where a concrete topic is discussed. However, this kind of analysis is not really useful when the topic is generic, such as it happens with CS and SDGs. In order to further explore topic discussions, we applied topic modelling to our dataset. Topic modelling is an unsupervised learning technique that aims to identify patterns and relationships among text documents [39] and widely adopted within a variety of domains such as bioinformatics, politics, transportation, etc. [40]. For this research, we have used Latent Dirichlet Allocation (LDA) as topic modelling method for understanding topics of discussion. Techniques offer the possibility of scaling up analyses to hundreds of thousands of tweets [41].

The process involved initially taking the entire corpus and cleaning the text to remove unnecessary elements such as stopwords and standard punctuations, including the # to convert hashtags into normal words. Furthermore, the process required the removal of specific keywords that were too frequent within the corpus to be distributed in most topics such as ‘citizenscience’, ‘citizen’, ‘science’, ‘sdg’ and ‘sdgs’. Upon cleaning the tweets, they were split into words, to be created into a dictionary. The process then created a term frequency inverse document frequency (TF-IDF), as the one mentioned in the previous section, based on the corpus, which was further used to create an LDA model. First we measured the coherence [42] of the different LDA models created according to the number of topics selected. For this test, we created models using from 5 topics to 29 topics and evaluated their coherence. This coherence analysis measures the degree of semantic similarity between high scoring words in the topic. This will allow us to select the appropriate number of topics. If we choose too few topics, every word will be placed in the same topics (not being those words coherent among them) and if we choose too many topics, the same words will appear in different topics, therefore complicating the interpretation of the data.

Figure 9 shows that the coherence increases steadily until we reach 12 topics and afterwards, there is not a significant increase in the performance. According to these results, we could choose 12, 17, 19, 25, 26 or 27 topics but since we are working with SDGs we decided to choose 17 topics, matching with the 17 different SDGs that we are looking for, aiming to find a correlation between a topic and a SDG.

Figure 10 presents a visualization of the distances between the 17 topics identified across the two principle component axes. This is a visualization of the topics in a two-dimensional space based on the words they comprise. Some of the topics overlap, which means that they have words in common. At the end, there are many tweets that discuss about very similar topics with some differences. For instance, in this concrete example in topic 7 there was discussion about climate change and the potential use of using data and machine learning to help with it, while in topic 6 climate change is discussed from the perspective of education and schools. The barchart on the right indicates the most salient terms across the entire dataset, where saliency indicates the distribution of topics weighted by the terms of overall frequency. While some of the salient terms are fairly generic (change, sustainable, agenda, project, etc), analysing uncommon terms provides an indication of the different terms that emerge from the analysis.

Exploration of the topics shows the following distribution of the key terms, together with a summary of the themes emerging.

- **Topic 1 keywords**: join, weobservereu, work, global, globalgoals, data, register, education, monitoring, forward. Theme: This topic reflects the discussion around
the project WeObserveEu, which is an ecosystem of citizen observatories for environmental monitoring.

- **Topic 2 keywords**: sustainable, globalgoals, innovation, tech, development, riskcentre, saeedbaygi, mahsamoulawi, techhealth, fintech. Theme: Impact of technology in sustainable development, taking into account mostly health issues.

- **Topic 3 keywords**: cs_sdg, conference, contribute, session, data, community, climatechange, learn, online, october. Theme: Citizen science SDG conference discussion around how to contribute data to the community to learn about climate change.

- **Topic 4 keywords**: planet, agenda, live, left, years, people, days, world, support, year. Theme: Agenda to follow to have a healthy planet in the future.

- **Topic 5 keywords**: climatechange, help, data, study, level, great, security, nocnews, love, come. Theme: Discussion started by user13 (see table 2) about how to make online data more accessible for all.

- **Topic 6 keywords**: support, participation, partnerships, commitment, strong, everyones, inclusiveness, relentless, supporters, lead. Theme: Topic around how everyone’s participation is important in sustainable development, mostly centered around water usage and climate change.

- **Topic 7 keywords**: openscience, user2, survey, join, research, opendata, important, international, developing, eu_h. Theme: Discussion about the importance of carrying out open science so research and findings can be accessible by everyone, either professional researchers or amateur ones. It was started by user2 from table 2.

- **Topic 8 keywords**: climatechange, quality, community, world, share, week, love, join, justice, related. Theme: Topic about the challenges that the world needs to tackle to tackle the climate change problem. There are discussions about finances, energy and health.

- **Topic 9 keywords**: user4, community, cs_sdg, free, like, climatechange, sustainability, great, facebook, help. Theme: SDGs in general. Most of the information comes from user 4 (table 2) who retweets a lot of tweets about SDGs.

- **Topic 10 keywords**: sustainability, work, apply, goal, book, user3, looking, goals, connections, globalgoalsun. Theme: Discussion around user3 (see table 2) about the use of technology to reach SDGs.

- **Topic 11 keywords**: cs_sdg, policy, data, link, good, details, advance, makers, work. Theme: Citizen science SDG conference discussion about what policies should be implemented around data.

- **Topic 12 keywords**: climatechange, gogreen, climate-justice, degreesc, climatestrike, savetheearth, climate, c_degrees, youthforclimate, ocean. Theme: Another topic on climate change focusing in this case in the importance of youth actions to reach climate justice. Climate justice frames climate change as an ethical and political issue, rather than one that is purely environmental or physical in nature.

