Universidad Rey Juan Carlos

TESIS DOCTORAL

Establishing a pipeline for social network analysis in Citizen Science: Integration into a data visualization platform for interactive and integrative analysis of discourse

Autor: ***Fernando Martínez Martínez***

Directores: ***Estefanía Martín Barroso, David Roldán Álvarez***

Programa de Doctorado en

Tecnologías de la Información y las Comunicaciones

2024

Agradecimientos

A los compañeros del proyecto CSTrack, con quien he aprendido muchísimo, he sacado buenas amistades y con quien espero coincidir en muchos más proyectos.

A David y Estefanía, por animarme y ayudarme a canalizar las ideas y llevarme a, espero, un estilo de escritura como mínimo aceptable. También por abrirme el mundo de la Universidad, donde espero estar muchos años con ellos. Gracias por todo.

A mi madre y a mi hermano, que me habrán visto horas y horas colgado del ordenador y muy posiblemente malhumorado. Poco a poco, todo acabará yendo bien. Os quiero.

A Patri, que la he tenido encima para que no decayera mi ánimo. Esto marca un punto nuevo y espero que sea un comienzo para los dos. Te quiero.

Y a ti papa, que no estás aquí para verlo, pero sé que te haría ilusión. Esto es para ti, te echo de menos cada día más.

Resumen

En las últimas décadas, los proyectos relacionados con Ciencia Ciudadana (en adelante CC) se han multiplicado y su calado en cuestiones de importancia para la investigación y la sociedad ha aumentado. La CC se puede definir como la participación de ciudadanos sin un trasfondo científico para que realicen tareas típicamente asociadas a investigadores, bajo la supervisión de profesionales enmarcados en un proyecto o institución.

La CC se ha visto favorecida en gran medida por la expansión de Internet y las tecnologías asociadas, gracias a las mejoras en conectividad y las posibilidades de captación de posibles participantes y expansión de sus noticias, hallazgos y actividades. En las últimas dos décadas, muchos proyectos, plataformas e instituciones han aumentado su presencia en Internet, buscando aumentar la participación de aquellos interesados, así como de mejorar la transmisión de información. Además, esto ha moldeado nuevas formas de participación online expandiendo los ámbitos de aplicación de esta práctica. Dentro de todas las tecnologías que han favorecido este crecimiento, son destacables las redes sociales.

Las redes sociales son plataformas web que promueven la conexión entre individuos, dando lugar a complejas redes que suelen estudiarse en forma de grafos. Además de la estructura formada por las conexiones entre individuos, las redes sociales presentan características propias relacionadas con las herramientas y contenidos ofrecidos en las plataformas web, como bien puede ser información sobre los individuos que las conforman o relaciones entre ellos empleando características de las plataformas. Estas características especiales convierten a las redes formadas en plataformas sociales en sujetos de investigación, permitiendo conocer el comportamiento de las personas, sus gustos, las modas que se generan y conocer el conocimiento que comparten. En la CC, las redes sociales se han aprovechado para compartir las iniciativas de esta práctica, dar a conocer noticias del ámbito, captar participantes y mantenerse informado sobre esta materia. Numerosos proyectos y participantes de CC se encuentran activamente compartiendo información en redes sociales, y una de las redes sociales que ha captado la atención de los investigadores sobre CC es Twitter.

Para la investigación de estas redes sociales se emplea lo que se conoce como análisis de redes sociales, una disciplina que existe desde hace más de 70 años pero que ha avanzado con la llegada de nuevas técnicas y tecnologías como la inteligencia artificial. En esta

tesis, se realiza el estudio de la comunidad de CC en Twitter, con el objetivo de desarrollar una línea de estudio estandarizada, dado que las investigaciones existentes solo suelen centrarse en la aplicación de técnicas aisladas en casos de estudio generales, y traer esa línea de análisis a una plataforma donde cualquier persona pueda realizar estos análisis sobre cualquier temática. De cara a validar los análisis, se plantean las dimensiones que debían poder ser analizadas con las técnicas, el contenido y la estructura de las redes formadas. Para obtener un conocimiento completo de estas dimensiones, nuestras técnicas debían ser capaces de ofrecernos la posibilidad de determinar si existen múltiples discusiones dentro de una colección de textos y extraer los idiomas en él, determinar el inventario léxico (hashtags, palabras más usadas, etc.), analizar y conocer los temas de discusión, e identificar los principales actores en esta discusión.

Estado del arte

La CC no tiene un origen claro en la historia, pero la intervención de ciudadanos en tareas científicas se ha registrado muchos siglos atrás. Esta disciplina o práctica tomó fuerzas a finales del siglo XX y desde entonces, cada vez más, vemos iniciativas de CC promovidas por gobiernos o la Unión Europea.

Históricamente, la CC ha estado muy ligada a las ciencias naturales, pero también encontramos proyectos de CC en otros ámbitos, como la salud y el cuidado de la salud, las humanidades o la educación. Además, uno de los principales elementos que ha traído a participantes a la CC es la posibilidad de obtener conocimientos y aprender sobre temas que consideren de interés para ellos (aprender prácticas científicas o conocimiento sobre algunos campos), junto con la participación en actividades que consideran de su gusto. Recientemente se ha observado la estrecha relación entre la CC y los objetivos de desarrollo sostenible (ODS), dado que las iniciativas de CC pueden ayudar a completar estos objetivos o a educar en relación con ellos.

Las nuevas tecnologías han sido impulsoras del crecimiento de esta práctica. La posibilidad de encontrar información sobre proyectos de CC en Internet de manera más sencilla para participar en ellos, la obtención de datos desde Internet, la participación online o compartir esa información son algunos ejemplos de su contribución. Las redes sociales se han convertido también en un elemento clave en este crecimiento. Se les considera el ambiente perfecto para compartir información de manera rápida y accesible y ser el entorno perfecto para atraer usuarios y crear comunidad y propiciar un ambiente colaborativo, uno de los dogmas principales de la CC.

Entre todas las redes sociales que existen, Twitter ha sido catalogada como una de las que más han contribuido a la CC. En esta red social cualquier usuario, organización, entidad o proyecto puede crearse una cuenta y compartir información, interactuar con personas afines e invitar a la participación en actividades o compartir datos. Toda la información que se comparte en estas plataformas es susceptible de ser analizada para obtener conocimiento en relación con las comunidades que se crean, lo que se conoce como análisis de redes sociales. Los usuarios de Twitter relacionados con CC utilizan esta red para comunicar sus descubrimientos, opiniones o para atraer nuevos participantes y promocionar sus proyectos.

El análisis de redes sociales es una disciplina que nace hace más de 70 años. A principios del siglo XX se establecen las bases teóricas de esta disciplina, en los años 30 comienzan sus primeros análisis básicos y se extiende en los años 50 alcanzando una gran relevancia en los años 70. Esta disciplina, con la llegada de las redes sociales en Internet, ha sido impulsada para conocer los comportamientos, gustos, intereses y la estructura que conforman los usuarios que se encuentran en ellas. Las principales redes sociales comparten ciertas características, y existen elementos específicos de cada una de ellas. En Twitter, las principales características que son objeto de estudio son la relación de seguimiento entre usuarios, los retweets, los likes, las menciones y el contenido de los tweets como los hashtags.

En este marco se desarrolla la presente tesis, en el análisis de la información relativa a CC en redes sociales y más específicamente en Twitter. El enfoque ha consistido en crear una línea de análisis útil para comprender los estados de la discusión en Twitter y que pudiera servir como método estándar de análisis gracias a su completitud de técnicas que se complementan entre ellas. Además, se quería abordar uno de los principales problemas encontrados en el análisis de redes sociales, el cual es la dificultad de replicabilidad de los estudios que se realizan. Para ello, se quería desarrollar una plataforma de visualización de datos que contuviera las técnicas de análisis de redes sociales de mayor utilidad para obtener una compresión completa de las comunidades, sus comportamientos, sus contenidos y sus participantes.

Metodología

Para llevar a cabo esta tesis se usó la aplicación Lynguo. Esta herramienta está diseñada para recoger tweets diariamente sobre la temática elegida, en nuestro caso sobre CC, aplicando un filtro específico de palabras relacionadas con esta práctica. Además, para enriquecer la información extraída con esta herramienta, se empleó la API de Twitter la cual permite obtener información detallada de los usuarios y completar el análisis. Con estos tweets queríamos llevar a cabo dos tareas principalmente, diseñar una línea de análisis completa para tweets y la creación de la plataforma de análisis. Ambos objetivos se motivan en el intento de ofrecer una solución a dos problemas comunes en el análisis de redes sociales, como hemos detallado anteriormente, la falta de replicabilidad en los estudios y la falta de aplicación de una línea de análisis completa. Para llevar a cabo esto, se debían seleccionar las principales técnicas empleadas en esta disciplina.

Las técnicas que finalmente se desarrollaron contenían análisis de contenido evaluando los hashtags en los textos, las palabras más usadas, la técnica conocida como Frecuencia del Término - Frecuencia Inversa de los Documentos, usada para catalogar las palabras usadas en función de su importancia, y finalmente el análisis de temáticas usando técnicas de *machine learning*. Además, se quería cumplimentar el análisis de contenido con el análisis estructural de redes sociales más clásico. Para ello, se evaluaron las redes que se formaban al emplear las características de Twitter como son el seguimiento, el retweet, las menciones y también la utilización de hashtags en el mismo texto. Con todas estas técnicas elegidas, se desarrolló la plataforma de visualización que podría ser de alta utilidad para un amplio rango de usuarios. Personas que quieran utilizarla para conocer más acerca de la comunidad de CC en Twitter, investigadores que quieran replicar análisis o realizar otros nuevos, proyectos o instituciones que quieran hacer un seguimiento al flujo de información o a los responsables públicos hacedores de políticas que quieran emplearlo para tomar decisiones basadas en evidencias.

Casos de estudio

Para poner a prueba la utilidad y completitud de la línea de análisis, se realizaron diversos estudios sobre temas de importancia para la CC. Cada estudio sirvió para mejorar e implementar nuevas técnicas de cara a refinar el análisis y obtener una línea de investigación lo más completa de los tweets analizados.

El primer estudio estuvo relacionado con el aprendizaje dentro de la CC en el cual se emplearon la primera versión de las técnicas. En este estudio, se incluyó un análisis de los hashtags más utilizados y una primera aproximación al análisis estructural de la red formada por los usuarios en base a los retweets. Se obtuvieron resultados satisfactorios, viéndose corroborados algunos de los comportamientos de la comunidad de CC, validando las técnicas usadas, pero que evidenciaron la necesidad de mejorar la línea de análisis.

En el siguiente estudio, se analizó la relación entre los ODS y la CC introduciendo nuevas técnicas, como el *topic modelling* o la clasificación de tweets empleando técnicas de *machine learning*. Además, se mejoró aquellas técnicas usadas en el primer estudio. La línea de análisis seguía siendo consistente en los resultados, las nuevas técnicas introducidas incluían nuevas dimensiones a la comprensión del estado de la discusión y se pudo observar cómo ciertos comportamientos y relaciones típicas de redes sociales aplicaban en este caso.

En el tercer caso de estudio, se volvieron a incluir nuevos métodos y refinamientos a los disponibles comprobando el campo de la salud y el cuidado de la salud. En este estudio, se incluyó un nuevo método de *topic modelling*, el análisis de la red formada por hashtags y también una de menciones. En este caso, se descubrieron comportamientos que se salen de la norma, como que los proyectos e instituciones son tanto o más activos retweeteando que las personas individuales, lo que difiere del comportamiento típico. Otro resultado interesante fue que a pesar de la cercanía de la pandemia de COVID-19 y en un contexto de salud, se encontró mayor porcentaje de discusión sobre otros temas.

Finalmente, se realizó un estudio destinado a comprobar si, dentro de las características de Twitter, la relación de seguimiento influía sobre el número de las otras interacciones. Este estudio fue realizado junto a otros colaboradores alemanes del proyecto CSTrack donde surge esta tesis. Se realizó una clasificación de usuarios en base a las características de su perfil mediante métodos de *machine learning*, obteniendo un porcentaje de acierto por encima del 92%. La aportación de esta tesis a este estudio consistió en modelar las redes de interacciones basadas en retweets, menciones y respuestas, así como una red de seguimiento para un perfil de usuario de CC como primera aproximación a una comprobación de la comunidad entera. Se obtuvieron resultados que parecían apuntar hacia la aceptación de la teoría de que el seguimiento influye en los retweets y en las menciones siendo mayor el número de estas interacciones si el usuario sigue a la cuenta objetivo, excepto para las respuestas a tweets, que ocurrían en mayor medida cuando no se seguía a la cuenta objetivo.

Tras finalizar todos los estudios, se pudo comprobar que las técnicas desarrolladas permitían comprender las dimensiones de las discusiones en Twitter. Gracias a las técnicas de análisis de hashtags, el filtrado de tweets, el análisis de idiomas y las comprobaciones de términos más frecuentes y términos más importantes se puede conocer al completo el contenido de los tweets. Además, empleando el *topic modelling* podemos extraer los temas de discusión y conocer sus similitudes y palabras más importantes. Todas estas técnicas aportaban información sobre el análisis de contenido. El análisis estructural de las redes ofrecía la comprensión de las distintas redes formadas en el conjunto de tweets, pudiendo ver cómo se distribuyen las comunidades, cómo se interconectan los usuarios con las menciones y los retweets, así como conocer los usuarios de mayor importancia basándonos en su actividad e interacciones, extrayendo las medidas de centralidad de las redes.

Discusión

La finalidad última de esta tesis era la inclusión de esta línea de investigación en una plataforma que permitiera realizar análisis de redes sociales sobre conjuntos de tweets. La creación de esta plataforma surge como respuesta a dos de los principales problemas que surgen del análisis de redes sociales: la falta de una línea de investigación estándar y aplicación de varias técnicas juntas, lo cual limita la comprensión de los tweets analizados, estudiando pequeñas dimensiones. Además, también se quería solucionar la falta de replicabilidad, ya que la mayoría de los estudios se centran en casos de estudio pequeños y no es fácil replicar los análisis siguiendo sus líneas de investigación. Los análisis realizados en esta tesis han servido para obtener conocimiento de todas las dimensiones principales cuando se realizan análisis de datos provenientes de redes sociales, y al estar integrados en una plataforma, la replicabilidad de los estudios es completa, solo es necesario volver a cargar los tweets con nuevos datos o cargar nuevos conjuntos de datos. También es importante remarcar que esta plataforma permite el acceso a estas investigaciones a cualquier interesado que no tenga conocimiento técnico, ampliando el rango de aplicación de estos estudios.

Conclusiones

Esta tesis ha generado dos publicaciones internacionales indexadas en JCR, una publicación internacional sin indexación JCR, y otras dos publicaciones en conferencias internacionales.

La plataforma diseñada ofrece como principales contribuciones, una línea de investigación que permite el conocimiento completo de los tweets analizados, una mayor accesibilidad a estos tipos de estudios, una replicabilidad completa de los estudios realizados para permitir comprobaciones en el tiempo o comparativas entre datos, así como la posibilidad de ir implementando mejoras y nuevas técnicas para nutrir aún más la línea de investigación.

Presenta algunas limitaciones, principalmente relacionadas con el uso de inteligencia artificial, pues ralentiza la carga de la plataforma. Como perspectivas futuras planteamos la mejora del servidor donde se aloja la plataforma para mejorar los tiempos de procesamiento, mejorar la clasificación de usuarios para permitir una mejor anonimización y poner pública la plataforma para que la puedan usar todos los usuarios.

Summary

In the last decades there has been an increment in the number of projects related to Citizen Science. Besides growing in number, their importance in relation to research and society has also augmented. Citizen Science can be defined as the participation of non-technical individuals in research processes under the supervision of professionals. Citizen Science has no clear origin in history, although there are many examples of individuals conducting research in past centuries. This discipline started to grow in the final years of the XX century, and since then it has not stopped growing. Citizen science initiatives are commonly seen promoted from institutions and projects, and even high stages of governance such as the European Union.

Although the Citizen Science has been traditionally linked to natural sciences, we find citizen science projects in many different areas such as health, social sciences, or education. One of the main characteristics of the Citizen Science that engages participants is the possibility of learning while being part of the projects while performing appealing activities. Lately, Citizen Science has been linked to the sustainable development goals (SDGs), due to the contribution this discipline can make to achieving the goals or educating about them.

New technologies have been a fostering element for Citizen Science. The possibility of searching on the Internet for projects in which participate, obtaining data, sharing information, or participating online are some examples of what is possible nowadays. Among all these technologies, social media is key element, as it is considered the perfect environment for information sharing and collaboration.

Among all the existing social networks, Twitter has been highlighted for its contribution. In this social networking platform any user, individual or organization, can share information, interact with other users, and invite to participate in existing projects. All the information that is shared can be analysed to gain knowledge about the behaviour of the communities, what is known as social network analysis. Users related to CS are found in Twitter communicating their discoveries or data, sharing opinions, or trying to engage participants in their projects.

Social network analysis is a discipline that was born more than 70 years ago, and, with the arrival of social media, it has grown in the past decades to investigate about the users, their content, and dynamics. One of the most used social media platforms in Citizen

Science is Twitter. It has some specific attributes such as the follow relation, retweets, likes, mentions and the content in the texts such as the hashtags.

In this framework we started to develop this thesis, with the aim to analyse the content present in social media about Citizen Science, and more specifically in Twitter. Our approach was to design a pipeline for analysis to completely understand the state of the discussion about CS in Twitter, and that could serve as a standardized set of analyses for social network analysis owing to their completeness regarding the techniques and how they complement each other. Besides, we wanted to approach one of the main issues in social network analysis, the lack of replicability of the studies that are performed. To do so, we decided to create a dashboard for visualization containing all the different techniques from the pipeline for analysis. These techniques were: content analysis checking the hashtags, most common words, TF-IDF, which stands for Term Frequency-Inverse Document frequency and calculates the importance of the words in the texts, and machine learning techniques such as topic modelling. Besides, we wanted to complement the content analysis with traditional structural analysis investigating the networks that the users form via the different interactions, retweets, mentions and hashtags in the same tweet.

To test the usefulness of our pipeline for analysis, we performed several case studies about topics closely related and of importance to Citizen Science. The first conducted analysis was about learning inside Citizen Science, in which we used our first iteration over the pipeline for analysis. We obtained satisfactory results although they evidenced that we needed to improve some techniques. In the second study about the SDGs and Citizen Science, we applied new techniques and improvements to the previously used. Our pipeline for analysis brought useful and consistent insights about this conversation, so we decided to keep improving the pipeline and add new techniques to complement what we had in a next study about health and healthcare in Citizen Science. Our last study consisted of research about the possible effects of static relations (follow) on dynamic relations (retweets, likes and mentions). The results seem to show evidence that the static relations affect the dynamic ones, unveiling a new aspect of the behaviour of the community.

With all the different techniques tested, we developed the platform of visualization. This platform was designed to contain all the different techniques we used in the studies, which would turn useful for a wide range of users. The conclusions that we draw from this thesis is that our platform fulfilled all the necessities we established for SNA. This platform can turn useful for a wide range of individuals. Non-technical individuals could use the platform to obtain knowledge about the Citizen Science community and researchers could use it to replicate the analyses we designed. Projects and institutions could see it as a useful tool to monitor their engagement and the interests of the community. Finally, policy makers could benefit from, such platform when in the decision-making process about policies, doing it based on evidence.

Table of Contents

1. Introduction ... 1

1.1. Motivation ... 1

1.2. Goals and proposal ... 6

1.3. Document structure ... 9

2. Related work ... 11

2.1 CS: History, present and future with the sustainable development ... 11

2.3 Technologies involved in the growth of CS ... 21

2.4 Social Network Analysis and CS ... 24

3. The SNA techniques and the implementation of the CS dashboard ... 29

3.1 The harvesting of tweets ... 29

3.2 The design of the pipeline for analysis ... 30

3.2.1 Content analysis ... 30

3.2.2 Structural analysis ... 34

3.3 The creation of the dashboard ... 38

4. Case Studies ... 51

4.1 Open Learning in Citizen Science ... 51

4.1.1. Motivation ... 51

4.1.2 Methodology ... 52

4.1.3 Results ... 53

4.1.4 Discussion ... 56

4.1.5 Conclusions ... 57

4.2 Understanding the discussion of CS around SDGs ... 58

4.2.1 Motivation ... 58

4.2.2 Methodology ... 59

4.2.3 Results ... 62

4.2.4 Discussion ... 80

4.2.5 Conclusions ... 82

4.3 An analytics approach to health and healthcare in CS communications on Twitter ... 82

4.3.1 Motivation ... 82

4.3.2 Methodology ... 83

4.3.3 Results ... 86

4.3.4 Discussion 102
4.3.5 Conclusions 104
4.4 Static interactions and their influence on dynamic interactions between Twitter accounts in a CS context 104
4.4.1 Motivation 104
4.4.2 Methodology 105
4.4.3 Results 108
4.4.4 Discussion 111
4.4.5 Conclusions 111
4.5 Conclusions from all case studies 112
5. Conclusions and future work 115
5.1 Contributions 115
5.2 Advantages and limitations 119
5.3 Current and future work 120
5.4 Publications 121
5.4.1 Journals 121
5.4.2 International Conferences 121
5.4.3 GitHub repositories 122
Acknowledgements 122
Ethical approval 122
References 123

Figure index

Figure 1: Number of publications catalogued as about CS (Álvarez, 2020) 17
Figure 2: The 17 Sustainable Development Goals in the Agenda 2030 19
Figure 3: Representation of coreness 37
Figure 4: Dash architecture 40
Figure 6: Main menu with the different analyses available 41
Figure 7: Filtering menu 42
Figure 8: Example of the visualization of all hashtags in a case study of SDGs 43
Figure 9: Example of the visualization for Rt hashtags in a case study of SDGs 43
Figure 10: Example of all hash time series 44
Figure 11: Example of RT hash time series 44
Figure 12: Wordcloud showing the most common terms in subset of example 45
Figure 13: Tables. RT example with anonymized username 45
Figure 14: Example of eable of degrees 46
Figure 15: Results of the sentiment analysis from a subset as example 46
Figure 16: Example of network of retweets 47
Figure 17: Example of communities 48
Figure 18: Visualization for activity per country 49
Figure 19: Location of the users from the CS community. 49
Figure 19: First study - Open learning - Results from the hashtag analysis 54
Figure 20: First study - Open learning - Network of retweets from the learning dimension 55
Figure 21: Second study – SDGs - Top 10 used hashtags in tweets 63
Figure 22: Second study – SDGs - Top 10 retweeted hashtags 64
Figure 23: Second study – SDGs - Top 10 used hashtags outside the retweets 65
Figure 24: Second study – SDGs - Temporal evolution of top 10 most used hashtags 65
Figure 25: Second study – SDGs - Temporal evolution of top 10 used hashtags outside the retweets 66
Figure 26: Second study – SDGs - Temporal evolution of top 10 most retweeted hashtags 67
Figure 27: Second study – SDGs - Example of wordcloud with the most used terms 67
Figure 28: Second study – SDGs - Example of TF-IDF analysis 68
Figure 29: Second study – SDGs - Topic coherence 69
Figure 30: Second study – SDGs - Topic modelling intertopic distance map with the frequency of the terms 70
Figure 31: Second study – SDGs - Tweets per topic 73
Figure 32: Second study – SDGs - Classifier confusion matrix 74
Figure 33: Second study – SDGs - Tweets assigned to each SDG 75
Figure 34: Second study – SDGs - Two-mode graph with SDGs by colour 76
Figure 35: Second study – SDGs - Followers vs following 77
Figure 36: Second study – SDGs - Network of retweets 78
Figure 37: Second study – SDGs - Scatterplot of indegree vs outdegree 80
Figure 38: Third study – Health and Healthcare - Workflow which is representative of the pipeline for analysis 84
Figure 39: Third study – Health and Healthcare - Most used hashtags in the complete dataset . 87
Figure 40: Third study – Health and Healthcare - Top 50 hashtags in the filtered dataset 88
Figure 41: Third study – Health and Healthcare - Top 10 most retweeted hashtags 89
Figure 42: Third study – Health and Healthcare - Temporal evolution of retweeted hashtags ... 89

Figure 43: Third study – Health and Healthcare - Temporal evolution of most used hashtags... 90
Figure 44: Third study – Health and Healthcare - Network of interconnected hashtags 91
Figure 45: Third study – Health and Healthcare - Intertopic distance map from the topic modelling.. 92
Figure 46: Third study – Health and Healthcare - Topic hierarchy... 96
Figure 47: Third study – Health and Healthcare - Topics overtime 97
Figure 48: Third study – Health and Healthcare - Top 10 users with more followers 97
Figure 49: Third study – Health and Healthcare - Top 10 accounts with more impact............... 98
Figure 50: Third study – Health and Healthcare - Network of mentions 99
Figure 51: Third study – Health and Healthcare - Network of retweets with communities in different colours calculated with Louvain method. 100
Figure 52: Structure of the network containing different interactions as edges on the left. On the right, schedule of flow of information... 107
Figure 53: Fourth study- Static vs Dynamic Interactions-Mean value for follow and the interactions by user type.. 110

Table index

Table 1: Different definition for CS and their key aspects, which differ between entities.......... 15
Table 2: First study- Open learning - Centrality measures of the most retweeted accounts 56
Table 3: Second study – SDGs - SDGs Topics, keywords, and definition. Some topics contain keywords like userX, this is due to the anonymization of individuals....................................... 72
Table 4: Second study – SDGs - Performance of the BERT classifier....................................... 75
Table 5: Second study – SDGs - Top 10 ranked by Indegree.. 79
Table 6: Second study – SDGs - Top 10 users ranked by Outdegree... 79
Table 7: Third study – Health and Healthcare - Topic, Keyword and Explanation of the topics from e-Health 93
Table 8: Third study – Health and Healthcare - Topics and keywords....................................... 95
Table 9: Third study – Health and Healthcare - Top 20 users ranked by Indegree 101
Table 10: Third study – Health and Healthcare - Top 20 users ranked by Outdegree............... 102
Table 11: Fourth study- Static vs Dynamic Interactions-Distribution of interactions by user type 109
Table 12: Fourth Study-Static vs Dynamic Interactions-Breakdown of type of interaction by user type and follow relation.. 110

1. Introduction

This chapter presents the motivation that guided us to develop the work presented in the thesis. This motivation settles the groundings and briefly introduces the background. Later we also present the objectives we aimed for alongside the future perspectives. The structure of the document is also detailed at the end of the chapter.

1.1. Motivation

Citizen Science (CS) is typically described as the participation of non-scientist individuals in scientific research via different processes (Vohland et al., 2021). In the past decades the CS phenomenon has grown inside the scientific community. The potential of this activity has been demonstrated when used in large scale projects, providing valuable publications and data, results, and conclusions with further research.

CS community is formed of scientists, citizen scientists, organizations, universities, and some other institutions. There are many definitions for CS, but it is often considered as the participation of non-scientists people in scientific projects, activities, or processes. In CS, several profiles of people gather to collect, transcribe, or analyse data, what some authors consider to be something beneficial (Hecker et al., 2018). These activities have significantly increased in the last decades, especially due to the Internet and web-based technology (Aristeidou & Herodotou, 2020). According to different studies, the use of these technologies improves communication and collaboration (Johnston et al., 2017; Newman et al., 2012a), and the social networking sites promote these actions in general and specifically for the CS community too (Hansen et al., 2011), but not only in social networking sites, we find CS initiative all across the Internet, which is a valuable data and information source.

Since the invention of the Internet, the flow of information, its creation, acquisition, and sharing has been quickly evolving. Access to and use of the Internet has become essential in some aspects of life, and connectivity has grown exponentially, and therefore the presence of data has increased alongside the people´s activity on the Internet.

The Internet (i.e., the connection between several computers using packet switching technology) dates to the 60s, when the ARPANET was created (Leiner et al., 2009). In the 90s, the Internet was primarily used to send emails, to transfer files and use limited software (Martínez-Domínguez & Fierros-González, 2022). Since then, the Internet experienced an exponential increase until 2001 and afterwards it started growing linearly, meaning that the number of users and the connections keep on increasing at a reduced rate, but still increasing (Dhamdhere & Dovrolis, 2008). One of the earliest statistical reports shows that half of all households have an Internet connection in the EU (Van Dijk, 2009). Later, in 2011 the number of devices connected to the Internet was higher than the number of humans that use them (Settembre, 2012), being the rise of the smartphones a crucial factor. According to Eurostat[1], in 2022, 93% of the European households and the global population had access to the Internet. Nowadays, in 2023, Internet allows multiple actions, e.g., blogging, picture sharing, reading news, and so on. Most of the simplest aspects of life are undertaken on the Internet and a 65.7% of the global population has access to the Internet[2].

The evolution of smartphones and other smart devices has also increased the connectivity and therefore the sharing, generating and acquisition of information. The number of smartphones in 2019 was estimated to be over 4 billion (Sahlström et al., 2019) and the number of smart devices in general to be over 41 billion by 2023 (Rahmani et al., 2023). The arrival of smartphones completely changed our life paradigm, turning these devices into a fundamental element in every day's life. Several studies have measured the time we spent using our mobile devices in the last years and it has experienced exponential growth (Sapacz et al., 2016). An example is the study of use of digital devices by people aged 15-74 in the EU during their working time[3]. The results show that 30% of employed people in EU use digital devices while working. Among all the factors that led to this situation, the rise of the social networking sites is especially remarkable, in fact we will go over this specific factor later in this document.

This high amount of people using Internet translates into the creation of a big amount of data, which can be used for different purposes including research (Birnbaum, 2004; M. Chen et al., 2020; Gosling & Mason, 2015). The information extracted from Internet is useful for research (Griffin et al., 2022), however this information lacks a structure, and this hinders the analyses. In fact, several studies have stated that the research conducted using data from the Internet could be improved, it still needs standardization and

[1] https://shorturl.at/pyAIT
[2] https://shorturl.at/BELV9
[3] https://shorturl.at/pDRU7

sometimes it is questioned by statisticians (Lefever et al., 2007), it has proven itself as useful and it has shed light on many aspects. Internet data has been successfully used in political polls to predict tendencies (Hargittai & Karaoglu, 2018), economic research (Edelman, 2012), geospatial analyses (Zook & Graham, 2007), demographic studies (Zagheni & Weber, 2015), social sciences (Askitas & Zimmermann, 2015), among others. The usefulness of Internet data has been already addressed, web-based analytics, spatial data, Internet surveys, etc. have proved themselves as valuable techniques in research. Research using online data comes with limitations (obligatory necessity to be connected to the Internet, lack of ability to perform this type of research, or problems regarding the non-random nature of populations used when collecting data as some examples), but allows access to global data being cost and time effective for researchers (Lefever et al., 2007).

There are some specific types of application in which people engage and share plenty of information of a diverse range. These are the social networks. The Cambridge dictionary defines a social network as a website or computer program that allows people to communicate and share information on the Internet using a computer or a mobile phone (Horst & Miller, 2020). Social networks have evolved from direct exchange of information to virtual gathering points that have become essential to business, organizations, public services, individuals, and governments.

Some people trace social networks back to the pre-Internet era with the invention of the telegraph machine (Rosenwald, 2017), others refer to the first emailing services (Kahlon et al., 2014) but most agree that the first social networks appeared in the late 90s and early 2000s. In those mentioned years, sites like Six Degrees, Friendster, Myspace, or LinkedIn appeared. By 2008 Facebook overtook all these websites, which was launched in 2004, becoming the most popular social network and introducing the dynamics of these websites for its future competitors.

In the following years, other social networks were created such as Reddit, Google+, Instagram or Twitter. Most recently Tik Tok emerges as the most used social network nowadays (Montag et al., 2021). By 2021, 4,26 billion people were social network users worldwide, projected to be increased to six billion by 2027 (Statista, s. f.) . The average user spends 144 minutes per day using social media and, although the most common use it is as a personal blog, there is plenty of information that is useful for scientific research.

In the last decade, the use of social network data has established as a customary practice due to the amount of data available in them. Social networks are accountable for 13% of the data and information shared daily on the Internet, an amount of data far from negligible. Studies using data from the social networks have been conducted all over the world in multiple fields. Besides, the analysis of the social networks is a common practice in business to retrieve information from trends, preferences or opinions from the users and make the products and services evolve and adapt to the customer´s needs.

Among all the available social networking sites, we focus specifically on Twitter, which is now called X. Twitter is a social network platform and microblogging service intended to share messages of less than 280 characters of any kind. During the time the studies of this thesis were developed, it was still called Twitter, so we will address this platform with its former name. It was proposed to be a short message service that debuted in July 2006. The emerge of this social network occurred between 2008 and 2009, during the USA presidential elections in which Barack Obama showed a dynamic activity in social networks and settling the behaviours of politicians in the next years in relation to social networking presence.

Twitter established itself as one of the most used social networks in the world (Burgess & Baym, 2022) and it is a powerful tool for information dissemination. Twitter users' share news about multiple topics such as science, economics, art, fashion, celebrities, and many others.

All the information shared daily on Twitter has proven valuable in scientific research. Numerous studies have used tweets (i.e., the messages written by the accounts in Twitter) to gain insight in multitudinous topics. Researchers have been able to access to these tweets by means of the Twitter API, a programming interface that allows users to communicate with the original software, and to download information from this platform.

There have been studies involving crisis management, the COVID pandemics (Xue et al., 2020), analysis of opinion (Bruzzese et al., 2022), climate change (Moernaut et al., 2022), sociology (Peters et al., 2022), fake news (Bodaghi & Oliveira, 2022), and others. Useful outcomes are obtained for either individuals, organizations, or political entities (Huszár et al., 2022; Wang et al., 2021; Wang & Yang, 2020). However, one important aspect is how restrictive performing these analyses could be for the general public, who may have no technical abilities to do so (Ebrahim, 2020).

In Twitter, participants from the CS community share results and create communication channels (Mazumdar & Thakker, 2020) and projects tend to establish their presence in the platform to disseminate their activities and outcomes (Wiggins & Crowston, 2011). Monitoring the activities of these profiles could turn crucial for government policies, investment and research and development (X. Li et al., 2022), especially considering the topics that cover the discussion about CS in Twitter and the fields and research areas that CS projects address (Preece, 2016). CS projects have been found to measure and analyse about sustainability, health, healthcare, biodiversity, medicine, public education and many others (Bonney et al., 2009).

The existing studies about the CS community in Twitter set the groundings with discoveries such as the fact that the CS community most common activity is retweeting, the proposition of tools to increase the visibility of CS projects in Twitter and some basic analysis of how the communities behave (Mazumdar & Thakker, 2020). All this motivated us to deepen in this community to gain full knowledge of the dynamics of the CS community in Twitter. Due to the growing importance of CS community, the European Union has funded several projects to investigate its behaviour and get a better knowledge of CS. One of these projects is the CSTrack project[4], explained in the next section, that has focused on analysing all the available information in Internet about CS to broaden the knowledge about CS using quantitative and qualitative methods (De-Groot et al., 2022).

Besides, we wanted to address one of the most crucial problems of the Twitter analysis, the lack of standardization and combination of methods (Liang & Fu, 2015). It is therefore why we consider necessary to palliate and to offer a solution to the previous challenge.

Conducting social network analysis (SNA from now on) for certain people could turn difficult, since this calls for a certain level of expertise in programming and API usage. Furthermore, most of the current SNA using Twitter data show mono-technique analysis with a problem in reproducibility (Liang & Fu, 2015). This was the motivation to develop the research done in this thesis. We will present how we applied the existing techniques for SNA using a platform in which non-technical individuals could apply these techniques to social networks for gaining knowledge from the discussion in Twitter and how we did studies to analyse different key topics inside the CS community. We will present a dashboard for conduct an analysis of Twitter data. This platform provides a standardized

[4] https://cordis.europa.eu/project/id/872522

pipeline for this kind of research, collects all the different technologies applied to SNA and offers easy and free access to SNA, and finally deepen in the dynamics and content from the CS community in Twitter.

1.2. Goals and proposal

The present thesis has three goals:

- G1: Gaining knowledge and deepen in the content created and shared by the CS community.
- G2: To provide a standardized pipeline for SNA, useful for multiple topics.
- G3: To provide a comprehensive and easy-to-use platform to perform SNA despite the background of the user, a platform that reunites all the techniques and that could be used in the future in more diverse topics too.

The idea of bringing together the different techniques that are traditionally applied to this field, alongside newer techniques including machine learning, sentiment analysis or new graph analysis approaches arose as a solution to the lack of a standard pipeline for analysis. Several studies, as it was reported before, present results obtained by applying single techniques that shed light on specific aspects, but not the whole panorama of the discussion. By applying a combination of techniques, we were able to better understand the discussion around different topics. Applying traditional methodologies such as hashtag analysis, most frequent terms count, analysis of retweets and exploratory analysis of the graphs formed by the users we get to know the dynamics inside the CS community. Adding more modern techniques such as machine learning for topic analysis and analysis of language distribution, sentiment analysis, geographical analysis, or mixed graphs we can understand underlying aspects of the community. All these together provide more complete knowledge of what is going on inside CS in Twitter.

This thesis was developed in the framework of the CSTrack project. This project was born with the aim to expand the knowledge about CS, not as a participant but through the eyes of an external observer. The project was funded by the European Commission under the Horizon 2020/SwafS program[5]. It had 9 partners from different countries with different expertise in fields such as computer science, data analysis, educational studies, or social studies. This collaboration was expected to strengthen the findings and to provide a

[5] https://cordis.europa.eu/project/id/872522

comprehensive outcome comprising all different points of view. Different methods were used in the project: literature review, web analytics, surveys, and analysis of discourse in social media. To gain knowledge, the analysis was divided into different levels of analytics, as each one will help to identify specific characteristics:

- Micro level: this was based on specific cases. In the micro level the CSTrack project analysed the communications in project web forums, like the "Chimp & See" project in Zooniverse. This helped to obtain detailed information from the participation and role-taking.
- Meso level: the CSTrack project created a database containing 4,500 CS projects from 56 CS global platforms using web crawling techniques. This database was used to apply, for example, semantic analyses or research areas identification.
- Macro level: this level was destined to analyse the discourse inside social media, specifically Twitter. Information from the CS community inside Twitter was harvested and used to detect trends, connections, actors, and other interactions. This level combined content analysis with structural network analysis.

Several case studies were analysed in each different level. For example, at micro level the CSTrack project performed an analysis about COVID-19. The COVID-19 has been a challenging event for scientists, companies, and individuals as it was necessary to identify medical solutions and health implications. This micro level analysis studied 25 CS projects. The results showed that the projects addressed three main issues: the tracking of the spread of the pandemic in the population, investigating the influence of COVID-19 on people's health, and the investigations of the virus' biology. The tasks performed by citizen scientists were answering online surveys, tracking self-data with wearables, and distributed computing. All projects, according to their description and related information, were accessible, targeted a wide audience and required no special skills to participate in them.

At the meso level, one example was the identification of research areas for the projects from the database present in Zooniverse, a number of 218 projects. The first thing to take into consideration was that most of the projects could be categorised as multi-disciplinary, but for the identification this was neglected. To assign the research areas, ESA (Explicit Semantic Analysis) approach was used, which combines statistical models with semantic information (in this case extracted from Wikipedia). The results shows that 67.4% of the projects were labelled as belonging to more than one research areas (as expected), so the multi or inter-disciplinarity in CS is a prevailing characteristic.

Finally, it is the macro level, in which this thesis is framed. One example of this macro level analysis was the sentiment analysis performed in tweets about climate change from the CS community and outside the CS community. The initial approach in this macro level was to develop different SNA techniques and develop a complete content analysis to be applied to specific cases, following the main trend in CS analysis (Cox et al., 2015; Simpson et al., 2014). Once we deepen in the different techniques available, we shifted our approach into the development of a pipeline for analysis combining SNA structural analysis with content analysis that could be utilized in every situation. The analysis comprised hashtag analysis, most common words count, TF-IDF analysis combined with machine learning approaches for topic detection and sentiment analysis, techniques that allowed us to identify trends or important issues for the community. The SNA was performed using the different interactions present in Twitter, which helped us to discover important actors in the CS community and their connections.

Then, we thought about developing an analytics dashboard in which researchers, policy makers, participants, and many other stakeholders could access and analyses this information. These improvements seek the replicability of the analyses and allow non-technical profiles access to this information. Once the dashboard was developed, we thought about our research questions to validate this approach:

- RQ1: Is there a multi-lingual and multi-topical conversation in the CS community?
- RQ2: What is the lexical inventory in these conversations, i.e., hashtags, most common words, etc.?
- RQ3: What are the main topics? Are these topics evolving in time?
- RQ4: Can we define main actors through the SNA structural analysis?

Answering these research questions is the goal of this thesis. By doing so, we expected to obtain a useful tool for a wide range of users that could help researchers to obtain quick knowledge, policy makers to identify trends and important topics, newcomers to CS to quickly engage into the community or project managers to follow up their evolution in Twitter, as some examples.

1.3. Document structure

This document is divided into 5 chapters. The content addressed in each chapter is as follows:

- **Chapter 1:** In the present chapter we expose the framework of this study, alongside the motivation behind it and the goals set to be achieved in the thesis.
- **Chapter 2:** Related research works will be explained, detailing the ideas behind the study and the most commonly used techniques in SNA.
- **Chapter 3:** This chapter contains a detailed description of the techniques we selected, and technologies applied in the research. From the harvesting of tweets to the application of the algorithms or creation and analysis of networks, a detailed review of the processes is presented. Also, the creation of the dashboard is presented including its architecture, list of techniques available, functionality, and examples of the results.
- **Chapter 4:** The different case studies performed to validate the pipeline for analysis and the techniques implemented in the dashboard are presented in this chapter. The first study analyses the state of the discussion about learning inside the CS community. The second study is focused on the analysis of the discussion about SDGs inside the CS community using at the first-time machine learning techniques. The third study analyses the discussion about health, e-health, and healthcare. The fourth study details the creation and analysis of mixed networks using static and dynamic links between users. Finally, the general conclusions from the different studies and a detailed review of the findings can be found.
- **Chapter 5:** In this chapter, the conclusions extracted from our work are presented as a summary of all the outcomes from the different analyses and studies performed alongside the analysis of the dashboard. Future perspectives are also included.

2. Related work

CS phenomenon has grown inside the scientific community. In a globalized world, in which everyone is connected, web-based platforms have fostered the collaboration, information sharing and knowledge acquisition, and CS has also been benefited by this. Social media platforms that have arisen since the invention of the Internet have turned into an idoneous place for CS communities. Although our knowledge of these communities, their content and dynamics is limited, and has not been deeply analysed.

This chapter will cover how the evolution of science brought CS and the history of this discipline, alongside different examples in ancient history that resemble CS. Besides, it will introduce the growth and change in the paradigm that has happened in CS thanks to different technologies, especially social media platforms, and the efforts done to analyse and comprehend these communities and, more important, the complete landscape around the conversation occurring inside this community.

2.1 CS: History, present and future with the sustainable development

It would be impossible to understand CS without science. It is an arduous task to track the origin of science with precision, although there are many examples of it that trace it back to past times of human existence.

Some researchers point that in classical antiquity we find examples of science such as the development of astronomy and mathematics or a glimpse of modern life sciences. All this development occurred owing to the observation of the world by those people with intellectual inquisitiveness. Although the modern scientific method did not exist, and of course neither did modern tools and methods. Some others state that it happened during the Renaissance, since many fields experienced growth due to the encouragement to research that this cultural period brought, alongside the human necessity of answering the vital questions. Thus, the start of Modern Science can sometimes be tracked back to the 1600s. In these early stages, what we call a “scientist” was not properly established yet, regular citizens dedicate their lives to experimentation started to create the figure of scientists. These individuals are nowadays believed to have been a sort of citizen scientist themselves. It is in the following years when the first scientific societies start to appear and, most importantly, the scientific method.

But modern science is said to have started at the beginning of last century, leading to great discoveries in different fields such as X-rays, radioactivity, the development of quantum physics, and many others. Many new research fields were created, and others evolved, and some specific event was what brought this: the revolution of the creation of computers.

Computers increased the amount of data that scientists handle and democratized the access to information and, therefore, science. Non-technical individuals got access to information fostering the collaboration not only between institutions (thanks to inventions like the Internet) but with external collaborators that many times are citizens. Collaborative science is not, of course, new, but its growth after World War II is noticeable. The eagerness to participate, the encouragement from institutions to make people participate led to the establishment of what we now know as CS.

CS is typically described as the participation of non-scientist individuals in scientific research via different processes (Vohland et al., 2021). Citizens offer their work force to conduct scientific research donating their computers to make calculations, annotating animal observations in the wild, classifying images, sharing air pollution information they measure (Strasser et al., 2018) , among others. However, there are many definitions for this practice (Vohland et al., 2021) and its origin is diffused. Some experts in CS differ in opinions about what it is CS and what it is not (Strasser et al., 2018), which will be later discussed when introducing the different definitions offered nowadays. Now, we will focus on its history and how in the last decades there has been a significant increase in the number of people participating in CS. For this growth, there are several factors that could have helped, which will be also discussed.

To locate CS in History, we have already addressed how in past times citizens got involved in science owing to their economical possibilities although this was not technically CS. There are examples of collaborative science in the past, but CS as we know it nowadays was still not “a thing”. One of the most antique examples dates back to old Japanese collaborative projects 1,200 years ago, like the recording of the cherry blossom (Kobori et al., 2016). Another specific group of citizens from the past that deserve to be highlighted are the Amateur Naturalists from the eighteenth and nineteenth century. For example, the Nature’s Calendar (started in 1797) (Amano et al., 2010), in the UK, is a good example of something similar to CS at the end of the eighteenth century. These people, that called themselves “men of science” or “Naturforscher” (White, 2016), were mainly unpaid and dedicated to the observation and annotation of natural phenomena, animals, etc. but like a “hobby”. Thus, although present citizen scientist also have this practice as a hobby, these ancient examples differ to the current interpretation since there were no interaction between the citizens and professional scientists (Haklay, 2013) (there were no professional scientists back then). During the nineteenth century these individuals were undertaking continued (Opitz et al., 2016) and it was after Darwin when the professionalization of science began (L’Hermite-Leclercq, 1987). One good example of CS alike project from this period is the National Weather Service Cooperative Observer Program of the USA, started in 1890 (Havens & Henderson, 2013).

