Gómez Sánchez-Seco et al. *EPJ Data Science* (2025) 14:31
https://doi.org/10.1140/epjds/s13688-025-00545-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Holistic approach to analysing debates on ecological sustainability over time on X

Javier Gómez Sánchez-Seco[1,2], Mary Luz Mouronte-López[1*] and Rosa M. Benito[2]

Handling Editor: Jisun An

*Correspondence:
maryluz.mouronte@ufv.es
[1]Higher Polytechnic School, Universidad Francisco de Vitoria, carretera Pozuelo a, Av de Majadahonda, Km 1.800, 28223 Madrid, Spain
Full list of author information is available at the end of the article

Abstract

Using network theory and data analysis, we study the messages on Twitter (X) about ecological sustainability over the period 2007-2022. With a global view of 70,311,541 messages we examined the sentiment, keywords and hashtags utilised, as well as the correlations between sentiment and both socioeconomic and environmental variables. In addition to the above, we carried out an in-depth analysis of the global interactions network (retweets, replies and quotes), with a special focus on the study of the community network (CNET) (with 4576 supernodes, and 9855 links). The sentiment shown in the text of the tweets was positive over the years in all analysed locations, although close to neutral. Keyword analysis detected terms present in tweets posted from various regions, showing global thinking in the world. The relationships between sentiment and variables examined were continent- and country-specific, identifying a stronger correlation with socioeconomic attributes. Regarding CNET, according to the study performed using adjacency and laplacian embeddings, as well as Chebyshev, Euclidean, Minkowski, and Manhattan distances, pairs of unconnected supernodes appeared to have more similarity in their connection patterns than pairs of connected supernodes, due to the topological structure of CNET which has a large number of peripheral nodes that are not connected to each other, but are connected to nodes with higher centrality. In agreement with the Jaccard coefficient, resource allocation index, Adamic Adar index, and preferential attachment score, there is little possibility of link formation between supernodes. Statistically the supernodes also exhibited high topological similarity. A few specific supernodes host most of the users, showing the highest centralities among those analysed. The basic structure of CNET, which maintained its key properties, was also examined. Strategies that promote communication between supernodes to achieve greater participation and diversity in discussions need to be further investigated.

Keywords: Sustainability; Twitter; Network Theory; Data analytics

1 Introduction

1.1 Overview of ecological sustainability. State of art

In recent years, ecological sustainability has emerged as a major global concern [1–4], although the mechanisms applied in order to achieve such sustainability in each geographical area vary considerably. Differences also exist in the ability of regions to adopt sustainable practices. Certain areas, such as United States, Western Europe or Saudi Arabia,

have made significant progress in the transition to renewable energy sources and efficient management of natural resources [5–7], but others, such as China, still have a long way to go to reconcile economic development with environmental preservation [8, 9]. The understanding of what sustainability means as well as deciding the necessary actions to be taken in achieving it is influenced by geographical, socioeconomic and cultural factors. Strategies adopted in one region may not be easily put into practice in another, as local conditions play a crucial role in the effectiveness of the efforts implemented [10, 11]. Exploring these regional differences not only helps to broaden the understanding of sustainability, but also highlights the need for particular approaches that are adapted to the characteristics of each context. The term sustainability, described by the ONU in 1987, as achieving a balance between the current and the future by ensuring that today's necessities are achieved without endangering those of the future [12].

Differences between regions considering socioeconomic variables have been studied in several works [13–15]. By analysing both similarities and differences between geographical regions considering variables that correspond to gender, education, economical and working world fields, we have shown that the highest dissimilarities in all regions corresponded to gender domain and the most significant communalities occurred in the rest of fields relying on the geographical area [16].

In addition, social networks have made it possible for human interactions to be immediate and global. Platforms such as X (formerly known as Twitter), constitute relevant data repositories to carry out analyses on global perceptions in real time [17–22]. The speed at which information spreads through these networks provides a mechanism for knowing the public opinion on social critical issues. Analysing conversations about ecological sustainability in different parts of the world enables us to not only reveal cultural differences, but also help in defining communication and awareness-raising strategies [23].