- **Topic 13 keywords**: data, environment, crowd, decade, waste, sobre, methods, climate action, para, geospatial. Theme: Topic that combines the impact of waste in the environment and the importance of having geospatial open data.
• **Topic 14 keywords**: opendata, wind, temperature, humidity, pressure, summary, badawczej, meteology, katowice, bdzin. Theme: Bot that tweets repeatedly about the weather conditions in Katowice.

• **Topic 15 keywords**: future, earth, biodiversity, conservation, bethechange, renewableenergy, ecology, actnow, degreesc, saveourplanet. Theme: Topic on how individual usage of energy impacts the conservation of our ecosystems.

• **Topic 16 keywords**: youtube, qoghgf, industry, good, datascience, director, covid, bigdata, wednesday, health. Theme: Another bot tweeting about general SDGs news.

• **Topic 17 keywords**: woods, trees, designed, involved, protect, pests, species, observatree, early, warning. Theme: Topic around the project ObservaTree, a tree health early warning system powered by CS.

To finish this analysis, it could be interesting to know how many tweets are assigned to each topic. Figure 11 shows that tweets are distributed nearly equally across the 17 topics that we have used to divide them. However, we can see that topic 9 has much more tweets than anyone else. We expected these results since user4 retweets a lot about anything related to SDGs. This assumption can be corroborated just by accessing the user’s timeline. In addition, these results show that there are SDGs that do not have their own topic, and one of them (SDG13, Climate Action) that is present in several topics.

![FIGURE 11. Number of tweets belonging to each topic.](image)

### B. SDG CLASSIFICATION

Once we learnt about the topics that are most commonly discussed in CS, it is time to evaluate what SDGs are the most addressed by CS discussions. To do so, we trained a BERT based classifier [43] to automatically assign tweets to a SDG. BERT stands for Bidirectional Encoder Representations from Transformers, designed to pretrain deep bidirectional representations from unlabelled texts by jointly conditioning on both left and right context in all layers. We used the Bert Based Multilingual Uncased pretrained model in order to address language issues. This model lays on the Bert based transformer, which consists of 12 layers (transformer blocks), 12 attention heads and 110 million parameters. To compile the model, we used the Adam optimizer with a learning rate of 2e-5 and an epsilon of 1e-08. We used Sparse Categorical accuracy for the metrics of the model.

To collect tweets for training and evaluating the algorithm, we first compiled 57,843 tweets that where directly downloaded from the Twitter API by using the keywords SDG1 to SDG17 (English version), ODD1 to ODD17 (French version) and ODS1 to ODS17 (Spanish version). We searched by those three languages because the are the ones that appear most often in our dataset. 80% of the tweets were used for training, while 20% of the tweets where used for testing. This division was based on random sampling.

In order to create the training dataset we ensured that the tweets from the analyzed dataset were removed from the training dataset in order to avoid overfitting. In addition, the training datasets only consists of unique tweets. To do so, we converted the retweets to the original tweets and then we removed the duplicates. With this, we avoided that the same tweet appears both in the training and the testing samples. Afterwards, we preprocessed the tweets from the training dataset. Apart from the tasks mentioned in section IV we also lemmatized and stemmed the text. Then, we trained the model by using the training dataset with the samplings mentioned in the previous paragraph and predicted the categories of the tweets of the dataset we are analysing in our work. Finally, we preprocessed the tweets as explained in section IV and carried out the predictions.

The confusion matrix in Figure 12 and the table 1 show the results of the developed Twitter SDG classifier. In the diagonal, we can see the percentage (normalized to 1) of successes when assigning a tweet to the correct category. For instance, if we have a look at the first row, it can be observed that 75% of tweets that correspond to the SDG1, are correctly assigned to that SDG. On the other hand, we can observe that 5% of the tweets that belong to SDG1, are assigned to SDG2. The table shows that it has a good performance in assigning tweets to its corresponding SDG, having most of them a F1-score higher than 0.70, taking into account the small size of the dataset. The overall F1-score is 0.82. However, the current performance can be improved in the future by adding further training data, so to be said, by gathering more tweets related to SDGs. This will improve the prediction for those categories, such as SDG9 for which we have a small set of tweets.

We applied the developed classifier to dataset mentioned in section IV. Figure 13 shows a graph with the number of tweets assigned to each SDG. We can see that there is a large amount of tweets classified under the category SDG13. SDG13 talks about taking urgent action to combat climate change and its impact. These results are in accordance with the results obtained after applying topic modelling to our dataset since there were 6 topics on climate change which means at the end that climate change is a hot topic in the CS scenario. The outputs of this model can help people to easily see the direction of CS discussions in Twitter.
C. SDGs DETECTION IN RETWEETS

Once all the tweets were assigned a SDG, the next step was to visualize a graph that represents how users retweet certain tweets and the category assigned to those tweets. The number of nodes obtained was 12,144 between users and tweets (7,916 users and 4,228 tweets). The way we obtained and displayed the results was via a bipartite (two-mode) graph. One set of nodes was composed of the users and the other one of the tweets they retweeted, both with unique values so the same user or tweet would not be displayed twice or more times. The high number of nodes obtained drove us to use the k-core algorithm provided by networkX module based on the k-core decomposition of graphs. The k-core decomposition is a perfect tool to visualise large scale networks [38] because it is based on the trimming of the network removing the least connected nodes in succession while the k increases and the visualization focuses on the central cores. Figure 14 presents a graphical representation of this. While we increase the k the least connected nodes on the periphery are removed and we keep the core with more connected nodes, obtaining a maximal connected subgraph moving to the central core (with the maximal k) by the end.