By the end of this century there were many people fully dedicated to science as a profession (L’Hermite-Leclercq, 1987; White, 2016) which evolved to the scientific profession we know nowadays. After the professionalization of this occupation, the collaboration and contribution to professional scientist from individuals continued in fields such as animal taxonomy, astronomy, botanic, geology, among others (Ayres, 2008; Secord, 1994; Strasser, 2012). Some authors state that in experimental sciences the contribution from “amateurs” was limited during this period, since the non-experienced individuals were not allowed to be part of this research (Coleman, 1977; Cunningham & Williams, 2002; Shapin, 2017). But, this situation has been reversed in the present times. Technologies have allowed citizens to be involved in processes to a certain point. However, we must not forget the limitations to contribute to empirical studies in laboratories, where the expertise in lab work is a must to be part of research.

Another group that must be pointed up are those people that participated (and continue to do it) in producing knowledge about environmental issues and biology monitoring air pollution, harmful molecules in food and air, measuring rainfall amount and many other activities (Fleming & Johnson, 2014; Wittner, 2009) or in the late twentieth century many contributions to health issues discussing and monitoring the emergence of diseases, pollutants in the environment or helping during the AIDS crisis (Brown, 1997, 2007). Furthermore, an interesting scenario that started to occur was the immeasurable contribution from women specifically that expanded the biomedical knowledge about women-related health via self-examination and annotation (Cornejo & Denman, 2005; Parry, 2016; Strasser et al., 2018). Also, in the twentieth century we find what is commonly referred as the first modern CS project, The Common Birds Census (1962) (Harris et al., 2015), but events like this were yet not considered Citizen Science, or at least was not categorized as CS, because despite all the examples the origin of this practice is yet not clearly established. Many of these actions happened without the supervision of a specialized institution, without scientific validation or are simply not categorized as CS since the term did not exist but gathered some of the characteristics to be considered as such.

What it is known is the origin of the term CS (although discussed), which is sometimes tracked back to 1989. In this year the term "Citizen Science" appeared in the MIT Technology Review. In this review, there was an article called "Lab for the Environment" (Kerson, Raymond, 1989) that introduced different lab practices that were described as follows: community-based laboratories that explore environmental hazards, laboratory work by Greenpeace, and Audubon's recruitment of volunteers in a "citizen science" programme. This CS definition involves the generation of scientific data, a wide area to be analysed and political importance of the issue, which seems to comply with all the definitions of CS.

Another definition of this term is from 1995 (Irwin, 1995). It differs to what we understand for CS nowadays, as he described CS as a science that is useful to citizens and performed by citizens, and therefore linked to the terms "science for the people" and "science by the people". This definition and its interpretation are in question, and another interpretation is owed to Richard Bonney, who points in a different direction. He defined CS as scientific projects in which citizens get involved providing data and, as a "reward", the obtains some scientific knowledge or skills(Bonney, 1996).

The increasing practice of CS has led several organizations and institutions to also offer their personal definition for CS. These definitions differ and cover different aspects of CS, or what they understand for CS. Some of the most relevant aspects can be found in Table 1.

	Wikipedia	**Oxford Dictionary**	**EU**	**UNESCO**	**European Citizen Science Association**	**Citizen Science Association (US)**
Year	2005	2014	2019	2023	-	-
Participants	Undertaken by general public	Undertaken by general public	Citizens	Undertaken by wide range of non-scientific stakeholders	Undertaken by general public	Undertaken by general public
Type of collaboration	Basic participation (i.e., monitoring and so)	Collaboration with scientist and institutions		Collaboration with scientist and institutions	Basic participation (i.e., monitoring and so)	
Level of collaboration			All steps of scientific process/research	All steps of scientific process		All steps of scientific process
Outcome			Innovation in research		Participation driven to develop CS	
Social contribution				Inclusiveness	Inclusiveness	
Contribution to policies					Improve decision making	

Table 1: Different definition for CS and their key aspects, which differ between entities.

Many different origins, many definitions and technicalities that change from one description or interpretation to another. As it can be seen, the definitions tend to highlight the collaboration, the non-professional profile of the participants and how they participate in scientific processes with the supervision of scientific institutions. So, what seems clear is that nowadays, most of the practitioners and CS researchers generally accept the following definition: CS as engaging the public in a scientific project (Bonney et al., 2014; Shirk et al., 2012; Silvertown, 2009). In this thesis, we follow this previous definition as it contains most of the important aspects. The definition is still a subject of controversy, and it will continue to be, but what cannot be denied is the enormous contribution that CS has meant when used to collect data and information useful for scientist, policymakers, and the public (Miller-Rushing et al., 2012; Silvertown, 2009).

Another key aspect about CS is that some authors and practitioners state that CS is somehow democratizing science. CS allows everyone to be part of science, an idea that many institutions, practitioners, and other stakeholder have embraced (Strasser et al., 2018). Little research has been done about this issue, analysing the typical profile of a participant in a CS project but there are some authors that have conducted this type of surveys (Curtis, 2015; Reed et al., 2013). This idea of democratization is directly linked to actual paradigm in relation to new technologies, which make participation more accessible.

We previously stated that CS has contributed to the scientific community in numerous fields. In the last decades plenty of CS projects have emerged and others have been contributing to the scientist with their collected data and other processes. The increment in CS projects is related to the increase of publications about CS shown in Figure 1. We find CS typically linked to environmental studies and natural sciences, with great projects such as the previously mentioned Bird Census, the Cornell Lab[6] with their bird counting projects, Butterfly count[7], collect weather data[8], and so on[9]. But, although maybe the participation in this kind of projects is easier for the citizens (Fraisl et al., 2022), there are CS projects in other fields. For example, another important area that attracts citizens is

[6] https://www.birds.cornell.edu/home/

[7] https://naba.org/

[8] https://shorturl.at/ijtw5

[9] https://shorturl.at/fnOP1

astronomy. NASA has launched some CS projects: the Exoplanet Watch[10] encourages participants to watch planets outside the solar system and monitor them to learn more about them. Stardust@home[11], aimed to discover interstellar dust impacts rewarding the participants by naming a particle after them, or Backyard Worlds: Planet Nine[12], to let people watch pictures from telescopes to see if someone find something interesting such as a new planet, star, etc.

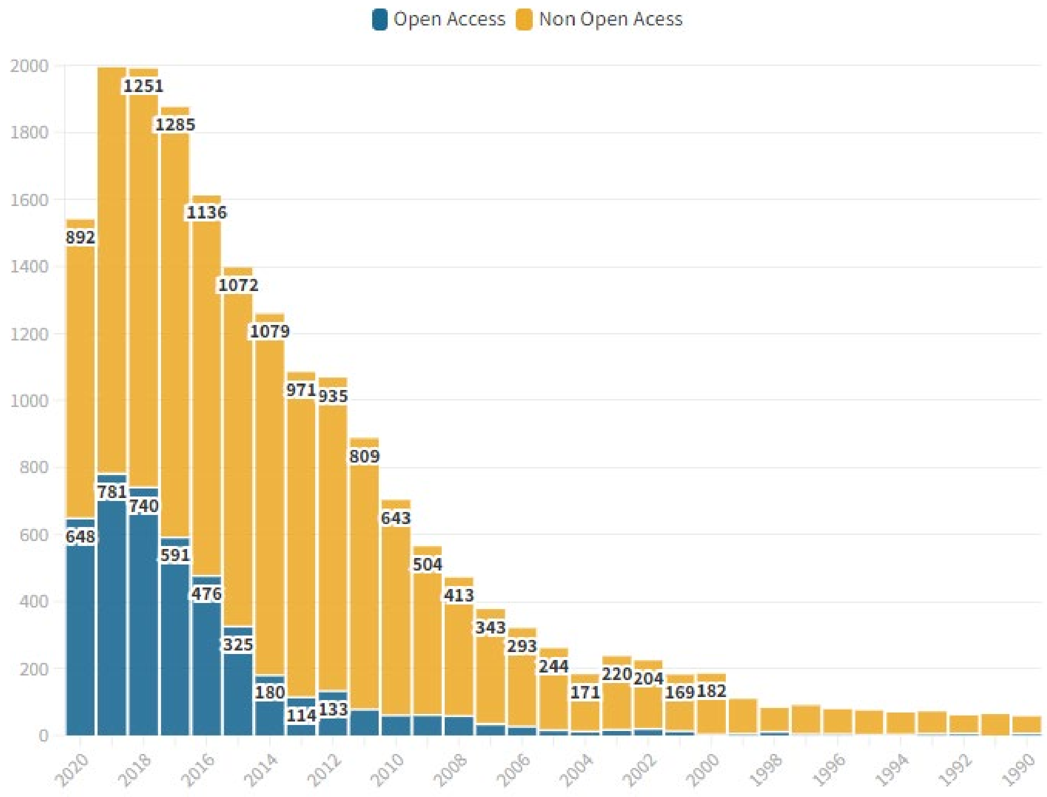


Figure 1: Number of publications catalogued as about CS (Álvarez, 2020)

We also find projects in humanities, a quite recent area of CS (Hedges & Dunn, 2018) in which participants mainly collect, transcribe, or annotate historical sources. This field has experienced growth owing to technologies such as online surveys, online data collection, transcription, artificial intelligence, digital libraries, and gamification (Dobreva, 2016; Tauginienė et al., 2020). All these technologies brought a differentiation in humanities creating the Digital Humanities, which combines computing with the disciples in humanities. Some interesting projects are Old Weather[13], in which citizens transcribe

[10] https://shorturl.at/dklB5

[11] https://shorturl.at/fuNU6

[12] https://shorturl.at/acjvL

[13] https://www.oldweather.org/

arctic and worldwide weather annotations from ships since the mid-19th century, or Transcribe Bentham[14], focused on the transcription of old and unstudied manuscripts from Jeremy Bentham, a British philosopher from the 19th century.

Another field with CS projects is health and healthcare. There are projects focused, again, on data collecting or data processing related to Parkinson disease (Boving et al., 2021), 3D mapping of retinal neurons[15], wildlife health (Lawson et al., 2015), use of antibiotics (Roberts, 2020) or analysis of communications during COVID-19 (Santana et al., 2023). Something important about CS and health is the controversies when defining a project as a CS project in health (Haklay et al., 2020) since there are many factors that can be discussed in relation to the participants' contribution. One example is, if the project can be considered CS if the participation is passive, the citizen is simply a patient or wearing a monitor (Ceccaroni et al., 2021). Another factor is the goal, if it is commercial, it does not lay under the expected outcome from CS (Ceccaroni et al., 2021). Or the organization, again if the projects is coordinated by a commercial entity, it is not done by a public organization as it is normally done in CS (Ceccaroni et al., 2021).

We can continue enumerating other areas that have CS projects (art history, oceanography, seismology …), so what it is clear is that participants can be involved in a diverse range of activities from different fields. These many projects launched by numerous institutions due to the increasing participation drives us to think that citizens find CS appealing and they are willing to participate in such activities.

The participants in CS are non-professionals, non-scientists, regular citizens involved in scientific processes that feel attracted by a specific area and start collaborating. Mainly, what attracts the participants is the possible outcome translated into learning, acquiring scientific knowledge (Kloetzer et al., 2021), and simply enjoy their hobbies while contributing to something else. The analysis of the motivations that drive participants to collaborate is mainly done via surveys and interviews (Schaefer et al., 2021) and have pointed that the main reasons are contributing to scientific research, an intrinsic interest in the topic, enjoyment, social interaction, eagerness to learn and willingness to help. All the different researchers that have conducted studies about the motivation are focused on isolated cases and a general analysis could benefit both the participants and organizers

[14] https://shorturl.at/zKW36
[15] https://eyewire.org/explore

(Vohland et al., 2021). Conducting universal studies about motivation could translate into improvement of the outcomes and benefits for participants. These outcomes, besides the previously explained learning outcome, are related to health, opportunity to socialise or empowerment (Jones, s. f.). Getting to know better the motivations and expectations will translate in better engagement, recruitment, retention, and evaluation (Land-Zandstra et al., 2021).

This growing trend in participation, aside from the motivation of the participants, unveils itself as relevant for a key issue nowadays, the Sustainable Development Goals (SDGs) proposed in the Agenda 2030. The SDGs are aimed towards the change of present policies to impulse the sustainable development, the improvement of current actions to make them viable, green and focused on every individual on Earth[16].

These SDGs are 17, interconnected between them, and each one addresses a specific issue from biodiversity to poverty or gender equality. The Agenda 2023 encourages the participation and collaboration of citizens, institutions, organizations, governments in order to shape a better future and these words resemble what CS is aimed for. In fact, some studies have focused on how the SDGs can build on CS and how CS can contribute to the SDGS (Shulla et al., 2020). Figure 2 shows the SDGs included in the Agenda 2030.

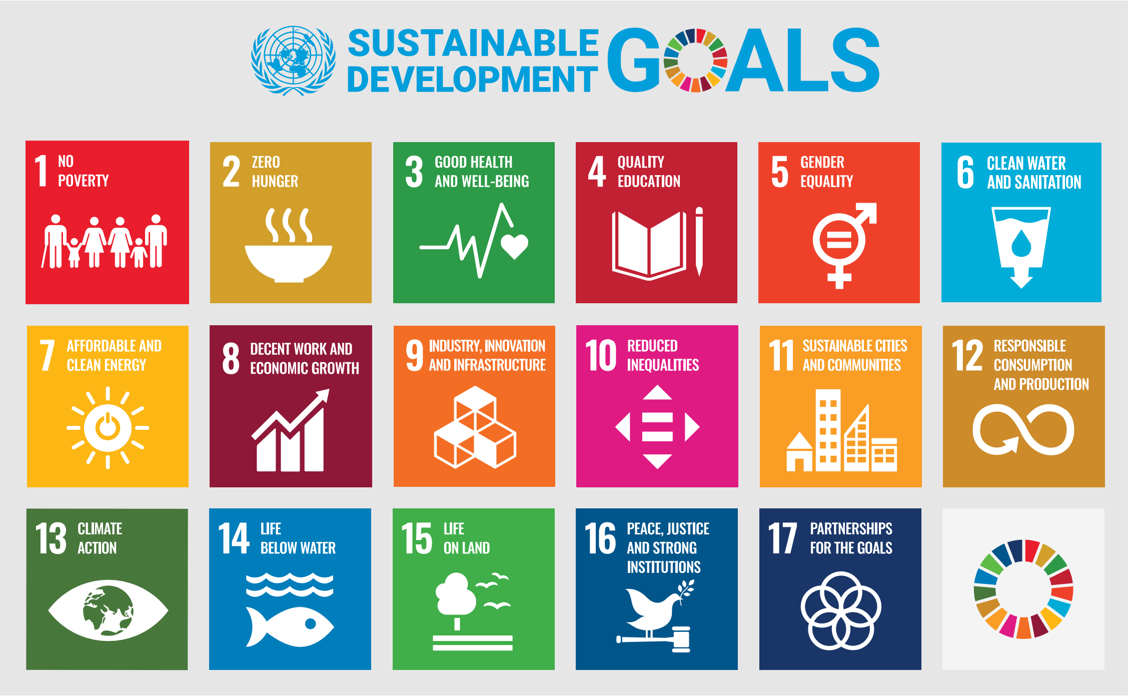

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Figure 2: The 17 Sustainable Development Goals in the Agenda 2030

[16] https://sdgs.un.org/goals

In Shulla´s et al. work (2020) they measured a contribution to SDG4 (Quality Education), SDG11 (Sustainable Cities and Communities), SDG13 (Climate Action), and SDG15 (Life on Land), several topics that we have already stated are connected to CS and that have been studied that CS contributes to these SDGs. Besides, in other studies, CS has also been measured to contribute to SDG3 and SDG6 (Bales et al., 2012; Chandler et al., 2017; Quinlivan et al., 2020; Sprinks et al., 2021).

Some more examples are the work conducted by Fritz et al. in 2019 (Fritz et al., 2019), in which they measured the importance of CS data as indicator for SDGs achievement or advance. Another example is the numerous contributions from citizen scientist to Zooniverse[17], a platform that allows people to participate in real research related to sustainable development (Wuebben et al., 2020). Other studies measured the contribution to SDG2 (Zero hunger), SDG3 (Good health and well-being), SDG6 (Clean water and sanitation), SDG11 (Sustainable cities and communities) and SDG12 (Responsible consumption and production) from many CS projects focused on monitoring the soil (Head et al., 2020).

Other studies state that the main contribution from CS to SDGs is due to the potential social transformation it brings and, therefore, it could help in achieving all the SDGs (Moczek et al., 2021). Many CS projects in the recent years are aimed towards the achievement of SDGs combining tech-based approaches (Sanabria-Z et al., 2022) which could even enlarge their impact in this subject. In fact, the EU recently stated in a report that CS projects in the framework of the H2020, the EU´s research and funding program, have been accomplished and that the support to these initiatives should continue due to their impact and potential regarding sustainable development (European Commission et al., 2020).

So, in general, CS is growing as a practice, and it is contributing to levels that even reach the policies that are shaping the future of the countries in sustainability. The factors that have led the growth of CS to this point have been addressed by different authors and, aside from the motivation of the participants, one of the main elements that triggered this growing is the rise of technologies.

[17] https://www.zooniverse.org/

2.3 Technologies involved in the growth of CS

One of the factors that led the growing of CS is the adoption of web-based technologies (Catlin-Groves, 2012). Web-based technologies are defined as network applications accessible over the Internet (blogs, discussion boards, conferencing sessions tools, online multimedia and mobile technologies, online games etc.) that enable individuals to connect to each other (Kyei-Blankson et al., 2016). The main tool integrated in the web technologies that has been useful to impulse the growth of CS and other areas are the web pages (forums, mailing, repositories, or social media) owing to their social component that allows the users to share information, getting in contact, etc. It must be noted that other digital technologies have also helped to CS, technologies such as mobile apps, sensors and others (Newman et al., 2012b), because they allow the remote collection of data and automatic sharing, interpretation of data in an easier way thanks to gamification, easier engagement using apps, and many other examples.

Web-based technologies have been useful for many different areas such as biomedicine (Lowe et al., 1996), business (Bajgoric, 2001), institutional use (Kelly, 2010), or surveying (Alvarez & VanBeselaere, 2005), but we will focus specially in how useful these technologies are for education (Jaffee, 2003) and collaboration.

There is a large number of studies in relation to using web technologies for education and learning using web (Al Kurdi et al., 2020; Aroyo & Dicheva, 2001; Kunicina et al., 2018, 2019). The development of different web applications and tools have led to a new paradigm in which education on many fields is accessible to everyone that has a computer and access to the Internet. A crucial event in recent history that also triggered this advance was the COVID-19 pandemic. During this time, remote teaching became a necessity and the lack of possibilities to attend to in-person courses led to the proliferation of online courses and other ways to acquire information (Ali, 2020; Bozkurt & Sharma, 2020). In fact, this new trend of online or hybrid learning and remote schooling seems to have future, although there is still a need for improvements and adaptations (Garbe et al., 2020; Mizunoya et al., 2020).

For citizen scientists, learning is a valuable and expected outcome from their participation in projects (Committee on Designing Citizen Science to Support Science Learning et al., 2018). Web-based tools and applications have been used to enhance learning from the participants and other stakeholders, while using them to enable new ways of participation.

In fact, these technologies have a promising future together with CS (Newman et al., 2012a). However, there is still need for more research and new methodologies to obtain a complete understanding of the learning gains from the participants to enhance the current methods (Jordan et al., 2012).

Diverse institutions have created forums and web pages to provide a complete list of projects and actions they undertake, allowing the potential participants to know about the CS projects they can collaborate with. There are online repositories, that act as databases, in which data can be uploaded so participants can contribute sharing their annotations, being Zenodo[18]and Zooniverse[19] great examples. This is an essential task done by citizen scientist, because web-based technologies and emerging technologies has turned citizen scientist into remote sensors (E. Li et al., 2019; Malthus et al., 2020; Stewart et al., 2020). These participants are able to fetch data from their current locations and share it with institutions thanks to these technologies (repositories, direct mailing, mobile apps, etc.). Other tools like instant messaging applications, online surveys, specific web applications that allow participants to perform online tasks have also contributed to the impact on CS and their growth in the past decades.

For instance, the EU-Citizen Science webpage[20] provides valuable information such as a list of projects in process in Europe or papers that could be interesting for stakeholders or a list of organizations that accept collaborators and platforms or users of interest. Zenodo is an online repository in which we can find articles, datasets, images or software. Citizen science (.gov)[21], NASA Citizen Science[22] and Citizen Science by EU[23] are more relevant examples. Alongside the other technologies described, the social networking sites or social media can be effective tools and have an impact in CS and be useful to measure this impact. Web 2.0 brought a new paradigm, a more collaborative, user driven, data driven and evolving environment which turned into most of the applications and tools we use nowadays (Donelan et al., 2010).

The social networks have been a key tool in the development and current popularity of CS (Catlin-Groves, 2012) and a game-changer in several aspects of human life. Social

[18] https://zenodo.org/
[19] https://www.zooniverse.org/
[20] https://eu-citizen.science/
[21] https://www.citizenscience.gov/#
[22] https://science.nasa.gov/citizen-science/
[23] https://eu-citizen.science/

networks can be defined as services aimed to connect individuals via the creation of public or semi-public profiles and the connection to other users (Adamic et al., 2003; Backstrom et al., 2006). These sites have attracted millions of users since their creation, which is dated by some authors back to 1997 with the launch of the space SixDegrees.com (Montag et al., 2021). Their popularity grew in the mid-2000s when MySpace[24] came out, a blogging site that allowed users to share their comments, ideas or thoughts, alongside other sites like Flickr[25] (photos) or LinkedIn[26] (labour relations). Approximately in 2006 Facebook[27] came out and settled the dynamics of modern social networking sites. Later, more social networks followed Facebook and the explosion of diverse social platforms began. YouTube[28], Twitter[29], Instagram[30], Snapchat[31] are some examples, and more recently TikTok[32] which has turned into the most used social network for new generations (Montag et al., 2021). In these networks, each user can visualize his own network and create connections when linking to other profiles or other profiles linking to the user (Boyd & Ellison, 2007). This feature, alongside some intraspecific features of each social networking site, turn these networking sites into a compelling place of study. Precisely, the study of social networks drove into the birth of a specific discipline, called Social Network Analysis (SNA).

Among all social networks, Twitter the one that has contributed to the growth of CS (Mazumdar & Thakker, 2020). It is a platform for projects to engage larger audiences (Liberatore et al., 2018). Twitter is a microblogging platform that started operating in 2006 and gained popularity in the early 2010s. It is based on the follow relation. Each user follows different profiles (individuals, organizations, businesses…) depending on the content they like, and each user is followed by others if they find their content interests. Also, the follow relation happens when users are close to each other's (friends, colleagues, institutions, and others). This follow relation creates a network structure to which SNA can be applied, unveiling many aspects (main actors and communities for example).

[24] https://myspace.com/
[25] https://www.flickr.com/
[26] https://www.linkedin.com/
[27] https://www.facebook.com/
[28] https://www.youtube.com/
[29] https://twitter.com/?lang=en
[30] https://www.instagram.com/
[31] https://www.snapchat.com/
[32] https://www.tiktok.com/en/

All these different analyses translate in interesting findings and insights from the communities inside the social media platforms, which can translate into policy decisions, information to improve the social media usage, discovery of trends or events, among others. Next, we will briefly describe SNA and how it is applied to research about CS.

2.4 Social Network Analysis and CS

SNA is normally defined as a set of social actors, which are called nodes, or members that are connected by different types of relations (Wasserman & Faust, 1994). SNA has existed for at least 70 years and precursors can be found in the XVIII century (Sigrist & Widmer, 2011). Traditionally, these analyses have not only focused on the behaviour of these actors (Bernard, 2005), but also the patterns they create by interacting and connecting. In the end, the social part, must not be neglected (Bernard, 2005). These members are studied as nodes and they can be many different elements: mails, blogs, pages, journal articles, classes, or individuals. These elements will eventually form a graph structure, so SNA relies on presenting this data as graphs or multi-graphs with the connections between elements being the edges of the graph. This is the reason why SNA is normally called a structural analysis and it is based on the graph theory. There are different types of networks depending on the type of connections between the entities:

- One Mode Network, interactions between one only group of actors (Wasserman & Iacobucci, 1991), or Two Mode Network, relations between two different groups (Wasserman & Iacobucci, 1991).
- Complete Network, focused on a complete set of nodes and their relations (Marsden, 1990).
- Ego Networks, the connections around a focal actor, or ego, to others (Arnaboldi et al., 2012). This will be extended in the next chapter of this document.

Another important part for the structural analysis is the linkage between entities (Zhang, 2010). These connections can be directed or undirected, so graphs can be directed or undirected (Fagiolo, 2007). These relations create paths between the nodes, and they are traditionally analysed via algorithms (Marin, 2011). Alongside the analysing of the structure, there are other measures related to it like *ties* or *density* (connections between users and the ratio between possible connections versus actual connections in a network, respectively), but these connections also provide each node of a specific feature called *centrality*, which is different numbers assigned to the nodes within a graph based on the infrastructure, and there are different ways to measure it. This analysis of centrality tends to unveil the most prominent actors in the networks and the most important measures are (Zhang, 2010). The most important indicators in centrality are the *degree*, *indegree*, *outdegree*, *betweenness* and *eigenvector*, which are measurements related to the number of nodes connected to the users, connections created to the users and the paths and flow between them. Besides analysing the structure, it turned important to also analyse the features each node present, as characteristics, which will unveil once more patterns, distribution and therefore information about life features of the entities (Bernard, 2005). Also, the rise of machine learning techniques, deep learning, artificial intelligence, has been helpful to get a better understanding of the social networks. Both SNA and these features will be covered more in detail in the next chapter. Social network sites follow this subjacent graph structure. Using SNA in social media help researches to find important actors inside these applications, information that turns important for a wide range of stakeholders.

Nowadays we can find multiple social networking sites offering diverse experiences inside them, social interaction, professional networking, sharing of personal information, and many other examples. Twitter is one of those that follow the subjacent graph structure, supported by the connections of the users between each other's. But not only via the follow relation, the actions performed in the platform also connect users in different ways, such as retweeting, mentioning, liking, or replying. As it mentioned before, Twitter has participated in the growth of CS, due to its characteristics and dynamics. The main activity undertaken in Twitter is writing tweets, texts up to 280 characters containing any kind of ideas, thoughts, information, and any other form of textual communication. Inside these tweets we can find hashtags, words preceded by a hash symbol (#) which are used to creates tags, links that can be clicked to find other tweets containing the same hashtag. This is one of the important features that are analysed performing content level analysis in SNA on Twitter (Small, 2011). This type of analysis helps to discover trends and important topics for the users.

Other important feature is retweeting, the action of reposting the tweet written by someone else and make it appear on your timeline (your tweet record at your profile) so that your followers can see it. Retweets are of great importance because it creates bridges for information flow, users can receive pieces of information despite not following the original account (Zhiheng Xu & Qing Yang, 2012). Analysing retweets is a common practice in SNA nowadays under the perspective of creating networks using the retweets information as links. Many received retweets indicates that one actor is popular and reaches a large audience.

Another important feature is likes or favourites. They are a simple indicator of how many people read your tweet and showed their agreement or liking to your content. Again, they are an indicator of engagement, being useful once more to discover important actors in the network. Additionally, there are many other useful indicators that can be used in SNA on Twitter: quotes (writing the username of another user inside the tweet to get their attention, to start a conversation, to indicate anything related to them…), replying (commenting the original tweet), addition of images or videos to the tweet, or the alternative text (description of the images or videos for visually impaired people) would be the main examples.

It is possible to retrieve all this information using the APIs that such platforms offer. An API is a set of functions and procedures designed to access features of an operating system, application, or other services[33]. The Twitter API is designed to collect tweets and information from the users, which can be used to perform all the previously explained analyses.

Many researchers have conducted different analyses using Twitter data in diverse areas. Hashtag analysis allows discovering the most popular ones used when discussing politics in Canada (Small, 2011), exploration of content in relation to specific hashtags (Negrón, 2019; Xiong et al., 2019), sporting events (Blaszka et al., 2012), political polarization (Lai et al., 2015), or medicine (Nawaz et al., 2022) among others. Use of retweet analysis is useful to understand the information shared relating to health (So et al., 2016), politics (Stieglitz & Dang-Xuan, 2012), analysis of fake news (Jang et al., 2019), or analysis of engagement (Soboleva et al., 2017) among others. Finally, more modern techniques such as topic modelling, a machine learning technique to discover topics of discussion in a set of texts have been used, to analyse COVID-19 crisis (Prabhakar Kaila & Prasad, 2020) or sentiment analysis with deep learning (Ruz et al., 2020) during crucial events.

Thus, since many different projects and individuals related to CS can be found on Twitter, researchers have used these techniques to get a better understanding about the CS community. Projects in this platform share news, information or opportunities to

[33] https://www.ibm.com/topics/api

collaborate. Individuals related to CS are also present in this platform, sharing their results, experiences or following the projects and institutions to keep up to date with the latest events and discoveries. All these different actors have been analysed using the SNA techniques, giving a bit more context about the CS community. For example, discourse analysis during COVID-19 crisis (Ilhan & Aydınoğlu, 2023), text analysis about pluvial flooding (See, 2019), analysis of learning outcomes from the CS community (Aristeidou & Herodotou, 2020), information quality analysis inside CS (Ahmed et al., 2020) or even programs to monitor specific cases inside the CS community (Harley & Kinsela, 2022). There are studies and literature associated to SNA and CS, although this is not as extensive as in other fields. Most of the cases are centred around specific isolated cases (Cox et al., 2015; Simpson et al., 2014), and there is a lack of complete and general analysis of the status of the discourse in Twitter. The best reference is the analysis performed by Mazumdar et al., who combined different techniques of analysis to try to understand the behaviour, interests, and status of discourse of the CS community in Twitter (Mazumdar & Thakker, 2020).

Nowadays the CS community in Twitter is centred around popular accounts which are mainly institutions and projects (Mazumdar & Thakker, 2020). The main behaviour is retweeting, and the discussion addresses a wide range of topics (Mazumdar & Thakker, 2020). There are several studies about SNA and CS but about isolated cases or specific fields of study, and more knowledge could be obtained by performing research about the whole community (Mazumdar & Thakker, 2020). The importance of CS nowadays, which was explored in previous sections, invites to deepen in this group, which could lead to a better understanding of it for sure, and a better approach for policy making, initiation in CS, management of projects inside social media, and many other actions.

With this knowledge in mind, this thesis began in the framework of the EU CSTrack project previously introduced. This thesis was developed in the macro-level of analysis about the discourse of CS in social media. According to the lack of standardization in SNA, we aimed to design a standard pipeline for SNA in Twitter applied to the CS community and to palliate the lack of replicability of the analyses. This thesis is focused on designing a platform that could allow everyone to perform SNA live on the Internet. The steps followed will be explained next, alongside a more detailed description of SNA, its characteristics and indicators, application and how the platform was designed and tested in different studies.

3. The SNA techniques and the implementation of the CS dashboard

This chapter introduces the process followed in this thesis to design the pipeline for SNA applied to the CS community in Twitter. The objective was to select a collection of content analyses to be applied according to the features of Twitter, along with the structural analysis of the resulting networks. Twitter contains different elements and relations between users that can help reveal important information from the community, as we specified in the state-of-the-art. In this chapter, we will cover how we acquired the tweets for the analysis, the selected techniques, how we applied structural analysis in SNA with a detailed background, to finally show how the dashboard is implemented.

3.1 The harvesting of tweets

All the different analyses explained need to be performed over a set of elements. In this case, the elements are tweets, and specifically tweets about CS. To obtain them we made use of the Lynguo tool[34] in most of case studies. In specific studies we used the Twitter API directly.

This tool was designed by the IIC (Instituto de Ingeniería del Conocimiento) as a social media monitoring tool. The tool applies filters to select different tweets, and, in our case, it used words related to Citizen Science, such as "citsci" or "citizen science". We collected 800,000 tweets (approximately 450,000 unique tweets since it is possible to find duplicates in the Tweets due to harvesting repeated texts) from the 30th of September 2020 until early September 2022. Lynguo downloads the tweets accompanied by different fields containing information from the tweet and the users. The different fields are as follows:

- Text: the actual tweet, what the user wrote.
- User: the username of the tweet´s author
- Date: date of creation of the tweet.
- Link: the web link to the tweet.
- Impact: this is a calculation from 0 to 100 that gives a score for the interest or repercussion of the tweet based on impressions and followers.

[34] https://lynguo.iic.uam.es/

- Opinion: this is a sentiment score ranging from -100 to 100, negative values for negative sentiment, positive for positive sentiment and 0 for neutral.
- Category: different values for the content of the text.
- Erased: states if the tweet is erased or currently in Twitter´s timeline.
- Mark: this is a classification that can be: CS All World (CS tweet outside Europe), CS EU (CS tweet from Europe) or some other classifications like CS higher education.
- Official: if the account is verified.

These tweets are stored as a CSV file, which we uploaded to a private MongoDB[35] database, in order to store the data more efficiently and access the different fields of the CSV for later analyses such as the extraction of followers. Some of the elements inside the data-frame are retweets, which specially labelled as "RT@username: [...]", the use of these retweets will be detailed later in the content analysis and in the network analysis sections.

From the fields text, date, and username, most of the analyses are performed, both content analysis and network analysis, since from these fields we obtain the important data.

3.2 The design of the pipeline for analysis

In this section, we will cover the different techniques we selected to perform content analysis, alongside their importance and purpose in SNA. Besides, we will introduce the different techniques used in the structural analysis of networks, also with a description of their purpose and what information do they bring in. Once all the techniques are covered, we will discuss how to bring them together, because applying all of them in research can result in a better comprehension and knowledge acquisition from the tweets.

3.2.1 Content analysis

Once we had the tweets, the knowledge acquired following our state-of-the-art review was put into practice. The idea was to combine the existing techniques for content analysis with the structural SNA. The most important features to analyse in tweets in relation to the content are the hashtags, most common words, the languages present in the tweets, and the topics addressed by the members of the community (Zhiheng Xu & Qing Yang, 2012).

[35] https://www.mongodb.com/es

All the different analyses we selected and lately performed, were designed in Python, and declared as functions to increase replicability for specific cases using data with the structure described in the previous section.

First, we decided to design an effective way to filter the tweets obtained from Lynguo, to perform case studies to validate the results and later allow the analysis of specific portions of the data, useful for policy makers and other stakeholders. This was achieved creating a filter function based on regex that finds the occurrences of keywords of our interest (according to the topic to be explored) inside the tweets. This filter was applied to the different cases studies to answer our RQ1.

To discover how many different languages were present in the dataset, we trained an NLP (Natural Language Processing) based multinomial Naive Bayes classifier using 10,337 texts in 17 different languages including English, French, Spanish, Portuguese, Italian, German and others.

As we explained in the section 2.4, Twitter contains several features of interest for researchers. One of them, related to the content of the tweet, is the use of hashtags (Ferragina et al., 2015). Hashtags are words preceded by a hash symbol (#) which work as tags or links to specific topics, finding all the tweets containing the same hashtags connected to each other's. Hashtags are normally used to label the topic of the tweet and to promote ideas or any other subject. So, our focus was to identify the hashtags inside the tweets about CS so that we could analyse their count and topics addressed. We extracted the occurrences of words preceded by a hash symbol in the tweets applying non case sensitive regex patterns. However, we had to consider the use of hashtags and therefore the topics that people write about are not constant in time (it could be used to promote a conference or an event, for example). Therefore, the use of hashtags could evolve in time, and to perform this kind of analysis time series of the use of hashtags were created. The dataset contains the dates of creation of the tweets, so to perform the time analysis of hashtags we extracted the hashtags as previously explained but accompanied by the date of creation of the tweet, forming tuples (hashtag, date) used to create the time series. Both, the visualization of hashtags and the time series were created using the Plotly package for Python[36], with examples of the resulting graphs later in section 3.3 The Creation of the Dashboard.

[36] https://plotly.com/

We focus now on retweets, which, as previously explained, are found in our dataset as "RT@username: the tweet", so we can easily discriminate which are original tweets and which retweets. If someone retweets a tweet, it is also retweeting the hashtags inside the tweet. This relation is important for information spreading. Measuring the retweets hashtags, we can get information about which topics are more shared than others. So, we also performed an analysis of retweeted hashtags, selecting only those tweets labelled as RTs and extracting the hashtags inside. We followed then the same procedure to create the time analysis of retweeted hashtags, creating the tuples containing the hashtags and dates for those tweets that were retweets.

Following with the content analysis, another feature that is commonly addressed is the count of most common words inside the texts (Majkowska et al., 2021). We selected all the nouns, verbs and adjectives inside the texts following NER (Named Entity Recognition) methods, used to determine the grammatical categories of the words. We removed every word preceded by a @ or a # to remove usernames and hashtags, alongside the most common stopwords (preposition, conjunctions, punctuation symbols, etc.) in different languages. The final count was displayed using the wordcloud[37] module from Python.

Additionally to this last analysis, TF-IDF analysis shows the relevance of these words inside the texts (Ramos, 2003). TF-IDF stands for Term Frequency- Inverse Document Frequency, and it is a statistical computation that determines the weight of the words in the corpus of texts (tweets in our case). It calculates the frequency of the word in the documents and multiplies it for the inverse frequency balanced by the number of documents. The result is a value between 0 and 1, being the values nearest to 0 those of higher weight or importance, and vice versa. This technique is especially indicated to discover words that could be used as keywords to obtain results when filtering.

Culminating the content analysis, we applied machine learning techniques to unveil the topics of discussion and to perform a specific analysis in relation to the SDGs, due to the importance of CS in this subject.

Topic Modelling is an unsupervised learning technique used to identify patterns between texts, which has been widely used in SNA. This technique allows the unveiling of what the users are talking about, helping researchers determine which topics are of general

[37] https://pypi.org/project/wordcloud/

interest or could be used to make policy decisions, or simply understand the concerns of the communities. Our first approach to the topic modelling was using the LDA algorithm (Jelodar et al., 2019). This process started removing unnecessary elements, such as symbols, punctuation, stopwords, from the tweets and cleaning the entire corpus. We converted the hashtags to plain words and removed some highly repeated words that added noise to the analysis, "Citizen Science" and other related words in our case. We also applied a TF-IDF analysis and a topic coherence analysis, destined to select the number of topics with higher performance. We extracted the keywords of each topic and the number of tweets per topic. To represent the information, we chose the Intertopic distance map. This visualization shows the topics as bubbles distributed in four axes, the closeness or distance between the bubbles mean their similarity or dissimilarity. Similar topics will be plotted close to each other or even overlapping, while topics about different issues will be distributed far from the others according to how different they are.

Later, we then changed to use BERTopic, a topic modelling technique from Google (Uthirapathy & Sandanam, 2023). The decision to change to this model was taken because of how easy is to tune BERT in comparison to LDA, besides allowing sentence embedding which is ideal for SNA. Sentence embedding is the technique in which the sentences are represented as numerical vectors, facilitating the calculation for computers since this analysis of similarity is based on the distance between these vectors.

The process of using BERT was like using LDA. It started applying the same cleaning process. Then, we selected the parameters for the model: multilingual model, represent the top 10 words inside the topics and the rest was set as default. All these verifications are done to obtain the best performance, which is measured first with the topic coherence. This metric evaluates through statistics and probability from the complete corpus the coherence of the topics that are extracted by the model. This means how interpretable and coherent for humans are the topics extracted based on the words contained in them and the relations between them (Syed & Spruit, 2017). In our case, BERTopic analyses this coherence automatically but allows the modification if needed. The other measure to evaluate the performance is the F1 score of the model, which ranges from 0 to 1 and being the higher, the better. The F1 score is the result Then we checked the minimum topic size, although BERTopic calculates the topic size automatically unlike LDA, we decided to test different sizes obtaining the best performance with a value of 10 minimum size for the topics. Since the number of topics will be high, we set the model to perform an automatic merge of topics to reduce the number, something that, again BERTopic calculates automatically. All the tunings of the model resulted in an overall F1 score over 0.81. To present the results we again extracted the keywords, the intertopic distance map and the number of tweets per topic.

One specific technique that we applied only in one case study was the classification of tweets by SDGs. We trained a BERT classifier to determine whether the tweets were about one or another SDG so that we could see the distribution of tweets and the most addressed SDGs then. The classification can be linked to the structural analysis as we analysed the diffusion of information related to the SDGs using the retweets, and it will be detailed in the case study 4.2 Understanding the discussion of Citizen Science around Sustainable Development Goals.

Applying all the previous techniques we expected to be able to get all the possible knowledge form the content inside the tweets, but we needed to combine this analysis with the structural one.

3.2.2 Structural analysis

Once we had selected the different analyses for the content inside the tweets, we needed to enrich the pipeline for analysis with the traditional SNA techniques. These techniques are based on the relations between the users, who are the nodes, the actors, which form in the end graphs or networks. These networks can be of different nature and in this thesis, we explored all of them. The networks can be:

- One Mode Network: this type involves a set of interaction between one only group of actors (Wasserman & Iacobucci, 1991).
- Two Mode Network: this type is based on the relations between two differentiated types of actors. An example would be the relations between profit organizations with non-profitable organizations (Hawe, 2004).
- Complete Network: intimately related with the next type, these networks are focused on a complete set of nodes and their relations without existing a central or focal actor in this community (Marsden, 1990).
- Ego Network: on the contrary, ego networks are the connections around a focal actor, or ego, to others. An example would be focusing on one single user of a social networking site and checking their friends, and typically the friends of the friends too (Arnaboldi et al., 2012).

As it was said before, the nodes in these networks are the users in the social media platform, and the connections that act as links between the nodes are the activities undertaken in the platform. In Twitter the most common actions are following and retweeting, but other activities such as mentioning someone, liking a tweet, replying to someone are also interesting connections between users. In this thesis we explored the networks formed by the follow relation, retweeting and quoting. These activities also present directionality, they have a source and a target, and this differentiation, this directionality or non-directionality of the connections, separates the networks in two categories: directed graphs or undirected graphs.

A directed graph contains a set of nodes and links or edges that follow a certain direction, forming the pair (i, j), which means i interacts with j and it does not exist in the opposite direction (although explicitly stated). An undirected graph contains a set of nodes too and edges that form pairs (i, j) but these links are bidirectional. Retweeting, following, and quoting are clearly directed actions, so the networks we formed and explored were directed graphs.

The connections between users provide each node of specific measurements of what is called centrality. An analysis of centrality is focused on exploring the connections between the actors, the nodes, and helps to determine popularity, activity, behaviour, and so on. The main measurements and their definitions are in the following list:

- Degree Centrality: the degree of a node is the number of connections that this node has. These connections are directed towards it or leaving the node, marking the difference between indegree (directed towards the node) and outdegree (leaving the node). This usually indicates popularity.
- Betweenness Centrality: betweenness is the number of times a node connects other pairs of nodes that have no direct connection between them. This measure indicates that this actor has control over the flow of resources (information, money, power, etc.) acting as a gatekeeper. This means that these nodes are influential, powerful, or popular.
- Closeness Centrality: this is based on the distance between nodes. If a node is closely connected to other nodes (no more than one step to the adjacent node) it means that is independent to make the resources flow by it itself.

The first network we explored was the one formed with the retweets. The users retweet tweets from other users and thus they create a connection. The tweets labelled as "RT@user: original tweet" helped us to create the edges. We extracted the column user and the @username without the @ and created tuples containing the user that retweeted and the user that was retweeted (a, b). Using the networkX[38] package from Python we created the directed graph and extracted the values for Indegree, Outdegree, Betweenness and EigenVector to help us identify the main actors in the platform.

We also explored the networks formed via quotations/mentions. The idea was similar as the previous one, selecting the user that created the tweet and extracting the @username inside the tweet. By doing so, we again created tuples containing the user that mentioned and the user that was mentioned (a, b). NetworkX package was again used to extract the same metrics.

One special behaviour that we wanted to check was the connection between hashtags. A user normally uses more than one hashtag inside a tweet, and the hashtag inside those tweets have relation between them, according to the topics or something specific. Creating a network of hashtags would help us identify related topics, enriching the topic modelling analysis with the intertopic distance. We took all the tweets and created pairs with the hashtags inside them, to later represent it in a graph to unveil the connections.

Furthermore, in Twitter the users follow each other or not, creating an unequal relationship between them. To extract this information, we took all the users inside our dataset and searched them with a specific query in the Twitter API, to retrieve the list of users that they follow and that are following them. As we were checking the connections inside the CS community and extracting all the followers could bring us users from outside, we filtered for only those inside the CS community. The task of retrieving all the followers list was arduous and time consuming, so for the case studies we focused on creating Ego-Networks or just collecting the number of followers.

According to the dynamics inside Twitter, we thought that the action of retweeting or mentioning or liking a tweet could be influenced by the fact of being a follower of that account or not. We performed a specific analysis to check the influence of a static interaction (following or being followed) on a dynamic one (retweeting, mentioning, or liking), which will translate into a better understanding of the dynamics inside the social

[38] https://networkx.org/

network platform and potential recommendations to users to enhance their engagement. This specific case will be covered in detailed in the next chapter with the case studies.

When calculating and representing graphs from social media platforms we tend to encounter one recurrent problem. Network containing high number of nodes, as the ones we are calculating, tend to show many low-connected nodes. These isolated nodes add noise to the image while being negligible for the analysis, so we made use of a technique to “trim” the networks and remove all those low connected nodes.

To remove these nodes, we used the k-core method from networkX. This method is especially well-suited for large scale networks (Alvarez-Hamelin et al., 2005). Given a graph, we can extract an I-core which is a subgraph in which the nodes have a minimum number of connections, a minimal degree or k (or higher). This is related to the concept of “coreness”, which can be described as the maximum core, or decomposition of the graph, containing a node X with a maximum k. Coreness leads to a hierarchical decomposition of the graphs shrinking the network removing nodes with low degrees, affecting the degree of other nodes (Batagelj & Zaversnik, 2003). Figure 3 shows a visual representation of the concept of coreness.