Network Theory can be used to analyse the conversations on ecological sustainability that happen in social networks [24–26]. It provides mechanisms to unravel the complex interactions between the elements of any social network, offers procedures to examine the network dynamic [27–29], introduces methods to know how information spreads in the network [30, 31], and provides mechanisms to analyse the interconnection between the users' communities [32–35]. The application of Network Theory to the study of messages posted on a social network from different locations makes it possible to detect emerging patterns and underlying structures that are not detectable through traditional methods of analysis [36, 37].

Due to the large amount of data available on Twitter, discussions on a wide variety of topics have been studied in depth from different points of view. Some of these themes are bot analysis [38–40], information dissemination [41], political opinion differences [42], identification of fake news [43], virilization of ideas [44], contagion studies [45, 46]. With regard to ecological sustainability, the following categories, among others, have been examined: environmental crises management [47], opinions and polarisation on climate change [48, 49], formation of network communities with regard to specific climate change topics [50], the conservation of animal species [51], in addition to the linguistic analysis of green consumerism [52], the monitoring of users in climate protests [53], and the use of blue zones in cities [54].

In previous research, we have studied opinions about the education system on Twitter [20]. A largely negative opinion was detected in many countries, although a positive

perception was also identified in some low-income locations. Additionally, we have proposed models of inferring opinions through the retweet networks both in bidimensional and multidimensional systems too. We suggested a general methodology to analyse and evaluate the existence of polarisation in social networks considering a bidimensional environment [55]. The validity of the method was tested by using it in debates on the political situation in Venezuela. We proposed some metrics for estimating the main polarisation forces in multidimensional contexts in addition to implement a technique to show the structure of the ideological space [56].

In relation to environmental problems we have analysed in detail the opinions on Twitter about climate change for the period from February 2021 to May 2021 using different techniques [22]. In particular we analysed the topics and sentiments in the messages containing the words `climate action`, `climate change`, `climate crisis`, `climate disaster`, `climate emergency`, `global warming`, and `greenhouse effect` [22]. Globally, by country and by gender, the most discussed topic was climate change activism. Concerning firms, sustainability was among some of the main topics addressed. Overall, the sentiment shown in the discussions was negative. However, some differences in the sentiment mostly expressed in the discussions associated with certain topics, were found by both country and gender [22].

In addition, a machine learning model that explained the formation of links between users was built considering the aforementioned interactions on Twitter about climate change from March 2021 to May 2021 [57]. Among all the models analysed Random Forest and Support Vector Machine models were the ones that exhibited the most optimal results, using as explanatory variables the features obtained through `node2vec` algorithm and several similarity metrics between nodes, respectively. Communities, degree, betweenness and other parameters of the interaction networks were also computed. We did not detect significant differences between interaction networks in the hashtags' probability distributions for the analysed communities [57].

With the aim of identifying which countries exhibited particular air quality characteristics, we have analysed the time series of the carbon and greenhouse gas emissions finding no relation with various socioeconomic indicators [21]. It was also found that some countries showed an unusual emissions pattern. Furthermore, the distribution of hashtags of the tweets published from these locations was distinct from the rest of regions. Various hashtags common to the examined countries and with the highest occurrence in the tweets posted from the location exhibiting an atypical emissions characteristic were related to the environment. Some hashtags that existed only in the tweets published from the mentioned countries with peculiar emission characteristics referred specifically to the necessity of cleaning up the environment [21]. Moreover, the differences and communalities in the behavior of humans and bots in Twitter conversations about ecological sustainability was considered [58]. We selected the tweets with the following terms: `green urban`, `pollution`, `renewable energy`, `sustainable agriculture`, `sustainable city`, `sustainable food`, `sustainable industry`, and `sustainable transport`, and found that while humans and bots showed mostly positive sentiment in the posted tweets, the bots exhibited a greater number of neutral messages. Using clustering techniques, we identify groups of users with different behaviours in both humans and bots in tweets and retweets. It was not possible to detect groups with specific behavioural patterns when considering only replies or quoted messages [58]. Accordingly,

a more detailed study on ecological sustainability should be carried out. Thus, this paper analyses Twitter conversations covering aspects such as agriculture, food, cities, industry and transportation, as well as energy and urban spaces, and focuses in examining the relation between users perceptions and socioeconomic and environmental variables.