The minimal k was deprecated because it returned the maximal number of nodes and it was difficult to get a good visualization, so we decided to show the second core obtained using a k with a value of 2. The resulting graph contains 2,261 nodes, in Figure 15 those nodes corresponding to users are displayed in a light blue and small size. The nodes corresponding to tweets are displayed in a bigger size and they have different colours depending on the SDG they belong to. In Figure 15 it can be observed that the predominant colour is green corresponding to SDG13, which corresponds to the SDG “Climate action”.

VI. NETWORK ANALYSIS

A. TO FOLLOW OR TO BE FOLLOWED

Prior to the benefit from more advanced analytics, the first analysis we want to show is the comparison ratio between followers and followings. Figure 16 shows the number of followers and followings of each user in our dataset, which contains the information about 10,186 unique users. In this scatter plot, each user is represented by a dot. The x axis represents the number of followers the user has, while the y axis represents the number of users that this user is following. Simple descriptive statistics show that 25% of the users have less than 242 followers and they follow less than 251 users while 50% of the users have between 242 and 2,476 followers and they are following between 251 and 1,849 users. Just as a particularity, there are 52 users that are not following anyone and 62 users that are not being followed by anyone.
Traditionally, it is considered that users that are highly followed are personalities and institutions whose influence and reputation is superior to individual users who subscribe to a large number of accounts without themselves being widely followed. After analysing our results we found that: a) there are 5,229 users (51.34%) that are followed by less users than they follow, b) 2,325 users (22.83%) of users that are followed by the same amount of users than they follow and c) 717 users (7.04%) that are follow two times more than they follow themselves. This means that from our whole dataset, d) only 1915 (18.79%) of the users are followed more than two times more than they follow themselves.

In the last two categories (c and d), we find users who are followed by more users than they follow themselves, generally because they occupy a privileged position in the field. It will be in those two categories where we will find some of the most prominent people in the field that, in our scenario, are people working in CS that create conversation around SDGs in Twitter. It would be interesting to extract the expertise and popularity of the followers of the users we are analysing, so we could bring to the table further discussion regarding quality and quantity. In this scenario we could debate about if it is more relevant to have fewer but more influential users rather than being follower a large amount of users that are not popular. In future sections we will raise this question again, but from the perspective of tweets and retweets.

**B. USER’S RETWEETS VISUALIZATION**

CS communities, as every other community in social media, tend to congregate around certain users. In our study we explored the links between the users in our database extracting the retweets given from one user to another. These individual links were represented in a directed graph, where every node acted for one user and the connection between nodes stood for the retweet given. We applied the same filters as in the previous analysis, besides we made use again of the \( k \)-core algorithm provided by the networkX package for Python. We obtained a graph with 1234 nodes (users) and 2,754 edges (connections/retweets between these users). The number of strong connected components in our graph is 1,234 and 8 weakly connected components. These nodes form 8 different subgraphs, being the bigger one a graph composed of 1,200 nodes, so almost all the users that were firstly obtained, meaning that they form a big, connected community. The graph with the higher number of nodes is presented in Figure 17, where the nodes with higher number of retweets are redder.

**C. GRAPH CENTRALITY MEASURES**

A visual representation of a network is not enough to understand the complete characteristics of it, so we provide a centrality analysis of this system. The analysis is based on the indegree, outdegree, betweenness and eigenvector values of each node [44]. The indegree is the number of retweets received, while the outdegree is the number of retweets given. The betweenness centrality represents the number of times a vertex (node) is present on the shortest path between two other vertices. Finally, the Eigenvector centrality assigns each vertex a score of authority that is based on the score of the vertices with which it is connected. In Table 2 the values of
these measurements for the user with the higher indegrees are shown.

The column Name shows the anonymized id of the user, the column Val shows the value itself of the different parameters and the column R tells us the position of that user in relation of the rest of the nodes. As it can be seen in Table 2, most of the accounts receive a high number of retweets but do not retweet that much, meaning that they do not share the information that other users provide or create. Anyway, there are three exceptions as four users (user1, user2, user4 and user6) from table 2 also give a substantial number of retweets to other users, their rank in outdegree is between the top 50.

In Table 3 we can observe the same situation but with the user with a higher score in outdegree. The scenario in this case is completely different, the users with higher outdegree are really low in the ranks of indegree. They retweet a lot of information regarding SDGs, but they are not retweeted. There are only two exceptions, the user4 (the user6 from table 2) and user11 (user39 in table 2) which are high in ranks of indegree. The user1, the first one giving retweets, it is an account fully dedicated to retweet other users.

Finally, we can observe an expected situation in every social media, the most popular accounts receive the most retweets but do not share the information from others. This is not a good behaviour when it comes to share information from others, not only own content. In any case, we see some accounts active in both ways, a good habit in order to distribute information and connect users by showing other accounts. In Figure 18 we display a visual representation of this phenomena. In the x axis we have the indegree and in the y axis the outdegree. The users who receive more retweets are allocated close to the x axis, most of the accounts are congregate in the low values of both axis. The users who retweet more are high in y values, close to zero in the x axis.