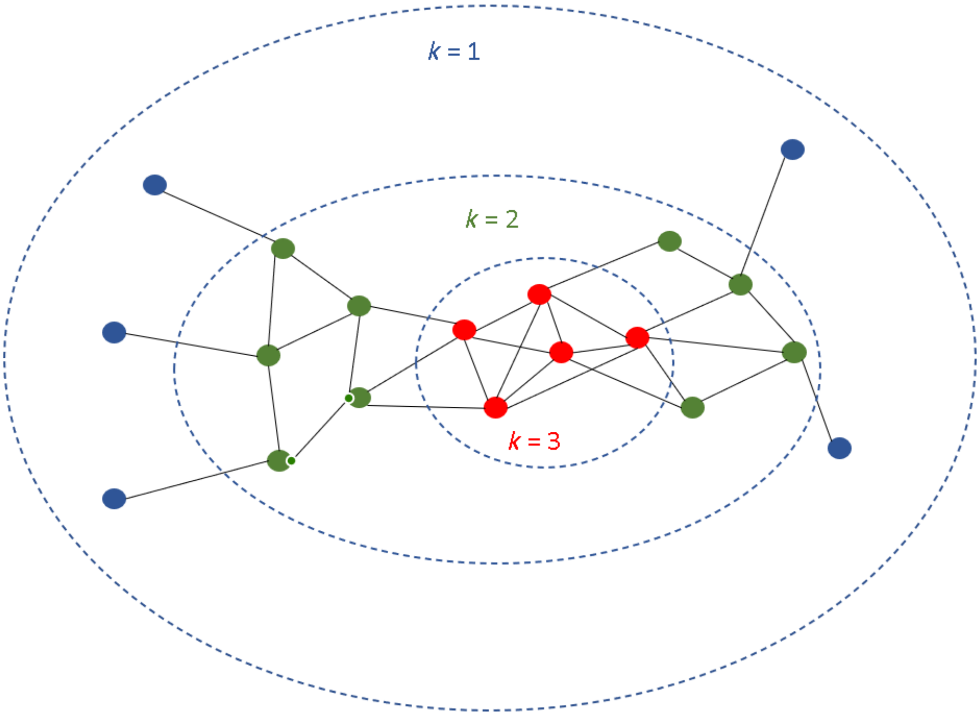


Figure 3: Representation of coreness

Once we created all the networks, we used them to explore the centrality measures from them. According to what it was stated about the CS community, the most common activity is retweeting (Mazumdar & Thakker, 2020), so we focused on extracting the values of centrality from the networks of retweets and followers.

Finally, another important phenomenon that happens in social networks is that users congregate around influential users forming communities. We also explored this cohesion in communities applying the Louvain method form the networkX module in Python.

Next, we explain how these techniques were implemented in the dashboard developed.

3.3 The creation of the dashboard

The idea of developing this dashboard came after conducting research about the use of Twitter and specially the post-analysis situation. As it is known now, most of the case are isolated case studies (Karami et al., 2020) and frequently focused on the application of one or two techniques at most (De França et al., 2023; Hagras et al., 2017). Besides, there is nowadays a problem with reproducibility, although most of the results from these studies are presented as papers in which instructions are included, it finally turns out that it is difficult to conduct or reproduce the same analysis (Mamo et al., 2023). Furthermore, there are few platforms for SNA in Twitter. We found examples of application to topic mining sites or fake news detection (Carvalho et al., 2017, 2017; Faustini & Covões, 2020). There are platforms used to analyse data from Twitter, but we found no platforms aggregating all the different SNA techniques.

Therefore, we created a web platform where several procedures occurred in the background, and it would offer visualizations that would show the information. This dashboard would combine different techniques and analyses to finally offer a complete understanding of the whole panorama of the discussion in Twitter about CS and it would provide the possibility to analyse specific topics. This platform would help both specialized and non-specialized stakeholders to deepen in the information from the CS community, facilitating the knowledge acquisition, policy making, understanding of the behaviours and so on. The architecture of the platform and the techniques included will be described in the next sections.

3.3.1 Architecture

The dashboard was created by means of the Dash module for Python, version 1.21.0 to integrate Plotly visualizations. The Dash module is an open-source library released under MIT license which serves as a tool for web-based applications that provide visualizations.

The dashboard contains the environment with the tweets and a MongoDB database with the information from the users. Using these tweets and information from the users, the different metrics are calculated, and the visualizations are plotted into the dashboard. The Dash module combines the Plotly visualizations and the web application. For a better understanding of the architecture and the relationship between elements check the Figure 45, a detailed diagram of the architecture.

To perform the analysis, the utils employ the tweets that are stored in a Pandas[39] data-frame. The information is extracted from these tweets and users, such as the location, profile information, impact, influence, and others. All this information is stored in a MongoDB database. Since there is no relation between tweets, we created a non-relational database.

The Pandas data-frame is loaded to the app's environment alongside the users' information from the MongoDB database to be used by the utils. The utils are functions that calculate an analyse the tweets and information from the users to create the different visualizations as explained in section 3.2 The design of the pipeline for analysis. These functions were developed as wrappers so we can change the framework of analysis using the same utils each time. The analyses that have been integrated in the platform are hashtag analysis, temporal analysis of hashtags, calculation of most common words and visualization in wordcloud, network degree analysis, sentiment analysis, retweet network visualization, community detection, geolocation of users and topic modelling.

[39] https://pandas.pydata.org/

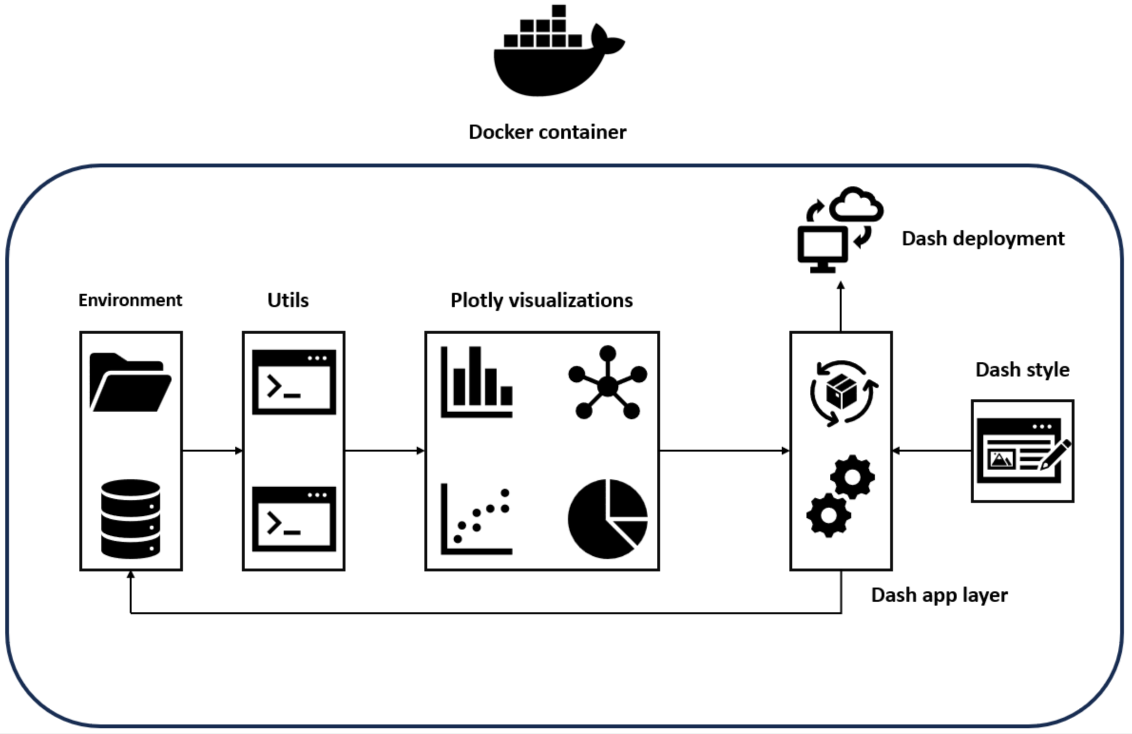


Figure 4: Dash architecture

All the visualizations are created using Plotly, due to the possibility of making them interactive in web environments and to make a pure Python environment with Dash app. Finally, the Dash app layer and style layer create the dashboard.

The methodology of the different techniques integrated in the dashboard, which were used for the different studies, will be explained next. Besides, the improvements of the techniques applied in the studies and those that are still not implemented in the dashboard, will be explained in their own sections alongside a reminder in each case study.

3.3.2 Dashboard analytics and visualizations

The dashboard contains on the left a main menu to select the different types of analyses. Each section of analysis contains different subsections, Figure 6 shows a visualization of the menu.

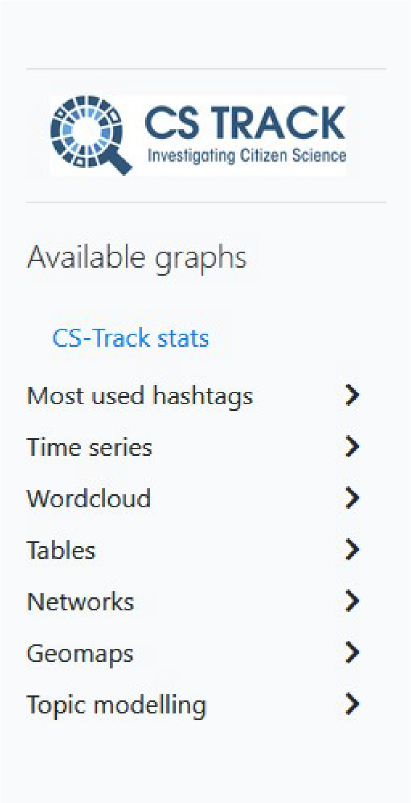


Figure 5: Main menu with the different analyses available

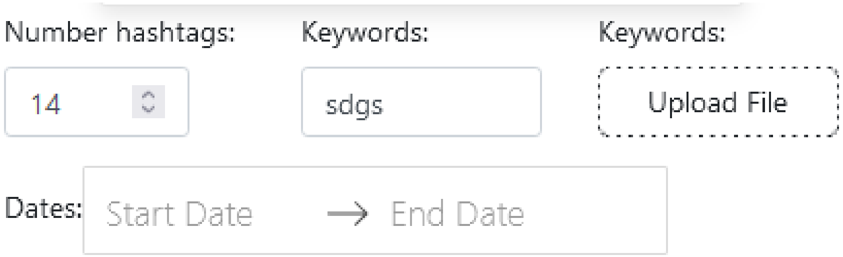
3.3.3 Filtering of the dataset

The dataset can be analysed in its fullness to obtain the complete state of the discussion, but for analysing specific topics we decided to include a filtering system. This filtering also was useful also to validate the techniques in case studies.

In each section, the filter can be applied to the dataset by writing a set of keywords (or uploading a text file containing them) and selecting a starting point and ending point for the filter to be applied in specific dates (if omitted, it applies the complete period).

The filtering menu contains, in this case, a selection of hashtags to be displayed, the keywords box, the keyword upload box and the boxes for date selection which display a calendar to select the dates. Figure 7 shows an example of filtering, in this case for the hashtag analysis.

The filter is based on the filtering functions we first designed when establishing the pipeline for analysis, which are integrated in the dashboard.



Number hashtags: 14

Keywords: sdgs

Keywords:

Figure 6: Filtering menu

3.3.4 CSTrack Twitter stats

The dashboard offers the possibility to check the stats of the CSTrack project´s Twitter account. This was developed since we found useful to monitor the account during the project.

3.3.5 Most used hashtags

The next section is “Most used hashtags”, which displays the calculation of the most frequently used hashtags in the dataset. This is calculated as explained before, in Python by extracting the words preceded by a hash symbol (#) inside the text field. We make two calculations: all hashtags (complete dataset) and retweeted hashtags (hashtags counted in the retweets exclusively).

The visualization is created with Plotly, a histogram than can be changed to display different numbers of hashtags in the screen. Two examples can be shown at the next two figures. Figure 8 shows results of the count of all hashtags in a case study about SDGs and Figure 9 shows the calculation done inside the retweets in the same case study. As we previously explained in section 3.3.3 Filtering of the dataset, the filters appear in the menu above the visualization.

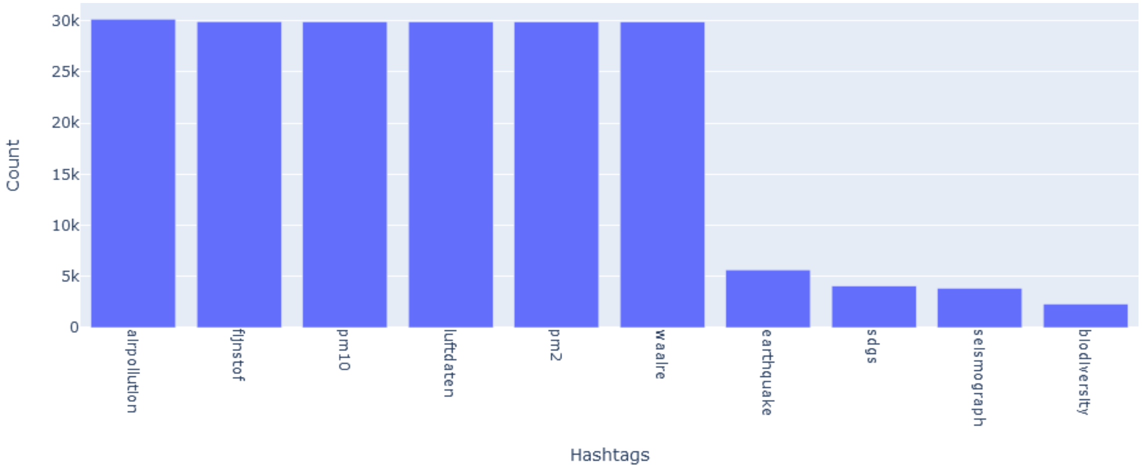


Figure 7: Example of the visualization of all hashtags in a case study of SDGs

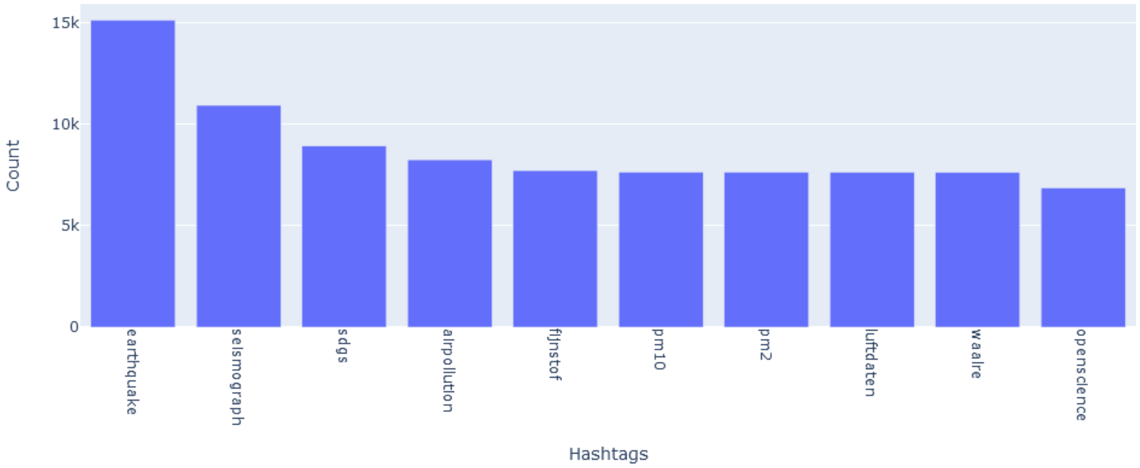


Figure 8: Example of the visualization for Rt hashtags in a case study of SDGs

3.3.6 Time series

We also provide in the dashboard an analysis of the evolution of usage of hashtags in time. This section is called "Time series" and it contains line graphs showing the count of hashtags in the dates from the dataset (the dates when the tweets were written). This kind of analysis allows a more detailed comprehension of the usage of hashtags and the discovery of events in time that trigger usage.

The time series is displayed using Plotly and calculated creating the graph with the time series option that Plotly offers. The hashtag is isolated from the text and stored in a different data-frame alongside another column corresponding to the date of the tweet where the hashtag comes from. This subset is the one used by Plotly.

As an example, Figure 10 shows the results of all hashtags time series analysis using health-related keywords. In the Y axis we see the count of hashtags and in the X axis the analysed dates.

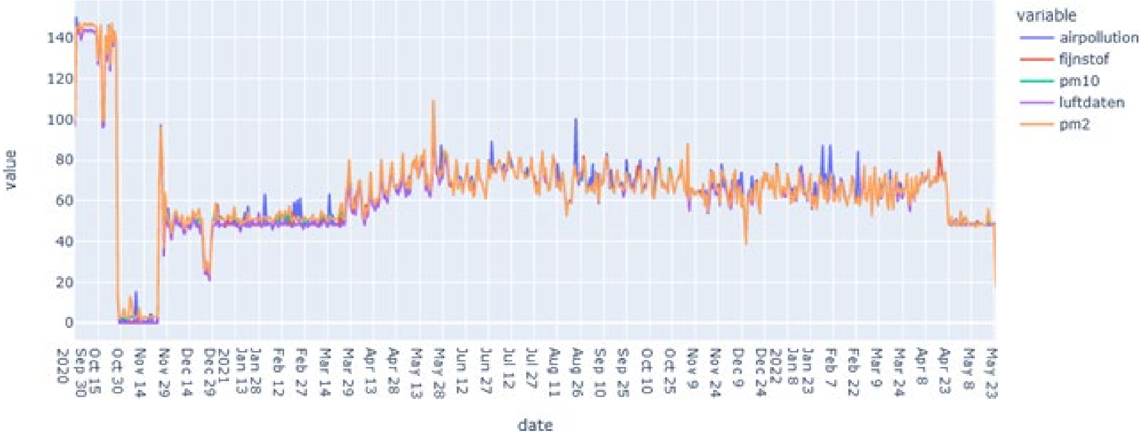


Figure 9: Example of all hash time series

Like the previous section, there is an option to display the results from the retweeted hashtags. Figure 11 shows the result in the same case with health-related keywords. In this case, we discovered a high usage of the hashtag SDGs due to the celebration of a conference in October 2020.

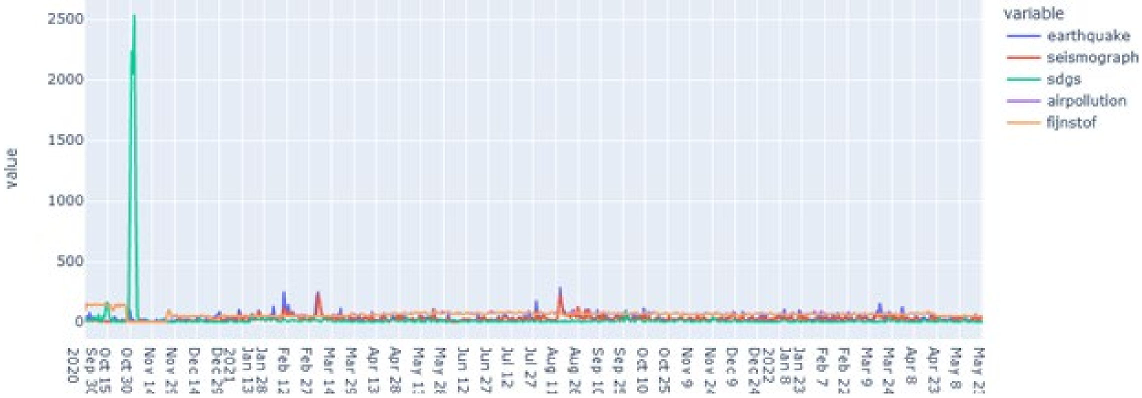


Figure 10: Example of RT hash time series

3.3.7 Wordcloud of most used terms

The following sections contains the analysis of most frequent terms inside the dataset. The count of the terms is done following what was explained in 3.2 The design of the pipeline for analysis. The final count of terms is displayed in a wordcloud by means of the wordcloud Python module. This wordcloud is embedded in the dashboard's environment. An example of visualization can be found in Figure 12.

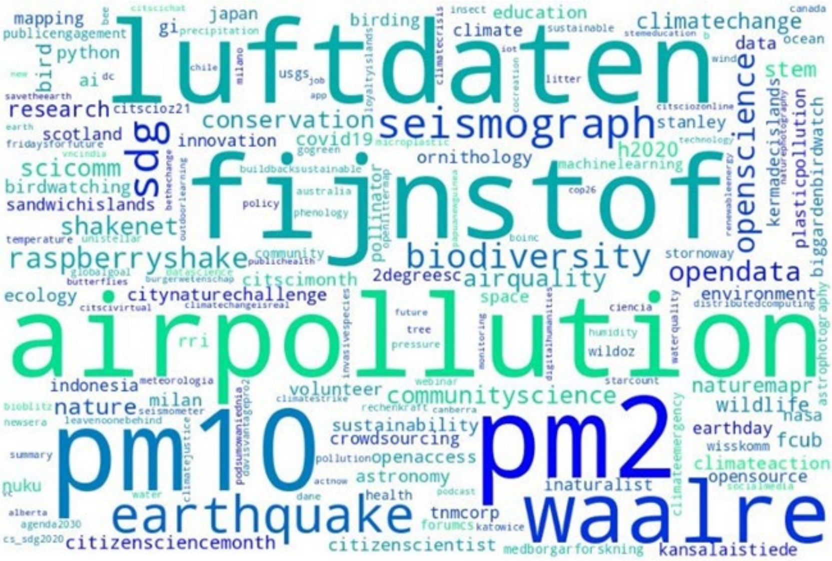
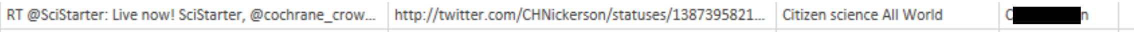


Figure 11: Wordcloud showing the most common terms in subset of example

3.3.8 Tables

The next section contains the results from the degree analysis and sentiment analysis. These results are displayed in tables. The degree analysis is based on the structural analysis done on the network formed by retweets, as it was explained in the section 3.2.2 Structural analysis. An example is Figure 13: RT example with anonymized username which shows that an anonymized username in the 4th field (since it is an individual) retweeted a tweet from SciStarter (RT@SciSarter: text from SciStarter account).

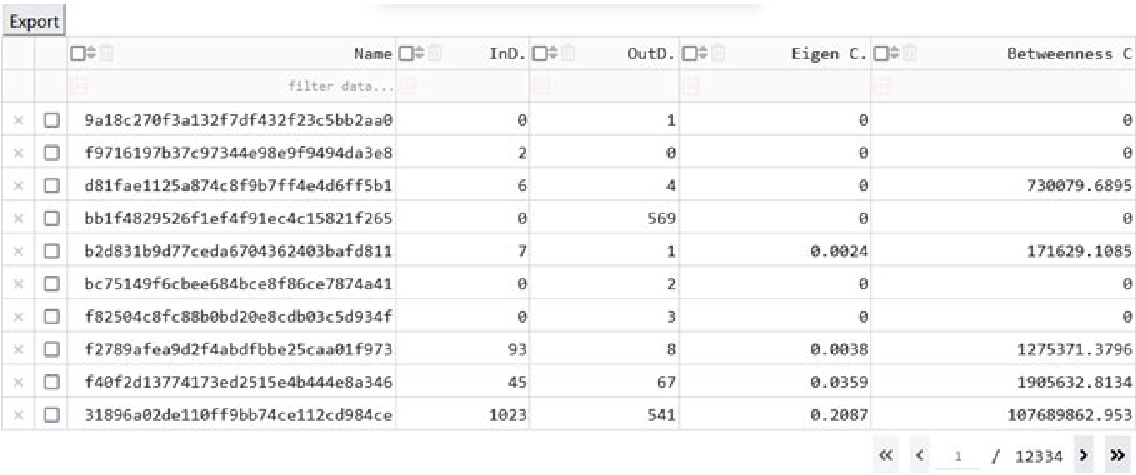


RT @SciStarter: Live now! SciStarter, @cochrane_crow...	http://twitter.com/CHNickerson/statuses/1387395821...	Citizen science All World	[redacted]

Figure 12: Tables. RT example with anonymized username

As explained before, using the networkX module for Python we created the resulting graph and calculated the values for the centrality measures. This centrality measures were Indegree, Outdegree, Betweenness and Eigenvector (Borgatti & Everett, 2006). This table allows the user to obtain a rank of influence of the actors of the conversation based on the centrality measures. Currently all the nodes are anonymized to protect the individuals. An example of a table of centrality can be found on Figure 14. The first field contains the anonymized name using the hash library, then the Indegree value (InD.), the

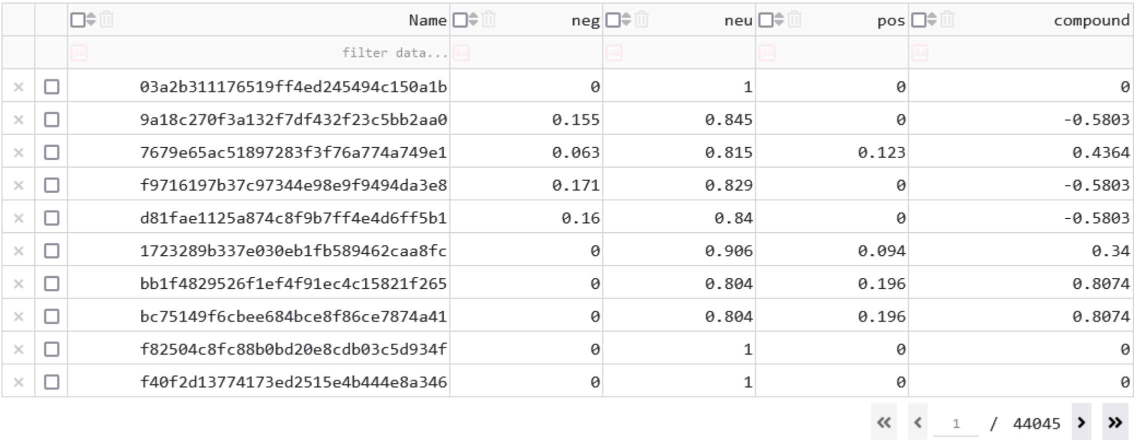
Outdegree (OutD.), Eigenvector (Eigen C.) and finally the Betweenness (Betweenness C.).



Name	InD.	OutD.	Eigen C.	Betweenness C
9a18c270f3a132f7df432f23c5bb2aa0	0	1	0	0
f9716197b37c97344e98e9f9494da3e8	2	0	0	0
d81fae1125a874c8f9b7ff4e4d6ff5b1	6	4	0	730079.6895
bb1f4829526f1ef4f91ec4c15821f265	0	569	0	0
b2d831b9d77ceda6704362403bafd811	7	1	0.0024	171629.1085
bc75149f6cbee684bce8f86ce7874a41	0	2	0	0
f82504c8fc88b0bd20e8cdb03c5d934f	0	3	0	0
f2789afea9d2f4abdfbbe25caa01f973	93	8	0.0038	1275371.3796
f40f2d13774173ed2515e4b444e8a346	45	67	0.0359	1905632.8134
31896a02de110ff9bb74ce112cd984ce	1023	541	0.2087	107689862.953

Figure 13: Example of eable of degrees

The next subsections inside "Tables" is the sentiment analysis, a specific analysis used in the dashboard. This was performed using the Vader Sentiment library for Python. Vader is a lexicon and rule-based sentiment analysis tool designed for social media, so it fitted our purposes perfectly. This tool analyses the texts based on a list of words associated to values for sentiment resulting in a score for negative, positive, and neutral sentiment and finally a combined score (Hutto & Gilbert, 2014). So, the sentiment was calculated for all the tweets and displayed in a table ordered from highest positive results to lowest. The values can range from -1 (negative), 0 (neutral), to 1 (positive). In Figure 15 we can see an example of the results from this analysis.



Name	neg	neu	pos	compound
03a2b311176519ff4ed245494c150a1b	0	1	0	0
9a18c270f3a132f7df432f23c5bb2aa0	0.155	0.845	0	-0.5803
7679e65ac51897283f3f76a774a749e1	0.063	0.815	0.123	0.4364
f9716197b37c97344e98e9f9494da3e8	0.171	0.829	0	-0.5803
d81fae1125a874c8f9b7ff4e4d6ff5b1	0.16	0.84	0	-0.5803
1723289b337e030eb1fb589462caa8fc	0	0.906	0.094	0.34
bb1f4829526f1ef4f91ec4c15821f265	0	0.804	0.196	0.8074
bc75149f6cbee684bce8f86ce7874a41	0	0.804	0.196	0.8074
f82504c8fc88b0bd20e8cdb03c5d934f	0	1	0	0
f40f2d13774173ed2515e4b444e8a346	0	1	0	0

Figure 14: Results of the sentiment analysis from a subset as example

3.3.9 Networks

The section networks has three different subsections: retweets network, two-mode graph, and communities. These networks are created again following the retweet relation as edges using the networkX module for Python. Instead of tables, what we have in these three subsections are visualizations of the graphs created using the webweb module for Python and embedded in the dashboard.

Figure 16 contains the visualization of the network of retweets, which is the same explained in the previous section. The webweb display shows all the nodes from the dataset or subset and the bigger and redder the node, the more Indegree they have.

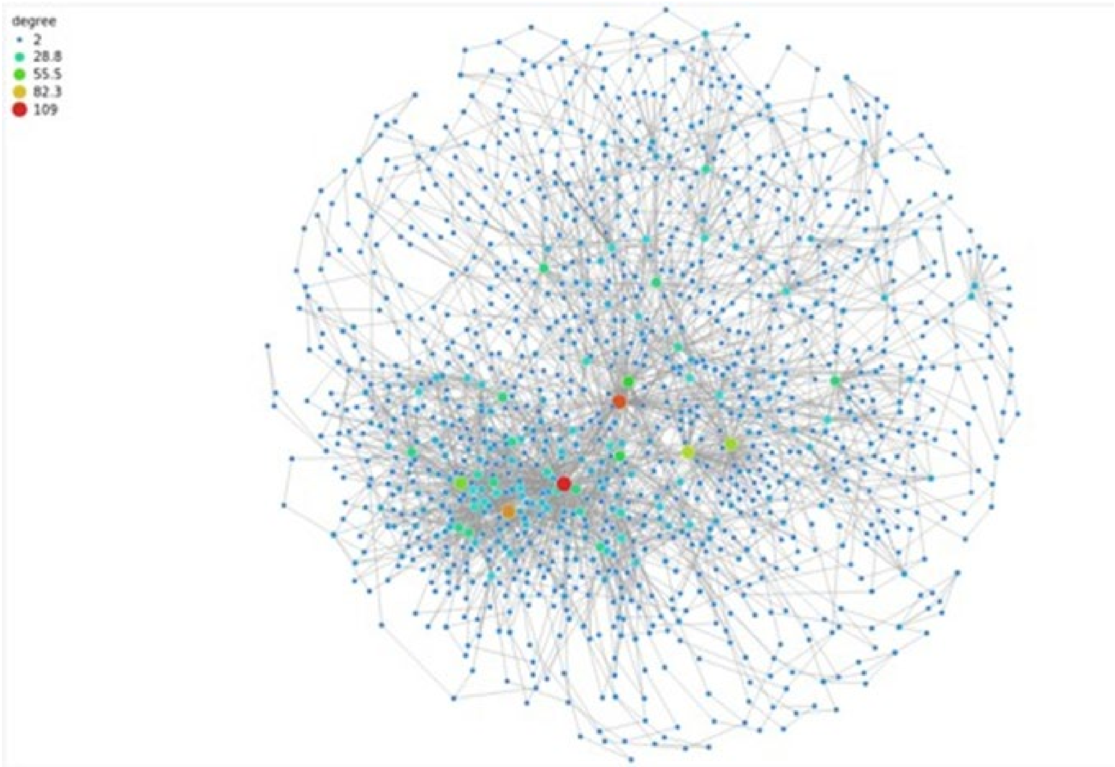


Figure 15: Example of network of retweets

In the next subsection we have the Two-mode graph. The concept of two-mode graph was previously described in 3.2.2 Structural analysis, the nodes are divided into two groups since there are different characteristics between them. In our case, nodes are divided into tweets and users and the connection is the act of retweeting. The bigger the node, the more retweets it received.

The last subsection is the “Communities”. Again, the connections are retweets, but the display is different. Every node is a community, and the bigger the node the more users it contains. The users congregating around certain users and interacting with them form the different communities, and the connections between communities are retweets given from one community to another. The calculation of the communities was done using the Louvain method from networkX. Figure 17 shows the resulting visualization, each node is a community. We can see how communities are connected between each other. The redder the node, the more users contained in that community, in the image red nodes have over 80 members, yellow and orange communities have over 62 members, and from blue to green maximum 42 members in the community. Webweb allows the tuning of the visualization changing the gravity and charge of the nodes, to expand the graph, changing the opacity of the connections to only see the nodes, or selecting specific nodes via writing their names, among other possibilities.

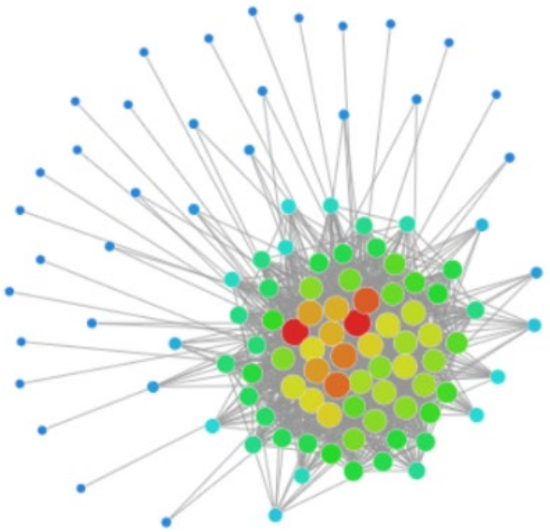


Figure 16: Example of communities

3.3.10 Geomaps

The last section is the "Geomaps" that shows the approximate location of the users from the dataset. To obtain this location we used again the Twitter API, obtaining the coordinates associated with the username. These coordinates were stored in MongoDB, and they are called by the Dash app layer to create the visualization.

The first visualization is the activity by country and continent, showing the number of tweets, retweets, or number of followers. An example can be found in Figure 18.

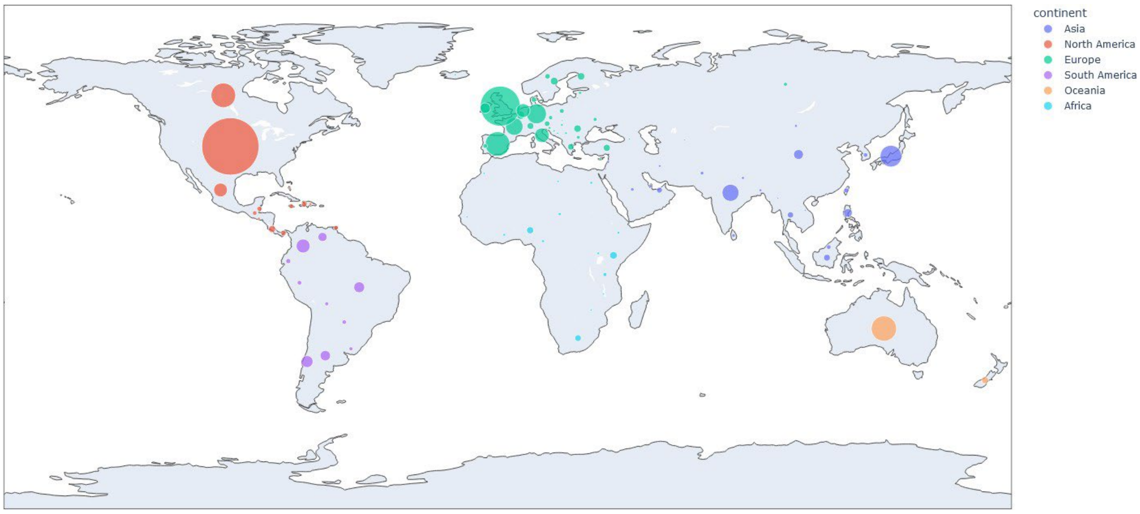


Figure 17: Visualization for activity per country.

The other visualization contains the approximate location of the users, as shown in Figure 19.

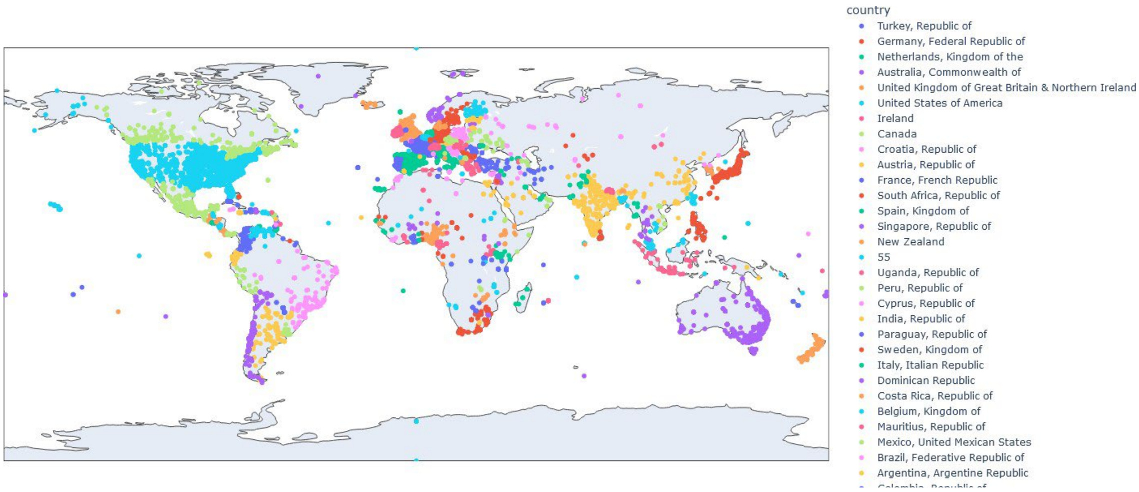


Figure 18: Location of the users from the CS community.

All the different analysis we designed were then put to test in different case studies to proof their potential and answer the research questions we established. These case studies were relevant to CS as we addressed important fields of this practice. In the next section the different cases will be covered along with a description of which techniques were used or were improved in that case. Also, we will discuss the contribution of the analysis when answering the research questions to validate this thesis.

4. Case Studies

Once we designed the pipeline for analysis, we put it to test and, therefore, test the usefulness of the platform in which the analyses are integrated. To test the pipeline, we addressed dimensions of importance for CS, and obtained useful insights from these topics. These insights from the topics translate into the answers to the research questions we established for the thesis.

In the following sections we will explore different case studies, with their motivation and specific methodology according to the timeframe when they were performed, alongside the state of the pipeline and the dashboard. Each study offered conclusions regarding our research questions, driving to general conclusions when putting all of them together.

4.1 Open Learning in Citizen Science

Many participants get involved in CS because they find it as a possibility to gain knowledge, to learn about science (Eitzel et al., 2017). Our first study focused on analysing the discussion on Twitter about learning when being part of CS. The first iteration over the pipeline occurred when analysing this learning dimension.

4.1.1. Motivation

As we know by now, the public participation in scientific research is a common practice in the last decades, therefore turning all these projects into CS projects (Kullenberg & Kasperowski, 2016). All the information that is generated by the participants is part of the scientific process as it is used to be analysed and extract conclusions (Hecker et al., 2018). Although, the exploitation of the information is on behalf of the researchers, the participants take part in these projects since it is an opportunity to learn (Bonney et al., 2016).

There is not only one way of learning in a CS project, there are three main ways described: i) first, the CS project is specifically designed for learning, meaning that the project is essentially designed to promote learning in all the stages; ii) another possibility is when projects initially are not intended to promote learning but they are redirected to offer some knowledge; iii) finally, there are CS projects in which all the practices are intended to promote learning in the stakeholders (Committee on Designing Citizen Science to Support Science Learning et al., 2018).

The web-based technologies have propelled new ways of learning, institutions, businesses, or other entities have taken advantage of social media to engage citizens and communities (Daume & Galaz, 2016). These platforms have arisen as a perfect via to promote collaboration, create connections, share findings and finally it could translate into learning inside the CS community (Aristeidou & Herodotou, 2020) (Liberatore et al., 2018). One of the main examples is the increasing activity in social networks from the CS scientists. In these social networks, these individuals share knowledge related to general understanding of the scientific processes or projects they are involved in. Other examples are the institutions that could act as facilitators of interactions, promoting the participation in online CS communities. This is directly related to scaffolding, which means that learners are supported and accompanied by during the learning process. Scaffolding could lead to deeper learning, but a highly scaffolded scenario may translate into situations in which informal participation could be restricted.

With this study we aimed at gaining understanding about the conversation related to the learning that happens inside the CS projects, with the networks formed in Twitter via interactions as means of analysis. To obtain this knowledge we wanted to check which were the most used hashtags, part of the lexical inventory helpful to determine topics of discussion. Along with hashtags, we wanted to determine who were the most important actors inside this conversation, and to do so we utilized the structural analysis techniques. The results obtained would be an initial approach to answer the research questions of this thesis.

4.1.2 Methodology

We downloaded 123,164 tweets from September 29, 2020 and January 27, 2021 related to CS in different languages such as English, Spanish, French, Dutch, Finish, among others.

Next, we next applied the filtering functions using words related to learning such as education, technology-enhanced-learning or personalized learning environments. This reduced the number of tweets to 11,761. These final set of tweets were used to analyse the hashtags and words in it to scrutinize the situation around learning in CS. The extraction of hashtags was undertaken via NLP extracting the words preceded by a hash symbol (#). Then, we isolated those tweets that were retweets, to count which hashtags are more retweeted.

Alongside this exploratory analysis of words and hashtags, we performed also performed the structural analysis focused on the network analysis. This was possible since the data from Twitter was processed to identify the links between users formed via the retweets and citations they give to others. These links were represented in a directed graph, in which every user is a node and. the edge the retweet or the citation.

From the filtered tweets related to learning, we obtained a network with 3,937 nodes and 2,634 edges. Besides, we applied measurements of centrality such as the in-degree, out-degree, eigenvector and betweenness, already explained in section 3.2.2 Structural analysis.

4.1.3 Results

To determine if there is a conversation about education, from the exploratory analysis we saw that only 9.5% of the total tweets of the period studied are about education topics, quite below the result we expected (25%) due to the close relation with education and learning in the CS projects.

Continuing with the pipeline for analysis, we analysed the content to gain knowledge about the topics, trends, and interests of the users. This lexical inventory inside the tweets can help as an overview of the interests of the community. Figure 19 shows the results of the hashtag analysis, showing the top 14 hashtags in those tweets about learning. The most used one is #SDGs, with 547 retweets. Then, we have the first education related hashtag, #education, with 418 retweets. The following hashtags were #outdoorlearning (226 retweets), #schools (148 retweets) and #homeschooling (128 retweets). The education related hashtags represent the 32.28% of the analysed hashtags, which led us to think that learning and education is not the focus in the information dissemination of the CS projects in Twitter.

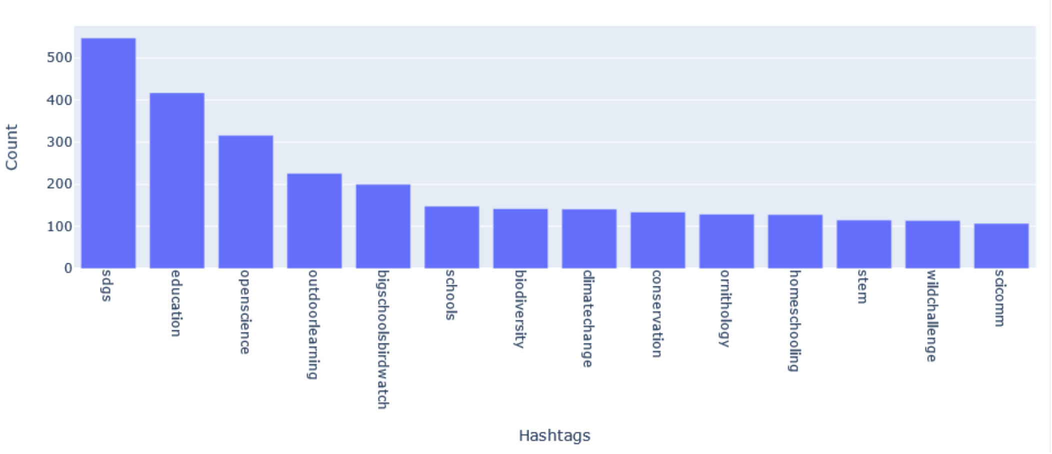


Figure 19: First study - Open learning - Results from the hashtag analysis

Next, we did the structural analysis of the networks in the platform. This analysis allows the discovery of the most prominent accounts, which is useful to determine who are the more active users or users producing more amount of interesting information. In this first case study, the network that was built using the retweets had 301 strongly connected components and 2,661 weakly connected ones. These nodes formed 301 different subgraphs being the biggest one a graph containing 1,419 nodes. This graph is shown in Figure 20, in which we see that the redder and bigger the node, the higher the degree or number of received retweets. In this graph we can see peripheral nodes with a high number of received retweets, but unconnected to others, this meaning that they are isolated users that are highly shared by others but not active retweeting. Besides, they are not central nodes, so they do not act as information spreaders as they are not connected to others. Those nodes in the centre of the graph received less than 60 retweets, but act as bridges in the information spreading.