1.2 Purposes and objectives of this research

The main goal of this paper is to analyse the time evolution of the perceptions of ecological sustainability by geographic area (country and continent) by analysing the large dataset of Twitter conversation used in our previous study [58]. In this way we focus on the most significant key areas in which research [59–61] and important international institutions [62–64] have shown that effective action should be taken to achieve a much more sustainable planet. Accordingly, our study will focus in the following key areas: agriculture, renewable energy, transport, pollution, and the remaining areas were collated according to the keywords that were employed to download the tweets. The relevance of these areas to the sustainability issue is indicated by the SDGs 9, 11 and 14 [65]. We also examine how the activity on Twitter was related to some of the most important events on ecological sustainability found in the dataset.

In addition to the above, we analyse the relationships of the annual median of the sentiment on ecological sustainability by geographical area with 45 socioeconomic and 48 environmental variables, which can provide important insights into the influence of these attributes on perceptions by geographic area.

We also build a network in which the nodes are the users and a link is formed when they interact with each other by retweeting, quoting, or replying. Due to the enormous size (8,840,532 nodes) of the network, in order to study its structure we perform a community network analysis using the Leiden method. Following on from this, the c-network is built (CNET) and is formed by supernodes that correspond to each community [66, 67]. There is a link between two supernodes if any of the users of these communities has made an interaction with each other. Regarding the structural analyses of the enormous interaction network, we implemented a profound and comprehensive study on the CNET. In particular, we perform the following analyses: (i) an examination of the connectivity between supernodes based on adjacency and laplacian embeddings, (ii) a study of the possibilities for link formation between supernodes (iii) an investigation into the structural properties of a part of CNET, which retains its key properties. The `NetBone` package, recently developed and implemented in Python [68] is used. (iv) an examination of CNET's basic topological properties through various recently proposed parameters to detect the importance of nodes [69, 70] (v) an analysis of the differences in the topological structures of supernodes. A young metric designed to evaluate the topological dissimilarity between graphs [71] is used.

Studying all of the above aspects together, we conducted a holistic study of the perceptions detected in Twitter discussions on ecological sustainability in 2007-2022. Specifically, we aim to answer the following research questions: (i) Does the observed global perception of ecological sustainability differ from that which is observed between geographical areas? What has been the sentiment at some of the most important events on ecological sustainability? (ii) Overall, and by geographical area, what are the most used hashtags and keywords? (iii) Is there a relationship between sentiment and socioeconomic and environmental variables? If it exists, can this relationship be established at the global

level, or can it only be determined at the regional level? (iv) What is the structure of the interaction network? How many user communities are there? What is the level of sentiment in those communities? How is CNET structured and which are its main characteristics and properties?

As we explained in detail in the sections labelled Results and Discussion, the most important findings of this research are: (i) Considering the 70,311,541 Twitter messages that were obtained, over the large time period of analysis and examined by geographical area, the sentiment detected was mostly positive, although it is very close to neutrality. The events under examination exhibited a similar pattern. (ii) A higher correlation of the median of the sentiment with socioeconomic factors than with environmental factors was identified in most of the geographical areas analysed, which only occurred in the regional context. (iii) A total of 4576 communities were identified that exhibited disparities in terms of median sentiment. However, again, these values were predominantly close to neutrality. Differences in the nature of each community's discourse were also observed, as evidenced by the use of specific hashtags. (iv) According to the metric used, the entire CNET and the CNET of every interaction type corresponding to retweets, quotes and replies exhibited analogous topological characteristics. Globally, the supernodes showed low structural dissimilarity according to metric and statistics examined. (v) The CNET network presented the conventional social network structure, containing a few supernodes with high centrality magnitudes but a large number of nodes with low centrality values. The ratio between the number of supernodes and links connecting them, as well as a relatively small radius corresponded to a network with high connectivity. Nevertheless, it was evident that a more extensive multilevel dialogue was required during the discussions. (vi) According to the evaluation metrics used, the skeleton of the entire CNET network appropriately reproduced its key properties. (vii) In CNET, a dissimilarity was identified between the centralities of the supernodes. The application of classical centrality estimations revealed that the most relevant communities were located in the core of the network. Despite this, the utilisation of recently suggested metrics, which consider both the flow of information and a broader perspective of the degree of supernodes, enabled the identification of additional key supernodes in the CNET structure.