VII. DISCUSSION

The increasing use of social media, with platforms such as Twitter, serves an important role in dissemination of scientific information [45]. When looking at CS and SDGs in Twitter, we see that in our dataset only 7.08% of CS tweets are related to SDGs during the gathering period. Although this percentage is small, it can probably be enlarged by adding more filters to the search query. While our research focuses on discussions observed on Twitter around this topic, it is worthwhile to recognise the challenges around social media and the relevance of misinformation, that also affects CS [46].

Our work shows that there is discussion around SDGs in the CS scenario, but the quality of the information analysed has not been addressed in our research. In fact, along this document we can observe that some bots arise as important subjects in the dissemination of information. This finding aligns with Marlow et al. work [47] in which it is stated that bots are having a significant impact upon discourses on Twitter.

It seems clear from the extracted dataset that the most important topic discussed on Twitter by users tweeting about CS is climate change. Due to the characteristics of CS projects, which are usually within the area of conservation and environmental sciences [48] it seems natural that when going to social networks, discussions deal with this two topics, which are usually close to climate change issues [49].

Based on our results it can be observed that #climate-change and #opencience are the most used and retweeted hashtags in Twitter, alongside with #sdg and #cs_sdg2020. When taking into account the rest of used and retweeted hashtags, what we seem to find is that the general discussion is apparently focused in nature, climate, energy and sustainability which are subjects where a good number of the CS projects can be found [49]. Moreover, climate change is a topic that is discussed from very different perspectives, ranging from more environmental and physical nature (quality of air, impact on phenology, etc) to a more political and ethical issue (climate justice) [50]. In section 11 we presented a topic modelling analysis in which 6 topics out of 17 where about climate change, further strengthening the previous affirmations. Climate change and CS are topics that has been discussed by many authors in the past, stating that CS can be important when contributing to climate change and promoting individual and collective actions that tackle that issue, provided that an appropriate theoretical background and a scientific methodology for gathering data is used [49], [51], [52].

On the basis of the results it can be said that it exists a relevant number of citizen scientists discussing about 2030 Agenda and SDGs in Twitter. Based on the results of the SDG detection, we can say that there are more than 12,000 users talking about SDGs, and this is only taking into account the retweets and that our filter process could be extended adding more keywords. Despite considering this as a big community, what it can be observed is that certain users orbit around certain topics (normally of the same matter). For example we can see groups of users around tweets categorised as belonging to SDG 13 and 11, which are SDGs related to climate that are quite common in CS discussions [48].

FIGURE 18. Scatterplot showing indegree vs outdegree.
Although, when having a look at the retweets network it seems that communities are formed around certain users. In this sense, it appears to be easier to categorise communities around users than around tweets. As in every social media, certain users tend to congregate people and monopolise retweets, likes, etc. Our case does not deviate from this trend, since we found this clustering around certain users [15]. We can observe two big communities around those two nodes in red in Fig. 17. Those nodes correspond to the accounts of user1 and the user2, being these accounts the two most retweeted accounts receiving almost double the retweets than the following accounts in the ranking.

The user1 from table 2 being the most retweeted one was an obvious possibility, the official account in Twitter for SDGs must be the point of reference for the interaction with the people when communicating SDGs [54]. After that, we find the user2, account of a project, which is making a great effort in communication around the SDGs. Then, at the rear of this two accounts, what we have is a list of users in which we can find several entities (user6, user14, user19...) but also individuals (user3, user7, user10...). According to the results, we see that the accounts with the highest number of retweets are the first seven in table 2, existing a big difference with the rest of the users. Three of these accounts are entities or projects, the rest are individuals. These people are highly connected to SDGs of Climate and Biodiversity(user5) and to SDGs related to AI and technologies (user3 and user7). Lastly we have user4, who is an activist from Japan who has
TABLE 3. Centrality measures of the accounts that retweet the most.