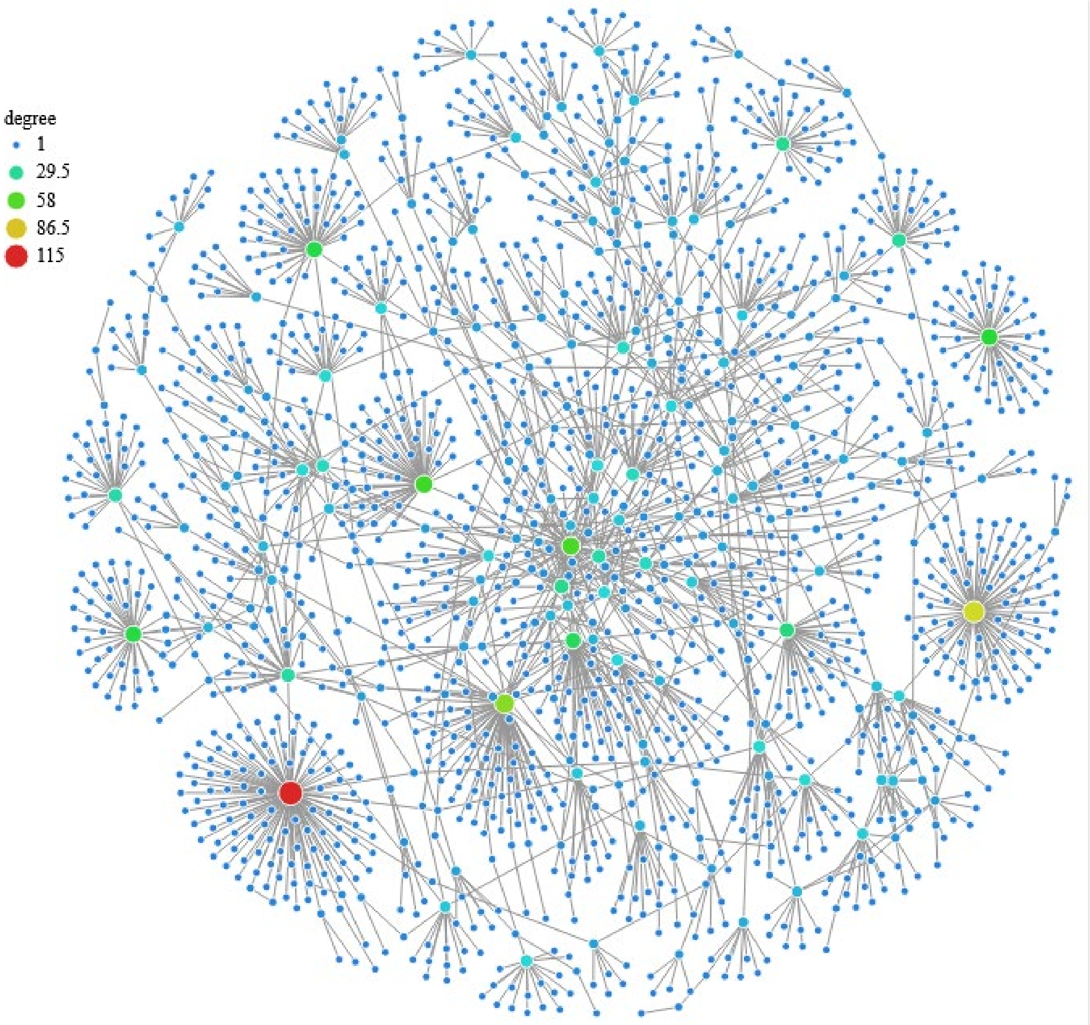


Figure 20: First study - Open learning - Network of retweets from the learning dimension

We previously explained that, from a graph, the centrality measures allow us to obtain a better understanding of the connections between the users and the structure of the network, unveiling the most prominent actor inside the network, therefore, those more active sharing or producing information. This is valuable information to comprehend the dynamics and for stakeholders interested in prominent accounts that act as information spreaders or users that produce content highly shared by others or monitoring conversations. These measures are obtained from the networks formed with the retweets. The data is in Table 2 , which shows the values for centrality for users with higher indegree, meaning that Table 2 is ordered to show the top 10 of users that received more retweets. Alongside the indegree, we show the Outdegree (number of retweets given), the Eigenvector which assigns to each node a score for authority in relation to the nodes it is connected, and the Betweenness which represents the number of times this node is in the shortest path between other nodes (an important measure to find information spreaders).

Name	InDeg		OutDeg		Eigen.		Betw.	
	Val	R	Val	R	Val	R	Val	R
RSPB_Learing	92	1	0	2460	1.80E07	115	0	1367
Heinz VHoenen	81	2	0	2388	1.19E07	117	0	1003
The Alice Roberts	77	3	0	2464	1.14E07	118	0	1386
NOAAeducation	69	4	0	2208	0.2006	7	0	109
Sofiessketches	56	5	0	2300	8.26E08	122	0	581
Nypl	43	6	0	2324	6.35E-08	125	0	773
Truejainology	42	7	0	2608	6.20E08	126	0	2165
F_pilla	41	8	2	75	3.81E05	83	5E06	39
Seabird_watch	36	9	2	163	0.2655	4	4E05	22
TaiwanBirding	34	10	0	2281	5.02E08	130	0	455

Table 2: First study- Open learning - Centrality measures of the most retweeted accounts

4.1.4 Discussion

This first approach to the pipeline for analysis offered preliminary insights to the state of the conversation about learning in the CS community in Twitter. Our first approach was to apply the filters and check the number of tweets related to open learning. The resulting percentage of tweets about this issue in relation to the total number, which was low according to what we expected. But we find conversation, so we get a partial answer to RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?) (Eleta & Golbeck, 2014).

The hashtag analysis show that the most used one is #SDGS, which is not a word directly related to education, but there is a specific SDG related to education: SDG4, quality education. Between the next results we found words related to education such as: #education, #outdoorlearning, #schools and #homeschooling. This drove us to think that the rest of the hashtags have and underlying relation with education that cannot be captured via a single hashtag analysis, we would need to deepen in the content. The content analysis performed using the count of hashtags was aimed to answer RQ2 (What is the lexical inventory in these conversations, i.e., hashtags, most common words, etc.?) and RQ3 (What are the main topics? Are these topics evolving in time?) of this thesis, to find the lexical inventory and get a first glimpse of the topics of interest in the set of tweets. Hashtags are an important feature to unveil the topics of discussion since users write them to label their tweets according to the topic of the text (Ferragina et al., 2015).

The structural analysis offered a good insight into the most important accounts. The graph we plotted helps distinguish the allocation of the users in those small communities formed via the interactions. This was an initiator of the idea to perform community analysis. The centrality measures gave good information about those users. The relation between InDeg and OutDeg for prominent accounts confirm the trend that happens in every social media. This trend is that normally influential accounts do not interact as much as the number of interactions they get (Zhang, 2010). However, these accounts of high Indegree, according to Betweenness and Eigenvector, are not positioned in high ranks of information flow, so it could be interesting to check other accounts with higher values to determine who are acting as information spreaders. This structural analysis gives information about the important actors inside this piece of the dataset, so the approach seems to work to answer to RQ4 (Can we define main actors through the SNA structural analysis?).

4.1.5 Conclusions

As a summary of this study, we applied the first designed techniques in an approach to answer the RQs. This case study show how education related topics are discussed inside the Twitter CS community. The conversation seems scarce in comparison to how important the learning dimension is in CS (Committee on Designing Citizen Science to Support Science Learning et al., 2018). Only 6 out of top 14 hashtags about learning are directly related to education, so with these results it seems like outdoor learning combined with the ornithology and birdwatch is the most prominent topic. It must be noted that SDGs, conservation, and biodiversity are highly used hashtags, so maybe there is a focus on learning about these topics, but more research would be needed.

The conversation seems to start with users that are highly retweeted but, these users, do not retweet other accounts that much so there could be interesting information from nonfamous users that is not being shared or reaching the whole community. Besides, users with lower values of received retweets are those more active in retweeting others.

This work shows how Twitter is used with educational purposes, but it seems to not be being used to its full potential. Besides, the replication of this study in the future could provide a time analysis and see if this situation is changing in time, or it could be applied to other topics.

4.2 Understanding the discussion of CS around SDGs

Once we had results from the first study, we improved the pipeline and addressed another topic, the SDGs. We already know how important CS is for the SDGs (Moczek et al., 2021). Different techniques were applied to analyse the state of discussion about SDGs in the CS community in Twitter.

4.2.1 Motivation

United Nation´s 2030 Agenda for Sustainable Development introduced the 17 SDGs in 2015. It has 169 associated targets, that were aimed at creating policies and drive the actions towards sustainable development (Department of Economic and Social Affairs, 2015). Several of the actions that the SDGs want to undertake can affect CS directly by encouraging participation, education, partnership, global citizenship, and others (Shulla et al., 2020).

Besides, CS activities can contribute to the implementation of the SDGs since this community works toward generating social cohesion, which is an important factor in the pursuit of the SDGs. Therefore, CS actions can not only bring advance in scientific knowledge (Committee on Designing Citizen Science to Support Science Learning et al., 2018) but it could also trigger societal transformation leading to reaching the goals purposed by the 2030 Agenda (Elliott & Rosenberg, 2019).

There are various studies measuring the impact and contribution from CS to the SDGs but, given the lack of a standardised practice to measure the impact of the CS community in relation to the SDGs (Wehn et al., 2021). In this study, we aimed to test our pipeline for analysis to gain knowledge to this respect. Measuring the impact of projects in social media has been a concern years before our study (Cox et al., 2015) (Davids et al., 2019), so we tried to supply a way to obtain this knowledge.

CS projects normally establish their accounts in Twitter and try to create communities and disseminate their information (Wiggins & Crowston, 2011). From previous studies we are aware of the dynamics of this community, as it was described in Mazumdar and Thakker work (2020) in which they stated that the main action performed by the users of this community was retweeting. They also tried to give tools and advice for these accounts to increase their presence and reach. With this background and given the fact that we did not find many studies regarding the relation to SDGs, with this study we aimed to offer a panoramic view of the state of the discussion about SDGs in the CS Twitter community. We explored the content analysis improving the analysis of hashtags including the time analysis, adding the language detection and the TF-IDF calculation, which will unveil the languages present in our dataset and the most important words in the collection of tweets, respectively. Also, we introduced the topic modelling analysis, which will give more information about the topics of discussion, a classification of tweets by SDGs using machine learning, destined to classify the tweets according to which SDG they address, and we also enlarged the structural analysis to unveil more dimensions of the social networks formed by the users and the features in Twitter.

4.2.2 Methodology

This study continues with the pipeline we first proposed in the Open Learning analysis provided by the techniques we implemented in the dashboard. First, we started the exploratory data analysis, with a data collection from the 30th of September 2020 to the 20th of June 2021. All these tweets come from the Lynguo platform which collects tweets containing keywords related to CS.

The number of tweets for this study ascended to 275,868 tweeted by 88,974 unique users. We had a total number of 176,728 retweets and 136,898 unique results for these retweets. Therefore having 39,830 duplicates, users retweeting the same tweet several times. We also had 4,441 replies. It is necessary to consider the possibility of people tweeting about CS without using keywords related to CS is a possibility. So, all these uncollected tweets could even enrich our findings.

All these tweets were passed through a second filter composed of SDG related words such as: SDG, Clean Water, Affordable and clean energy, and other words related to the 17 SDGs in order to retrieve all the tweets related to this subject. This resulted in 19,543 tweets by 10,186 unique users. The distribution of original and not duplicated tweets vs retweets was: 5,960 unique and 13,583 retweets.

Once the data were cleaned and prepared, we applied different techniques to analyse them. To identify the languages in the dataset we trained an NLP based model. We used a Multinomial naïve Bayes classifier trained with a dataset containing 10,337 texts in 17 different languages.

We next performed the hashtag analysis. In this study we wanted to introduce a new level of refinement, so we divided the hashtag analysis into 3 levels, to explore all the distributions of hashtags. The first level is the general count of the hashtags, combining those appearing in the retweets and in the normal tweets, which will serve as a comparison point when checking the hashtags about SDGs. Then we isolated those tweets starting with RT@X, the retweets, and counted the hashtags inside them. We did the same for those tweets that were not retweets. Then, we presented the results in bar graphs created with Plotly.

Alongside the new level of analysis for the hashtags, we first used the temporal analysis of hashtags, a technique that allows us to see the evolution of usage of these terms. Given the same 3 levels of analysis, we then extracted the hashtags alongside the date of publication or retweet and plotted the count of the results in time using plotly once more.

We complemented the hashtags' results with the most frequent terms analysis in our tweets. Most common words are a good measurement, since it allows the analysis of frequent terms in relation to the topics of discussion, or even can help determine which terms can be filtered out. To do the analysis of most frequent terms, we separated all the words in the tweets and cleaned the hashtags and usernames (preceded by # or @). We cleaned the most common stopwords (conjunctions or prepositions) from the languages present in our dataset. Finally, we presented the results in a wordcloud.

The relevance of the words in our dataset is measured using a statistical technique called TF-IDF (Term Frequency-Inverse Document Frequency) previously explained in previous chapter. Therefore, we can know the weight of these words in the different documents (tweets). The results were also displayed in a wordcloud.

Next, we wanted to further analyse the discourse, to obtain a better understanding of the important subjects for the members of the community, and to do so, it is interesting to explore the subjects or topics the users address. In this study, we applied topic modelling and specifically, a Latent Dirichlet Allocation model, as it was explained before, to a cleaned corpus of tweets. Applying topic modelling offers a better insight to the topics of discussion, which can be complemented with the analysis of hashtags and most common words and TF-IDF. Alongside applying this technique, we also applied a topic coherence analysis, meaning that we used different LDA models with different topic sizes and we measured the similarity of those topics extracted by the models, aiming to obtain the best performance from the algorithm and obtain the best results for the extracted topics. We tested the different models with different topic sizes, which is commonly called topic coherence test, and we finally decided to use a model destined to extract 17 topics to match the number of SDGs since this topic size was the highest in performance. Once we performed the analysis with the model, we extracted the keywords from the topics and represented the topic modelling distance visualization and calculated the number of tweets by topic.

Another exploration was the SDG classification, we classified the tweets according to which SDG they were addressing. This technique was specially designed for this case study so here we introduce its methodology. We trained a BERT classifier, which stands for Bidirectional Encoder Representations from Transformers, and it is designed to perform classifications using deep neural networks architecture processing all the text at once to capture all the relationships between words and phrases. We used the Bert Bases Multilingual Uncased pretrained model due to the multiple languages in our dataset. To compile it, we used the Adam optimizer using a learning rate of 2e-5 and an epsilon of 1e-08, alongside Sparse Categorical accuracy for the metrics, as these are the recommended settings to obtain the best performance out the fine-tuning of BERT according to what authors recommend (Sun et al., 2019).

The training was performed using 57,483 tweets downloaded directly from the Twitter API. They were obtained using a filter to retrieve tweets containing SDG1 to SDG17, ODD1 to ODD17 (in French) or ODS 1 to ODS17 (in Spanish), being this a quite restrictive query to collect tweets clearly linked to each SDG and avoid misclassifications. These three languages were the most predominant in our filtered dataset with tweets about SDGs, as it will be presented in the next section with the results of the study. 80% of the tweets were the training sample and the 20% remaining were used to test the model. All the tweets were processed to avoid duplicates, lemmatized, stemmed and the retweets were converted to the original tweet.

Once we had the classification to started with the network analysis. First, we took the classified tweets by SDG, and we checked the number of users that retweeted them. We had a total number of 12,144 nodes between users and tweets (7,196 users and 4,228 tweets). Its graphical representation was trimmed using the k-core method explained in previous chapter. This network was a two-mode network in which the nodes belong to two different groups, in this case tweets and users that retweeted them.

Another network analysis we performed was based in the following relation between users, since this connection between users also provides important information such as most important actors based in the number of followers, or the composition of the communities checking the ratio of followers and followed accounts. To get this information we used the Twitter API and downloaded the number of followers and following users to each user in our dataset. We had 10,186 unique users and represented the numbers in a scatterplot.

Finally, since retweeting is one of the most important behaviours in CS, we analysed the users with more retweets received and given forming a graph. Using the network package, we created a directed graph containing 1,234 nodes and 2,754 edges (trimmed using the k-core algorithm). To end the study, we calculated the centrality measures of the previous graph. Indegree, Outdegree, Eigenvector and Betweenness were calculated for the users.

Some of the analyses we used in this case study were an evolution of those we applied in our first case study, and some were new approaches and new techniques designed to provide better answers to the RQs.

4.2.3 Results

The first exploratory analysis about if there is multitopical and multilingual CS conversation in Twitter showed that the tweets containing keywords related to the SDGs represented a 7.08% of the total collected tweets. This was below our expectations since we understand that 2030 Agenda and SDGs are important subjects nowadays.

When we got the results from the language detection, we had a 78.2% of tweets in English, 18.2% in French, 2% Spanish, 1% German, 0.2% Portuguese, 0.6% Dutch, 0.1% Swedish, 0.4% Italian and 0.5% Danish. In the light of the results, it seems clear that the most used language in this discussion was English. This result was not surprising given the importance of English as business, scientific, institutional, and international language.

Continuing with the content analysis, the first analysis was to check which hashtags were being used inside this conversation to unveil what topics are important for the CS community when discussing about SDGs. The combined analysis (hashtags in the retweets and in normal tweets) gave us as result that #sdgs was the most used one with more than 8000 appearances, as it is expected. Due to the number of appearances was so hight that we left it out of the visualization to check the next hashtags. #climatechange, #openscience, #2degreesc (a company that fights global warming) and #cs_sdg2020 are the most cited hashtags. The rest of the hashtags were related to climate change. The distribution of hashtags analysis is displayed in Figure 21.

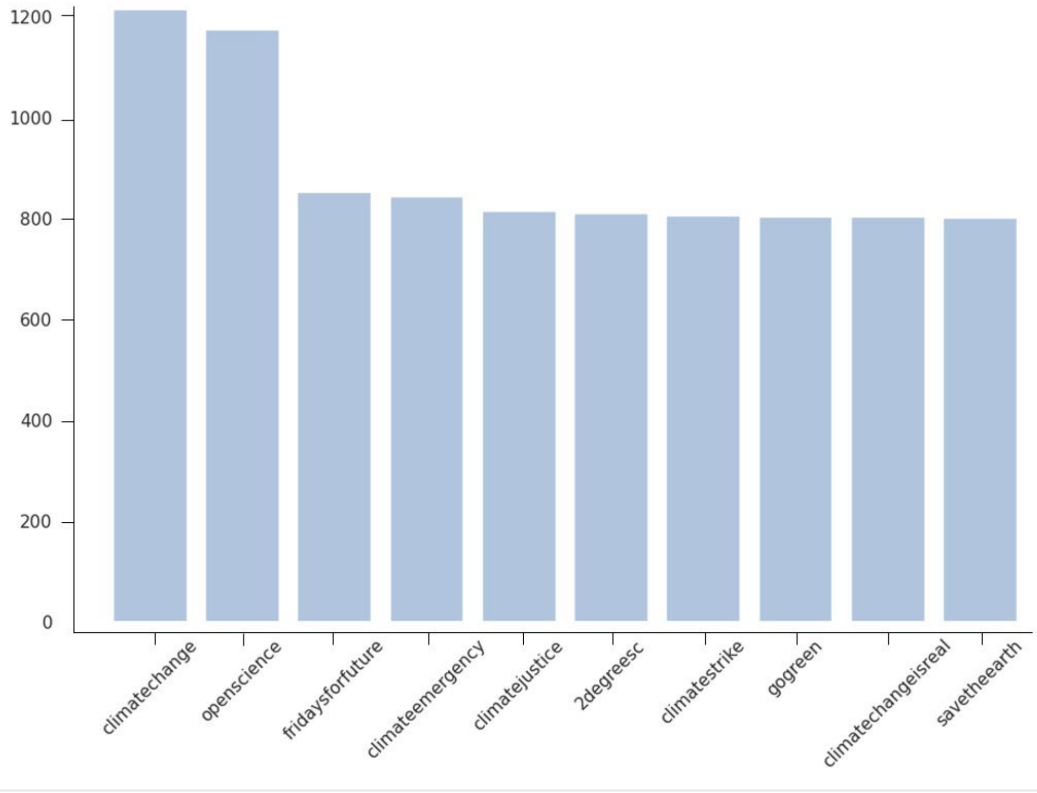


Figure 21: Second study – SDGs - Top 10 used hashtags in tweets

When we checked the retweets individually, we had a similar picture (see Figure 22). The first one was #sdgs with more than 8,000 retweets followed by #climatechange with 1,200 retweets, #openscience (almost 1,200 retweets), and #fridaysforfuture (800), which is a youth-led climate movement which protest against the lack of action on the climate crisis. The rest of most used hashtags in retweets are related to climate change again.

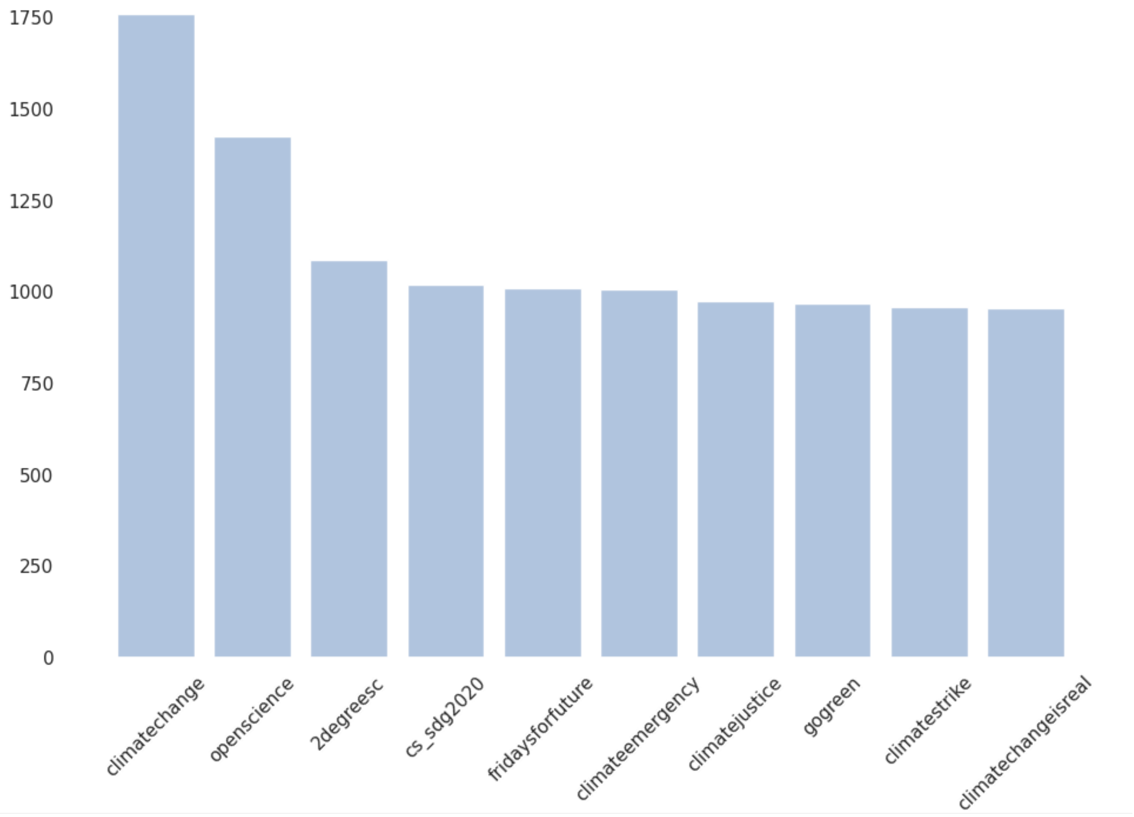


Figure 22: Second study – SDGs - Top 10 retweeted hashtags

The hashtags used outside the retweets again showed a similar picture (see the distribution in Figure 23). #sdgs continued to be the first one followed by #climatechange, #2degreesc, #openscience.

Climate change inside the SDGs discussion was an important topic, in the lexical inventory #sdgs appears many times, and the other hashtags have a relation.

Once we knew the most used hashtags and retweeted hashtags, we decided to perform the analysis of their usage and retweeting through time. This type of analysis allows to discover events in time that could trigger a higher usage of words or hashtags. According to the timeframe of collection of tweets, we analysed the evolution from late September 2020 until June 2021.

The first results when we checked the complete set of tweets showed that from September until October 2020 the hashtag #cs_sdg2020 was the most used one but suddenly the hashtag #sdg overtakes the first position for a short period of time in early November. In fact, the high number of appearances of #sdg in this period makes it the most used one, but the temporal analysis shows that it was only in a short period as Figure 24 shows, probably due to a high number of conferences (EST, SDG-PSS, OECD) and updates about SDGs occurring those days.

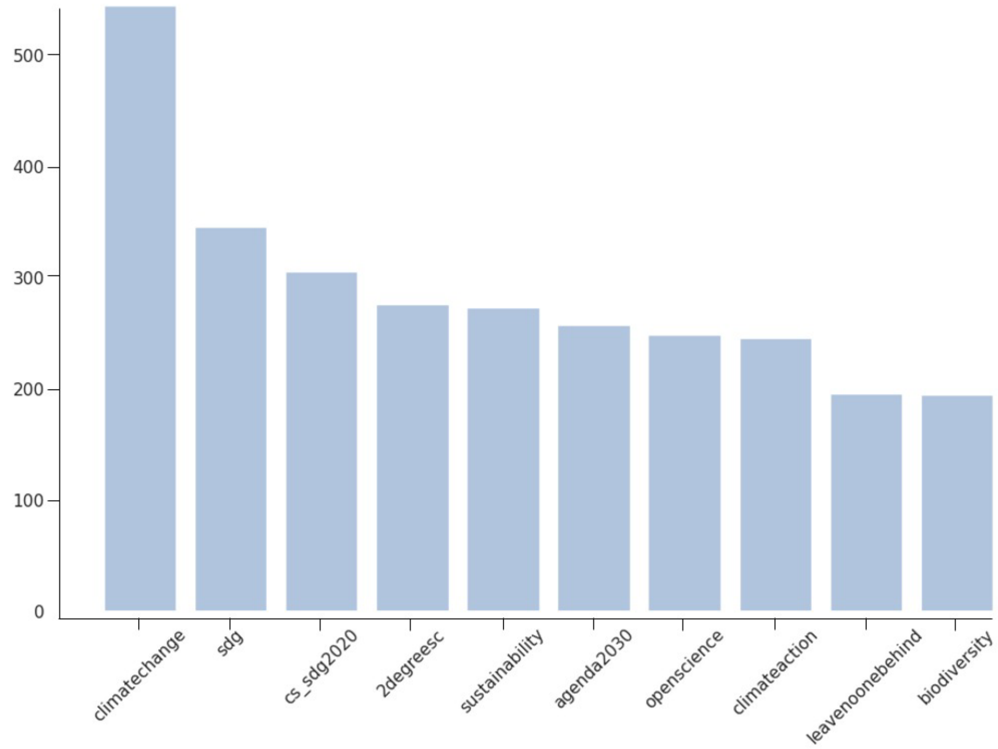


Figure 23: Second study – SDGs - Top 10 used hashtags outside the retweets

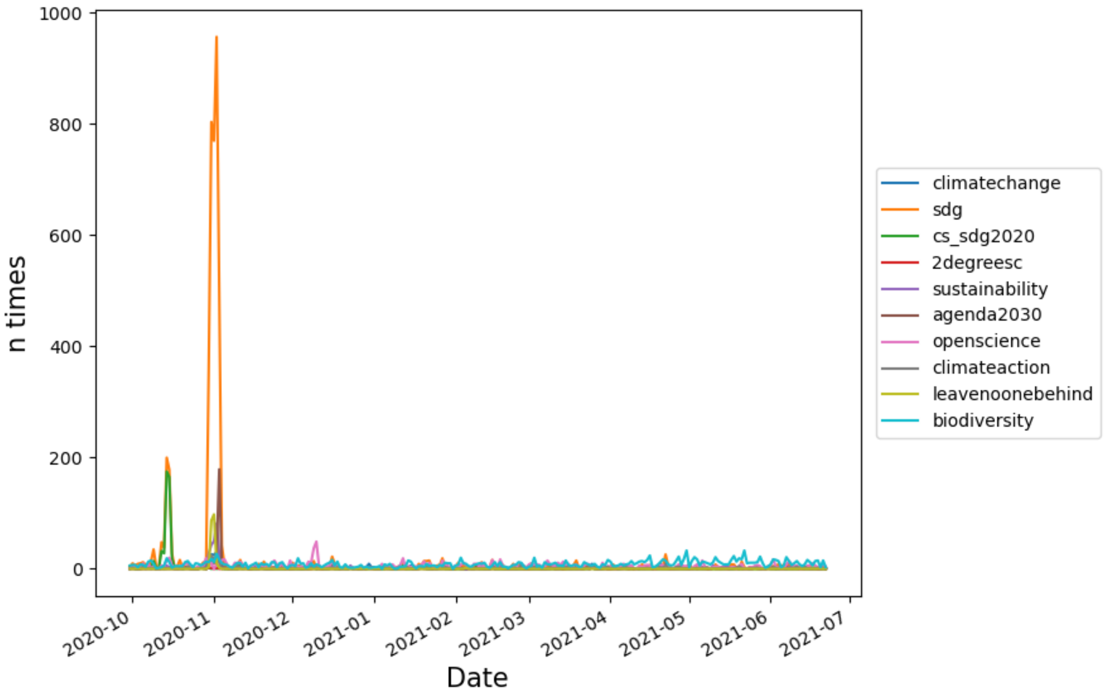


Figure 24: Second study – SDGs - Temporal evolution of top 10 most used hashtags

The high peak of #sdg does not allow a proper visualization of the graph, we then analysed the evolution of the hashtags without the retweets (meaning pure usage therefore). Figure 25 shows again the high peak of #sdg and we also see a high peak for #agenda2030. From November 2020, the most used hashtag afterwards is #biodiversity, being its usage more stable in time than the others although it appears with the least occurrences in the top 10.

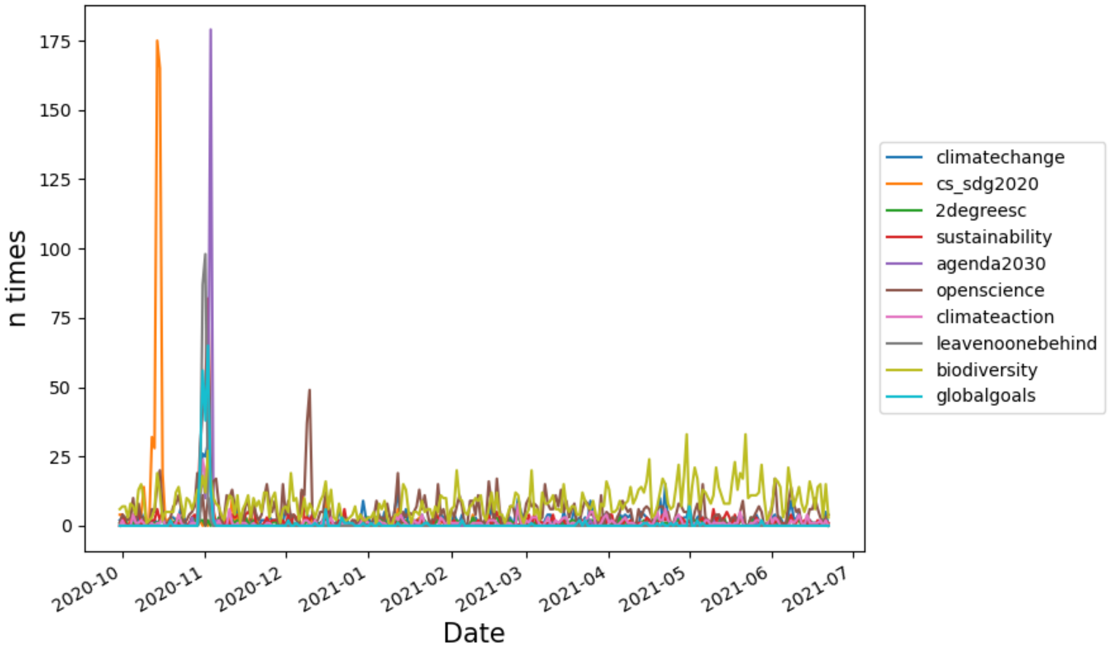


Figure 25: Second study – SDGs - Temporal evolution of top 10 used hashtags outside the retweets

Lastly, we checked the situation of the retweets of hashtags. Figure 26 shows that, right before #sdgs again, which is by far the most retweeted one, we have #climatechange.

Hashtags are an important part of the tweets, but it is interesting to check the other words used in the set of tweets. In the wordcloud from Figure 27, we can see that Climate, Citizen science, help, join, or change are the most used ones because they are displayed in bigger size. Climate awareness of the users seems evident as the word Climate appears as one of the most used ones. There are words related to other topics like inclusiveness, policy, health, social or work.

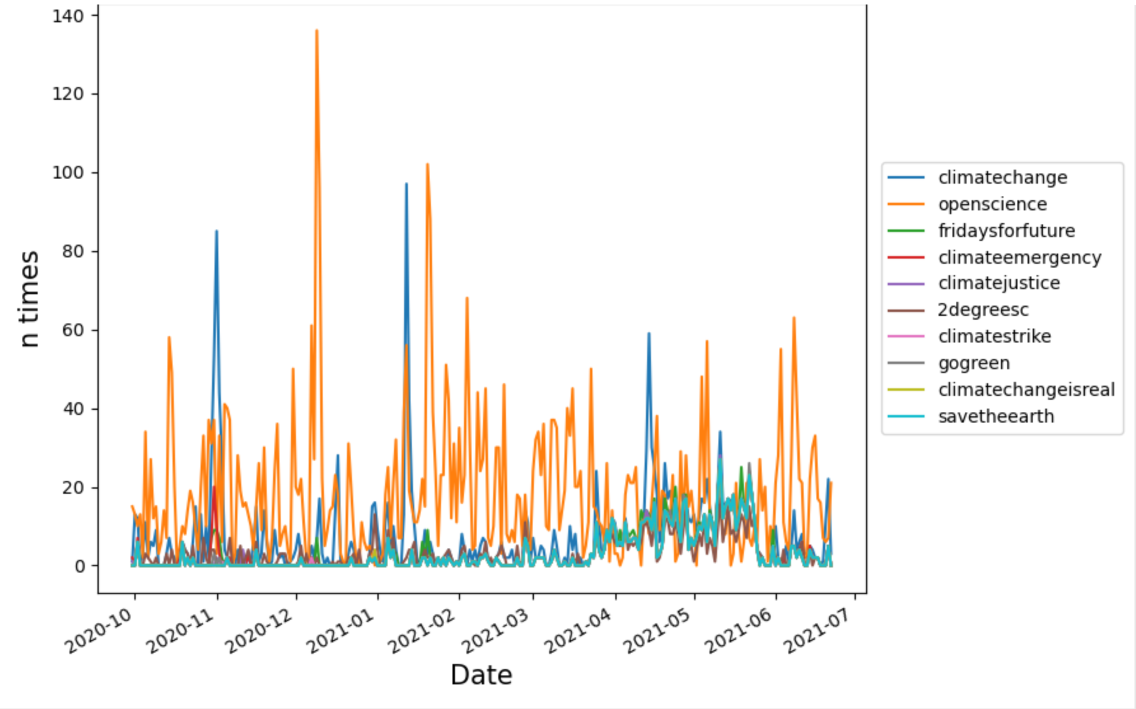


Figure 26: Second study – SDGs - Temporal evolution of top 10 most retweeted hashtags

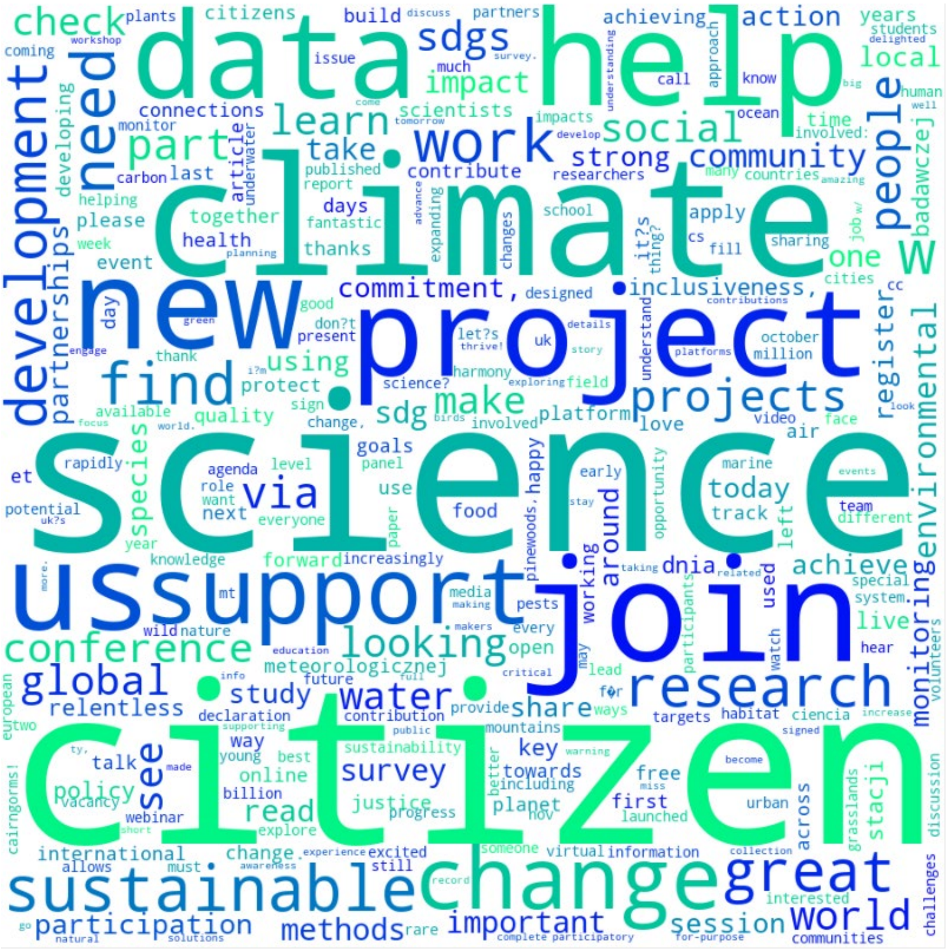


Figure 27: Second study – SDGs - Example of wordcloud with the most used terms

With the most used words unveiled, we checked their importance inside the collection of tweets using TF-IDF analysis. Figure 28 shows that words like social media, mediaquality, or citizens appear in our word cloud as main results. Then we can see words like world cities day, charity or harmony and many others related to climate, biodiversity, and development. Once more the importance of the topic climate change is clear form this analysis. TF-IDF analysis is an interesting approach to discover keywords for different subjects.

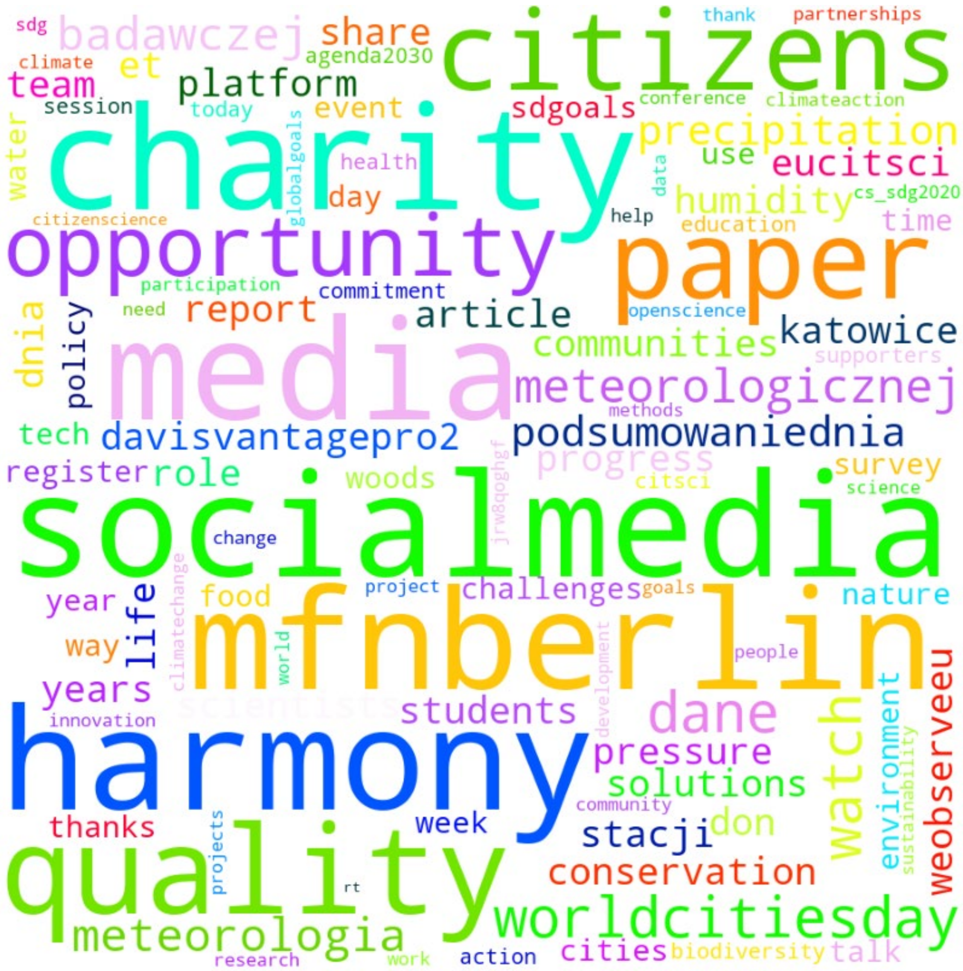


Figure 28: Second study – SDGs - Example of TF-IDF analysis

The lexical inventory was calculated with the previous analyses, but the content analysis was not finished. Hashtags and most common words alongside TF-IDF drove us to think that climate change, in the framework of the SDGs, was the predominant topic. To check if that was true, we performed the topic modelling analysis. Before getting the results, we performed the topic coherence analysis and, as it was mentioned before in 3.2.1 Content analysis, the results shows that we could choose any topic size above 12 as it can be seen in Figure 29, showing the results of the topic coherence test, in the Y axis we see the score for the topic modelling and in the X axis the number of topics in each test. The highest score results with a topic size of 17 and with topic sizes over 25. Thus, we decided to choose 17 topics like the number of SDGs.

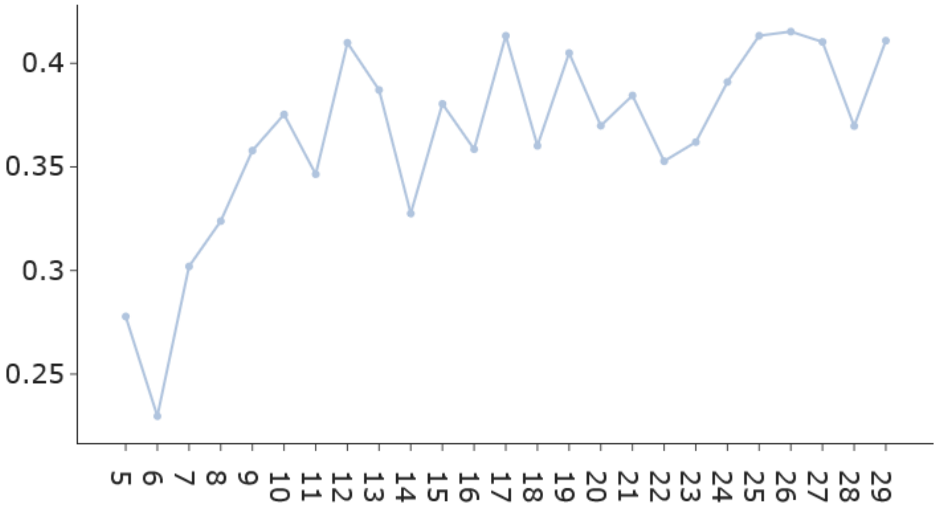
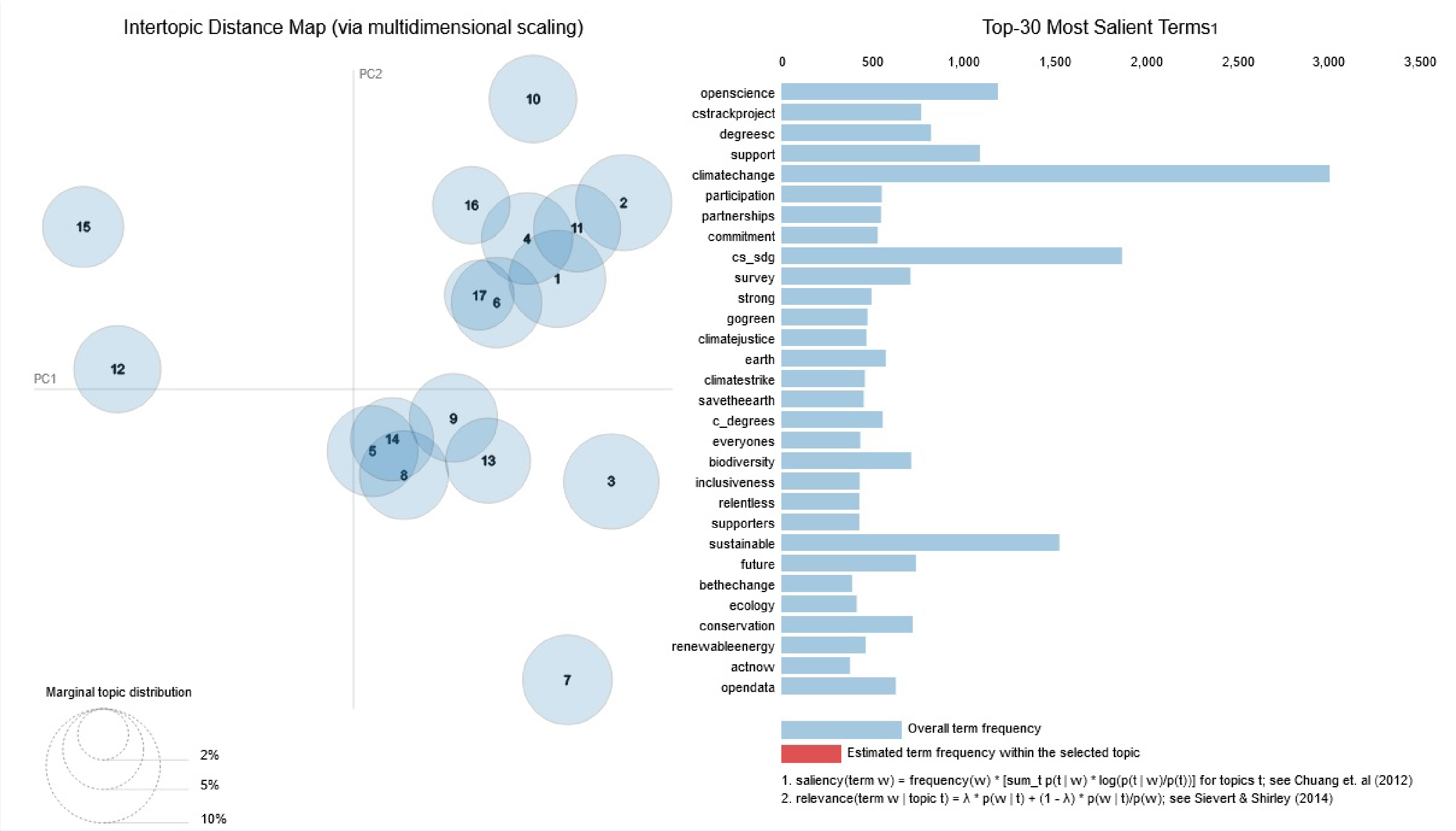


Figure 29: Second study – SDGs - Topic coherence

Once the model extracted the 17 topics, we represented the similarities between topics alongside the main keywords. Figure 30 shows the intertopic-distance map, in which we see that some topics overlap, meaning that they have words in common. As an example, we see that topic 6 is about climate change discussion from the perspective of education and some topics are distributed close to this meaning, topic 1 (environmental monitoring), 17 (a tree health early warning system powered by CS) or 4 (Agenda to follow to have a healthy planet in the future) for example, so some words are in common. We can compare it to topics like 7 (Open science) or 15 (energy impacts on the conservation of our ecosystems), we see that there is a big difference between them. On the right side of Figure 30, we see the most salient terms, which are the most important or noticeable. Some of the are quite generic but analysing them and finding uncommon terms is what marks the difference between topics.