2 Data and methods

2.1 Description of the datasets

The dataset of this study correspond to tweets related to environmental sustainability, published from January 2007 to December 2022, acquired using the Twitter API (and twarc2), which filters and downloads tweets based on specific keywords. The selected keywords to download tweets were: `green urban`, `pollution`, `renewable energy`, `sustainable agriculture`, `sustainable city`, `sustainable food`, `sustainable industry`, and `sustainable transport`. It has to be noticed that the dataset includes original tweets, retweets, replies and quotations. Each element of the dataset contains several attributes, such as the hashtags used, the text and timestamp of the message, as well as the identification of the interactions that happened subsequently and are associated with it. Also included is data on the sender's account, the virality of the message and, if available, the location from which the message was posted. This information is distributed in the following fields, which were provided by T-Hoarder tool [20–22, 57, 58, 72]: `author`, `id tweet`, `id user`, `lang`, `quoted id`, `relation`, `replied id`, `retweeted id`, `text`, `user replied`, `user retweeted`,

`user quoted`. This dataset was employed in the analysis of differences between humans and bots in discussion about ecological sustainability [58]. In order to detect the country from which each tweet was published, as in [20–22] we use geopy tool [73], using the Open Street Map (OSM) geolocation service nominatim [74, 75] and the location field of the user profile.

The dataset contains a total of 70,311,541 messages all of them written in english (22,908,407 original tweets, 1,264,184 quotes, 39,907,678 retweets and 6,231,272 replies) which corresponded to 13,095,195 unique users. After processing, it was possible to identify the location of a total of 2,844,265 tweets (see Supplementary Material Document, Table S12 and Table S13).

In addition to Twitter data, we utilise the repository WorldBank|Data|Indicators [76], which contains, for the period 1960-2022, 93 variables corresponding to 266 countries (45 attributes referring to socioeconomic and 48 to environmental domains). However, only the period 2007-2022 and 148 countries were used in the analysis (see Supplementary Material Document, Table S1) as no localised tweets were identified for the rest.

2.2 Sentiment analysis and correlations with other variables

Each message included in the dataset has a sentiment which was computed utilising `TextBlob` [77] [20, 22, 58]. This sentiment varies between −1 and 1 values, corresponding to the most positive and negative magnitudes possible, respectively. Based on the sentiment of each individual message, we proceeded to determine the sentiment by country and continent. From 2007 to 2022, a time series of sentiment was generated on a monthly basis, which allowed us to obtain a detailed description of sentiment time evolution. The proportion of positive, negative and neutral tweets was also calculated, providing a metric on the sentiment at each location. It has to be noticed that in this section we only consider original tweets. This was because in all other types of interactions the sentiment might not be reflective of the initial perception of a user who has been influenced by their interactions with other users. In contrast, in the study concerning interactions, which is detailed in a later section, only those sentiments included in posts involving quotes, replies and retweets were taken into consideration.

In addition to the above, with the purpose of exploring the relationships between the expressed sentiment and socioeconomic and environmental variables, a multiple linear regression model (lm in `R` [78]) was implemented [79]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon, \tag{1}$$

For each geographical area (country or continent), Y represents the predicted annual sentiment, $X_1 \cdots X_n$ symbolise the yearly explanatory variables (both socioeconomic and environmental). β_0 corresponds to the value of Y, if the explanatory variables are equal to zero, ε denotes the residual error and $\beta_1 \cdots \beta_n$ symbolise the regression coefficients.