| Name | InDeg | OutDeg | Eigen. | Betw. |
|------|-------|--------|--------|-------|
| User1 | 2312  | 280.1  | 25.49  | 2312  |
| User2 | 2902  | 73.2   | 25.49  | 2902  |
| User3 | 2900  | 69.3   | 25.49  | 2900  |
| User4 | 120.6 | 68.4   | 4.72R+15 | 1.83E+15 |
| User5 | 7699  | 58.5   | 25.49  | 7699  |
| User6 | 2263  | 50.6   | 25.49  | 2263  |
| User7 | 2195  | 45.7   | 25.49  | 2195  |
| User8 | 2795  | 43.8   | 25.49  | 2795  |
| User9 | 2134  | 42.9   | 25.49  | 2134  |
| User10| 2328  | 32.10  | 25.49  | 2328  |
| User11| 32.39  | 30.11  | 3.63B+16 | 8.4E+14 |
| User12| 17.97  | 25.12  | 1.03E+16 | 4.43E+16 |
| User13| 240.1  | 25.13  | 5.08E+15 | 1.21E+16 |
| User14| 2178  | 24.14  | 25.49  | 2178  |
| User15| 176.24  | 14.15  | 6.62E+15 | 2.91E+15 |
| User16| 6658  | 22.16  | 25.49  | 6658  |
| User17| 219.1  | 20.17  | 8.79E+15 | 50844409.6 |
| User18| 1719  | 21.18  | 8.15E+13 | 74907018136 |
| User19| 220.19  | 20.19  | 5.13E+16 | 2.73E+15 |
| User20| 108.16  | 20.20  | 1.22E+15 | 5.62E+11 |
| User21| 2106  | 19.21  | 25.49  | 2106  |
| User22| 2941  | 18.22  | 25.49  | 2941  |
| User23| 2159  | 17.23  | 25.49  | 2159  |
| User24| 36.31  | 17.24  | 1.79E+16 | 3.02E+16 |
| User25| 7667  | 16.25  | 4.31E+16 | 1.23E+16 |
| User26| 58.16  | 16.26  | 1.51E+16 | 7.76E+15 |
| User27| 316.16  | 16.27  | 1.67E+15 | 62653958087 |
| User28| 2262  | 14.28  | 25.49  | 2262  |
| User29| 227.2  | 14.29  | 3.75E+15 | 2.09E+16 |
| User30| 2.950  | 16.306  | 1.63E+16 | 1.22E+16 |
| User31| 1.855  | 14.31  | 25.49  | 8553  |
| User32| 7.78  | 13.32  | 8.07E+15 | 6273170165 |
| User33| 7.313  | 13.33  | 2.47E+16 | 5.54E+14 |
| User34| 15.118  | 12.34  | 1.17E+16 | 22001082389 |
| User35| 2310  | 12.35  | 25.49  | 2310  |
| User36| 55.19  | 11.36  | 1.94E+15 | 1.71E+15 |
| User37| 6.42  | 11.37  | 3.86E+15 | 971821596 |
| User38| 18.92  | 11.38  | 7.00E+15 | 99781253160 |
| User39| 2435  | 11.39  | 25.49845017 | 2435  |
| User40| 23.289  | 11.40  | 25.49  | 2892  |
| User41| 4.132  | 11.41  | 137829577.5 | 15785125986 |
| User42| 24.57  | 10.42  | 6.70E+15 | 4.33E+15 |
| User43| 1.159  | 10.43  | 1.27E+16 | 6305035774 |
| User44| 1.214  | 9.44  | 25.49  | 2142  |
| User45| 1.194  | 9.45  | 8.53E+14 | 40305297217 |
| User46| 2157  | 9.46  | 25.49  | 2157  |
| User47| 2238  | 9.47  | 25.49  | 2238  |
| User48| 26.51  | 9.48  | 1.15E+16 | 3.38E+15 |
| User49| 8.259  | 9.49  | 3.79E+15 | 673E+14 |
| User50| 14.137  | 9.50  | 8306847124  | 217  |

been involved in Sustainable Development Goals since Kyoto Conference in 1997. In social media, individuals leading the dissemination of information has been studied and observed for a long time [45], [55], in our network we have several examples that corroborate that.

Regardless of the existence of these big accounts that receive retweets and unite people, we find the same phenomena as in other studies. A normal situation in Twitter is that those who receive high numbers of retweets do not correspond by retweeting others. In the first position of those who retweet others we find a special account, a bot account which only retweet other users, user1 in table 3, with a high impact [47]. This is a special case so we focused on companies, projects and individual accounts. We found an interesting case in the account of the user4, in table 3, a conference account which is high in the rank of retweeted accounts and owns a top 4 rank in retweeting others. In spite of this remarkable situation we explained, the rule about not retweeting that much is fulfilled. Besides, we can see how the rich club effect in networks is also applied in this case, as those big accounts are perfectly connected and inside a community in that network [56].

VIII. CONCLUSION

Our research explores how CS discuss around SDGs in Twitter discussions. Although tweets about SDGs account for only around 7% of the total number of tweets analyzed, we could present an overall view of many of the elements that are part of the conversation such as popular words, topics and people. We explored the different SDGs that are mentioned
in CS tweets, reaching the conclusion that climate change is one of the most discussed topics by CS in Twitter. Different perspectives have been observed when talking about climate change, ranging from data gathering to promoting policies. Discussions usually happen around certain users, that are retweeted a lot. However, if we are looking for a more informative purpose we should have a look at the users that retweet a lot, trying to find a balance between retweeting and being retweeted.

With this analysis we tried to shed light into how CS works with SDGs providing an insight into the discussions in Twitter. However, it is important to note the limitations of this research. The first one, is that Twitter is not the only social network use in CS to communicate, so other platforms such as Facebook can be included in future studies in order to gather more complete insights. Our analysis did not aim to analyse the impact of specific tweets nor users, but to explore conversations as a whole, extracting topics and exploring the networks created through those conversations. Future works will include the analysis of topics over time, to see how the conversation evolve in Twitter. In addition, multilabelling of tweets will be taken into account in future work since it is possible to find a tweet associated to several SDGs. One of the positive points of this study, is that it can be replicated in the future in order to see the evolution of the conversations, so to be said, if the focus remains on climate change or if it shifts towards any other SDG. Therefore, a long-term analysis will be always plausible, since it only depends on the used dataset, which will grow over time.

REFERENCES

[1] R. Jordan, A. Crall, S. Gray, T. Phillips, and D. Mellor, “Citizen science as a distinct field of inquiry,” *Bioscience*, vol. 65, no. 2, pp. 208–211, Feb. 2015.