We also explored all the topics one by one and their keywords, which are provided in Table 3, in which we can see for example Topic 6 with its keywords: support, participation, partnerships, commitment, strong, everyones, inclusiveness, relentless, supporters, lead. And an explanation of the topic: Topic around how everyone's participation is important in sustainable development, mostly centred around water usage and climate change.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 30: Second study – SDGs - Topic modelling intertopic distance map with the frequency of the terms

Topic	Keyword	Definition
Topic 1	join, weobservereu, work, global, globalgoals, data, register, education, monitoring, forward	This topic reflects the discussion around the project WeObserveEu, which is an ecosystem of citizen observatories for environmental monitoring.
Topic 2	sustainable, globalgoals, innovation, tech, development, riskcentre, saeedbaygi, mahsamoulavi, techhealth, fintech	Impact of technology in sustainable development, considering mostly health issues.
Topic 3	cs_sdg, conference, contribute, session, data, community, climatechange, learn, online, october	Citizen science SDG conference discussion around how to contribute data to the community to learn about climate change
Topic 4	planet, agenda, live, left, years, people, days, world, support, year	Agenda to follow to have a healthy planet in the future.
Topic 5	climatechange, help, data, study, level, great, security, nocnews, love, come.	Discussion started by user13 (see Table 5) around how to make online data more accessible for all.
Topic 6	support, participation, partnerships, commitment, strong, everyones, inclusiveness, relentless, supporters, lead.	Topic around how everyone's participation is important in sustainable development, mostly centred around water usage and climate change.
Topic 7	openscience, user2, survey, join, research, opendata, important, international, developing, eu_h.	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 5.
Topic 8	climatechange, quality, community, world, share, week, love, join, justice, related.	Topic about the challenges that the world needs to tackle to takcle the climate change problem. There are discussions about finances, energy and health.
Topic 9	user4, community, cs_sdg, free, like, climatechange, sustainability, great, facebook, help	SDGs in general. Most of the information comes from user 4 (Table 5) who retweets a lot of tweets about SDGs
Topic 10	sustainability, work, apply, goal, book, user3, looking, goals, connections, globalgoalsun	Discussion around user3 (see Table 5) about the use of technology to reach SDGs.
Topic 11	cs_sdg, policy, data, link, good, details, advance, makers, work.	Citizen science SDG conference discussion about what policies should be implemented around data.
Topic 12	climatechange, gogreen, climatejustice, degreesc, climatestrike, savetheearth, climate, c_degrees, youthforclimate, ocean.	Another topic on climate change focusing on 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	data, environment, crowd, decade, waste, sobre, methods, climate action, para, geospatial.	Topic that combines the impact of waste in the environment and the importance of having geospatial open data.
Topic 14	opendata, wind, temperature, humidity, pressure, summary, badawczej, meteorology, katowice, bdzin.	Bot that tweets repeatedly about the weather conditions in Katowice.

Topic 15	future, earth, biodiversity, conservation, bethechange, renewableenergy, ecology, actnow, degreesc, saveourplanet	Topic on how individual usage of energy impacts the conservation of our ecosystems.
Topic 16	youtube, qoghgf, industry, good, datascience, director, covid, bigdata, wednesday, health.	Another bot tweeting about general SDGs news.
Topic 17	woods, trees, designed, involved, protect, pests, species, observatree, early, warning.	Topic around the project ObservaTree, a tree health early warning system powered by CS.

Table 3: Second study – SDGs - SDGs Topics, keywords, and definition. Some topics contain keywords like userX, this is due to the anonymization of individuals

Finally, it was interesting to check how many tweets we had per topic. Figure 31 shows that the tweets were distributed nearly equally apart from topic 9. This was as result of user4 retweeting a lot of content related to SDGs. Besides, what we see is that there were SDGs that did not have its own topic, except for SDG13 Climate Action which was part of the discussion in several topics.

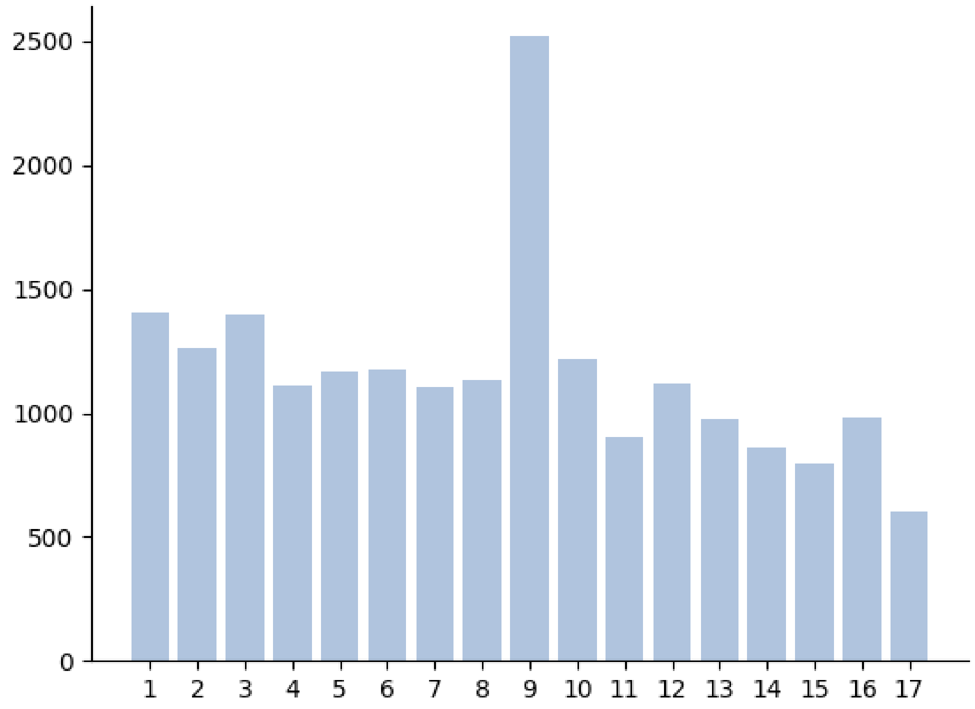
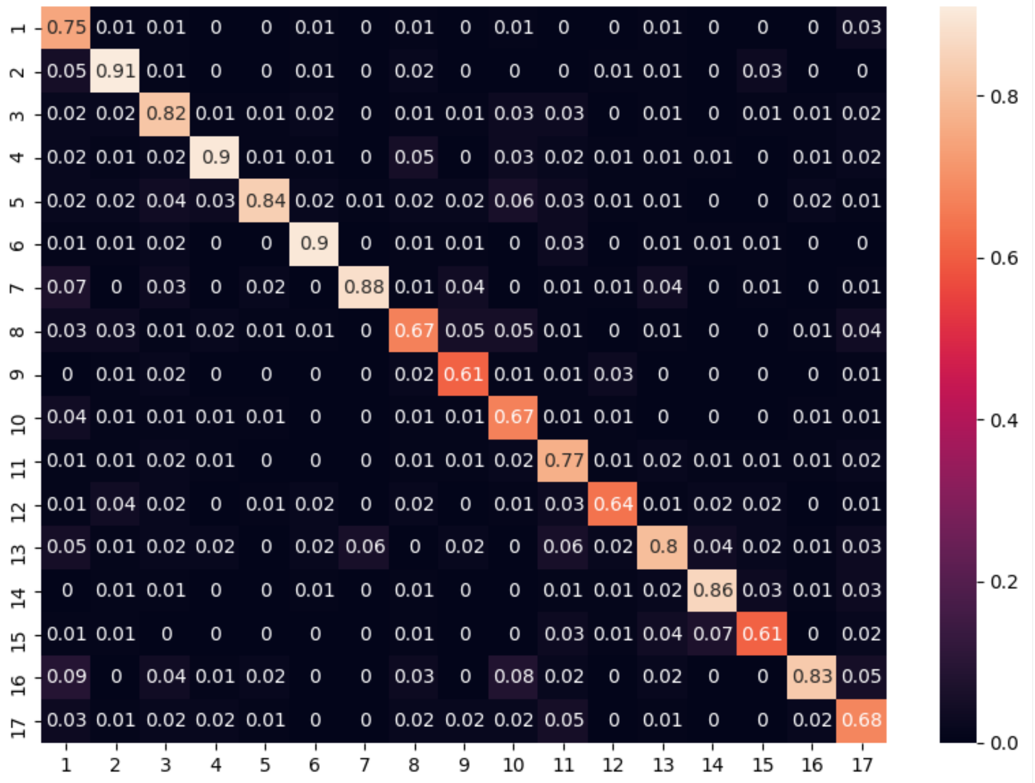


Figure 31: Second study – SDGs - Tweets per topic

Once we unveiled the topics of discussion, it was checked which SDGs were more addressed by the CS community on Twitter. This analysis, as stated before, was specific for this case study but it is aligned with the content analysis. As we previously explained, to perform this task we used the BERT classifier. This classifier once it was trained, offered us a performance that we are going to explain with the results from the confusion matrix and the score.

The confusion matrix in Figure 32 shows in the diagonal that, for example, 75% of the tweets about SDG1 are correctly assigned while a 5% of these tweets are wrongly assigned to SDG2 as it can be seen in row 2 column 1. Most of the SDGs have an F-score higher than 0.7 which is a good score given the small size of the training part (any score between 0.70 and 0.90 are good results (Zou et al., 2016)). The overall F-score was 0.82. This performance could be improved just by adding more texts to train the model. Table 4 shows the itemised information for precision by SDG.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.75	0.01	0.01	0	0	0.01	0	0.01	0	0.01	0	0	0.01	0	0	0	0.03
2	0.05	0.91	0.01	0	0	0.01	0	0.02	0	0	0	0.01	0.01	0	0.03	0	0
3	0.02	0.02	0.82	0.01	0.01	0.02	0	0.01	0.01	0.03	0.03	0	0.01	0	0.01	0.01	0.02
4	0.02	0.01	0.02	0.9	0.01	0.01	0	0.05	0	0.03	0.02	0.01	0.01	0.01	0	0.01	0.02
5	0.02	0.02	0.04	0.03	0.84	0.02	0.01	0.02	0.02	0.06	0.03	0.01	0.01	0	0	0.02	0.01
6	0.01	0.01	0.02	0	0	0.9	0	0.01	0.01	0	0.03	0	0.01	0.01	0.01	0	0
7	0.07	0	0.03	0	0.02	0	0.88	0.01	0.04	0	0.01	0.01	0.04	0	0.01	0	0.01
8	0.03	0.03	0.01	0.02	0.01	0.01	0	0.67	0.05	0.05	0.01	0	0.01	0	0	0.01	0.04
9	0	0.01	0.02	0	0	0	0	0.02	0.61	0.01	0.01	0.03	0	0	0	0	0.01
10	0.04	0.01	0.01	0.01	0.01	0	0	0.01	0.01	0.67	0.01	0.01	0	0	0	0.01	0.01
11	0.01	0.01	0.02	0.01	0	0	0	0.01	0.01	0.02	0.77	0.01	0.02	0.01	0.01	0.01	0.02
12	0.01	0.04	0.02	0	0.01	0.02	0	0.02	0	0.01	0.03	0.64	0.01	0.02	0.02	0	0.01
13	0.05	0.01	0.02	0.02	0	0.02	0.06	0	0.02	0	0.06	0.02	0.8	0.04	0.02	0.01	0.03
14	0	0.01	0.01	0	0	0.01	0	0.01	0	0	0.01	0.01	0.02	0.86	0.03	0.01	0.03
15	0.01	0.01	0	0	0	0	0	0.01	0	0	0.03	0.01	0.04	0.07	0.61	0	0.02
16	0.09	0	0.04	0.01	0.02	0	0	0.03	0	0.08	0.02	0	0.02	0	0	0.83	0.05
17	0.03	0.01	0.02	0.02	0.01	0	0	0.02	0.02	0.02	0.05	0	0.01	0	0	0.02	0.68

Figure 32: Second study – SDGs - Classifier confusion matrix

SDGs	precision	recall	f1-score	support
1	0.88	0.75	0.81	262
2	0.88	0.90	0.89	611
3	0.96	0.80	0.87	588
4	0.96	0.89	0.92	982
5	0.82	0.91	0.86	851
6	0.93	0.93	0.93	641
7	0.89	0.90	0.89	1065
8	0.71	0.81	0.76	398
9	0.78	0.62	0.69	144
10	0.72	0.74	0.73	205
11	0.69	0.90	0.78	353
12	0.85	0.75	0.80	296
13	0.82	0.85	0.83	981
14	0.88	0.89	0.88	492
15	0.80	0.67	0.73	298
16	0.91	0.90	0.91	860
17	0.74	0.75	0.74	396

Accuracy	-	-	0.82	9423
Macro avg	0.79	0.77	0.78	9423
Weighted avg	0.82	0.82	0.82	9423

Table 4: Second study – SDGs - Performance of the BERT classifier

In Figure 33 we can see how the distribution of tweets by SDGs was, seeing that there was a huge number of tweets assigned to SDG13, which matched our previous findings since climate change seemed to be one of the main topics of discussion.

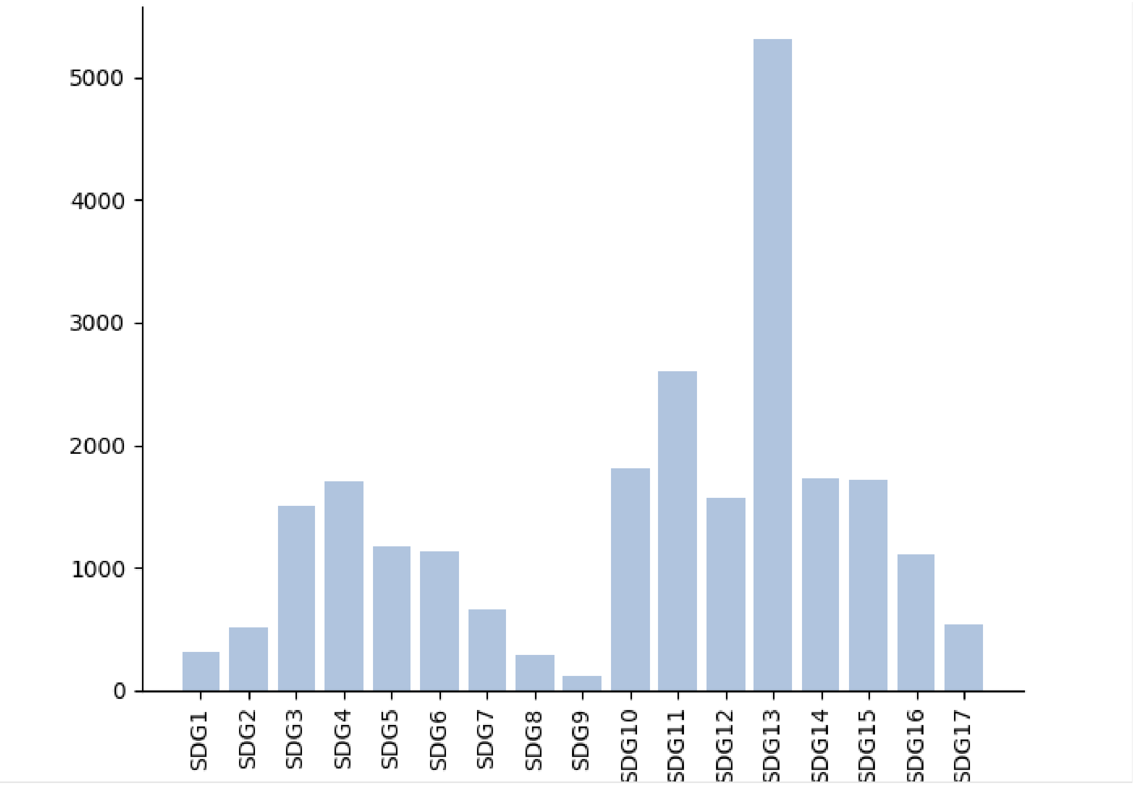


Figure 33: Second study – SDGs - Tweets assigned to each SDG

With all the tweets assigned to an SDG, we analysed how users retweet certain topics and in this specific case the SDGs, bringing the content analysis and the structural analysis together. By creating the graph that we explained in the Methodology of this case study, we connected the tweets to their retweeters and plotted the graph by means of k-core trimming. This helped us to reduce the number of nodes, remove small world networks and visualize an understandable graph. The core we decided to plot was core 2, to remove isolated nodes with just one connection. In Figure 34 we can see the graph we obtained and how the bigger the nodes are, the more retweets these tweets got. Besides, each tweet is coloured in a different colour according to which SDG it belongs to. According to the results the bigger nodes belong to tweets classified as SDG13, again Climate Action.

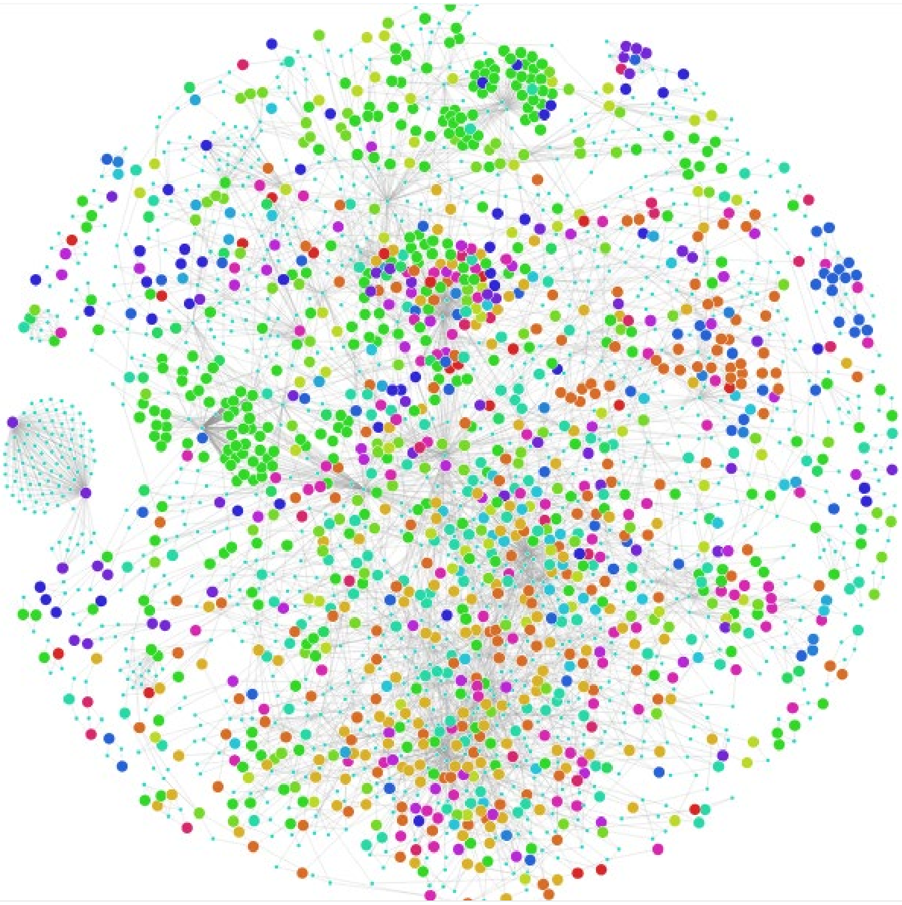


Figure 34: Second study – SDGs - Two-mode graph with SDGs by colour

In this case study, we first approached the follow relation between the users, extracting the number of followers. The number of followers is somewhat in between the content, the structural and a feature of each user that allows the characterization of main actors. Via the Twitter API, we extracted the number of followers and users they follow from the users present in our dataset related to SDGs. With the scatterplot we present them in Figure 35. In this representation, we want to show the ratio between following and followers. The X axis contains the number of followers the user has, and the Y axis the number of users they follow. Simple descriptive analysis tells that 25% of the users have less than 242 followers and they follow less than 251 users. 50% of the users have between 242 and 2,476 followers and the follow between 251 and 1,849 users.

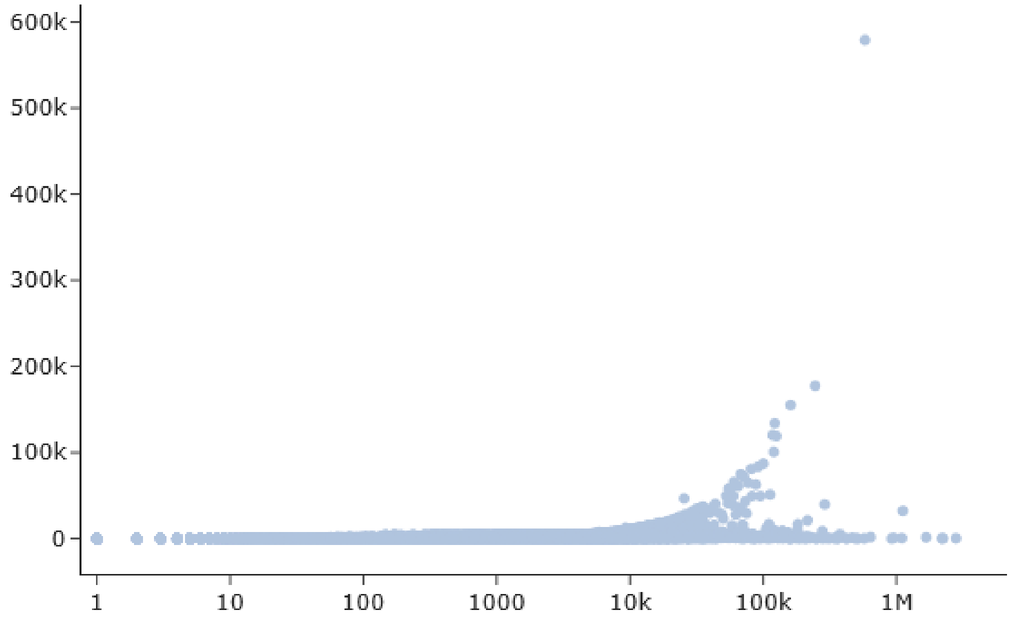


Figure 35: Second study – SDGs - Followers vs following

We found four profiles: 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 and finally d) only 1915 (18.79%) of the users are followed more than two times more than they follow themselves.

Categories c) and d) contain users that are highly followed, and it is in these two categories where we would find those prominent profiles working on creating content and share information in relation to CS and SDGs. Once we performed the retweet analysis, we were able to give a name to those important actors in this conversation.

CS community is no different from others and tend to congregate around specific users, users that produce interesting or specific information and one the most common interactions to occur between the users is the retweet. The retweets from our dataset were represented in a directed graph, being the nodes the users and the edge the given retweet. This graph was once more constructed using NetworkX, trimmed via the *k-core* algorithm and plotted using the webweb package. The results, as previously stated, showed 1,234 nodes and 2,754 edges. There were 1,234 strongly connected components and 8 weakly connected. We plotted the biggest subgraph showed in Figure 36 in which we see that the higher the number of received retweets the redder the nodes were. We see some central nodes with the higher number of retweets, which are 3 institutions, positioned in the middle of the graph since they are the most connected ones and centre of the activity of received retweets. We also see many unconnected nodes, which generally are individuals, meaning that they are not highly retweeted.

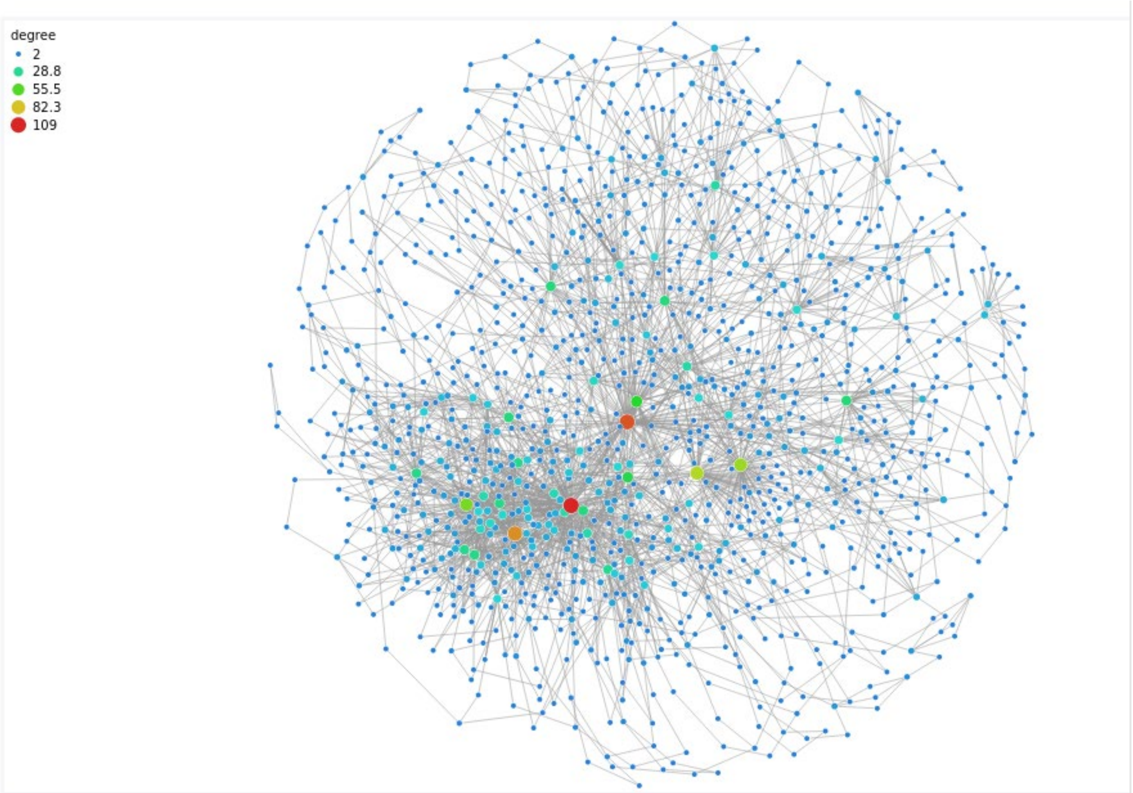


Figure 36: Second study – SDGs - Network of retweets

To complete understand the characteristics of the previous network, we provided the results from the centrality measures which are important to understand the connections in the graph and helps unveiling the main actors in the activity of retweeting, or the nexus between users that act as information spreaders. This information is of high importance for stakeholders, helping them determine who are the users that facilitate information spreading or those that are content creators with impact in the community. This analysis is based on the indegree, outdegree, betweenness and eigenvector values for the nodes. As a reminder, the indegree is the number of received retweets, the outdegree the number of given retweets, the betweenness is the number of times a node is in the shortest path between other nodes and, finally, the eigenvector is a score given to a node which sets an authority rank based on the number of nodes to which the node is connected.

We show the list of the top 10 users by indegree in Table 5. As it can be seen, users higher in the rank tend to not share information from others rather than just receiving retweets.

User	InDeg		OutDeg		Eigen.		Betw.	
	Val	R	Val	R	Val	R	Val	R
User 1	240	1	25	13	5.08E+15	196	1.21E+16	2
User 2	227	2	14	29	3.75E+15	2	1.09E+16	3
User 3	145	3	1	1801	154526982.6	273	0	1796
User 4	132	4	11	41	137829577.5	274	15785125986	112
User 5	124	5	0	9203	1136975.893	412	0	7057
User 6	120	6	68	4	4.72E+15	1	1.83E+15	1
User 7	102	7	2	436	281729571.1	264	22784914678	194
User 8	88	8	2	329	269518.6183	454	2793342905	187
User 9	84	9	0	7934	24683850670	252	0	515
User 10	84	10	2	428	2003821.205	323	4447440076	157

Table 5: Second study – SDGs - Top 10 ranked by Indegree

Table 6 shows the results the other way around, the users are ranked by higher outdegree (those who retweet others the most first). This table shows that these users do not receive that many retweets, although we see two accounts high in both ranks (user4, user6 in the previous table, and user11, user 39 in Table 5). User1 is ranked first since it is a bot account that only retweets information.

User	InDeg		OutDeg		Eigen.		Betw.	
	Val	R	Val	R	Val	R	Val	R
User 1	0	2312	280	1	25.49	2312	0	610
User 2	0	2902	73	2	25.49	2902	0	1355
User 3	0	2900	69	3	25.49	2900	0	1352
User 4	120	6	68	4	4.72E+15	1	1.83E+15	1
User 5	0	7699	58	5	25.49	7699	0	7375
User 6	0	2263	50	6	25.49	2263	0	551
User 7	0	2195	45	7	25.49	2195	0	472
User 8	0	2795	43	8	25.49	2795	0	1175
User 9	0	2134	42	9	25.49	2134	0	402
User 10	0	2328	32	10	25.49	2328	0	633

Table 6: Second study – SDGs - Top 10 users ranked by Outdegree

We also provide a scatterplot to analyse the indegree vs outdegree distribution that can be seen in Figure 37. In this figure we can see the indegree in the X axis and the outdegree in the Y axis. Most users are around values of indegree and outdegree between 0 and 50, and only few users are above 100 of indegree (number of retweets received) and just 1 user above 100 of outdegree (given retweets), in this case a bot that only retweets. This is a typical behaviour in social networks.

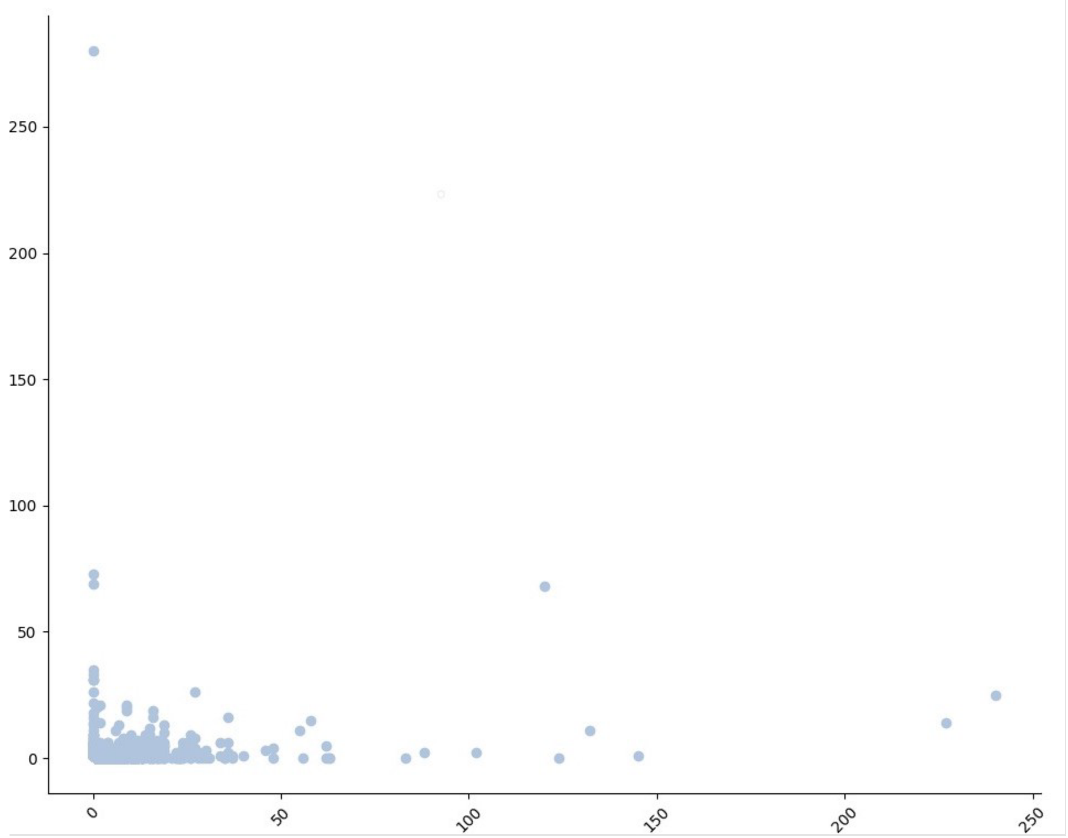


Figure 37: Second study – SDGs - Scatterplot of indegree vs outdegree

4.2.4 Discussion

This second case study explored how the SDGs are discussed in Twitter by the CS community, as a new step in our analysis of this social media which has an important role in scientific information dissemination (Tsubokura et al., 2018).

In our complete dataset we find a 7.08% of tweets about SDGs, which seems small but could be enlarged by adding new filters, since there may be more words useful to retrieve more tweets about this subject or that have not used the words we selected as filter. This work showed that there is discussion about SDGs, besides we found some bots sharing information and retweeting, which is aligned with previous studies that state that bots are having an impact on discourses on Twitter (Marlow et al., 2020). Although we obtain once more an answer to RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?), since we find conversation about SDGs, the results seem to be low.

However, this conversation about SDGs is, in comparison, larger than the conversation about learning.

From the results it seems clear that climate change is the most addressed issue in this discussion. This result is natural due to the characteristics of CS projects, which are normally enclosed in conservation or environmental sciences (Manske, 2021).

The analysis showed that #climatechange and #openscience are the most used and retweeted hashtags, and when checking the rest of hashtags, the conversation seems to evolve around sustainability, climate, energy, and nature. Most of the lexical inventory is in relation to these topics, inventory extracted and therefore answering to the RQ2 (What is the lexical inventory in these conversations, i.e., hashtags, most common words, etc.?). When checking the results from the topic modelling, this finding is reinforced since 6 out of 17 topics are related to climate change. Climate change and CS have been linked since long ago and many authors have discussed how CS can contribute to tackle climate change (Cooper et al., 2014; Groulx et al., 2017; Hurlbert & Liang, 2012). Defining these topics serves as an answer to RQ3 (What are the main topics? Are these topics evolving in time?).

Besides climate change, it seems that many users are discussing about the 2030 Agenda. The content analysis evinces the most addressed topics when combining the different techniques, so this conversation seems to be multitopical and in different languages, answering completely to RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?).

We can see how users orbit around certain topics, normally related, but it is easier to categorise communities around users as we did with the network of retweets. The results do not deviate from the trend, users congregate around certain users that monopolises retweets, likes, comments, etc (Mazumdar & Thakker, 2020).

Something everyone expects when analysing profiles in social media is that users that are highly followed are personalities, institutions or businesses with high influence and reputation, while individuals tend to have less followers than these accounts. Therefore, we see the same phenomenon as in other networks, certain users receive a high number of interactions, but they do not contribute that much interacting with others. Due to the nature of CS, which encourages cooperation, this is presented to us as a lost chance to build reciprocity in the community. This improved structural analysis does not only show who are the main actors and answers RQ4 (Can we define main actors through the SNA structural analysis?), but also provides information about the behaviour of the users.

4.2.5 Conclusions

As a summary, we explored how the discussion about SDGs was on Twitter, and although only a 7% of the total tweets were about this issue, we presented an overall view of many elements of this conversation. We reached the conclusion that climate change seems to be the most discussed topic by the CS community, from gathering data to promoting policies. Some other initial beliefs were confirmed once the analysis was done, such as the congregation around certain users or the rich club effects.

It is important to note, once more, that there are limitations to such study. New keywords could be added to further analyse the issue or other social media networks could be included to be analysed. However, our goal was not to analyse the impact of specific tweets or users, but to give a general view of the situation which was achieved. Besides, one of the positive points of this study is the possibility to easy replicate the analysis in time to see the evolution of the topics, a situation that reinforces our main idea in this thesis which is to provide a pipeline to analyse discussions and a platform to do so.

4.3 An analytics approach to health and healthcare in CS communications on Twitter

CS uses Twitter as a channel of communication that serves for sharing results, distributing news or updates and as an engaging tool. Many different topics are discussed, and, in this study, we focused on one of the most addressed topics lately, which is health and healthcare, another field closely related to CS (Boving et al., 2021). In this study we added more techniques such as an improved TF-IDF calculation or the network of mentions and the network of hashtags. We also did improvements to those techniques we had by that time, moving to a new topic modelling technique or adding more content to the structural analysis.

4.3.1 Motivation

Health became one of the most important topics in the recent years and it is especially due to the COVID-19 pandemic. In the context of health, we understand healthcare as the improvement or maintenance of health or well-being by means of different methods such as diagnosis, treatment, or cure of diseases. Both concepts are easily interpretable as important and there are many projects that aim to gather information about people´s health or even dedicated to communicating discoveries on this regard. Given the importance of both topics, their influence or presence in CS is to be expected.

CS is normally linked to environmental issues, biology, or conservation but there are many projects related to health, healthcare or health research (Lee et al., 2020). There are several examples of relation between CS and health in history, such as women´s health studies or HIV-related activities (Callon & Rabeharisoa, 2008). Nowadays, many CS projects base their activities on health-related issues, from data collection or processing, in the study of the Parkinson disease or 3D mapping of neurons to wildlife health like the Monkey health project.

From previous studies, we also see that Twitter has been used to analyse data from areas such as use of antibiotics, analysis of communications on COVID-19 or monitoring the perception of the users of 2009 H1N1 pandemic (Paul & Dredze, 2021; Scanfeld et al., 2010; Shahi et al., 2021).

Social media platforms are not only important to gather data from them but they can also reshape the idea of healthcare we nowadays have (Hawn, 2009). Some authors believe that Twitter is the preferred way of communicating in medical practices since it creates a collaborative environment for researchers and patients, but the misinformation is an important enemy (Pershad et al., 2018).

According to all this background, we believed that it should be possible to use Twitter to learn about the intersection of CS projects and activities with health and healthcare. We used our techniques and processes to gather information and obtain insights from this discussion with the goal of answering to the RQs we established for this thesis: finding a multitopical and multilingual conversation, discover the topics of discussion alongside the lexical inventory, and defining the main actors in this conversation. This will lead to the validation of the pipeline for analysis and a proof of the usefulness of the dashboard developed.

4.3.2 Methodology

The present study was based on that continuous improvement of the processes and SNA techniques que decided to implement to create the pipeline for analysis. The tweets that we analysed here were harvested from the Lynguo tool, ongoing from 30th of September 2020 until 7th of September 2021 when we downloaded them.

In the initial dataset, the one containing all the tweets related to Citizen Science thanks to the filter that Lynguo applies, we had 365,609 tweets, being 149,227 original tweets and the rest retweets. We found 103,432 unique users and 44,687 tweets mentioning one or more users.

But this study was about health and healthcare, so we applied a filter based on keywords related to this topic. These keywords were extracted from a selection of keywords about this issue such as healthcare, health, electronic health record (EHR), ehealth and others, alongside their homologous in other languages). The origin of these keywords is the dataset of taxonomy that the National Uniform Claim Committee (NUUC) has designed. In the categories "Grouping" and "Classification" we found interesting keywords related to health and healthcare. We decided to use these keywords since the NUUC provides a standardized dataset to be used in electronic environments when researching in this sector.

We gathered a total number of 359 keywords destined to be used as filter to collect tweets from our initial dataset that were discussing health and healthcare related topics. All the different techniques were applied to this final filtered dataset. For this specific study we presented the methodology in a workflow chart to provide a better understanding of how the analysis was performed, which we can be seen in Figure 38.

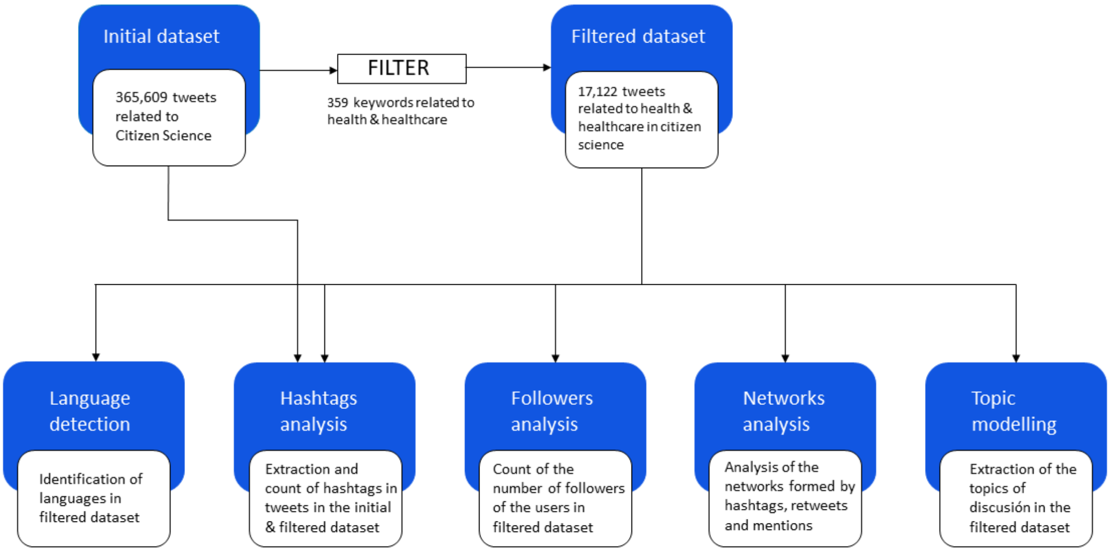

Initial dataset
365,609 tweets related to Citizen Science
FILTER
359 keywords related to health & healthcare
Filtered dataset
17,122 tweets related to health & healthcare in citizen science
Language detection
Identification of languages in filtered dataset
Hashtags analysis
Extraction and count of hashtags in tweets in the initial & filtered dataset
Followers analysis
Count of the number of followers of the users in filtered dataset
Networks analysis
Analysis of the networks formed by hashtags, retweets and mentions
Topic modelling
Extraction of the topics of discusión in the filtered dataset

Figure 38: Third study – Health and Healthcare - Workflow which is representative of the pipeline for analysis

To discover the languages in our dataset, we applied the previously explained NLP-based multinomial Naive Bayes Classifier. This model was trained using 10,337 texts in:

Danish, English, Italian, Tamil, Turkish, Malayalam, Kannada, Russian, Portuguese, Hindi, Dutch, Spanish, French, German, Swedish, Arabic and Greek.

To proceed with the hashtag analysis, we followed the same procedure we used in previous studies. We extracted the words preceded by a hash symbol (#) and counted their occurrences in the general dataset and inside and outside the retweets. We also analysed their temporal evolution. As a new feature, we analysed the co-occurrences of hashtags, meaning that we represented the connections of hashtags inside a same tweet which was presented in a network that allowed semantic and topical relations analysis.

In this research, we analysed different networks formed with different types of connections between users and features from Twitter. To perform the network analysis, we used the networkX package from Python, webweb package for visualization and *k-core* algorithm from networkX.

The first network was the previously mentioned formed by the connection between hashtags. Each tweet features normally more than one hashtag so; therefore, they will be connected to each other forming a network that allows a context analysis.

Another network was created when we analysed the mentions. Mentioning in Twitter is writing the name of other user in the tweet preceded by @, which is one of the main forms of interactions between users alongside the retweets. To extract these mentions we searched for tweets containing @+username. The edges of this network were created connecting the username from the dataset (the one that wrote the tweet) to the username found inside the tweet. The network we created was a bipartite directed graph, this decision was taken because bipartite graphs divide the nodes in two categories (source – target) in the direction of the mention.

The last one was the network of retweets. The process of creation was like the one of mentions, but instead of searching for @+username, we selected those tweets starting with “RT@username”. To create the edges, we linked the username that retweeted (User column in the dataset) to the @username in the retweet. Again, we decided to present it as a bipartite directed graph for the same reason stated before.

To all the networks in this study, we applied the method of *k-core* decomposition of graphs to provide better representations. We also introduced the community detection method, in this case supported by the Louvain method (Blondel et al., 2008).

In the case of the network of retweets, we also calculated the centrality measures, aimed to analyse the “influence” of the users based on the indegree, outdegree, eigenvector and betweenness. As a reminder, indegree shows the number of received retweets, outdegree the number of given retweets, eigenvector sets an authority score and betweenness shows how many times that node is present in the shortest path to another (nodes with high betweenness tend to be considered as bridges in information flow).