Additionally, to determine the most relevant variables corresponding to each country, Principal Component Analysis (PCA) was carried out. This made it possible to reduce the dimensionality of the dataset by identifying the most significant attributes to prevent over-fitting. For each attribute investigated, when there was no value for any of the 248 countries considered in the study over a particular year, that year was removed from the analysis.

2.3 Keywords analysis

Keyword analysis was implemented utilising the Twitter Keyword Graph (*TKG*) approach [80], in which terms form networks and their relevance were computed through centrality metrics. This method had shown good results in our previous research on the analysis of human and bot messages [58]. A short description of the performed procedure can be found in the Supplementary Material Document, section S1.2.

2.4 Interaction networks

2.4.1 Network building. Community identification

The interaction network was constructed, with each user constituting a node and the interactions (retweets, replies and quotes) between users representing links, where two users are linked if they have interacted with each other. In this way, a single network of interactions across all years of our dataset is created. In order to obtain further details regarding the large size of this network (8,840,532 nodes and 28,403,555 links), we performed a community analysis using the Leiden algorithm [81] which provided optimal results in several previous research [57, 82, 83]. The network was divided into 4576 communities, which will be used as supernodes to condense the information and facilitate the analysis of the network. The sentiment corresponding to each supernode was calculated as the median of the sentiment of interactions between users included in the supernode.

As a means to obtain more information on the communities, we drew histograms of the following metrics: number of nodes and links, mean degree, maximum and minimum degree. In addition to mean, maximum and minimum pagerank [84], and degree assortativity [85, 86].

2.4.2 Analysing the community network structure

A CNET was also built [66, 67, 87, 88], where the communities symbolised supernodes and the links between them were weighted connections, a link between two supernodes is established if a link existed between any of the users in their respective communities. The weight of the link is the sum of the weights of all the interactions between the supernodes in question. In this way, it was possible to represent the original network in a more manageable graph, which retained the key information. To examine the topological characteristics of CNET, various statistical properties were calculated, such as, maximum and minimum degrees of nodes, average degree, diameter, degree assortativity [89, 90], average length of the shortest paths, degree [91] and closeness [92, 93] and betweenness centrality [94–97].

In addition to the above, with the purpose of obtaining an enhanced comprehension of the importance of nodes we computed two recently proposed metrics: an extended version of HCC (hybrid characteristic centrality) [70] and the K-risk metric [69]. The extended version of HCC proposed by Liu and Zheng [70] takes into account both the HCC of the node and of its neighbours, i.e. it integrates the extended degrre and the E-shell. A configurable parameter, δ, is also considered in the metric. It takes values in the interval [0,1] and has the purpose of describing the relationship between extended and the traditional degrees (for $\delta = 1$ both are equivalent) [70]. In this analysis, a value of δ equal to 0.1 was considered, with the aim of obtaining a value with a significant difference in the traditional degree. We also used the recently proposed K-risk metric, which was suggested by García-Algarra et al. [69] as an alternative to classical centrality estimation parameters.

The metric, based on the analysis of the information flow in the network when certain nodes are suppressed, is derived from the conventional K-kernel decomposition [98]. In particular, it is defined as the sum of the differences in distance between a node located in a k-kernel and all those with which it is connected that are in a lower k-kernel, weighted by the k-indexes. An additional parameter is also considered in the calculation as a mechanism for taking into account ties between nodes that are in distinct K-shells [69]. We considered a value of 0.01 for this parameter, that is the same used in the original work [69].