[2] S. Hecker, R. Bonney, M. Haklay, F. Hölker, H. Hofer, C. Goebel, M. Gold, Z. Makuch, M. Ponti, A. Richter, L. Robinson, J. Iglesias, R. Owen, T. Peltola, A. Sforzi, J. Shirk, J. Vogel, K. Vohland, T. Witt, and A. Bonn, “Innovation in citizen science–perspectives on science-policy advances,” *Citizen Sci.: Theory Pract.*, vol. 3, pp. 1–14, Jun. 2018.

[3] S. Droeger, “Just because you paid them doesn’t mean their data are better,” in *Proc. Citizen Sci. Toolkit Conf.*, 2007, pp. 13–26.

[4] C. Kullenberg and D. Kasperowski, “What is citizen science?–A scientometric meta-analysis,” *PLoS ONE*, vol. 11, no. 1, 2016, Art. no. e0147152, doi: 10.1371/journal.pone.0147152.

[5] J. Silvertown, “A new dawn for citizen science,” *Trends Ecol. Evol.*, vol. 24, no. 9, pp. 467–471, Sep. 2009.

[6] M. Arestiadou and C. Herodotou, “Online citizen science: A systematic review of effects on learning and scientific literacy,” *Citizen Sci.: Theory Pract.*, vol. 5, no. 1, pp. 1–12, Apr. 2020.

[7] S. F. Johnston, B. Franks, and S. Whitelaw, “Crowd-sourced science: Societal engagement, scientific authority and ethical practice,” *J. Inf. Ethics*, vol. 26, no. 1, pp. 49–65, 2017.

[8] G. Newman, A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston, “The future of citizen science: Emerging technologies and shifting paradigms,” *Frontiers Ecol. Environ.*, vol. 10, no. 6, pp. 308–304, Aug. 2012.

[9] H. A. Triebenzer, B. A. Knuth, Y. C. Yuan, and J. L. Dickinson, “Internet-based social networking and collective action models of citizen science,” in *Citizen Science: Public Participation in Environmental Research*. Ithaca, NY, USA: Cornell Univ. Press, 2012, pp. 214–225.

[10] D. L. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Burlington, MA, USA: Morgan Kaufmann, 2010.

[11] X. Li, Q. Xie, J. Jiang, Y. Zhou, and L. Huang, “Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology,” *Technol. Forecasting Social Change*, vol. 146, pp. 687–705, Sep. 2019.

[12] E. C. Malthouse, M. Haenlein, B. Skiera, E. Wege, and M. Zhang, “Managing customer relationships in the social media era: Introducing the social CRM house,” *J. Interact. Marketing*, vol. 27, no. 4, pp. 270–280, Nov. 2013.

[13] L. P. L. K. Stanislawská, S. R. R. K. Jana Poláková, and G. Gresham, “Customer experience with organic food: Global view,” *Emirates J. Food Agricult.*, vol. 30, no. 11, pp. 918–926, Dec. 2018.

[14] P. Walsh, G. Clavio, M. D. Lovell, and M. Blaszkza, “Differences in event brand personality between social media users and non-users,” *Sport Marketing Quart.*, vol. 22, no. 4, pp. 214–223, 2013.

[15] S. Mazumdar and D. Thakker, “Citizen science on Twitter: Using data analytics to understand conversations and networks,” *Future Internet*, vol. 12, no. 12, p. 210, Nov. 2020.

[16] *Transforming Our World: The 2030 Agenda for Sustainable Development*, United Nations, New York, NY, USA, 2015.

[17] K. Shulla, W. L. Filho, J. H. Sommer, A. L. Salvia, and C. Borgemeister, “Channels of collaboration for citizen science and the sustainable development goals,” *J. Cleaner Prod.*, vol. 264, Aug. 2020, Art. no. 121735.

[18] D. Fraisl, J. Campbell, L. See, U. Wehn, J. Wardlaw, M. Gold, I. Moorthy, R. Arias, J. Piera, J. L. Oliver, J. Masó, M. Penker, and S. Fritz, “Mapping citizen science contributions to the UN sustainable development goals,” *Sustainability Sci.*, vol. 15, no. 6, pp. 1735–1751, Nov. 2020.

[19] J. D. Sachs, “From millennium development goals to sustainable development goals,” *Lancet*, vol. 379, no. 9832, pp. 2206–2211, Jun. 2012.

[20] *Learning Through Citizen Science: Enhancing Opportunities by Design*, National Academies of Sciences, Engineering, and Medicine, Washington, DC, USA, 2018.

[21] K. C. Elliott and J. Rosenberg, “Philosophical foundations for citizen science,” *Citizen Sci.: Theory Pract.*, vol. 4, no. 1, p. 9, Mar. 2019.

[22] S. Fritz, L. See, T. Carlson, M. M. Haklay, J. L. Oliver, D. Fraisl, and R. Mondardini, “Citizen science and the United Nations sustainable development goals,” *Nature Sustainability*, vol. 2, no. 10, pp. 922–930, 2019.

[23] J. Sprinks, S. M. Woods, S. Parkinson, U. Wehn, H. Joyce, L. Ceccaroni, and M. Gharesifard, “Coordinator perceptions when assessing the impact of citizen science towards sustainable development goals,” *Sustainability*, vol. 13, no. 4, p. 2377, Feb. 2021.