Next, we analysed the topics of discourse. To do so, we used the topic modelling technique, an unsupervised learning technique that identifies patterns between texts. In this study, the topic modelling analysis was performed using the BERT model, explained before. The way of applying this model was similar to the procedure we followed for the LDA topic modelling: cleaning the texts of stopwords, punctuation and symbols, removing the retweets, selected the parameters (multilingual model, 10 top words and default for the rest). The exception came with the minimum topic size, which can affect the results, because we tried different values (100, 50, 20 and 10) obtaining the best performance in topic similarity and score for the model with 10. The final number of topics was too high but expected. Therefore, we used the reduce topic number feature that BERT has. The results were displayed in an intertopic map, and we extracted the keywords for each topic.

4.3.3 Results

We first applied the filter and we found that 17, 122 tweets were about health and healthcare. This represents a 4.7% of the total tweets. In this subset we can find 7,964 unique users and 6,331 unique tweets. We also found 2,960 mentions inside these tweets.

Besides, with the NLP based model to detect languages we found that the main languages were English (73.7%), German (13.1%), French (5.62%), Spanish (5.13%), Dutch (1.73%), Italian (0.37%), Portuguese (0.16%), Swedish (0.13%) and Danish (0.07%). English being the most used one was something to expect as this language is the most used one in science dissemination and science. The pipeline for analysis was able to detect tweets about another different field apart from SDGs and learning and the tweets about this topic were multilingual.

Beginning with the content analysis to extract the lexical inventory, we extracted all the words preceded by a hash symbol. The total count was 1,059,463 hashtags, from which

30,288 were unique. Isolating the retweets, we found 582,330 hashtags inside them (19,789 unique) and 477,454 outside the retweets (29,453 unique).

From the original dataset (unfiltered one) we see the results of most used hashtags in Figure 39 (hash initial dataset), which shows the top 10 most used hashtags, which are #earthquake followed by #sdgs. We can see next #seismograph and #openscience but with 2000 less entries and #communityscience, #biodiversity, #opendata, or #scicomm. #earthquake and #seismograph come from a bot account that tweets about earthquake recordings. In fact, we also cleaned some more hashtags that came from bots posting about air quality because of the high numbers of hashtags they used that did not let appreciate the hashtags used by individuals and institutions.

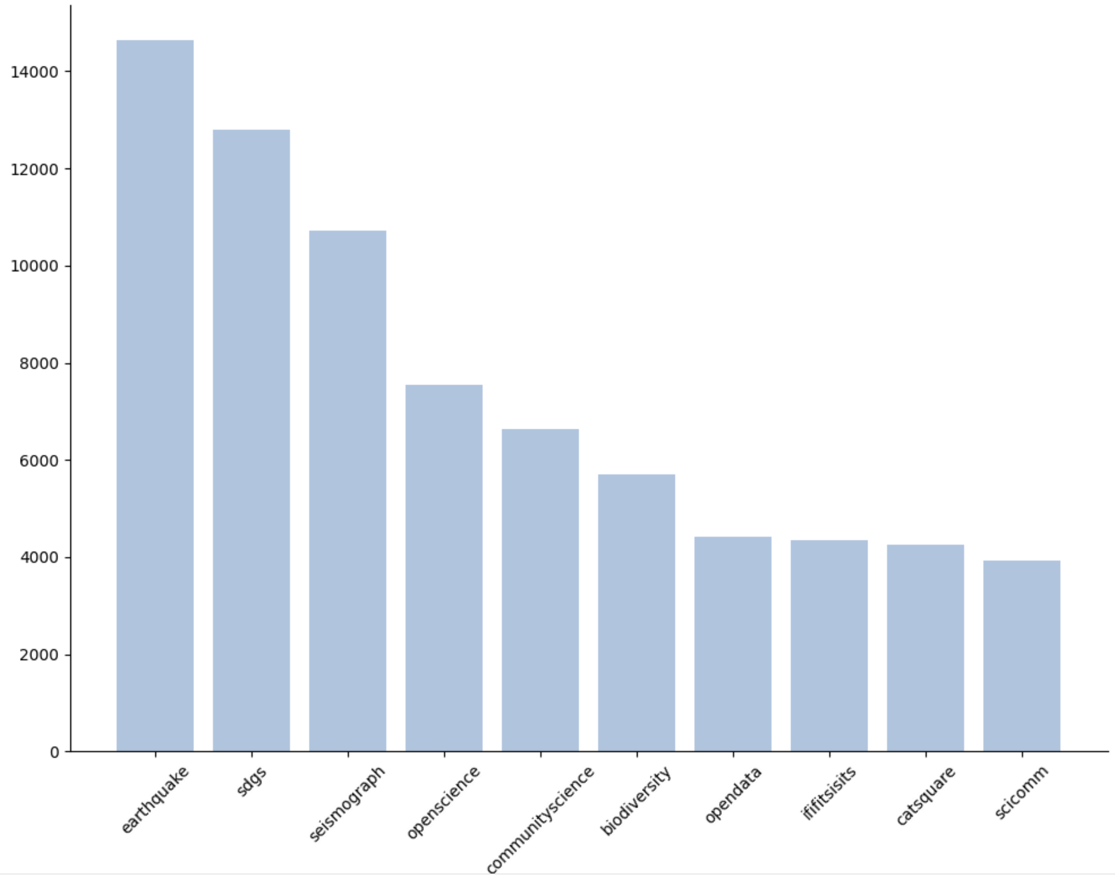


Figure 39: Third study – Health and Healthcare - Most used hashtags in the complete dataset

Focusing on the filtered dataset using the health-related words, we obtained 50,349 hashtags (4,135 unique). In the retweets we found 36,938 with 2,749 unique results. Outside the retweets we found 7,525 hashtags (2,262 unique). The results from the analysis of the complete filtered dataset are presented in Figure 40, with the top 50 hashtags found on it. In this figure, the bigger the square, the greater number of appearances that hashtag has, besides the colour scale indicates more appearances to less appearances from blue to yellow. We can see SDGS with 878 uses, or health with 612. On the other end, research has 124, or cancer 180.

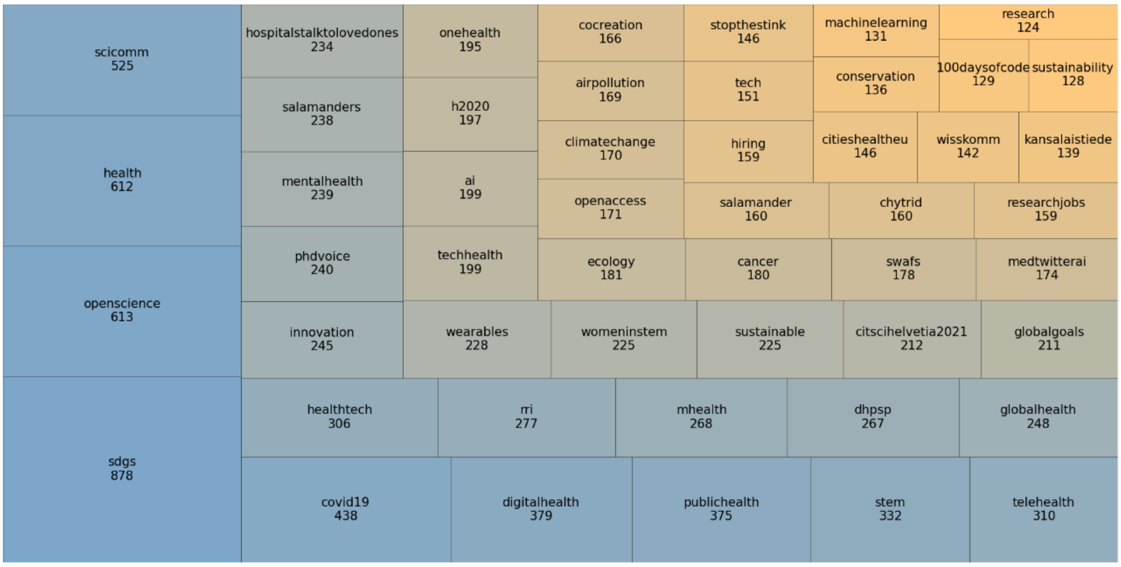


Figure 40: Third study – Health and Healthcare - Top 50 hashtags in the filtered dataset

Once we know how many we had, we checked the most used ones outside the retweets. The results show that #sdgs was the most used one followed by #openscience, #health and #scicomm. Then we find #covid19, #digitalhealth or #publichealth, a similar situation as shown in Figure 40. Checking the retweets, the situation is similar as Figure 41 shows, being #sdgs the most retweeted one by far. The next hashtags are the same but in different order, #digitalhealth is more retweeted than #covid19 for example.

As we explained, the tweets were collected from September 2020 to September 2021, so the use or retweeting of hashtags could variate in time. To analyse this situation, we performed as the previous case study the temporal analysis of hashtags.

Figure 42 shows the results of the temporal analysis of retweeted hashtags. The X-axis shows the dates and the Y-axis the occurrences (retweets in this case). #Health has the higher value throughout the time, followed by #openscience. We see some peaks for #dhpsp, #covid19 or #telehealth.

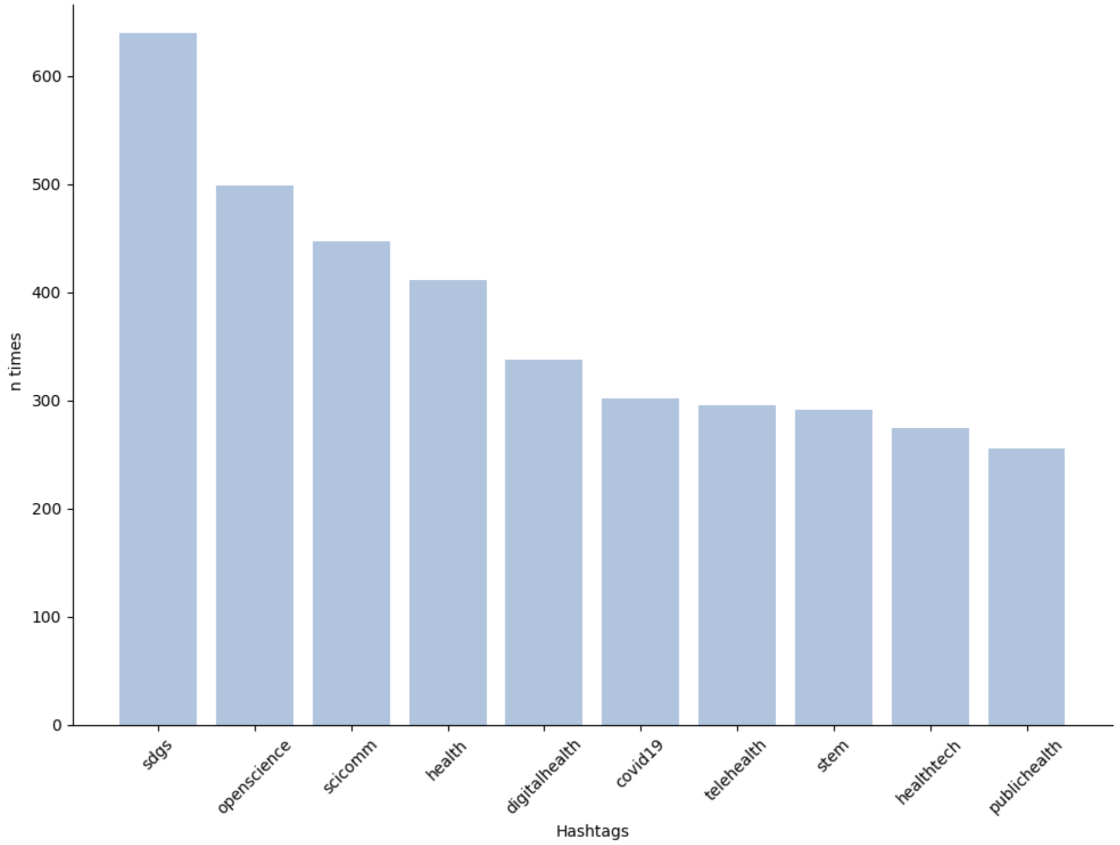


Figure 41: Third study – Health and Healthcare - Top 10 most retweeted hashtags

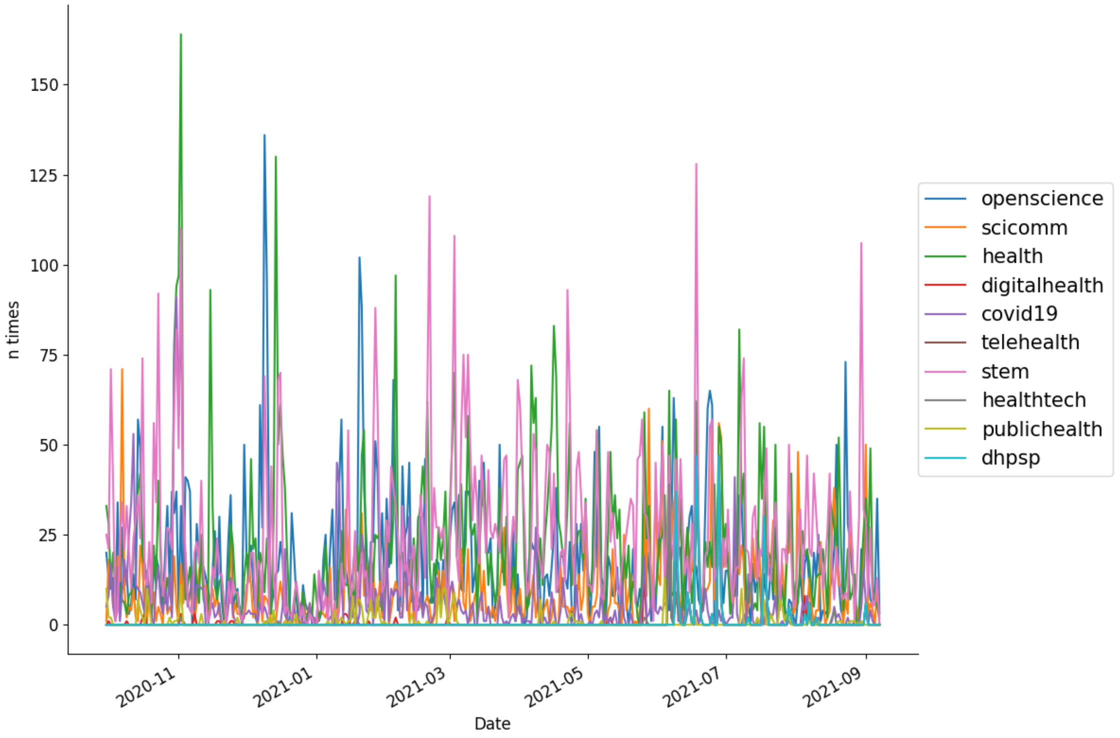


Figure 42: Third study – Health and Healthcare - Temporal evolution of retweeted hashtags

When it comes to use of hashtags, the temporal evolution can be seen in Figure 43. #health is the most used one by far once more, but there is a specific point in time during late January or early February that the hashtag #citscihelvetia2021has a high peak due to the celebration of this conference on January 14th and 15th. After that, #health continues to dominate the usage.

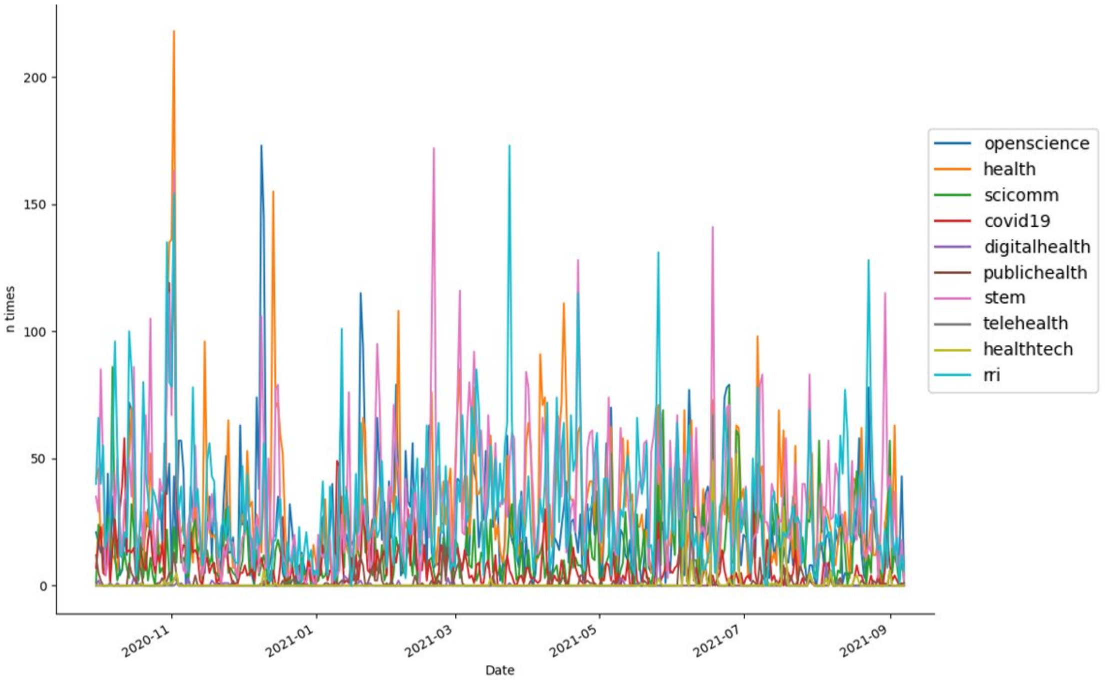


Figure 43: Third study – Health and Healthcare - Temporal evolution of most used hashtags

As we previously explained, one of our main ideas in this thesis was to bring the content analysis and structural analysis, so we investigated how content features from Twitter could be translated into networks. Thus, we introduced the network of hashtags, analysing the connection of hashtags inside the same tweet, thus allowing us to get some contextual understanding of their usage. The network we formed contained 24,934 edges and 5,183 nodes. This network had 23 cores and it was trimmed to show highly connected hashtags giving us a visualization containing 47 hashtags shown in Figure 44. The redder and bigger the node, the more times the hashtag appears inside tweets connected to others. This means that redder and bigger hashtags are commonly used alongside others and can serve as an indicator of important super-topics that are accompanied by other hashtags related to the topic, and also the connections between these topics.

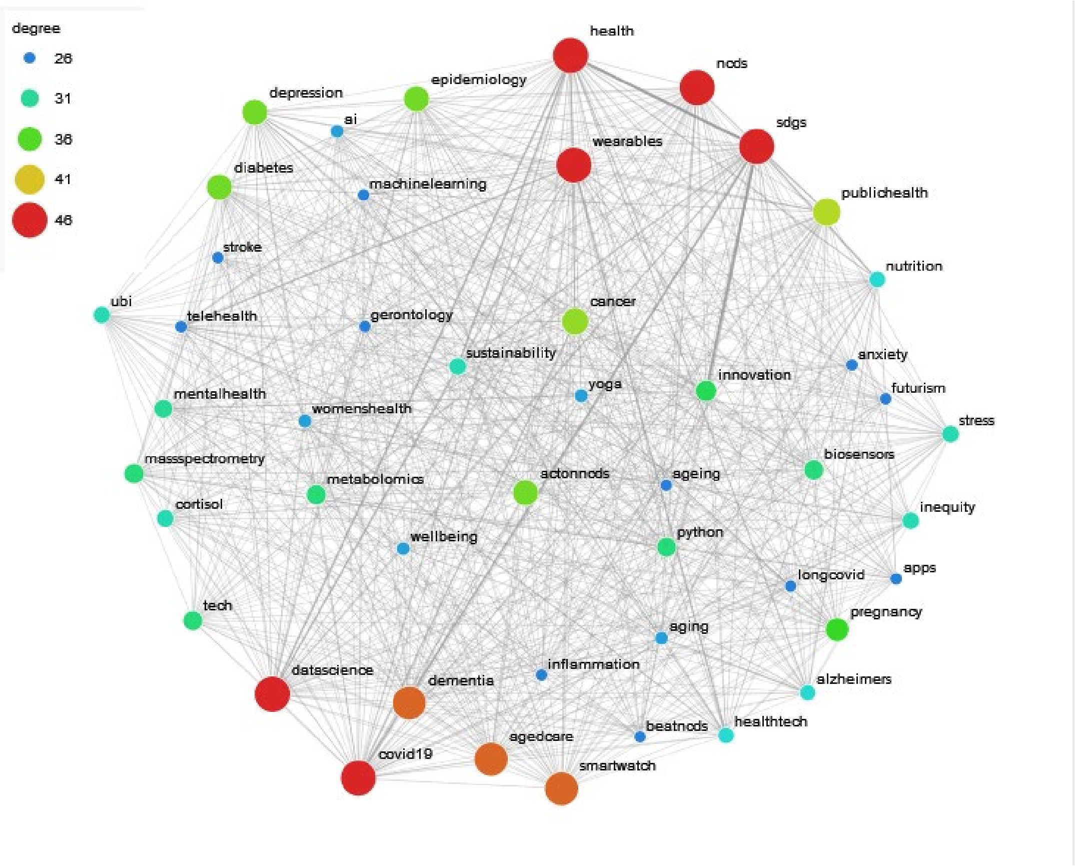


Figure 44: Third study – Health and Healthcare - Network of interconnected hashtags

The main nodes are #sdgs, #health, #ncds (non-communicable diseases), #wearables, #datascience and #covid19. In the network we see how areas like health, healthcare, mental health, technologies, IT, development goals, and medicine are highly connected.

Once more, following the pipeline for analysis, to complete the content analysis we checked the topics addressed by the users. Applying topic modelling we checked the intertopic distance to determine how many topics we had in our set of tweets. Figure 45 shows the results of the intertopic distance. This visualization shows the extracted topics represented in a graph as circles according to their size (number of tweets inside each topic), and the closeness between topics is a measure of their similarity. In the figure we have 2 groups of topics of high similarity and then a different group containing 2 topics in the bottom of the graph. To get a better understanding of the different topics, we providen explanation of the topics is in Table 7.

We find topics directly related to health and healthcare, for example, Mosquito Alert, Mental disorders, DHPSP, Mental disorders awareness, COVID-19, or Games for genomics among others, but all in different areas of health.

DHPSP would be the one more related to healthcare. Mental health is quite present in the different topics. There are other topics that address health in a more indirect way. Something that was surprising is finding a specific topic about COVID-19 although the number of tweets in the dataset about this issue was not large.

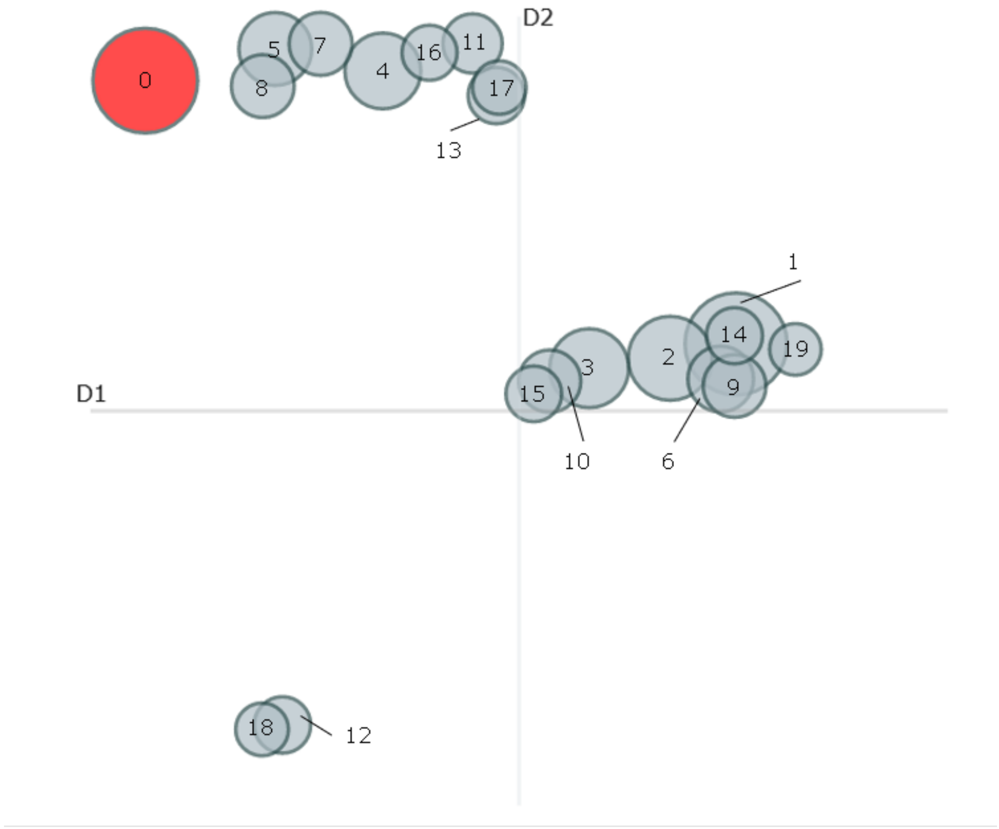


Figure 45: Third study – Health and Healthcare - Intertopic distance map from the topic modelling

Topic	**Keyword**	**Theme**
Topic 0	MosquitoAlert	This topic is about diseases and health problems related to mosquitoes.
Topic 1	Mental disorders	In this topic, we see discussion around brain and mental disorders.
Topic 2	100 days of code	This topic is about the challenge 100 days of code, where projects and individuals tweet every day about their progress coding using different hashtags such as #digitalhealth or #womeninstem.
Topic 3	Research of rare diseases	Here the discussion is about projects that research on rare diseases within SDGs and 2030 Agenda
Topic 4	Water sanitation	Another topic about SDGs, in this case about water sanitation.
Topic 5	Health of forest and animals	This topic is about wildlife and probably about monitoring animals and environments. This correlates with SDG 15 “Life on Land” and with the many CS projects that monitor animals.
Topic 6	BibchatDE	This topic is about the initiative Bibchat, which promotes information exchange between professors, libraries, professionals, etc.
Topic 7	Healthy diet	This topic is about healthy diets and sustainable production, a reference to SDG2 and 12.
Topic 8	DHPSP	We have conversation from two project accounts: DHPSP and HospitalTalkToLovedOnes. This discussion is about healthcare and technology and its use to improve the treatment to patients and loved ones.
Topic 9	Health of children	It seems like the discussion is about the health of children in the journeys to the countryside concerning bugs, climate and temperature.
Topic 10	Healthy diet in schools	Another topic about healthy diets but in this case in the schools´ canteens.
Topic 11	Cowslips	In this topic we find a conversation focused on the soil and plants with the project lookforcowslips participation, which monitors this specific plant.
Topic 12	Mental disorders awareness	It addresses the fundings and awareness of mental disorders.
Topic 13	Mixed subjects	This topic contains French language mainly and it is the least clear topic because it addresses such things as vaccines, bugs, conservation and Africa.
Topic 14	Pollution and Health	This topic is about pollution and its consequences for health, another topic connected to SDGs, in this case SDG11.
Topic 15	Environmental problems	This topic addresses environmental problems in general.
Topic 16	Pollution and sustainable cities	This topic is about sustainability and pollution in cities, with projects such as CitiesHealthEU, ecowaste_info or Urban_mind participating.
Topic 17	COVID-19	We clearly see a discussion around COVID-19 in this topic
Topic 18	Games for genomics	This conversation is about cancer and genomics and how games can be used to study genomic factors in cancer. Topic started by the project @Genigma3D
Topic 19	Alzheimer monitoring	Disucssion about monitoring patients with Alzheimer via smartphone apps.

Table 7: Third study – Health and Healthcare - Topic, Keyword and Explanation of the topics from e-Health

Besides the list and explanation of topics we also list in Table 8 the complete set of keywords for each topic.

Topic	**Keywords**
Topic 0: Mosquito Alert	mosquitos, mosquito, enfer- medades, mosquito alert, tick, diseases, heikkihelle, researchers, mosquitoes, mosquitoborne
Topic 1: Mental disorders	brain, mental, community, games, moncktonlaw, consent, user1 (anonymized user), site, comment, neurocognitive
Topic 2: 100 days of code	100daysofcode, machinelearning, citieshealtheu, wearables, womeninstem, globalhealth, digitalhealth, healthtech, telehealth, scienceetcite
Topic 3: Research of rare diseases	europe, eucitsciproject, citieshealtheu, share4rare, collaborations, Exchange, design, agenda, policymakers, place
Topic 4: Water sanitation	water, waterbugblitz, citieshealtheu, rivers, river, waterbug, openness, wetlands, catchmentsmatter, lake
Topic 5: Health of forest and animals	salamanders, salamander, user2, snap, fungus, photo, wildlife, tadpole, savebutterflies, communityscience
Topic 6: BibchatDE	bibliotheken, bibchatde,libraries, transkribieren, citsci geek, citscioz, medievaltwitter, germanistik, twittercampus, humanismus
Topic 7: Healthy diet	food, planetaryhealthdiet, great- foodtransformation, greatreset, dietary, foodcanfixit- path, wildlife, world, cityunihealth, foodpolicycity
Topic 8: DHPSP	dhpsp, hospitalstalktolovedones, digitalhealth, telehealth, patients, healthtech, hospi- tal, globalhealth, hospitals, hospitalstalkt
Topic 9: Health of children	journeys, bugsmatter, buzz dont tweet, schoolholidays, trips, summer, insect, populations, climate, temperatures
Topic 10: Healthy diet in schools	teachers, students, user3, join zoe, children, school, insights, lehr, schools, w jahr
Topic 11: Cowslip	soil, love plants, survey, cowslips, habitat, plantlife, switzerland, climate- change, lookforcowslips, compost
Topic 12: Mental disorders awareness	funding, public, isglobalorg, mental, publicengagement, participatory, media, citieshealtheu, funded, inspiringstories
Topic 13: Mixed subjects	chercheur, franais, fondateur, 19782007, scientifique, lhumanit, vacciner, insekten- hotels, conservacin, africa
Topic 14: Pollution and Health	sensors, particles, liquid, solids, satellites, harmful, breathe, data, mentalhealthaware- nessweek, connectwithnature
Topic 15: Environmental problems	bees, mason, backyards, seismograph, volunteers, populations, earthquake, gartenschlfer, thursenvironmental, wtpblue
Topic 16: Pollution and Sustainable cities	toxicsfree, pollution, quality, urban mind proj, consumption, wellbeing, data4good, ecowaste info, cities, citieshealth

Topic 17: Covid-19	vaccines, vaccination, vac- cine, coronavirus, vacuna, vacunarse, vaccinated, datascienceagainstcovid19, covid19vaccination, bha- vanamreligujcostdstgovtof
Topic 18: Games for genomics	game, cancer, cells, genomic, researchers, crgenomicah2020, genigma3dcitizenscience, gamers, gaming, trials
Topic 19: Alzheimer monitoring	smartphones, alzheimers, foldin- gathome, phone, apps, tracking, developer, applica- tions, smartphone, monitor

Table 8: Third study – Health and Healthcare - Topics and keywords

Besides, to determine the relation between topics, we checked the topic hierarchy, showed in Figure 46. We see three groups of related topics. In Figure 46, we see on the right 6 six topics with more relation to environmental fields that could affect health. The second group on the right is more related to e-Health, with topics about games and smartphones. The green group seem to be related to projects and education, with finally an isolated topic, topic 17, about COVID-19 and vaccines.

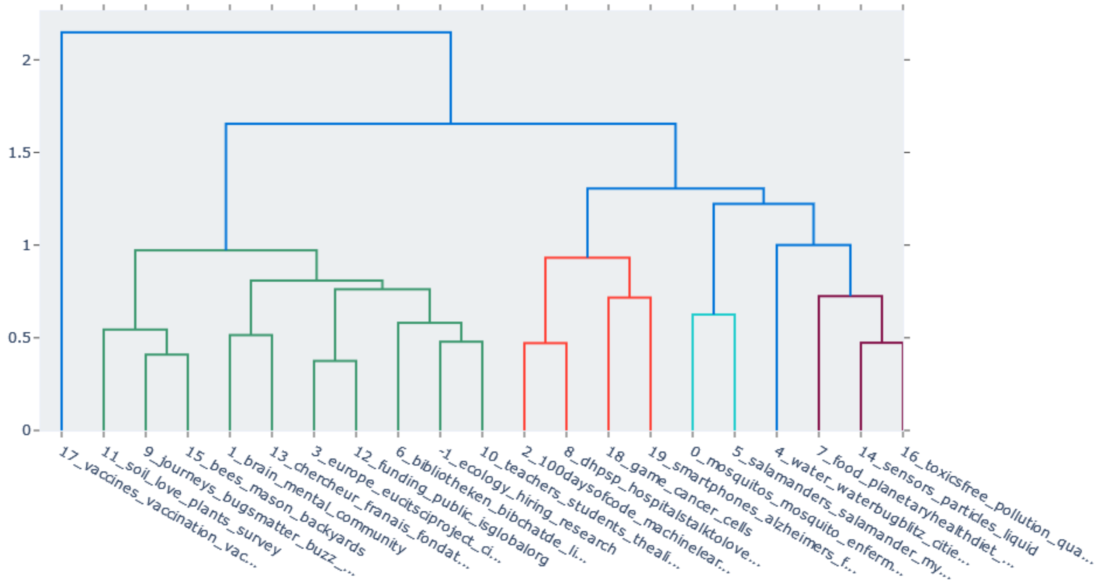


Figure 46: Third study – Health and Healthcare - Topic hierarchy

We also checked the number of tweets per topic, finding that most of them belong to Mosquito Alert and Mental disorders, with around 1750 each one. 100 days of code, Research of rare diseases and Water sanitation followed the other two with 1000 each. The rest of topics had around 600 tweets each one.

A new analysis we introduced here was the analysis of topics over time. This is depicted in Figure 47 and, as it can be seen, Mosquito Alert has the highest frequency over time alongside Mental disorders which has two high peaks surpassing Mosquito Alert. The highest peak of Mosquito Alert in early March 2021 could have been caused by the conference Mosquito Control. The peaks of Mental disorders could have happened because of the celebration of several webinars and conferences about mental health such as Mental Health considering Covid-19 Virtual Summit. Conferences and such events increase the use of hashtags, or the number of tweets written, but there are many other factors that could cause these increments. The rest of topics have a stable evolution through time.

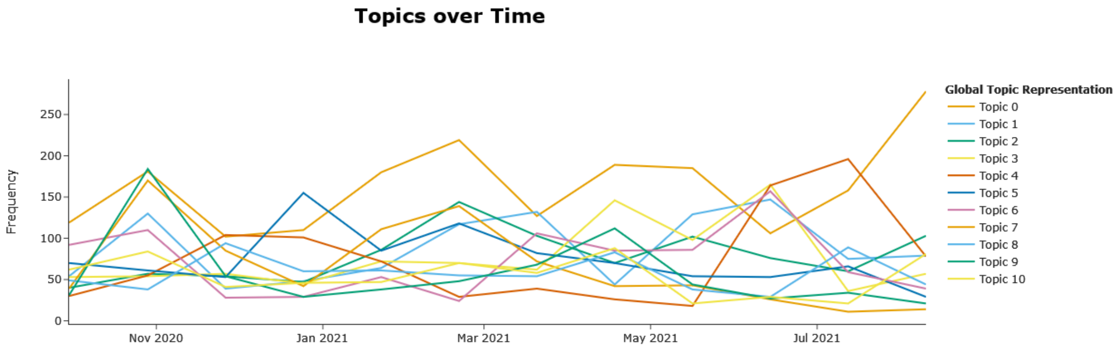


Figure 47: Third study – Health and Healthcare - Topics overtime

With the content analysis finished, we moved to the structural analysis to determine who were the main actors in this conversation. As in every social network, in Twitter we tend to consider that users are influential or more important if they have a high number of followers. To study this situation, we extracted the number of followers from the Twitter API. In Figure 48 we see the top 10 accounts with more followers but in a logarithmic scale, because some accounts have very high numbers of followers, and it affects the visualization.

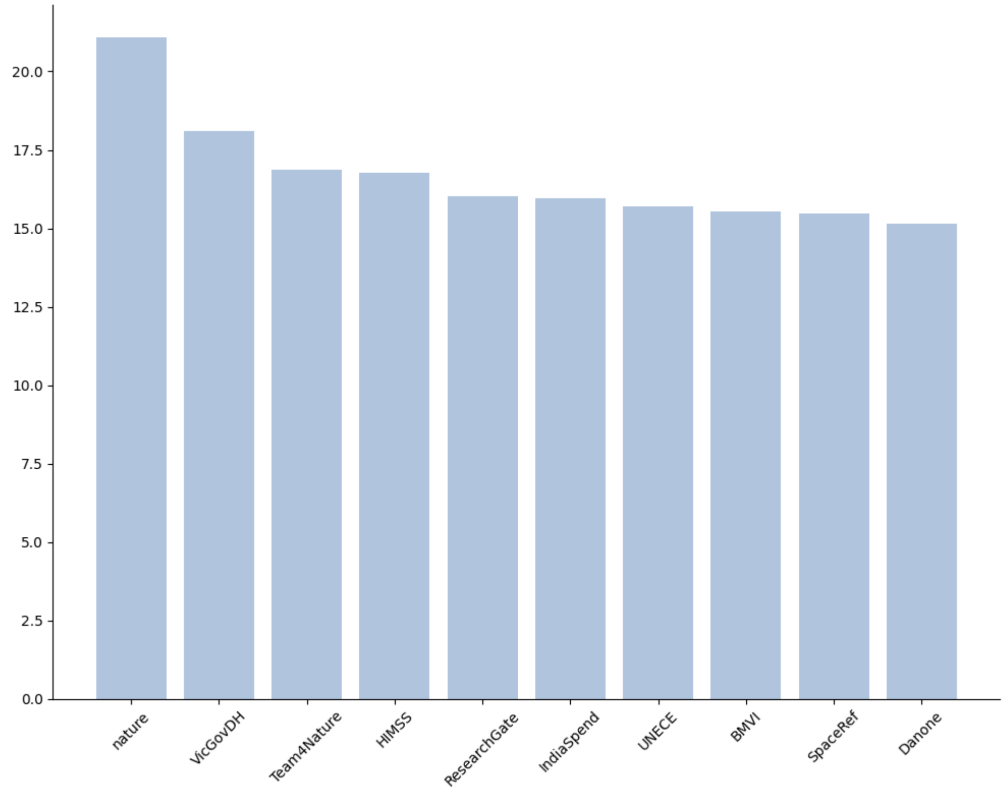


Figure 48: Third study – Health and Healthcare - Top 10 users with more followers

Related to this metric, we can check the impact of these accounts, which is calculated in proportion to the number of followers on a logarithmic scale and its value is between 0 and 100. This impact indicator helps to determine the influence of the accounts and in Figure 49 we can see the results. @nature, @NASAEarth and @CDCemergency have a similar impact, although nature has more followers, the rest of the top accounts are around 70 in impact factor.

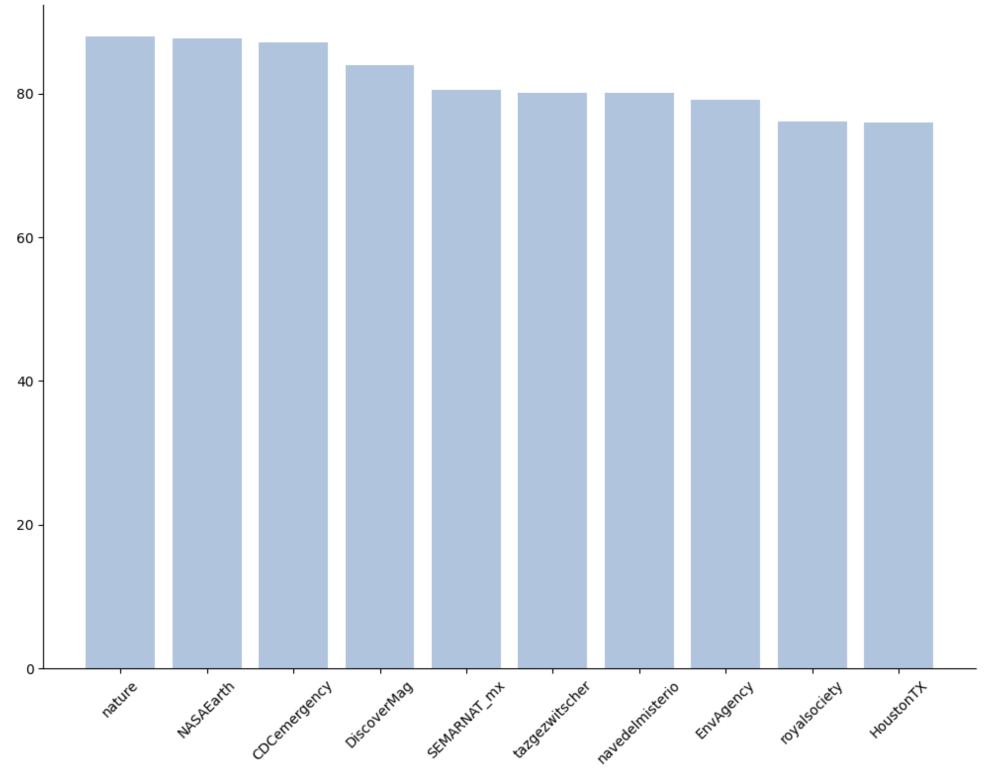


Figure 49: Third study – Health and Healthcare - Top 10 accounts with more impact

Besides the impact, another metric that could help unveil the main actors in this conversation about health and healthcare, is the number of received mentions. To analyse this, what we did is to create a network of mentions, which were extracted from the dataset connecting the user that wrote the tweet to the username they mentioned inside it.

We saw that 1,413 users mentioned others. We had a total number of 3,442 users mentioned, contained in 4,842 edges and 4,501 nodes that were trimmed by means of *k-core* algorithm once more.

Figure 50 shows the top 50 more mentioned accounts and, as we expected, these were project accounts while individuals are less than half of the top 50 nodes (only 11 users). We saw that CitSciOz was the most mentioned and mentioning account, followed by GAINBioblitz, Ideas 4 Change and SCOREwaterEU.

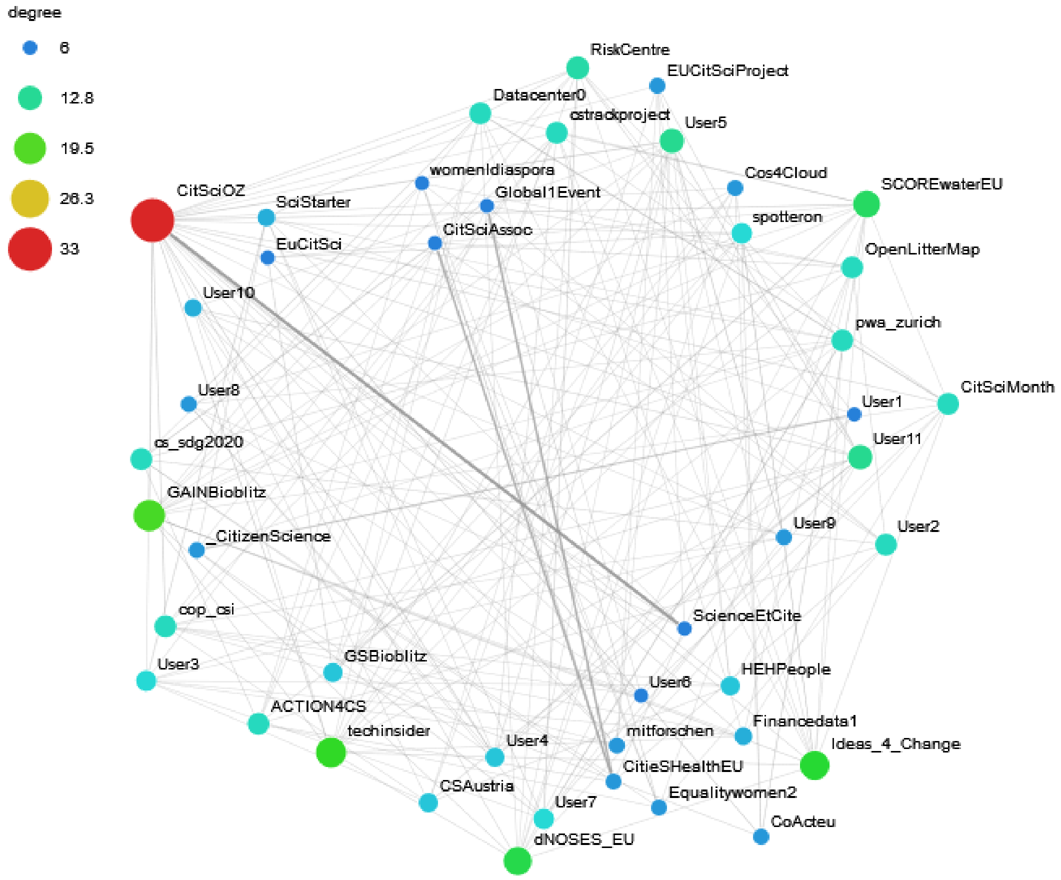


Figure 50: Third study – Health and Healthcare - Network of mentions

At this point, we had improved and enriched the structural analysis with these new metrics (cites, impact, followers), but the analysis of retweets is a foundation for structural analysis in Twitter. Thus, we then counted the retweets inside our dataset, and we obtained a graph with 6,808 users that retweeted others, and 1,293 users that were retweeted. For this analysis we also applied community detection, so the results showed 35 communities around users that are highly retweeted. These results are depicted in Figure 51 where we can see all the nodes and the communities they belong to are represented in different colours.

The biggest community, in dark blue, contains 116 users and the main accounts inside that community are CitieSHealthEU and EUCitSciProject. Analysing the accounts that receive more retweets inside each community, we can check the information they share, thus unveiling the topic of discussion inside that community. For example, the biggest community was found to be discussing about health problems and urban impact on health.