Furthermore, in order to gain a better understanding of CNET connectivity, the embedding techniques (both adjacency and laplacian) were used [99]. Dimensions equal to 5, 10, 15, 20 and 25 were taken and the most optimal of them was determined following the method named Normalised embedding loss function, which is described in [100]. In this research, we carried out the embedding process by the following steps:

(i) A vector $vdim = \{5, 10, 15, 20, 25\}$ was built. For each pair i, j, with $i = j - 1$ and j varying from 2 to 5, the following steps were repeated 10 times:

1 For all nodes, the embedding $vd[i]$ and $vd[j]$ were obtained, and x nodes were randomly selected from the total nodes.
2 For x nodes and $vd[i]$, $vd[j]$ dimensions, the Error $E_{vd[i],vd[j]}$ was calculated as in [100] (where this error is named normalised embedding loss function).

(ii) For the pair $vd[i]$, $vd[j]$, the obtained average error, $\overline{E_{vd[i],vd[j]}}$, in 10 runs was calculated.

(iii) Finally, the dimension $vd[j]$ with a $\overline{E_{vd[i],vd[j]}} < 0.01$ was considered.

Once the optimal dimension of the embedding was obtained, the feature vectors for both connected and unconnected pairs of nodes u, v, were calculated, X_u, X_v. After that, the following distances between X_u and X_v were computed: Chebyshev [101], Euclidean [102], Minkowski (with $p = 0.5$) [103] and Manhattan [103]. The aforementioned distances were then normalised according to the min-max criterion [104] and summed. Finally, the result obtained was again normalised following the same principle. Various histograms of this normalised global distance for both connected and unconnected pairs of nodes were plotted.

Furthermore, we compute the possibility of link formation between pairs of nodes to better understand the connectivity patterns, relationship dynamics and structural principles that govern the network, as well as the possibility of establishing new connections depending on the node to be treated. Jaccard coefficient [105], resource allocation index [106], Adamic Adar index [107], and preferential attachment score [108] were estimated. These metrics were normalised through the min-max criterion and histograms of each of them were generated.

We also obtained a skeleton of CNET utilising the `NetBone` package [68], with the intention of obtaining the basic functional structure of the network, eliminating unnecessary links while maintaining information flow with minimal information loss.

Identical metrics and connectivity analysis to those computed in CNET were implemented for CNET skeleton. The appropriateness of the method was checked through the use of various filters such as Noise Corrected, Global Threshold, Marginal Likelihood and Locally Adaptive Network Sparsification [68]. The following metrics were utilised: `Node Fraction`, `Edge Fraction`, `Weight Entropy` (a weighting method that measures

value dispersion in decision-making [109]), `Average Degree`, `LCC Size` (Largest Connected Component), and `Density` [68].

2.5 Analysis of supernodes

Considering the frequency distribution that corresponded to the number of users included in the supernodes, they were classified into the percentiles [0, 0.97), [0.97, 0.99), [0.99, 1]. The number of nodes contained in each of them was in the range [3, 109], [113, 3250], [3515, 1,487,676] respectively.

The following procedure was run 200 times for each percentile: (i) Two supernodes were successively randomly selected with replacements, (ii) The dissimilarity metric proposed by Schieber et al. [71] was computed utilising as graphs the networks and corresponds to the selected supernodes. As authors explain, in the definition of the metric, the network node dispersion notion and the distance distribution corresponding to each node included in the graphs are taken into consideration. Moreover, attention is paid to α-centrality for each network and its complement graph, as well as the Jensen–Shannon divergence [71]. Three configurable parameters are also taken into account in the metric description. According to the authors, the best results are provided for the values 0.45, 0.45 and 0.10 [71]. In this research we used these magnitudes for comparing the supernode structure. After running the 200 experiments, mean, standard deviation, median, minimum and maximum values were calculated.

A global analysis was also performed by running 1000 trials. In each trial, two supernodes from the total set were successively randomly selected with replacements. In the procedure, both the percentile and the supernode located on it were chosen randomly. The same values for the configurable factors were taken, and analogous statistical parameters were computed.