[24] U. Wehn, M. Gharesifard, L. Ceccaroni, H. Joyce, R. Ajates, S. Woods, A. Bilbao, S. Parkinson, M. Gold, and J. Wheatland, “Impact assessment of citizen science: State of the art and guiding principles for a consolidated approach,” *Sustainability Sci.*, vol. 16, no. 5, pp. 1683–1699, Sep. 2021.

[25] T. Phillips, N. Porticella, M. Constat, and R. Bonney, “A framework for articulating and measuring individual learning outcomes from participation in citizen science,” *Citizen Sci.: Theory Pract.*, vol. 3, no. 2, p. 3, Aug. 2018.

[26] J. Cox, E. Y. Oh, B. Simmons, C. Lintott, K. Masters, A. Greenhill, G. Graham, and K. Holmes, “Defining and measuring success in online citizen science: A case study of zooniverse projects,” *Comput. Sci. Eng.*, vol. 17, no. 4, pp. 28–41, Jul. 2015.

[27] D. Brossard, B. Lewenstein, and R. Bonney, “Scientific knowledge and attitude change: The impact of a citizen science project,” *Int. J. Sci. Educ.*, vol. 27, no. 9, pp. 1099–1121, Jan. 2005.

[28] J. C. Davids, M. M. Rutten, A. Pandey, N. Devkota, W. D. van Oyen, R. Prajapati, and N. van de Giesen, “Citizen science flow–an assessment of impact in citizen science,” *Citizen Sci.: Theory Pract.*, vol. 3, no. 2, p. 3, Aug. 2018.

[29] A. Liberatore, E. Bowkett, C. J. MacLeod, E. Spurr, and N. Longnecker, “Social media as a platform for a citizen science community of practice,” *Citizen Sci.: Theory Pract.*, vol. 3, no. 1, p. 3, Mar. 2018.

[30] M. M. Skoric, Q. Zhu, D. Goh, and N. Pang, “Social media and citizen engagement: A meta-analytic review,” *New Media Soc.*, vol. 18, no. 9, pp. 1817–1839, Oct. 2016, doi: 10.1177/1461444815616221.

[31] W. Zhang, T. J. Johnson, T. Seltzer, and S. L. Bichard, “The revolution will be networked: The influence of social networking sites on political attitudes and behavior,” *Social Sci. Comput. Rev.*, vol. 28, no. 1, pp. 75–92, 2010, doi: 10.1177/0894439309335162.

[32] M. Grandjean, “A social network analysis of Twitter: Mapping the digital humanities community,” *Cogent Arts Humanities*, vol. 3, no. 1, Dec. 2016, Art. no. 1171458.
[33] K. Subbian and P. Melville, “Supervised rank aggregation for predicting influencers in Twitter,” in Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE Third Int. Conf. Social Comput., Oct. 2011, pp. 661–665.

[34] E. Vainikka and J. Huhtamäki, Tvittien Politiikka–Poliittisten Viestinnän Sisäisintä Twitterissa. Media & Viestintä, Helsinki, Finland, 2015, pp. 165–183.

[35] T. Miyazaki, A. Rahimi, T. Cohn, and T. Baldwin, “Twitter geolocation using knowledge-based methods,” in Proc. Workshop W-NUT: 4th Workshop Noisy User-Generated Text (EMNLP), 2018, pp. 7–16.

[36] A. Wiggins and K. Crowston, “From conservation to crowdsourcing: A typology of citizen science,” in Proc. 44th Hawaii Int. Conf. Syst. Sci., Piscataway, NJ, USA, Jan. 2011, pp. 1–10.

[37] S. Daume and V. Galaz, “‘Anyone know what species this is?’–Twitter conversations as embryonic citizen science communities,” PloS ONE, vol. 11, no. 3, 2016, Art. no. e0151387.

[38] J. I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, and A. Vespignani, “K-core decomposition: A tool for the visualization of large scale networks,” 2005, arXiv:0504107. Accessed: Jul. 12, 2021. [Online]. Available: https://arxiv.org/abs/cs/0504107

[39] L. Bolelli, Š. Ertek, and C. L. Giles, “Topic and trend detection in text collections using latent Dirichlet allocation,” in Proc. Workshop W-NUT: 4th Workshop Noisy User-Generated Text (EMNLP), 2009, pp. 767–780.

[40] P. Bruneau, E. Brangbour, S. Marchand-Maillet, R. Hostache, M. Chini, R.-M. Pelich, P. Matgen, and T. Tamisier, “Measuring the impact of natural hazards with citizen science: The case of flooded area estimation using Twitter,” Remote Sens., vol. 13, no. 6, p. 1153, Mar. 2021.

[41] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[42] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in Proc. 8th ACM Int. Conf. Web Search Data Mining, Feb. 2015, pp. 399–408.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[44] S. P. Borgatti and M. G. Everett, “A graph-theoretic perspective on centrality,” Soc. Netw., vol. 28, no. 4, pp. 466–484, 2006.

[45] M. Tsukobura, Y. Onoue, H. A. Torii, S. Suda, K. Mori, Y. Nishikawa, A. Ozaki, and K. Uno, “Twitter use in scientific communication revealed by visualization of information spreading by influencers within half a year after the Fukushima daiichi nuclear power plant accident,” PloS ONE, vol. 13, no. 9, Sep. 2018, Art. no. e0203594.