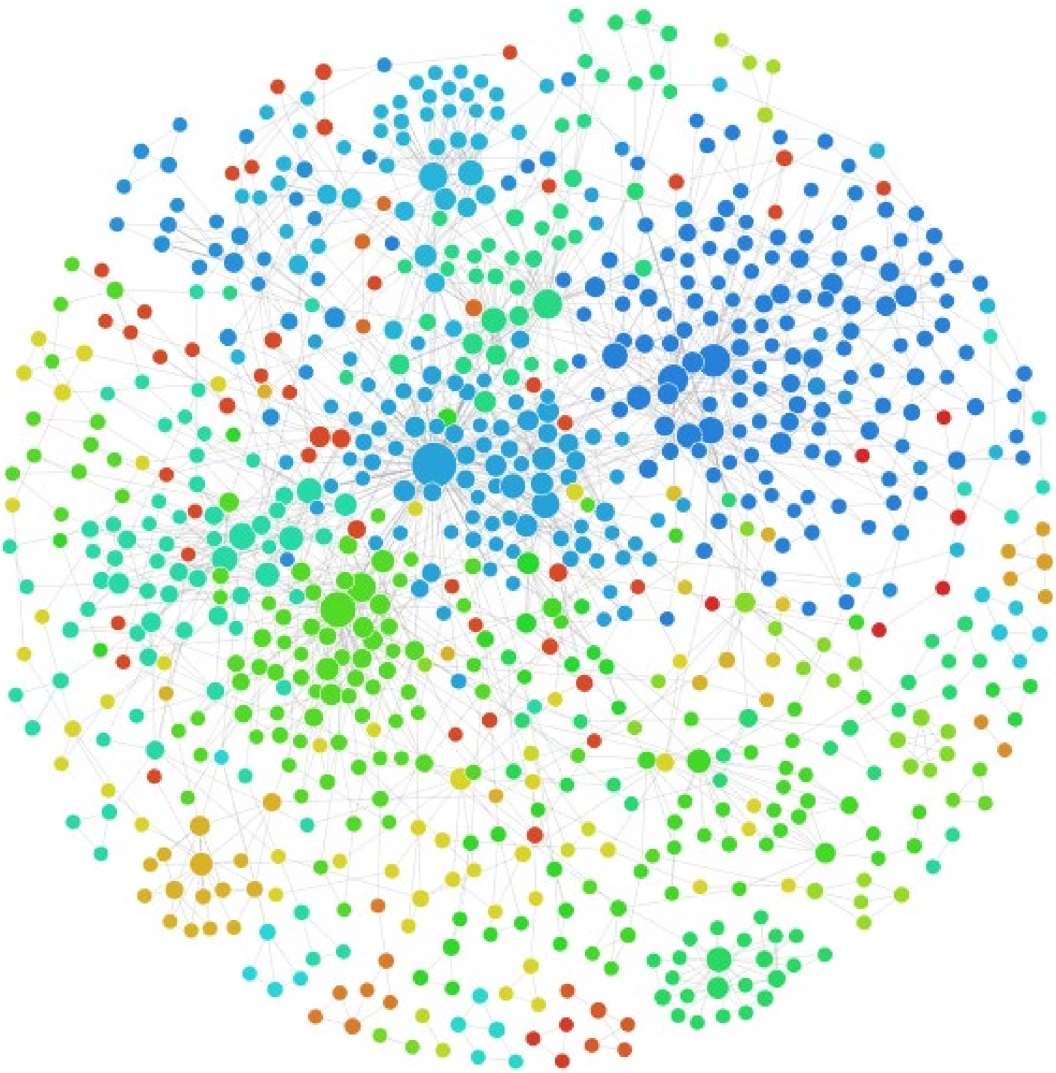


Figure 51: Third study – Health and Healthcare - Network of retweets with communities in different colours calculated with Louvain method.

We found other communities discussing such diverse topics as: community 1 discussed about Citizen Science in Germany and the connection between MfNBerlin and other projects. A community from Austria discussing Citizen Science and health, community 16 around biomedicine or community 7 which discussed about surveillance of mosquitoes in Australia. In future research, it could be interesting to also analyse these communities alongside characteristics such as country, language, institutions, etc.

In order to complete this information extracted, we then focused on the centrality analysis. Again, as a reminder, the centrality measures we checked were the indegree (which represents the number of received retweets), the outdegree or the number of retweets given and the betweenness (the number of times a node is present in the shortest path between nodes, acting in this context as bridges for information flow).

Table 9 shows who were the most retweeted users in this analysis, as they are ranked by Indegree. Once more, most users present in this table are projects, only four accounts belong to individuals (User1 to User4). The most retweeted account is CitieSHealthEU, whose role is monitoring the impact of Urban Environments on People´s Health.

Name	InDeg		OutDeg		Betw.	
	Val	R	Val	R	Val	R
CitieSHealthEU	96	1	30	3	0.06	1
Mitforschen	48	2	6	44	0.015	5
CSAustria	37	3	6	48	0.007	12
User1	34	4	4	101	0.0	96
EUCitSciProject	32	5	1	579	0.001	46
SciStarter	30	6	38	2	0.023	3
ORION_opensci	22	7	2	482	0.002	37
InfoGujcost	21	8	3	121	0.0	97
DHPSP	21	9	4	81	0.0	106
MozzieMonitors	21	10	6	41	0.008	10
User2	19	11	3	161	0.001	53
Love_plants	19	12	1	608	0.0	107
HEHPeople	18	13	3	166	0.001	51
User3	18	14	1	634	0.004	20
pwa_zurich	17	15	12	10	0.009	7
ScienceEtCite	17	16	6	43	0.004	17
CitSciMonth	16	17	23	4	0.01	6
_CitizenScience	16	18	9	23	0.008	11
User4	14	19	3	123	0.003	27
SLUBdresden	14	20	1	563	0.001	42

Table 9: Third study – Health and Healthcare - Top 20 users ranked by Indegree

Table 10 shows the results ranked by Outdegree, so the most retweeting accounts. In this case, we found just a bit more individuals (6 users compared to 4 individuals in the other ranking) being the rest project accounts. Besides, 3 of these other accounts are bots, whose only task is to retweet information (so obviously their values for Outdegree would be high). An interesting situation appeared in this case. Usually, individuals are more active when retweeting others, but in this case, we have projects very active in retweeting.

Name	InDeg		OutDeg		Betw.	
	Val	R	Val	R	Val	R
OpenSciTalk	0	662	39	1	0.0	537
SciStarter	30	6	38	2	0.023	3
CitieSHealthEU	96	1	30	3	0.06	1
CitSciMonth	16	17	23	4	0.01	6
B0tSci	0	722	16	5	0.0	630
CitSciOZ	9	43	16	6	0.031	2
Mosquito_Alert	14	22	15	7	0.009	9
RRIpeater	0	584	14	8	0.0	409
CitSci_Geek	2	264	14	9	0.001	56
User5	0	751	12	13	0.0	677

User6	0	788	12	14	0.0	744
SDGsbot	0	829	12	15	0.0	820
User7	5	89	12	12	0.002	31
User8	0	502	12	11	0.0	287
pwa_zurich	17	15	12	10	0.009	7
User9	0	594	11	16	0.0	421
User10	0	717	10	17	0.0	620
ScicommBot	0	769	10	18	0.0	711
cs_sdg2020	13	26	10	19	0.004	21
EuCitSci	7	55	9	20	0.009	8

Table 10: Third study – Health and Healthcare - Top 20 users ranked by Outdegree

4.3.4 Discussion

The foundation of this study is how Twitter has established itself as platform for dissemination of scientific information. In our case, we analysed the presence of discussion about CS and health and healthcare related topics. We found that 4.7% of the tweets we downloaded in one year were about this issue, an amount that could be enlarged by adding more keywords and that it will increase as time passes. The results from the filtering, combined with the language detection, mean that the pipeline for analysis is once again able to answer to RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?), being the conversation about CS in health and healthcare multitopical and multilingual.

It is important to note that we found several bots, which share large amount of information, but they do not check that information. We did not measure quality of the information, but we are aware of the importance of fighting against misinformation which also affects CS.

It seems from the results, that the most discussed topic are SDGs, which address health and healthcare too. There are goals inside the SDGs that are directly related to this topic, such as SDG6 (Clean water and sanitation), since unhygienic water can affect people´s health, or SDG3 (Good health and well-being), directly connected as it addresses health problems and well-being. Some others are not so directly connected but the interrelation can be sensed. For example, SDGs like SDG2 (No hunger), famine is a health problem, or SDG15 (Life on land), every living thing has health problems.

The hashtag analysis reinforced our belief that SDGs were the main topic, since the most used and retweeted hashtag was #SDGs. We calculated that 7% of the total appearances of the hashtag #SDGs in the complete (unfiltered) dataset were inside this health discussion. The rest of hashtags used inside this conversation suggest a variety of themes. What was surprising was the absence of a hashtag called #healthcare. Besides, we found interesting the low number of tweets related to COVID-19 by the time the study was conducted, although is one of the main used and interconnected hashtags. Another important topic arose from the topic modelling analysis, the Mosquito Alert, followed by topics such as Mental disorders, 100 days of code, research of rare disease or Water sanitation (directly connected to SDG 6).

Combining all the results, it seems like all the discussion about health and healthcare in Twitter inside the CS community occurs in the context of SDGs, a hot topic in Twitter as it has been previously analysed (Grover et al., 2021). We can link to SDGs the following topics: Mosquito Alert, Mental disorders, 100 days of code, Healthy diet, DHPSP, Mental disorders awareness, Pollution and health, Pollution and sustainable cities, COVID-19, Gamer for genomics, and Alzheimer monitoring. Most of these topics are also related to what the "who" monitors in relation to health and healthcare. So the existence of this conversation inside the framework of SDGs seems reasonable, while the users discuss about evolve around general health-related topics in this framework. We can see how some illnesses, processes or disorders are addressed by the users, such as dementia, depression, diabetes, anxiety, cancer or COVID-19.

We must not forget that we also had other topics that are more related to wildlife and other living things´ health, which also affects human health and therefore is understandable the presence of these topics. An important note, SDGs was not used as a keyword for the filtering of tweets, which reinforces our belief that this topic is of high importance and the main framework of discussion. The combination of all the techniques for content analysis seems to be helpful to gain a complete understanding of the topics addressed by the users, therefore being the pipeline able to answer to RQ2 (What is the lexical inventory in these conversations, i.e., hashtags, most common words, etc.?) and RQ3 (What are the main topics? Are these topics evolving in time?).

When analysing the mentions and retweets to unveil the dynamics inside the community and to discover the main actors in the conversation, thus answering to RQ4 (Can we define main actors through the SNA structural analysis?), we had results that do not deviate from a typical scenario in social media. Most mentioned accounts and most retweeted ones were institutions, projects, and organizations. Information spreading seems to be scarce according to the numbers of retweets received and given, especially if we point the low contribution on retweeting by individuals, who are known to be the main contribution to this dynamic.

4.3.5 Conclusions

This study was not aimed to analyse the influence of specific tweets but to shed light into the state of the conversation about health and healthcare in the CS community inside Twitter. Again, the study presents its limitations could analysed information from other social platforms to complement the data from Twitter, but this microblogging space has proven itself as a valuable source of information. It must be noted that the techniques applied in this case study also present they limitations. However, something to highlight is the possibility to replicate this analysis, not only in this same topic applying improved or new techniques, but also in any other research about other issues. This allows a long-term analysis that enriches the findings in time.

As an outcome from this study, we could translate the findings into policy recommendations. The findings could turn valuable for social media users, managers of online communities, policy makers, institutions, and many other stakeholders. Focusing on what concerns people about health and healthcare discovered via this analysis, policies could be applied to improve underrepresented sectors or strengthen those of general interest to people.

4.4 Static interactions and their influence on dynamic interactions between Twitter accounts in a CS context

In this study we wanted to test one hypothesis, whether the follow relation can influence other relations such as retweeting or mentioning. We performed a specific analysis which could translate into more content for the dashboard. First, all the users in the dataset were analysed and a classification model was trained and tested. Finally, we modelled a combined network with static and dynamic interaction to check the inference of the follow relation in other relations.

4.4.1 Motivation

The democratization of knowledge production and scientific collaboration has received a considerable contribution from the CS community. CS nowadays makes use of several social media platforms and web-based technologies that allows public collaboration (Newman et al., 2012a) (Nov et al., 2014) (Robson et al., 2013). Twitter facilitates the collaboration and communication between individuals and institutions, besides, the actual "citizens" produce a substantial amount of information and knowing the actual role they play and how they contribute to the knowledge creation has been matter of study since for some time (Krukowski et al., 2022).

To unveil the dynamics and quantify the contribution made by the two different profiles in this microblogging system, we aimed to combine machine learning, natural language processing, and network analysis, which are proven standard techniques that shed light on the characteristics and structure of the communications on Twitter (Mazumdar & Thakker, 2020).

This work applies an analysis based on the classification of users into two categories (individuals and institutions) using machine learning models, a solid foundation to our analysis (McCorriston et al., 2021) (Oentaryo et al., 2015) since is a common practice to better understand the relations between users (Kreutz & Daelemans, 2022) (Oentaryo et al., 2015). The classification of users is also of interest due to the differences in behaviour between these profiles (Oentaryo et al., 2015) and this binary distribution has already been addressed with high rates of success (McCorriston et al., 2021) (Oentaryo et al., 2015) using words used in tweets, average word length, users´ biography and others turned into embeddings (Kreutz & Daelemans, 2022).

Then we can use network analysis to study the interactions. In social media, two different types of interactions can be described (Tang & Liu, 2010). First, we can find "static" interactions since Twitter allows the follow relation, and "dynamic" interactions which are the retweets, replies, likes, or quotes. This analysis is aimed to provide valuable insights and recommendations for stakeholders involved in CS, since obtaining a better understanding about how dynamic interactions may be determined by the follow relation, can help to improve the impact and engagement of the accounts, and understand how the information spreads. During this case study, our aim was to find the most relevant features that help make a good classification of users. Also, we wanted to find a comprehensive and correct way to represent both static and dynamic interactions and analyse if these static interactions have an influence over the dynamic ones.

4.4.2 Methodology

In this section we will cover the different task performed to obtain the data, classify the users, and measure the performance of the classifier. Besides, we will also explain the steps followed to measure the dynamic and static interactions and how we represented them together and analysed the influence of the static interactions.

4.4.2.1 Data collection

Our colleagues of the CSTrack project from Duisburg University oversaw the data collection and classification of users. They used the tweets collected by the Lynguo tool for two years, from the 30th of September 2020 to the 26th of July 2022, based on CS related keywords for extraction. Then, they retrieved the information of those tweets by means of the Twitter API and limited the tweets to those in English to facilitate the processing. The resulting dataset contained 352,683 tweets and 114,488 users. The tweets from the Twitter API were accompanied by different data fields such as: username, text, contained entities (hashtags and so on), public metrics (retweets, likes, etc.) and profile picture and biography.

4.4.2.2 Classification of users

In the first step, our colleagues from Duisburg (from now on: they) undertook was the annotation. They manually labelled 2,448 randomly selected users as “personal” account or “institutional” account based on their biography. The resulting number was 1,779 personal accounts and 665 institutions. This unbalance is expected to occur in such a platform (Yan et al., 2013).

Next, they worked on the feature engineering. Based on the data fields: protected, verified, followers_count, following_count, listed_count, tweet_count and number of entities (McCorriston et al., 2021) (Yan et al., 2013). They also used a Python module to detect faces in the profile pictures to create another feature called face[40]. Moreover, they checked the names of the users and contrasted the data with a database of international names to create the feature namelist. Besides, they used the biography to transform it to embeddings using a BERT model with pretrained model all-MiniLM-L12-v2[41], a sentence-transformers model used to map sentences and paragraphs when using texts in the training of a classifier. The total number of features was extended to 348 and labelled as 1 or 0. The first analysis showed that there were differences between both types of accounts.

[40] https://pypi.org/project/face-recognition/

[41] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

For the final classification of the complete database, they used different models: Logistic Regression, Random Forests, TabNet and XGBoost (Arik & Pfister, 2020) (Breiman, 2001) (T. Chen & Guestrin, 2016) in Python. The database was split into 70/30 for training and measured based on the precision, recall and F1 (Manning et al., 2008).

4.4.2.3 Network Extraction

The role of the author of this thesis and his supervisors (from now on we) supervised the dynamic networks creation. These networks were created by means of the networkX package from Python. We created +a directed multigraph with the users being the nodes and the edges the different dynamic interactions that were extracted from the set of tweets. These three interactions were: retweeting, mentioning, and quoting.

The type of edge was stored as an attribute. The direction of the edges was calculated as if user u retweets, mentions, or replies to user v the edges is (u, v). The resulting network contained 107,413 nodes and 258,704 edges. Figure 52 shows a schedule of how the structure of the network was and how the flow of information occurs, since the direction of the dynamic interaction is opposite to the direction of the edge.

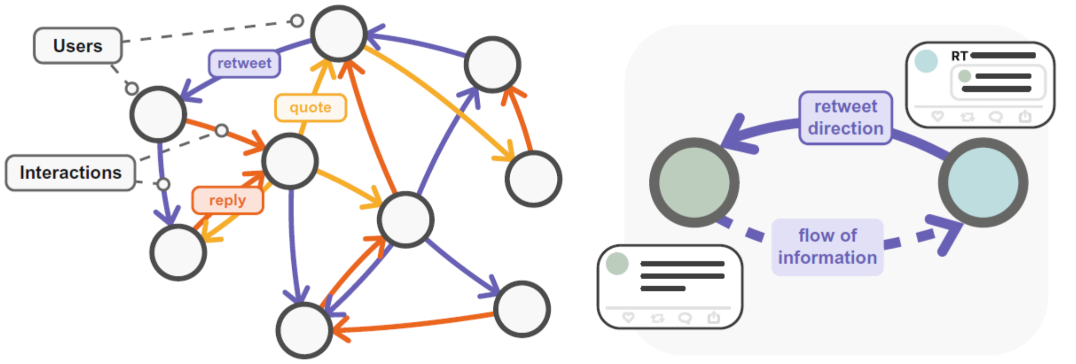


Figure 52: Structure of the network containing different interactions as edges on the left. On the right, schedule of flow of information.

To evaluate the influence of the static interactions on the dynamic ones, we created another network based on the follow-relation. This information was obtained by means of the Twitter API, which provides functions to extract the list of followers and followees (users that the source follows). The immense number of relations that we would need to manage presented itself as the main limitation (especially due to the time needed to extract all the users), so in this study we aimed to check ego-networks. An ego-network contains all the nodes that are linked to a central node (the ego) and containing links to neighbours, also called 1.5-neighborhood.

We decided to analyse the ego-network around SciStarter, one of the main actors in the CS community on Twitter, which was also analysed in 4.3 An analytics approach to health and healthcare in CS communications on Twitter. So, by means of the Twitter API we extracted all the followers of SciStarter, all the users that SciStarter follows and the 1.5-neighborhood (the users that the followers and followees follow and are followed).

The resulting static networks served as the groundings for the combined network we created next, the one containing both types of interactions. To do so, we filtered the dynamic interactions to only retain those users present in SciStarter´s ego-network. The last step was to label the final nodes as “Follower” or “Not follower” to examine the information flow according to this relation. The number of unique interactions got reduced to 44,229 unique values.

To measure the information flow and interaction behaviour, we again used the centrality measures. We revised the degree centrality (number of adjacent users), indegree (received interactions) and outdegree (interactions done by the user).

4.4.3 Results

Our colleagues started the analysis with the classification of users. First, they checked the performance of the classifier using and not using the biography of the accounts. The performance of the classifiers with and without the biography embeddings was over 0.8 and F1 values over 0.89.

The scores with embeddings were clearly higher resulting in a 92.5% of users correctly classified. The chosen model was Logistic Regression with a 0.939 accuracy and 0.948 F1 score, leading to a distribution of 30,214 (26.4%) institutions and 84,2564 (73.6%) institutions in the complete dataset. The dataset of users from the dynamic interactions contained 27,124 (25.3%) institutions and 80,289 (74.7%) personal accounts. The dataset belonging to the ego-network users contained 5,596 (39.0%) institutions and 8,769 (61.0%) personal accounts.

Once the classification was done, the information about the type of users was ready to check the influence of the static interactions. We started the network analysis and first analysed the degrees across the network. The average degree for all dynamic interactions was 4.87, calculated using the degree for each node and calculating the statistical mean, ranging from 1 to 7,567. This distribution shows a heavy-tailed appearance, meaning that most of the users are not so active while a small number of them are highly interacting.

Then, we checked the degree by interaction and revealed that institutions had higher indegree than personal accounts, meaning that they received more retweets pointing at them as information sources. Personal accounts had slightly higher outdegrees, they retweet more than they receive. This points towards the standard situation in social media on which institutions act as information sources and individuals find that information interesting and share that information. These values are especially representative in retweets, so we checked the distribution of interactions by type of account (source, target).

Most of the interactions, as shown in the Table 11, occur between personal accounts and institutional accounts, being the institutions the target of the interaction. As we expected, retweets were the most common interaction (Mazumdar & Thakker, 2020). It also seems like institutions also retweet other institutions helping information spreading.

User type			**Interaction type**		
Source	**Target**	**Total**	**Retweeting**	**Quoting**	**Replying**
Institution	Institution	65,358 (25%)	51,661 (79%)	8,527 (13%)	5,170 (8%)
Institution	Personal	28,314 (11%)	23,527 (83%)	3,396 (12%)	1,391 (5%)
Personal	Institution	86,236 (33%)	73,109 (85%)	11,856 (14%)	1,271 (1%)
Personal	Personal	78,779 (31%)	60,261 (77%)	9,749 (12%)	8,769 (11%)

Table 11: Fourth study- Static vs Dynamic Interactions-Distribution of interactions by user type

To check the influence of the static interactions in the dynamic interactions, we created the forementioned ego-network containing the followers and followees of SciStarter alongside the 1.5-neighborhood.

In the first analysis presented in Figure 53, institutions showed again a higher number of followers compared to personal accounts, more precisely almost three times higher in both cases (static and dynamic).

We then decided to check the source and target of the different types of interactions based on the characteristic we previously stored for each node stating if they were “follower” or “Not follower”. The distribution is showed in Table 12, seeing again that retweeting keeps on being the most common interaction, specifically retweeting and institution because the user follows it. Retweeting an account that is not being followed is the second most common interaction. The results show that institutions in this ego-network are slightly more active in retweeting. An interesting finding is that when it comes to replies, it is more common to reply to someone that is not being followed.

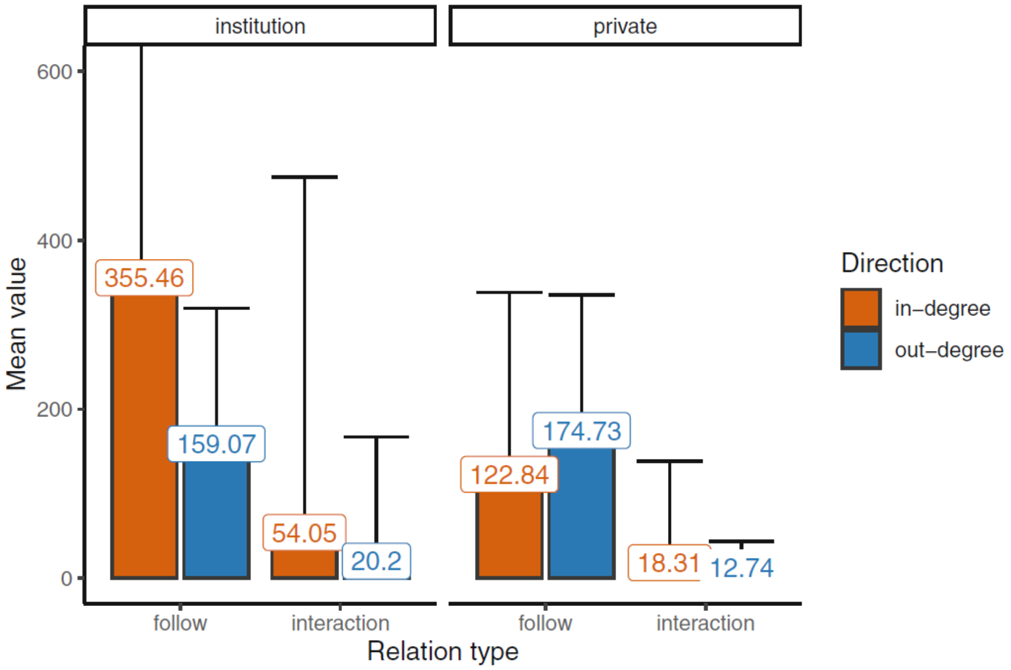


Figure 53: Fourth study- Static vs Dynamic Interactions-Mean value for follow and the interactions by user type

User type	Type of interaction	Follow relation	Count	Proportion
Institution	retweeted by	Follower	19,246	43.5%
	retweeted by	Non follower	3,398	7.7%
	quoted by	Follower	2,266	5.1%
	quoted by	Non follower	659	1.5%
	replied to	Follower	253	0.6%
	replied to	Non follower	1,749	4.0%
Personal	retweeted by	Follower	10,984	24.8%
	retweeted by	Non follower	2,001	4.5%
	quoted by	Follower	1,975	4.5%
	quoted by	Non follower	553	1.3%
	replied to	Follower	309	0.7%
	replied to	Non follower	834	1.9%

Table 12: Fourth Study-Static vs Dynamic Interactions-Breakdown of type of interaction by user type and follow relation

4.4.4 Discussion

This case study aimed to determine whether the characteristics from the Twitter profiles were useful when classifying user types and, once this classification was done, if the static interactions between users influence the dynamic interactions.

The high scores from the classifiers led us to believe that the data stored in the profiles is useful when undertaking the task of classifying. Besides, adding extra features as the picture's information and the embedding from the biographies increase the performance of the models and their accuracy.

We also observed differences between the behaviours of both types of users, moreover, there were differences between the different dynamic interactions. Retweeting was the most common type of interaction for both personal and institutional accounts. Besides, we saw that the retweets were mainly given by personal users acting as "information spreaders", which aligns with previous findings (McCorriston et al., 2021). This insight aligns with the search of main actors in this thesis, RQ4 (Can we define main actors through the SNA structural analysis?), since we see how institutions are mainly content created and individuals act as information spreaders.

Using the ego-network and 1.5 neighbourhood to examine the influence of the static topology on the dynamic interactions, we were able to see that retweets occur primarily because the source account follows the target account. However, we found that retweeting an account that is not being followed is the second most common interaction, what led us to believe that this information that is retweeted comes from one of those 1.5-neighbours. Quoting tends to appear mainly when the source account is being followed. This information also highlights how the main actors in the conversation behave, institutions are highly followed and their followers interact with them, while we see more interactions without following interactions for individuals. This also provides a new dimension of understanding for our RQ4 (Can we define main actors through the SNA structural analysis?).

4.4.5 Conclusions

This simple yet accurate and novel method helped us to model the interactions occurring inside Twitter and the cause why the interactions occur in these communities. As in previous studies, we once more find retweeting to be the most common activity inside this community, and how the institutions are those with which individuals interact the most. These findings could help stakeholders to better distinguish between the types of users that interact with them, understand how information flow occurs and therefore apply policies to improve science communication.

Of course, this study comes with limitations. In future work we could move to a non-binary classification, including more types of account leading to a more complex profiling and helping to get a deeper understanding of the dynamics and moving to a higher scale analysis when performing SNA. However, this analysis could trigger a better comprehension of the structure and behaviours inside social media and improve information dissemination. Another technique that could turn interesting and helpful could be content analysis combined with the previous techniques.

4.5 Conclusions from all case studies

The different case studies we performed were destined to answer our first four RQs while designing the pipeline for analysis. Analysing the complete dataset and filtering using different filters to find conversation about distinct topics would help answer the RQ1. Applying the filters, we were able to find conversations about SDGs, learning, health and healthcare. The tweets from these conversations represented a 21.28% of the total tweets about CS. However, this number is not representative since the tweets were analysed in different moments. The analysis performed about learning was done in an early stage of the collection of tweets, while the number of tweets about SDGs and health and healthcare are comparable. The total number of tweets about SDGs was 275.868 while about health and healthcare was 365.609, with a difference of three months between studies. What it seems clear from these analyses is that SDGs is indeed an important topic, as it was stated before (Shulla et al., 2020). We find SDGs as a hashtag in the first three studies, implying that a conversation about SDGs is present in all three cases or even that the conversation occurs in the framework of SDGs. The filters seem to work as we can find different topics, so we get a partial answer to RQ1, since the filter works, and we find multi-topical conversations.

To completely answer RQ1 we performed the language detection. In all the first three case studies we find different languages in the dataset, especially because we provided words in different languages to the filter. If we had just used words in English we would find even less results in other languages, only being helped by the hashtags to increase the acquisition of tweets (hashtags are used indistinctly in different languages, especially if they are acronyms). In the light of these results, we can state that we answered RQ1, and we find a multi-topical and multilingual conversation in the CS community in Twitter. But there is something that must be considered, the collection of tweets about CS is done via a very restrictive query searching for CS related words. Thus, we may find tweets about CS without using some of the words in the filter, which would enrich our data, therefore not being something counterproductive but an opportunity to iterate over this approach. Besides, the search of topics inside the CS community could also be extended using more words for the filters. This is aligned with the necessity of iterating over the data, a standard that is not widely followed nowadays although being a foundation in data analysis (Srivastava & Hopwood, 2009).

RQ2 and RQ3 are related to the content inside the collection of tweets. They are directly connected since the acquisition of the lexical inventory helps determining the topics of discussion, as hashtags are used to label the tweets, most common words are also helpful to determine what terms are most addressed by the uses, and TF-IDF shows those terms of importance in the collection of texts and not only those that are most used. When checking the results from the hashtag analysis the most prominent finding is that SDGs is the most used hashtag in all the cases, meaning that is a topic of high importance for the member of the CS community. The hashtags present in both the case study about learning and SDGs show how important is the environment and the nature sciences, due to the high number of hashtags related to these fields. Besides, in the list of hashtags from the health study we also found hashtags related to animals, ecology, and sustainability, demonstrating once more how strongly linked CS and the nature sciences are.

From the hashtag analysis, it is important to highlight how useful the temporal analysis turned out to be. Seeing the differences in usage and retweeting in different periods helps discovering happenings or event that trigger the usage. The best example is how the conference citscihelvetia2021 triggered the usage of that hashtags and then quickly dropped. Or, how interesting would be to discover why in early November 2021 there was that high usage and retweeting of the hashtag #sdgs. This specific technique would be important for stakeholders.

It is also important to highlight that including the TF-IDF is an added value to the most common words analysis. We see clear differences between the most used words and the

TF-IDF terms in the SDGs case study, evincing that there are underlying words that could be more important or at least a good addition to analyse the topic.

RQ3 gets its answer when combining the previous analysis with the topic modelling. BERT as our model for topic modelling showed a good performance and we were able to determine the topics of discussion inside the health analysis. Although LDA also performed well, the easiness of BERT made it our choice for the pipeline for analysis.

When addressing the SNA structural analysis, we saw that adding the network of hashtags is a perfect combination of content and structural analysis, helping us determine which topics are related. This is enriched with the hierarchical analysis of topics. Besides, addressing specific cases could lead to more combined analysis such as the one we did with the retweeting of SDGs. Getting a general state of the discussion about a certain topic could trigger ideas in researchers and stakeholders.

The analysis of mentions and retweets, alongside the centrality measures, was of use to discover the main actors in the different conversations. The statement about retweets being the most used interaction inside Twitter was proved, aligned with the discoveries from Mazumdar (Mazumdar & Thakker, 2020). But analysis the centrality measures showed a divergence in the general trend of popular accounts not interacting as much as the individuals do. We found many projects active in retweeting and mentioning. In the study about health, we had 39 projects in the top 50 accounts that mention other users. Besides, projects were more active in retweeting than the individuals in this case too. So, applying this pipeline for analysis could help determine in which fields reciprocity between projects and individuals is ideal and in which should be improved to enhance the information flow.

Our last case study was also helpful to, once more, confirm that retweets are of high importance in this community. But moreover, that the follow relation is essential in the distribution of interactions. All the interactions tend to occur due to an existing follow relation, except for the replies, which happened more often when not following. Integrating this analysis inside the pipeline for analysis would be a good addition for the future, to analyse specific cases.

5. Conclusions and future work

In this section we will discuss the contribution of this thesis to research reviewing the key aspects of the dashboard and its usefulness in the different case studies.

5.1 Contributions

This thesis has been aimed at providing a standardized pipeline for analysis of social networks and to integrate that analysis into a platform in which anyone could replicate any analysis and gain complete knowledge of the texts of study. The standard pipeline for analysis was designed to palliate the lack of a standard process to analyse data from social networks, since researchers tend to apply isolated techniques in specific case studies, neglecting the possible information obtained from complementary techniques (Liang & Fu, 2015). Also, integrating the analysis into a platform was our way to palliate another specific problem in SNA, the lack of reproducibility of the different studies (Mamo et al., 2023). The problems in reproducibility may come from difficulties to replicate the exact same techniques and processes used by researchers, to even the impossibility of performing the analysis due to the lack of skills, so with the platform we wanted to offer a solution to both possibilities.

To measure the completion of our idea, we proposed different goals to be achieved and different research questions to validate our goals. Our first goal, G1: Gaining knowledge and deepen in the content created and shared by the CS community, was the objective we should achieve when selecting, designing, and adjusting the different techniques for SNA. If the different techniques provide a complete understanding of the CS community in the different levels, content and structure, our pipeline for analysis was completed and therefore achieving our second goal, G2: To provide a standardized pipeline for SNA, useful for multiple topics. This pipeline for analysis would fall in the same problems of reproducibility and difficulty of usage if it was simply presented and detailed, so we formulated another goal, G3: To provide a comprehensive and easy-to-use platform to perform SNA despite the background of the user, a platform that reunites all the techniques and that could be used in the future in more diverse topics too. To validate our G1 and G2, we designed the different case studies about the CS community on Twitter.

To validate the obtaining of a complete understanding of the CS community and designing a pipeline for analysis useful for multiple topics we formulated four research questions regarding the different dimensions of the SNA, the content, and the structure of the networks. To access to different subjects and get understanding about the content and topics shared by the community, we designed a filter by keywords based on regex to extract coincidences of words in the corpus of texts. As shown in the different case studies, 4.1 Open Learning in Citizen Science, 4.2 Understanding the discussion of CS around SDGs, and 4.3 An analytics approach to health and healthcare in CS communications on Twitter, we were able to extract subsets of data about different themes. This filter by keywords offers a way to extract pieces of information in a corpus of texts and a partial answer to our RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?). We also selected a technique to extract the languages of the dataset, with which we were able to determine the different languages present in our dataset and that could be used for any set of texts. With these two techniques we obtain the answer for the RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?).

But we deepened in the content analysis and designed other techniques to extract the most used and retweeted hashtags, an important feature in SNA (Nawaz et al., 2022; Small, 2011). It allows the discovery of important events via the temporal analysis of hashtags, helping the community to maybe unveil new events they were not aware of or finding specific happenings worth investigating, therefore answering the RQ2 (What is the lexical inventory in these conversations, i.e., hashtags, most common words, etc.?) and also providing information about RQ1 (Is there a multi-lingual and multi-topical conversation in the CS community?) and RQ3 (What are the main topics? Are these topics evolving in time?). To give some examples, using the hashtag analysis we were able to determine the most used hashtags analysing Open learning (4.1 Open Learning in Citizen Science), in the analysis of eHealth and healthcare (4.3 An analytics approach to health and healthcare in CS communications on Twitter) and studying the relation of CS with the SDGs (4.2 Understanding the discussion of CS around SDGs). In these case studies we were able to determine that SDGs is an important topic in all the scenarios, so the CS community discusses and uses hashtags in the framework of the SDGs. Also, we find hashtags such as sustainability or SDGs, but also most common words like sustainable or sustainability. We also designed a way to extract the most used words, and seeing the most common words we can also have a glimpse of the most addressed topics or important ones, besides, knowing highly repeated words can help the discovery of bots, unnecessary terms for analysis or even events or happenings in time when combined with temporal analysis. This is also another way to answer RQ2 (What is the lexical inventory in these conversations, i.e., hashtags, most common words, etc.?) and an answer for RQ3 (What are the main topics? Are these topics evolving in time?). Alongside the analysis of hashtags and the count of most used words, we selected topic modelling techniques to extract the most addressed themes by the community, an important part of the content analysis (Hagras et al., 2017; Prabhakar Kaila & Prasad, 2020; Uthirapathy & Sandanam, 2023). Using the topic modelling we were able to extract the different subjects of conversation when the CS community on Twitter tweets about SDGS (4.2 Understanding the discussion of CS around SDGs) and about health and healthcare (4.3 An analytics approach to health and healthcare in CS communications on Twitter). We did not just extract the topics, but we also implemented the analysis over time, so we were able to determine that CS community discusses mainly about climate change and SDGs related to climate when writing about the SDGs, and also that even when tweeting about health and healthcare it seems to be in the framework of SDGs and that the conversation about the mosquito alert is the most stable in time only being passed by the discussion about water sanitation in July 2021, possibly due to an event related to it. With the topic modelling we were able to determine the most important subjects in any set of texts, so we answer our RQ3 (What are the main topics? Are these topics evolving in time?).

Another dimension in the content analysis we decided to check was the sentiment analysis. Checking the sentiment, it is possible to reformulate our contributions in the form of tweets to gain better impact using expressions linked to positive sentiments. Besides, sentiment analysis allows the discrimination of the type of content created in a specific topic, normally news and research are linked to neutral sentiments, while personal opinions are linked to positive or negative results (Al-Rubaiee et al., 2016). This specific analysis offers more context about the topics and the information shared, helping to determine if the users are reacting positively or negatively to specific topics, or if the information shared is neutral. This is way ahead of a simple answer to RQ3 (What are the main topics? Are these topics evolving in time?), but some additional information.

To achieve our G1 (Gaining knowledge and deepen in the content created and shared by the CS community) and therefore our G2 (To provide a standardized pipeline for SNA, useful for multiple topics), we also implemented the structural analysis, to investigate the relations between users. Across the results presented in the case studies, we used the analysis of retweets with which we could see the structure of the accounts involved in that conversation allowing the interpretation of the information flow (Prabhakar Kaila & Prasad, 2020; Scanfeld et al., 2010). Alongside the visual analysis of the networks, we provide the centrality values, useful to interpret the connections and evaluate which accounts are those that create information, which are active in spreading information and which are those important nodes acting as bridges or linkers between communities based on the eigenvector and betweenness. All these analyses are important to answer the RQ4 (Can we define main actors through the SNA structural analysis?), as we can see in the case studies we were able to determine most influential accounts, accounts that retweet the most and accounts that act as bridges in the information spreading, which is an important outcome for stakeholders that may want to find these prominent accounts or those accounts that have the strength to actively share information.

With all the different techniques we presented we were able to obtain a complete understanding of the CS community, as we tested with the case studies, achieving our G1 (Gaining knowledge and deepen in the content created and shared by the CS community). Besides, we proved the usefulness of our techniques and how they can serve as a pipeline for analysis, completing our G2 (To provide a standardized pipeline for SNA, useful for multiple topics). As it can be seen, all the different techniques deployed in the dashboard serve to answer all the different research questions formulated to validate the pipeline for analysis, thus it is useful to gain knowledge about the CS community in Twitter. All these techniques can be used and obtain a visualization without technical knowledge, the results are just some clicks away.

Something important is that the dashboard is currently designed for CS, but it will serve as a tool for any subset of tweets about any topic just by having the same data structure. With a simple data transformation any set of tweets can be used and obtain the same insights.

Since the usage of the dashboard is aimed to be easy and comprehensive, which is our G3 (To provide a comprehensive and easy-to-use platform to perform SNA despite the background of the user, a platform that reunites all the techniques and that could be used in the future in more diverse topics too), we believe that users may not need technical knowledge and possibly making it usable for a wide range of audiences. Based on the potential audiences, we defined three levels of contribution:

- Any individual can use the dashboard to explore a determined topic, search for themselves in Twitter and check their environment, gain knowledge about CS, and start following accounts that could be of interest, and many other situations.
- Researchers can also use the dashboard to replicate the SNA analysis we designed to obtain insights helpful enough to answer their questions.
- Project managers or institutional accounts can use the dashboard to monitor their community and position inside the CS community. Also, keep up to date with the interests of the CS users in Twitter to refine their engagement and content.
- In a higher level of importance, this dashboard was also aimed at providing useful insights from the CS community to potential policy makers. This complete analysis of the community could be used to determine which initiatives should be funded based on their importance for the community. Moreover, it can be used to determine if any specific topic is underrepresented to not fund initiatives about it or, on the contrary, try to raise the awareness of that specific matter.

Lastly, this type of information is usually not of public access, so our goal is to be able to deploy this dashboard in a public IP, aligned with the idea of democratization and collaboration promoted by the CS community and the SDGs.

5.2 Advantages and limitations

Most of the advantages of the dashboard and therefore the pipeline for analysis have already been addressed, but of course such type of analysis comes with some limitations. In this thesis, we are working on providing a solution for all those limitations we have detected.

One of the main advantages is the possibility of performing analysis for the complete CS community or for a specific topic inside of it. The limitations come with the filtering, as more keywords can be used in any analysis and some tweets about the topics can be neglected. This situation could happen to any user of the dashboard, or it could have happened to us in the case studies. Analysing the most common words and the TF-IDF can help to palliate this situation, including more detected keywords easily to the filter fields. Besides, another important advantage is the possibility to replicate the analysis as many times as needed, which will also serve as a solution for the previous explained issue.

In relation to the content analysis, it seems clear that the techniques provide good results in retrieving the lexical inventory and the topics of discussion. Since this process is reliant on AI, to satisfactorily execute it a machine of high performance is needed. Although the algorithms have been prepared to correctly function with just the CPU, to obtain a better performance it would be recommendable to utilize GPU.

Another important advantage is the possibility to perform structural analysis and obtain a clear visualization of the most important users. The main limitation here is the preservation of personal information and data protection. We need the dashboard to be able to offer names and information from institutions and projects while preserving the information of the individuals. We are currently working on this, since in our last case study we developed a highly accurate user profiling that will serve to anonymize individuals.

Once more, we must point out that the main advantage of the dashboard with the pipeline for analysis is the possibility to replicate the analysis in time, repeating the same analysis to see if any changes occurred. This highlights the last limitation, the access to the Twitter API. Since lately the access to the Twitter API has been involved in controversy because it has been removed the free access to the API in some situations and granting free access to other developers in specific cases. To perform the continuous analysis of tweets of CS we will search for the better alternative according to the situation of the API. It is also important to say that the dashboard will be useful for any type of set of tweets that follow the data structure, which will make the dashboard useful although the current situation of the API.

5.3 Current and future work

We are currently working on a classification of the complete set of users to provide a better result of the accounts inside the sentiment, degree, and network calculations. Besides, we will include more techniques to the dashboards as currently we are missing the TF-IDF analysis, the network of hashtags and network of mentions.

In the future, we will work on the improvement of the visualizations to make them more appealing and probably include more types of graphs such as the network of hashtags or the network of mentions, which also provide important information about the content and the main actors in the conversation. In relation to networks, we will include the combined network with static and dynamic interactions, which will complete the structural analysis providing important information about the behaviours of the users.

An important future perspective is deploying the dashboard in a public IP inside the Universidad Rey Juan Carlos´ computational cluster. This cluster allows the creation of powerful virtual machines, perfect to calculate machine learning methods fast. Besides, we will substitute the sentiment analysis for another sentiment analysis performed using machine learning.

Lastly, the dashboard will be updated to perform analysis with other sets of tweets that follow the data structure we defined for the pipeline for analysis. When adding this functionality, we would need to test whether the dashboard works adding this new CSV with the same structure before deploying the dashboard with this functionality.

5.4 Publications

This thesis has five publications, three in journals and two in international conferences. Furthermore, there is one paper under review process, in which we explain the architecture and the usage of the dashboard, submitted to a JCR journal. There is another paper in elaboration process about the application of a BERT model for sentiment analysis. We also created two GitHub repositories to store the dashboard and the techniques.

5.4.1 Journals

- Martínez-Martínez, F., Roldán-Álvarez, D., Martín, E., & Hoppe, H. U. (2023). An analytics approach to health and healthcare in citizen science communications on Twitter. *Digital Health*, *9*, 20552076221145349. Impact factor: 3,9. Quartile: Q2. Category: Medical informatics.
- De-Groot, R., Golumbic, Y. N., Martínez Martínez, F., Hoppe, H. U., & Reynolds, S. (2022). Developing a framework for investigating citizen science through a combination of web analytics and social science methods—The CS Track perspective. *Frontiers in Research Metrics and Analytics*, *7*, 988544.
- Roldán-Álvarez, D., Martínez-Martínez, F., Martín, E., & Haya, P. A. (2021). Understanding discussions of citizen Science around sustainable development goals in Twitter. *IEEE Access*, *9*, 144106-144120. Impact factor: 3,476. Quartile: Q2. Category: Computer Science, Information systems.

5.4.2 International Conferences

- Krukowski, S., Martínez-Martínez, F., & Hoppe, H. U. (2023, August). Differential Characteristics and Collaborative Interactions of Institutional and Personal Twitter

Accounts in a Citizen Science Context. In *International Conference on Collaboration Technologies* (pp. 68-83). Cham: Springer Nature Switzerland.

- Roldán-Álvarez, D., Martínez-Martínez, F., & Martín, E. (2021, July). Citizen science and open learning: A Twitter perspective. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 6-8). IEEE.

5.4.3 GitHub repositories

- CSTrack_URJC: https://github.com/FernanSLN/CSTrack_URJC. This repository contains the designed techniques and each iteration to obtain results for the case studies.
- CSTrack_Docs: https://github.com/davidrol6/CSTrack_Docs. This contains the dashboard and documentation about the dashboard and techniques.

Acknowledgements

This work has been funded by the European Community's call H2020-SwafS-2019-1 (H2020-EU.5.c) under grant agreement no. 872522.

Ethical approval

This thesis received the approval through the Ethics Committee of the Universidad Rey Juan Carlos, Spain, with the reference linked to the project CS-TRACK. Expanding our knowledge on citizen science through analytics and analysis: ENM61/ 201303202209622.

References

Adamic, L., Buyukkokten, O., & Adar, E. (2003). A social network caught in the Web. *First Monday*, *8*(6). https://doi.org/10.5210/fm.v8i6.1057

Ahmed, W., Vidal-Alaball, J., Downing, J., & López Seguí, F. (2020). COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *Journal of Medical Internet Research*, *22*(5), e19458. https://doi.org/10.2196/19458

Al Kurdi, B., Alshurideh, M., & Salloum, S. A. (2020). Investigating a theoretical framework for e-learning technology acceptance. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(6), 6484. https://doi.org/10.11591/ijece.v10i6.pp6484-6496

Ali, W. (2020). Online and Remote Learning in Higher Education Institutes: A Necessity in light of COVID-19 Pandemic. *Higher Education Studies*, *10*(3), 16. https://doi.org/10.5539/hes.v10n3p16

Al-Rubaiee, H., Qiu, R., & Li, D. (2016). The Importance of Neutral Class in Sentiment Analysis of Arabic Tweets. *International Journal of Computer Science and Information Technology*, *8*(2), 17-31. https://doi.org/10.5121/ijcsit.2016.8202

Álvarez, D. R. (2020, diciembre 31). *Evolution of academic publications in citizen science* (World) [Text]. Https://Cstrack.Eu/; Citizen Science | CS Track Project. https://cstrack.eu/format/reports/evolution-academic-publications-citizen-science/

Alvarez, R. M., & VanBeselaere, C. (2005). Web-Based Survey. En *Encyclopedia of Social Measurement* (pp. 955-962). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00390-X

Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., & Vespignani, A. (2005). *k-core decomposition: A tool for the visualization of large scale networks*. https://doi.org/10.48550/ARXIV.CS/0504107

Amano, T., Smithers, R. J., Sparks, T. H., & Sutherland, W. J. (2010). A 250-year index of first flowering dates and its response to temperature changes. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1693), 2451-2457. https://doi.org/10.1098/rspb.2010.0291

Arik, S. O., & Pfister, T. (2020). *TabNet: Attentive Interpretable Tabular Learning* (arXiv:1908.07442). arXiv. https://doi.org/10.48550/arXiv.1908.07442

Aristeidou, M., & Herodotou, C. (2020). Online Citizen Science: A Systematic Review of Effects on Learning and Scientific Literacy. *Citizen Science: Theory and Practice*, *5*(1), 11. https://doi.org/10.5334/cstp.224

Arnaboldi, V., Conti, M., Passarella, A., & Pezzoni, F. (2012). Analysis of Ego Network Structure in Online Social Networks. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 31-40. https://doi.org/10.1109/SocialCom-PASSAT.2012.41

Aroyo, L., & Dicheva, D. (2001). AIMS: Learning and teaching support for WWW-based education. *International Journal of Continuing Engineering Education and Lifelong Learning*, *11*(1/2), 152. https://doi.org/10.1504/IJCEELL.2001.000390

Askitas, N., & Zimmermann, K. F. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, *36*(1), 2-12. https://doi.org/10.1108/IJM-02-2015-0029

Ayres, P. (2008). Imperial nature. Joseph Hooker and the practices of Victorian science. *Annals of Botany*, *102*(4), 657-658. https://doi.org/10.1093/aob/mcn144

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 44-54. https://doi.org/10.1145/1150402.1150412

Bajgoric, N. (2001). Information Technologies for Virtual Enterprise and Agile Management. En *Agile Manufacturing: The 21st Century Competitive Strategy* (pp. 397-416). Elsevier. https://doi.org/10.1016/B978-008043567-1/50021-8

Bales, E., Nikzad, N., Quick, N., Ziftci, C., Patrick, K., & Griswold, W. (2012). Citisense: Mobile Air Quality Sensing for Individuals and Communities. Design and deployment of the Citisense mobile air-quality system. *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*. 6th International Conference on Pervasive Computing Technologies for Healthcare, San Diego, United States. https://doi.org/10.4108/icst.pervasivehealth.2012.248724

Batagelj, V., & Zaversnik, M. (2003). *An O(m) Algorithm for Cores Decomposition of Networks*. https://doi.org/10.48550/ARXIV.CS/0310049

Bernard, H. R. (2005). The development of Social Network Analysis: A Study in The Sociology of Science. *Social Networks*, *27*(4), 377-384. https://doi.org/10.1016/j.socnet.2005.06.004

Birnbaum, M. H. (2004). Human Research and Data Collection via the Internet. *Annual Review of Psychology*, *55*(1), 803-832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Blaszka, M., Burch, L. M., Frederick, E. L., Clavio, G., & Walsh, P. (2012). #WorldSeries: An Empirical Examination of a Twitter Hashtag During a Major Sporting Event. *International Journal of Sport Communication*, *5*(4), 435-453. https://doi.org/10.1123/ijsc.5.4.435

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bodaghi, A., & Oliveira, J. (2022). The theater of fake news spreading, who plays which role? A study on real graphs of spreading on Twitter. *Expert Systems with Applications*, *189*, 116110. https://doi.org/10.1016/j.eswa.2021.116110

Bonney, R. (1996). *Citizen Science: A lab tradition*. *Living Bird*(15), 7-15.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977-984. https://doi.org/10.1525/bio.2009.59.11.9

Bonney, R., Phillips, T. B., Ballard, H. L., & Enck, J. W. (2016). Can citizen science enhance public understanding of science? *Public Understanding of Science*, *25*(1), 2-16. https://doi.org/10.1177/0963662515607406

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next Steps for Citizen Science. *Science*, *343*(6178), 1436-1437. https://doi.org/10.1126/science.1251554

Borgatti, S. P., & Everett, M. G. (2006). A Graph-theoretic perspective on centrality. *Social Networks*, *28*(4), 466-484. https://doi.org/10.1016/j.socnet.2005.11.005

Boving, A. T., Shuster, C. L., Walls, T. A., & Brothers, T. (2021). Personal digital health in Parkinson's disease: Case histories and commentary. *DIGITAL HEALTH*, *7*, 20552076211061925. https://doi.org/10.1177/20552076211061925

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Bozkurt, A., & Sharma, R. C. (2020). *Emergency remote teaching in a time of global crisis due to CoronaVirus pandemic*. https://doi.org/10.5281/ZENODO.3778083

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brown, P. (1997). Popular Epidemiology Revisited. *Current Sociology*, *45*(3), 137-156. https://doi.org/10.1177/001139297045003008

Brown, P. (2007). *Toxic Exposures: Contested Illnesses and the Environmental Health Movement*. Columbia University Press. https://doi.org/10.7312/brow12948

Bruzzese, S., Ahmed, W., Blanc, S., & Brun, F. (2022). Ecosystem Services: A Social and Semantic Network Analysis of Public Opinion on Twitter. *International Journal of Environmental Research and Public Health*, *19*(22), 15012. https://doi.org/10.3390/ijerph192215012

Burgess, J., & Baym, N. K. (2022). *Twitter: A Biography*. NYU Press.

Callon, M., & Rabeharisoa, V. (2008). The Growing Engagement of Emergent Concerned Groups in Political and Economic Life: Lessons from the French Association of Neuromuscular Disease Patients. *Science, Technology, & Human Values*, *33*(2), 230-261. https://doi.org/10.1177/0162243907311264

Carvalho, J. P., Rosa, H., Brogueira, G., & Batista, F. (2017). MISNIS: An intelligent platform for twitter topic mining. *Expert Systems with Applications*, *89*, 374-388. https://doi.org/10.1016/j.eswa.2017.08.001

Catlin-Groves, C. L. (2012). The Citizen Science Landscape: From Volunteers to Citizen Sensors and Beyond. *International Journal of Zoology*, *2012*, 1-14. https://doi.org/10.1155/2012/349630

Ceccaroni, L., Woods, S. M., Sprinks, J., Wilson, S., Faustman, E. M., Bonn, A., Greshake Tzovaras, B., Subirats, L., & Kimura, A. H. (2021). Citizen Science, Health, and Environmental Justice. En K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 219-239). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_12

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280-294. https://doi.org/10.1016/j.biocon.2016.09.004

Chen, M., Gao, C., Song, M., Chen, S., Li, D., & Liu, Q. (2020). Internet data centers participating in demand response: A comprehensive review. *Renewable and Sustainable Energy Reviews*, *117*, 109466. https://doi.org/10.1016/j.rser.2019.109466

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Coleman, W. (1977). *Biology in the Nineteenth Century: Problems of Form, Function and Transformation*. Cambridge University Press.

Committee on Designing Citizen Science to Support Science Learning, Board on Science Education, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine. (2018). *Learning Through Citizen Science: Enhancing Opportunities by Design* (R. Pandya & K. A. Dibner, Eds.; p. 25183). National Academies Press. https://doi.org/10.17226/25183

Cooper, C. B., Shirk, J., & Zuckerberg, B. (2014). The Invisible Prevalence of Citizen Science in Global Research: Migratory Birds and Climate Change. *PLoS ONE*, *9*(9), e106508. https://doi.org/10.1371/journal.pone.0106508

Cornejo, E., & Denman, C. A. (2005). Into Our Own Hands: The Women-s Health Movement in the United States, 1969-1990 (Morgen). *Transforming Anthropology*, *13*(1), 70-71. https://doi.org/10.1525/tran.2005.13.1.70

Cox, J., Oh, E. Y., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, G., & Holmes, K. (2015). Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects. *Computing in Science & Engineering*, *17*(4), 28-41. https://doi.org/10.1109/MCSE.2015.65

Cunningham, A., & Williams, P. (2002). *The Laboratory Revolution in Medicine*. Cambridge University Press.

Curtis, V. (2015). *Online citizen science projects: An exploration of motivation, contribution and participation*. https://doi.org/10.21954/OU.RO.0000A4FF

Daume, S., & Galaz, V. (2016). “Anyone Know What Species This Is?” – Twitter Conversations as Embryonic Citizen Science Communities. *PLOS ONE*, *11*(3), e0151387. https://doi.org/10.1371/journal.pone.0151387

Davids, J. C., Rutten, M. M., Pandey, A., Devkota, N., Van Oyen, W. D., Prajapati, R., & Van De Giesen, N. (2019). Citizen science flow – an assessment of simple streamflow measurement methods. *Hydrology and Earth System Sciences*, *23*(2), 1045-1065. https://doi.org/10.5194/hess-23-1045-2019

De França, F. O., Di Genova, D. V. B., Penteado, C. L. C., & Kamienski, C. A. (2023). Understanding conflict origin and dynamics on Twitter: A real-time detection system. *Expert Systems with Applications*, *212*, 118748. https://doi.org/10.1016/j.eswa.2022.118748

De-Groot, R., Golumbic, Y. N., Martínez Martínez, F., Hoppe, H. U., & Reynolds, S. (2022). Developing a framework for investigating citizen science through a combination of web analytics and social science methods—The CS Track perspective. *Frontiers in Research Metrics and Analytics*, *7*, 988544. https://doi.org/10.3389/frma.2022.988544

Department of Economic and Social Affairs. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*. https://sdgs.un.org/2030agenda

Dhamdhere, A., & Dovrolis, C. (2008). Ten years in the evolution of the internet ecosystem. *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, 183-196. https://doi.org/10.1145/1452520.1452543

Dobreva, M. (2016). Collective Knowledge and Creativity: The Future of Citizen Science in the Humanities. En S. Kunifuji, G. A. Papadopoulos, A. M. J. Skulimowski, & J. Kacprzyk (Eds.), *Knowledge, Information and Creativity Support Systems* (pp. 565-573). Springer International Publishing. https://doi.org/10.1007/978-3-319-27478-2_44

Donelan, H. M., Kear, K. L., & Ramage, M. (2010). *Online Communication and Collaboration: A Reader*. Routledge.

Ebrahim, R. S. (2020). The Role of Trust in Understanding the Impact of Social Media Marketing on Brand Equity and Brand Loyalty. *Journal of Relationship Marketing*, *19*(4), 287-308. https://doi.org/10.1080/15332667.2019.1705742

Edelman, B. (2012). Using Internet Data for Economic Research. *Journal of Economic Perspectives*, *26*(2), 189-206. https://doi.org/10.1257/jep.26.2.189

Eitzel, M. V., Cappadonna, J. L., Santos-Lang, C., Duerr, R. E., Virapongse, A., West, S. E., Kyba, C. C. M., Bowser, A., Cooper, C. B., Sforzi, A., Metcalfe, A. N., Harris, E. S., Thiel, M., Haklay, M., Ponciano, L., Roche, J., Ceccaroni, L., Shilling, F. M., Dörler, D., … Jiang, Q. (2017). Citizen Science Terminology Matters: Exploring Key Terms. *Citizen Science: Theory and Practice*, *2*(1), 1. https://doi.org/10.5334/cstp.96

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, *41*, 424-432. https://doi.org/10.1016/j.chb.2014.05.005

Elliott, K. C., & Rosenberg, J. (2019). Philosophical Foundations for Citizen Science. *Citizen Science: Theory and Practice*, *4*(1), 9. https://doi.org/10.5334/cstp.155

European Commission, Warin, C., Delaney, N., & Tornasi, Z. (2020). *Citizen science and citizen engagement: Achievements in Horizon 2020 and recommendations on the way forward*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/05286

Fagiolo, G. (2007). *Directed or Undirected? A New Index to Check for Directionality of Relations in Socio-Economic Networks* (arXiv:physics/0612017). arXiv. https://doi.org/10.48550/arXiv.physics/0612017

Faustini, P. H. A., & Covões, T. F. (2020). Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, *158*, 113503. https://doi.org/10.1016/j.eswa.2020.113503

Ferragina, P., Piccinno, F., & Santoro, R. (2015). On Analyzing Hashtags in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), Article 1. https://doi.org/10.1609/icwsm.v9i1.14584

Fleming, J. R., & Johnson, A. (Eds.). (2014). *Toxic airs: Body, place, planet in historical perspective*. University of Pittsburgh Press.

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P.-Y., Danielsen, F., Hitchcock, C. B., Hulbert, J. M., Piera, J., Spiers, H., Thiel, M., & Haklay, M. (2022). Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, *2*(1), Article 1. https://doi.org/10.1038/s43586-022-00144-4

Fritz, S., See, L., Carlson, T., Haklay, M. (Muki), Oliver, J. L., Fraisl, D., Mondardini, R., Brocklehurst, M., Shanley, L. A., Schade, S., Wehn, U., Abrate, T., Anstee, J., Arnold, S., Billot, M., Campbell, J., Espey, J., Gold, M., Hager, G., … West, S. (2019). Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability*, *2*(10), Article 10. https://doi.org/10.1038/s41893-019-0390-3

Garbe, A., Ogurlu, U., Logan, N., & Cook, P. (2020). Parents' Experiences with Remote Education during COVID-19 School Closures. *American Journal of Qualitative Research*, *4*(3). https://doi.org/10.29333/ajqr/8471

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*(1), 877-902. https://doi.org/10.1146/annurev-psych-010814-015321

Griffin, M., Martino, R. J., LoSchiavo, C., Comer-Carruthers, C., Krause, K. D., Stults, C. B., & Halkitis, P. N. (2022). Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity*, *56*(4), 2841-2852. https://doi.org/10.1007/s11135-021-01252-1

Groulx, M., Brisbois, M. C., Lemieux, C. J., Winegardner, A., & Fishback, L. (2017). A Role for Nature-Based Citizen Science in Promoting Individual and Collective Climate Change Action? A Systematic Review of Learning Outcomes. *Science Communication*, *39*(1), 45-76. https://doi.org/10.1177/1075547016688324

Grover, P., Kar, A. K., Gupta, S., & Modgil, S. (2021). Influence of political leaders on sustainable development goals – insights from twitter. *Journal of Enterprise Information Management*, *34*(6), 1893-1916. https://doi.org/10.1108/JEIM-07-2020-0304

Hagras, M., Hassan, G., & Farag, N. (2017). Towards Natural Disasters Detection from Twitter Using Topic Modelling. *2017 European Conference on Electrical Engineering and Computer Science (EECS)*, 272-279. https://doi.org/10.1109/EECS.2017.57

Haklay. (2013). Neogeography and the Delusion of Democratisation. *Environment and Planning A: Economy and Space*, *45*(1), 55-69. https://doi.org/10.1068/a45184

Haklay, M., Motion, A., Balázs, B., Kieslinger, B., Greshake Tzovaras, B., Nold, C., Dörler, D., Fraisl, D., Riemenschneider, D., Heigl, F., Brounéus, F., Hager, G., Heuer, K., Wagenknecht, K., Vohland, K., Shanley, L., Deveaux, L., Ceccaroni, L., Weißpflug, M., … Wehn, U. (2020). *ECSA's Characteristics of Citizen Science*. https://doi.org/10.5281/ZENODO.3758668

Hansen, D. L., Schneiderman, B., & Smith, M. A. (2011). *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann.

Hargittai, E., & Karaoglu, G. (2018). Biases of Online Political Polls: Who Participates? *Socius*, *4*, 2378023118791080. https://doi.org/10.1177/2378023118791080

Harley, M. D., & Kinsela, M. A. (2022). CoastSnap: A global citizen science program to monitor changing coastlines. *Continental Shelf Research*, *245*, 104796. https://doi.org/10.1016/j.csr.2022.104796

Harris, S. J., Massimino, D., Newson, S. E., Eaton, M. A., Balmer, D. E., Noble, D. G., Musgrove, A. J., Gillings, S., Procter, D., & Pearce-Higgins, J. W. (2015). The Breeding Bird Survey 2014. *BTO Research Report*, *673*.

Havens, K., & Henderson, S. (2013). Citizen Science Takes Root. *American Scientist*, *101*(5), 378. https://doi.org/10.1511/2013.104.378

Hawe, P. (2004). A glossary of terms for navigating the field of social network analysis. *Journal of Epidemiology & Community Health*, *58*(12), 971-975. https://doi.org/10.1136/jech.2003.014530

Hawn, C. (2009). Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. *Health Affairs*, *28*(2), 361-368. https://doi.org/10.1377/hlthaff.28.2.361

Head, J. S., Crockatt, M. E., Didarali, Z., Woodward, M.-J., & Emmett, B. A. (2020). The Role of Citizen Science in Meeting SDG Targets around Soil Health. *Sustainability*, *12*(24), Article 24. https://doi.org/10.3390/su122410254

Hecker, S., Bonney, R., Haklay, M., Hölker, F., Hofer, H., Goebel, C., Gold, M., Makuch, Z., Ponti, M., Richter, A., Robinson, L., Iglesias, J. R., Owen, R., Peltola, T., Sforzi, A., Shirk, J., Vogel, J., Vohland, K., Witt, T., & Bonn, A. (2018). Innovation in Citizen Science – Perspectives on Science-Policy Advances. *Citizen Science: Theory and Practice*, *3*(1), 4. https://doi.org/10.5334/cstp.114

Hedges, M., & Dunn, S. E. (2018). *Academic crowdsourcing in the humanities: Crowds, communities and co-production*. Chandos Publishing, an imprint of Elsevier.

Horst, H. A., & Miller, D. (2020). *Digital Anthropology*. Routledge.

Hurlbert, A. H., & Liang, Z. (2012). Spatiotemporal Variation in Avian Migration Phenology: Citizen Science Reveals Effects of Climate Change. *PLoS ONE*, *7*(2), e31662. https://doi.org/10.1371/journal.pone.0031662

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), e2025334119. https://doi.org/10.1073/pnas.2025334119

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

Ilhan, A. O., & Aydınoğlu, A. U. (2023). Data Wars During COVID-19 Pandemic in Turkey: Regulatory Science, Trust, Risk, and Citizen Science. En V. Göçoğlu & N.

Karkin (Eds.), *Citizen-Centered Public Policy Making in Turkey* (pp. 289-309). Springer International Publishing. https://doi.org/10.1007/978-3-031-35364-2_16

Irwin, A. (1995). *Citizen science: A study of people, expertise, and sustainable development* (1. publ). Routledge.

Jaffee, D. (2003). Virtual Transformation: Web-Based Technology and Pedagogical Change. *Teaching Sociology*, *31*(2), 227. https://doi.org/10.2307/3211312

Jang, Y., Park, C.-H., & Seo, Y.-S. (2019). Fake News Analysis Modeling Using Quote Retweet. *Electronics*, *8*(12), 1377. https://doi.org/10.3390/electronics8121377

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

Johnston, S. F., Franks, B., & Whitelaw, S. (2017). Crowd-sourced science: Societal engagement, scientific authority and ethical practice. *Journal of Information Ethics*, *26*(1), Article 1.

Jones, G. (s. f.). *Citizen Science and Environmental Monitoring-Final Report*.

Jordan, R. C., Ballard, H. L., & Phillips, T. B. (2012). Key issues and new approaches for evaluating citizen-science learning outcomes. *Frontiers in Ecology and the Environment*, *10*(6), 307-309. https://doi.org/10.1890/110280

Kahlon, M., Yuan, L., Daigre, J., Meeks, E., Nelson, K., Piontkowski, C., Reuter, K., Sak, R., Turner, B., Weber, G. M., & Chatterjee, A. (2014). The Use and Significance of a Research Networking System. *Journal of Medical Internet Research*, *16*(2), e46. https://doi.org/10.2196/jmir.3137

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access*, *8*, 67698-67717. https://doi.org/10.1109/ACCESS.2020.2983656

Kelly, B. (2010). A deployment strategy for maximising the impact of institutional use of Web 2.0. En *Web 2.0 and Libraries* (pp. 95-122). Elsevier. https://doi.org/10.1016/B978-1-84334-346-2.50005-3

Kerson, Raymond. (1989). Lab for the environment. *MIT Technology Review*, *92*(1).

Kloetzer, L., Lorke, J., Roche, J., Golumbic, Y., Winter, S., & Jõgeva, A. (2021). Learning in Citizen Science. En K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 283-308). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_15

Kobori, H., Dickinson, J. L., Washitani, I., Sakurai, R., Amano, T., Komatsu, N., Kitamura, W., Takagawa, S., Koyama, K., Ogawara, T., & Miller-Rushing, A. J. (2016). Citizen science: A new approach to advance ecology, education, and conservation. *Ecological Research*, *31*(1), 1-19. https://doi.org/10.1007/s11284-015-1314-y

Kreutz, T., & Daelemans, W. (2022). Detecting Vaccine Skepticism on Twitter Using Heterogeneous Information Networks. En P. Rosso, V. Basile, R. Martínez, E. Métais, & F. Meziane (Eds.), *Natural Language Processing and Information Systems* (Vol. 13286, pp. 370-381). Springer International Publishing. https://doi.org/10.1007/978-3-031-08473-7_34

Krukowski, S., Amarasinghe, I., Gutiérrez-Páez, N. F., & Hoppe, H. U. (2022). Does Volunteer Engagement Pay Off? An Analysis of User Participation in Online Citizen Science Projects. En L.-H. Wong, Y. Hayashi, C. A. Collazos, C. Alvarez, G. Zurita, & N. Baloian (Eds.), *Collaboration Technologies and Social Computing* (Vol. 13632, pp. 67-82). Springer International Publishing. https://doi.org/10.1007/978-3-031-20218-6_5

Kullenberg, C., & Kasperowski, D. (2016). What Is Citizen Science? – A Scientometric Meta-Analysis. *PLOS ONE*, *11*(1), e0147152. https://doi.org/10.1371/journal.pone.0147152

Kunicina, N., Zabasta, A., Bruzgiene, R., Dubauskiene, N., Patlins, A., & Ribickis, L. (2019). Student Engagement in Cross-Domain Innovation Development and Its Impact on Learning Outcomes and Career Development in Electrical Engineering. *2019 IEEE Global Engineering Education Conference (EDUCON)*, 661-668. https://doi.org/10.1109/EDUCON.2019.8725269

Kunicina, N., Zabasta, A., Nikiforova, O., Romanovs, A., & Patlins, A. (2018). Modern tools of career development and motivation of students in Electrical Engineering Education. *2018 IEEE 59th International Scientific Conference on Power and Electrical*

Engineering of Riga Technical University (RTUCON), 1-6. https://doi.org/10.1109/RTUCON.2018.8659905

Kyei-Blankson, L., Blankson, J., Ntuli, E., & Agyeman, C. (Eds.). (2016). *Handbook of Research on Strategic Management of Interaction, Presence, and Participation in Online Courses:* IGI Global. https://doi.org/10.4018/978-1-4666-9582-5

Lai, M., Bosco, C., Patti, V., & Virone, D. (2015). Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1-9. https://doi.org/10.1109/DSAA.2015.7344884

Land-Zandstra, A., Agnello, G., & Gültekin, Y. S. (2021). Participants in Citizen Science. En K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 243-259). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_13

Lawson, B., Petrovan, S. O., & Cunningham, A. A. (2015). Citizen Science and Wildlife Disease Surveillance. *EcoHealth*, *12*(4), 693-702. https://doi.org/10.1007/s10393-015-1054-z

Lee, K. A., Lee, J. R., & Bell, P. (2020). A review of Citizen Science within the Earth Sciences: Potential benefits and obstacles. *Proceedings of the Geologists' Association*, *131*(6), 605-617. https://doi.org/10.1016/j.pgeola.2020.07.010

Lefever, S., Dal, M., & Matthíasdóttir, Á. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574-582. https://doi.org/10.1111/j.1467-8535.2006.00638.x

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2009). A brief history of the internet. *ACM SIGCOMM Computer Communication Review*, *39*(5), 22-31. https://doi.org/10.1145/1629607.1629613

L'Hermite-Leclercq, P. (1987). Histoire de la vie privée, sous la direction de Ph. Ariès et G. Duby. Tome II: De l'Europe féodale à la Renaissance, volume dirigé par G. Duby. *Bulletin Monumental*, *145*(2), 222-223.

Li, E., Parker, S. S., Pauly, G. B., Randall, J. M., Brown, B. V., & Cohen, B. S. (2019). An Urban Biodiversity Assessment Framework That Combines an Urban Habitat Classification Scheme and Citizen Science Data. *Frontiers in Ecology and Evolution*, *7*, 277. https://doi.org/10.3389/fevo.2019.00277

Li, X., Xie, Q., & Huang, L. (2022). Identifying the Development Trends of Emerging Technologies Using Patent Analysis and Web News Data Mining: The Case of Perovskite Solar Cell Technology. *IEEE Transactions on Engineering Management*, *69*(6), 2603-2618. https://doi.org/10.1109/TEM.2019.2949124

Liang, H., & Fu, K. (2015). Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science. *PLOS ONE*, *10*(8), e0134270. https://doi.org/10.1371/journal.pone.0134270

Liberatore, A., Bowkett, E., MacLeod, C. J., Spurr, E., & Longnecker, N. (2018). Social Media as a Platform for a Citizen Science Community of Practice. *Citizen Science: Theory and Practice*, *3*(1), 3. https://doi.org/10.5334/cstp.108

Lowe, H. J., Lomax, E. C., & Polonkey, S. E. (1996). The World Wide Web: A Review of an Emerging Internet-based Technology for the Distribution of Biomedical Information. *Journal of the American Medical Informatics Association*, *3*(1), 1-14. https://doi.org/10.1136/jamia.1996.96342645

Majkowska, A., Migdał-Najman, K., Najman, K., & Raca, K. (2021). Identification of the Words Most Frequently Used by Different Generations of Twitter Users. En K. Jajuga, K. Najman, & M. Walesiak (Eds.), *Data Analysis and Classification* (pp. 27-47). Springer International Publishing. https://doi.org/10.1007/978-3-030-75190-6_3

Malthus, T. J., Ohmsen, R., & Woerd, H. J. V. D. (2020). An Evaluation of Citizen Science Smartphone Apps for Inland Water Quality Assessment. *Remote Sensing*, *12*(10), 1578. https://doi.org/10.3390/rs12101578

Mamo, N., Azzopardi, J., & Layfield, C. (2023). The myth of reproducibility: A review of event tracking evaluations on Twitter. *Frontiers in Big Data*, *6*, 1067335. https://doi.org/10.3389/fdata.2023.1067335

Manning, C. D., Raghavan, P., & Schütze, H. (2008, julio 7). *Introduction to Information Retrieval*. Higher Education from Cambridge University Press; Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Manske, S. (2021, marzo 15). *Are citizen science projects multi-disciplinary research activities?* (World) [Text]. Https://Cstrack.Eu/; Citizen Science | CS Track Project. https://cstrack.eu/format/graphical-article/cs-projects-multi-disciplinary-research-activities/

Marin, A. (2011). Social network analysis: An introduction. *The Sage Handbook of Social Network …*. https://www.academia.edu/3587118/Social_network_analysis_An_introduction

Marlow, T., Miller, S., & Roberts, J. T. (2020). *Twitter Discourses on Climate Change: Exploring Topics and the Presence of Bots* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/h6ktm

Marsden, P. V. (1990). Network Data and Measurement. *Annual Review of Sociology*, *16*(1), 435-463. https://doi.org/10.1146/annurev.so.16.080190.002251

Martínez-Domínguez, M., & Fierros-González, I. (2022). Determinants of internet use by school-age children: The challenges for Mexico during the COVID-19 pandemic. *Telecommunications Policy*, *46*(1), 102241. https://doi.org/10.1016/j.telpol.2021.102241

Mazumdar, S., & Thakker, D. (2020). Citizen Science on Twitter: Using Data Analytics to Understand Conversations and Networks. *Future Internet*, *12*(12), 210. https://doi.org/10.3390/fi12120210

McCorriston, J., Jurgens, D., & Ruths, D. (2021). Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 650-653. https://doi.org/10.1609/icwsm.v9i1.14672

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, *10*(6), 285-290. https://doi.org/10.1890/110278

Mizunoya, S., Dreesen, T., Akseer, S., Brossard, M., Dewan, P., Giraldo, J.-P., Kamei, A., Ortiz, J., & Molom, O. (2020). *Promising practices for equitable remote learning Emerging lessons from COVID-19 education responses in 127 countries*.

Moczek, N., Voigt-Heucke, S. L., Mortega, K. G., Fabó Cartas, C., & Knobloch, J. (2021). A Self-Assessment of European Citizen Science Projects on Their Contribution to the UN Sustainable Development Goals (SDGs). *Sustainability*, *13*(4), Article 4. https://doi.org/10.3390/su13041774

Moernaut, R., Mast, J., Temmerman, M., & Broersma, M. (2022). Hot weather, hot topic. Polarization and sceptical framing in the climate debate on Twitter. *Information, Communication & Society*, *25*(8), 1047-1066. https://doi.org/10.1080/1369118X.2020.1834600

Montag, C., Yang, H., & Elhai, J. D. (2021). On the Psychology of TikTok Use: A First Glimpse From Empirical Findings. *Frontiers in Public Health*, *9*, 641673. https://doi.org/10.3389/fpubh.2021.641673

Nawaz, F. A., Barr, A. A., Desai, M. Y., Tsagkaris, C., Singh, R., Klager, E., Eibensteiner, F., Parvanov, E. D., Hribersek, M., Kletecka-Pulker, M., Willschke, H., & Atanasov, A. G. (2022). Promoting Research, Awareness, and Discussion on AI in Medicine Using #MedTwitterAI: A Longitudinal Twitter Hashtag Analysis. *Frontiers in Public Health*, *10*, 856571. https://doi.org/10.3389/fpubh.2022.856571

Negrón, J. B. (2019). #EULAR2018: The Annual European Congress of Rheumatology—a Twitter hashtag analysis. *Rheumatology International*, *39*(5), 893-899. https://doi.org/10.1007/s00296-019-04249-0

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012a). The future of citizen science: Emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298-304. https://doi.org/10.1890/110294

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012b). The future of citizen science: Emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298-304. https://doi.org/10.1890/110294

Nov, O., Arazy, O., & Anderson, D. (2014). Scientists@Home: What Drives the Quantity and Quality of Online Citizen Science Participation? *PLoS ONE*, *9*(4), e90375. https://doi.org/10.1371/journal.pone.0090375

Oentaryo, R. J., Low, J.-W., & Lim, E.-P. (2015). Chalk and Cheese in Twitter: Discriminating Personal and Organization Accounts. En A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *Advances in Information Retrieval* (Vol. 9022, pp. 465-476). Springer International Publishing. https://doi.org/10.1007/978-3-319-16354-3_51

Opitz, D. L., Bergwik, S., & Tiggelen, B. V. (2016). *Domesticity in the Making of Modern Science*. Springer.

Parry, M. (2016). Jennifer Nelson. *More Than Medicine: A History of the Feminist Women's Health Movement . The American Historical Review*, *121*(2), 606-607. https://doi.org/10.1093/ahr/121.2.606

Paul, M., & Dredze, M. (2021). You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 265-272. https://doi.org/10.1609/icwsm.v5i1.14137

Pershad, Y., Hangge, P., Albadawi, H., & Oklu, R. (2018). Social Medicine: Twitter in Healthcare. *Journal of Clinical Medicine*, *7*(6), 121. https://doi.org/10.3390/jcm7060121

Peters, A., Rohr, L., & Squires, L. (2022). Examining social media in the online classroom: Postsecondary students' Twitter use and motivations. *International Journal of Social Media and Interactive Learning Environments*, *6*(4), 328. https://doi.org/10.1504/IJSMILE.2022.124787

Prabhakar Kaila, D. R., & Prasad, D. A. V. K. (2020). *Informational Flow on Twitter – Corona Virus Outbreak – Topic Modelling Approach* (SSRN Scholarly Paper 3565169). https://papers.ssrn.com/abstract=3565169

Preece, J. (2016). Citizen Science: New Research Challenges for Human–Computer Interaction. *International Journal of Human-Computer Interaction*, *32*(8), 585-612. https://doi.org/10.1080/10447318.2016.1194153

Quinlivan, L., Chapman, D. V., & Sullivan, T. (2020). Validating citizen science monitoring of ambient water quality for the United Nations sustainable development goals. *Science of The Total Environment*, *699*, 134255. https://doi.org/10.1016/j.scitotenv.2019.134255

Rahmani, H., Shetty, D., Wagih, M., Ghasempour, Y., Palazzi, V., Carvalho, N. B., Correia, R., Costanzo, A., Vital, D., Alimenti, F., Kettle, J., Masotti, D., Mezzanotte, P., Roselli, L., & Grosinger, J. (2023). Next-Generation IoT Devices: Sustainable Eco-Friendly Manufacturing, Energy Harvesting, and Wireless Connectivity. *IEEE Journal of Microwaves*, *3*(1), 237-255. https://doi.org/10.1109/JMW.2022.3228683

Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries*.

Reed, J., Raddick, M. J., Lardner, A., & Carney, K. (2013). An Exploratory Factor Analysis of Motivations for Participating in Zooniverse, a Collection of Virtual Citizen Science Projects. *2013 46th Hawaii International Conference on System Sciences*, 610-619. https://doi.org/10.1109/HICSS.2013.85

Roberts, A. P. (2020). Swab and Send: A citizen science, antibiotic discovery project. *Future Science OA*, *6*(6), FSO477. https://doi.org/10.2144/fsoa-2020-0053

Robson, C., Hearst, M., Kau, C., & Pierce, J. (2013). Comparing the use of social networking and traditional media channels for promoting citizen science. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1463-1468. https://doi.org/10.1145/2441776.2441941

Rosenwald, M. S. (2017). *Before Twitter and Facebook, there was Morse code: Remembering social media's true inventor—The Washington Post*. https://www.washingtonpost.com/news/retropolis/wp/2017/05/24/before-there-was-twitter-there-was-morse-code-remembering-social-medias-true-inventor/

Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, *106*, 92-104. https://doi.org/10.1016/j.future.2020.01.005

Sahlström, F., Tanner, M., & Olin-Scheller, C. (2019). Smartphones in classrooms: Reading, writing and talking in rapidly changing educational spaces. *Learning, Culture and Social Interaction*, *22*, 100319. https://doi.org/10.1016/j.lcsi.2019.100319

Sanabria-Z, J., Alfaro-Ponce, B., González Peña, O. I., Terashima-Marín, H., & Ortiz-Bayliss, J. C. (2022). Engagement and Social Impact in Tech-Based Citizen Science Initiatives for Achieving the SDGs: A Systematic Literature Review with a Perspective on Complex Thinking. *Sustainability*, *14*(17), Article 17. https://doi.org/10.3390/su141710978

Santana, E., Bernardo, J., Donici, I., Valente, R., Pedro, B., Almeida, I., Silva, S., Alegre, C., Loureiro, T., & Silva, R. (2023). An analysis of science communication about COVID-19 vaccination in Portuguese online news media. *Journal of Science Communication*, *22*(05). https://doi.org/10.22323/2.22050202

Sapacz, M., Rockman, G., & Clark, J. (2016). Are we addicted to our cell phones? *Computers in Human Behavior*, *57*, 153-159. https://doi.org/10.1016/j.chb.2015.12.004

Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, *38*(3), 182-188. https://doi.org/10.1016/j.ajic.2009.11.004

Schaefer, T., Kieslinger, B., Brandt, M., & van den Bogaert, V. (2021). Evaluation in Citizen Science: The Art of Tracing a Moving Target. En K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 495-514). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_25

Secord, A. (1994). Science in the Pub: Artisan Botanists in Early Nineteenth-Century Lancashire. *History of Science*, *32*(3), 269-315. https://doi.org/10.1177/007327539403200302

See, L. (2019). A Review of Citizen Science and Crowdsourcing in Applications of Pluvial Flooding. *Frontiers in Earth Science*, *7*, 44. https://doi.org/10.3389/feart.2019.00044

Settembre, M. (2012). Towards a hyper-connected world. *2012 15th International Telecommunications Network Strategy and Planning Symposium (NETWORKS)*, 1-5. https://doi.org/10.1109/NETWKS.2012.6381667

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, *22*, 100104. https://doi.org/10.1016/j.osnem.2020.100104

Shapin, S. (2017). *Leviathan and the air-pump: Hobbes, boyle, and the experimental life*. Princeton University Press.

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M. E., & Bonney, R. (2012). Public Participation in Scientific Research: A Framework for Deliberate Design. *Ecology and Society*, *17*(2), art29. https://doi.org/10.5751/ES-04705-170229

Shulla, K., Leal Filho, W., Sommer, J. H., Lange Salvia, A., & Borgemeister, C. (2020). Channels of collaboration for citizen science and the sustainable development goals.

Journal of Cleaner Production, *264*, 121735. https://doi.org/10.1016/j.jclepro.2020.121735

Sigrist, R., & Widmer, E. D. (2011). Training links and transmission of knowledge in 18th Century botany: A social network analysis. *Redes : revista hispana para el análisis de redes sociales*, *21*, 0347-0387.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*(9), 467-471. https://doi.org/10.1016/j.tree.2009.03.017

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. *Proceedings of the 23rd International Conference on World Wide Web*, 1049-1054. https://doi.org/10.1145/2567948.2579215

Small, T. A. (2011). WHAT THE HASHTAG?: A content analysis of Canadian politics on Twitter. *Information, Communication & Society*, *14*(6), 872-895. https://doi.org/10.1080/1369118X.2011.554572

So, J., Prestin, A., Lee, L., Wang, Y., Yen, J., & Chou, W.-Y. S. (2016). What Do People Like to "Share" About Obesity? A Content Analysis of Frequent Retweets About Obesity on Twitter. *Health Communication*, *31*(2), 193-206. https://doi.org/10.1080/10410236.2014.940675

Soboleva, A., Burton, S., Mallik, G., & Khan, A. (2017). 'Retweet for a Chance to…': An analysis of what triggers consumers to engage in seeded eWOM on Twitter. *Journal of Marketing Management*, *33*(13-14), 1120-1148. https://doi.org/10.1080/0267257X.2017.1369142

Sprinks, J., Woods, S. M., Parkinson, S., Wehn, U., Joyce, H., Ceccaroni, L., & Gharesifard, M. (2021). Coordinator Perceptions When Assessing the Impact of Citizen Science towards Sustainable Development Goals. *Sustainability*, *13*(4), 2377. https://doi.org/10.3390/su13042377

Srivastava, P., & Hopwood, N. (2009). A Practical Iterative Framework for Qualitative Data Analysis. *International Journal of Qualitative Methods*, *8*(1), 76-84. https://doi.org/10.1177/160940690900800107

Statista, S. (s. f.). *Number of worldwide social network users 2027 | Statista*. Recuperado 13 de noviembre de 2023, de https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Stewart, C., Labrèche, G., & González, D. L. (2020). A Pilot Study on Remote Sensing and Citizen Science for Archaeological Prospection. *Remote Sensing*, *12*(17), 2795. https://doi.org/10.3390/rs12172795

Stieglitz, S., & Dang-Xuan, L. (2012). Political Communication and Influence through Microblogging—An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior. *2012 45th Hawaii International Conference on System Sciences*, 3500-3509. https://doi.org/10.1109/HICSS.2012.476

Strasser, B. J. (2012). Collecting Nature: Practices, Styles, and Narratives. *Osiris*, *27*(1), 303-340. https://doi.org/10.1086/667832

Strasser, B. J., Baudry, J., Mahr, D., Sanchez, G., & Tancoigne, E. (2018). "Citizen Science"? Rethinking Science and Public Participation. *Science & Technology Studies*, 52-76. https://doi.org/10.23987/sts.60425

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? En M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese Computational Linguistics* (pp. 194-206). Springer International Publishing. https://doi.org/10.1007/978-3-030-32381-3_16

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165-174. https://doi.org/10.1109/DSAA.2017.61

Tang, L., & Liu, H. (2010). Community Detection and Mining in Social Media. En *Synthesis Lectures on Data Mining and Knowledge Discovery* (Vol. 2). https://doi.org/10.2200/S00298ED1V01Y201009DMK003

Tauginienė, L., Butkevičienė, E., Vohland, K., Heinisch, B., Daskolia, M., Suškevičs, M., Portela, M., Balázs, B., & Prūse, B. (2020). Citizen science in the social sciences and humanities: The power of interdisciplinarity. *Palgrave Communications*, *6*(1), Article 1. https://doi.org/10.1057/s41599-020-0471-y

Tsubokura, M., Onoue, Y., Torii, H. A., Suda, S., Mori, K., Nishikawa, Y., Ozaki, A., & Uno, K. (2018). 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*, *13*(9), e0203594. https://doi.org/10.1371/journal.pone.0203594

Uthirapathy, S. E., & Sandanam, D. (2023). Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model. *Procedia Computer Science*, *218*, 908-917. https://doi.org/10.1016/j.procs.2023.01.071

Van Dijk, J. A. G. M. (2009). The Digital Divide in Europe. *The Routledge Handbook of Internet Politics*.

Vohland, K., Land-zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (Eds.). (2021). *The Science of Citizen Science*. Springer Nature. https://doi.org/10.1007/978-3-030-58278-4

Wang, Y., Hao, H., & Platt, L. S. (2021). Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on Twitter. *Computers in Human Behavior*, *114*, 106568. https://doi.org/10.1016/j.chb.2020.106568

Wang, Y., & Yang, Y. (2020). Dialogic communication on social media: How organizations use Twitter to build dialogic relationships with their publics. *Computers in Human Behavior*, *104*, 106183. https://doi.org/10.1016/j.chb.2019.106183

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (1.ª ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511815478

Wasserman, S., & Iacobucci, D. (1991). Statistical modelling of one-mode and two-mode networks: Simultaneous analysis of graphs and bipartite graphs. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 13-43. https://doi.org/10.1111/j.2044-8317.1991.tb00949.x

Wehn, U., Gharesifard, M., Ceccaroni, L., Joyce, H., Ajates, R., Woods, S., Bilbao, A., Parkinson, S., Gold, M., & Wheatland, J. (2021). Impact assessment of citizen science: State of the art and guiding principles for a consolidated approach. *Sustainability Science*, *16*(5), 1683-1699. https://doi.org/10.1007/s11625-021-00959-2

White, P. (2016). The Man of Science. En B. Lightman (Ed.), *A Companion to the History of Science* (1.ª ed., pp. 153-163). Wiley. https://doi.org/10.1002/9781118620762.ch11

Wiggins, A., & Crowston, K. (2011). From conservation to crowdsourcing: 44th Hawaii International Conference on System Sciences, HICSS-44 2010. *Proceedings of the 44th Annual Hawaii International Conference on System Sciences, HICSS-44 2010*. https://doi.org/10.1109/HICSS.2011.207

Wittner, L. S. (2009). *Confronting the Bomb: A Short History of the World Nuclear Disarmament Movement*. Stanford University Press.

Wuebben, D., Romero-Luis, J., & Gertrudix, M. (2020). Citizen Science and Citizen Energy Communities: A Systematic Review and Potential Alliances for SDGs. *Sustainability*, *12*(23), Article 23. https://doi.org/10.3390/su122310096

Xiong, Y., Cho, M., & Boatwright, B. (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement. *Public Relations Review*, *45*(1), 10-23. https://doi.org/10.1016/j.pubrev.2018.10.014

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(11), e20550. https://doi.org/10.2196/20550

Yan, L., Ma, Q., & Yoshikawa, M. (2013). Classifying Twitter Users Based on User Profile and Followers Distribution. En H. Decker, L. Lhotská, S. Link, J. Basl, & A. M. Tjoa (Eds.), *Database and Expert Systems Applications* (Vol. 8055, pp. 396-403). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40285-2_34

Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, *36*(1), 13-25. https://doi.org/10.1108/IJM-12-2014-0261

Zhang, M. (2010). Social Network Analysis: History, Concepts, and Research. En B. Furht (Ed.), *Handbook of Social Network Technologies and Applications* (pp. 3-21). Springer US. https://doi.org/10.1007/978-1-4419-7142-5_1

Zhiheng Xu & Qing Yang. (2012). Analyzing User Retweet Behavior on Twitter. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 46-50. https://doi.org/10.1109/ASONAM.2012.18

Zook, M. A., & Graham, M. (2007). Mapping DigiPlace: Geocoded Internet Data and the Representation of Place. *Environment and Planning B: Planning and Design*, *34*(3), 466-482. https://doi.org/10.1068/b3311

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*, *5*, 2-8. https://doi.org/10.1016/j.bdr.2015.12.001