[46] T. R. Katappally, “A global digital citizen science policy to tackle pandemics like COVID-19,” J. Med. Internet Res., vol. 22, no. 5, May 2020, Art. no. e19357.

[47] T. Marlow, S. Miller, and J. T. Roberts, “Twitter discourses on climate change: Exploring topics and the presence of Bots,” SocArXiv, Feb. 2020.

[48] S. Manske. (2021). Are Citizen Science Projects Multi-Disciplinary Research Activities?: CS-Track e-magazine. [Online]. Available: https://cs-track.eu/graphical-article/cs-projects-multi-disciplinary-research-activities/

[49] C. B. Cooper, J. Shirk, and B. Zuckerberg, “The invisible prevalence of citizen science in global research: Migratory birds and climate change,” PloS ONE, vol. 9, no. 9, Sep. 2014, Art. no. e106508.

[50] D. Schlossberg and L. B. Collins, “From environmental to climate justice: Climate change and the discourse of environmental justice,” WIREs Climate Change, vol. 5, no. 3, pp. 359–374, May 2014.

[51] A. H. Hurlbert and Z. Liang, “Spatiotemporal variation in avian migration phenology: Citizen science reveals effects of climate change,” PloS ONE, vol. 7, no. 2, Feb. 2012, Art. no. e31662.

[52] M. Groulx, M. C. Brisbois, C. J. Lemieux, A. Winegardner, and A. H. Hurlbert and Z. Liang, “Spatiotemporal variation in avian migration phenology: Citizen science reveals effects of climate change,” PloS ONE, vol. 7, no. 2, Feb. 2012, Art. no. e31662.

[53] A. Landherr, B. Friedl, and J. Heidemann, “A critical review of centrality measures in social networks,” Medit. J. Social Sci., vol. 39, no. 3, pp. 213–215, 2013, doi: 10.1016/j.pubrev.2013.02.005.

[54] C. Mhamdi, “Transgressing media boundaries: News creation and dissemination in a globalized world,” Medit. J. Social Sci., Sep. 2016, doi: 10.5901/mjss.2016.v7n5p272.

[55] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, “Detecting rich-club ordering in complex networks,” Nature Phys., vol. 2, no. 2, pp. 110–115, Feb. 2006, doi: 10.1038/nphys209.

DAVID ROLDÁN-ÁLVAREZ received the Ph.D. degree from Universidad Rey Juan Carlos, in July 2017. He is currently working as a Doctoral Assistant Professor at the Escuela Técnica Superior de Ingeniería de Telecomunicación, Universidad Rey Juan Carlos, combining his teaching work with participation in research projects. He has participated as a Researcher and a Developer in three R+D+i projects and transfer to industry (BlueThinking of the Orange Foundation, Clipper of JISC, and DEDOS of ISBAN) and five research projects, of which three are national (TIC and TIN programs) and two are international (FP7, 2012-2015 and Swafs 2019-2022). His interests include teaching innovation through technology to improve both the performance of teachers and the performance of students.

FERNANDO MARTÍNEZ-MARTÍNEZ received the joint master’s degree in bioinformatics and bio-statistics from the Universidad de Barcelona and the Universidad Oberta de Catalunya. His current work is at the Escuela Técnica Superior de Ingeniería de Telecomunicación, Universidad Rey Juan Carlos, as a Programmer Technician associated to the European CS-Track project. He has participated as a Researcher and a Developer with the Josep Carreras Leukaemia Research Institute (IJC) in the research project evaluation of the contribution of polymorphic transposons to infertility and other phenotypes in mice, in 2020. His interests include machine learning and AI to innovate in the research field, alongside with data analysis of medical, biological, and social features.

ESTEFÁNIA MARTÍN was the Vice Principal of promotion and research (Subdirectora de Promoción e Investigación) with the URJC School of Computer Science (Escuela Técnica Superior de Ingeniería Informática), from December 2015 to January 2020. She is currently an Associate Professor of technology-enhanced learning at the LITE Laboratory, Universidad Rey Juan Carlos (URJC). She leads BlueThinking, an application that allows the person with ASD to learn programming and improve their executive functions; DEDOS project, which provides authoring tools for creating educational activities on multiple devices, and ClipIt, a video-based social network platform that has been developed in the context of the European Union project JuxtaLearn. Furthermore, she is CoP of CS-Track project focused on Citizen Science. Her research interests include HCI, people with disabilities, CSCL, e-learning, blended learning, video-based learning, recommender systems, and authoring tools.

PABLO A. HAYA received the bachelor’s and master’s degrees in telecommunications engineering from the Universidad Politécnica Madrid and the Ph.D. degree in computer science engineering, and telecommunications from the Universidad Autónoma of Madrid, in 1999 and 2006, respectively. He is currently the Head of the Social Business Analytics (SBA) Group, Instituto de Ingeniería del Conocimiento (IIC). He is focused on providing solutions that analyze individual and social human behavior by means of state-of-the-art technologies, such as natural language processing, social network analysis, and machine learning. He has published more than 70 journal articles, book chapters, and conference proceedings. Currently, his research interests include social computing and human–computer interaction with a particular focus on their application into social media and technology-enhanced learning.

** **