3 Results

3.1 Analysis of Twitter sentiment on sustainability and correlation with other variables

We have analysed the sentiment about sustainability in a total of 2,844,265 tweets posted by 828,844 users corresponding to the six continents North America (NA), South America (SA), Europe (EU), Asia (AS), Africa (AF) and Oceania (OC). In this study we have analysed the sentiment for different locations both at the level of continent and country, and the time evolution of the sentiment over a period of 15 years in order to analyse the general tendency of the sentiment on sustainability in the last years. In all continents and countries examined, the number of tweets with a positive sentiment was higher than the number of negative and neutral tweets. Table 1 depicts the total number of tweets and the total sum of tweets exhibiting positive, negative and neutral sentiments by continent. This suggests a general trend towards an optimistic perception in sustainability-related conversations on Twitter.

Examining the monthly time series, sentiment was positive over the years, but showed values close to neutrality. Most individuals did not seem to be aware of environmental problems. Specifically, in Europe as can be seen in Fig. 1, which shows the results for 10 representative European countries, positive sentiment was in the interval [0, 0.2] during the studied time period but this is not the case for other regions. For example in Africa and South America, the low number of tweets led to a radicalisation of sentiment.

Table 1 Total number of tweets and number of tweets with positive, negative and neutral sentiment on sustainability by continent. North America (NA), South America (SA), Europe (EU), Asia (AS), Africa (AF) and Oceania (OC)

Continent	Total	No. positives	No. neutrals	No. negatives	No. unique users
NA	1,302,388	844,003 (64%)	149,722 (11%)	308,663 (25%)	366,683
EU	793,965	493,036 (62%)	102,721 (12%)	198,208 (26%)	207,975
AS	425,303	259,547 (61%)	52,048 (12%)	113,708 (27%)	143,130
OC	153,119	95,411 (62%)	21,301 (14%)	36,407 (24%)	39,856
AF	133,998	84,120 (63%)	19,821 (15%)	30,057 (22%)	58,627
SA	35,492	22,740 (64%)	4338 (12%)	8414 (24%)	12,573

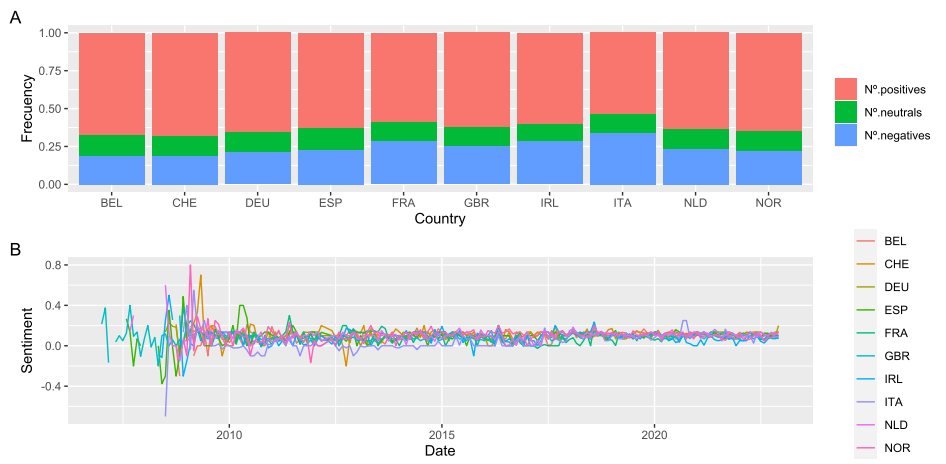

Figure 1 Sustainability sentiment analysis for the 10 countries with the highest impact in Europe: A) frequency of positive, negative and neutral tweets by country, B) Time series showing the temporal evolution of the sentiment on sustainability by country

Table 2 Multiple linear regression metrics for the continents with the most tweets. EU: Europe, NA: North of American and AS: Asia. See supplementary material document Table S2 for the performance metrics corresponding to the multiple linear regression models for each socioeconomic and environmental variable for these 3 continents

Metric	EU	NA	AS
Residual standard E	0.01178	0.002807	0.01958
Multiple R-squared	0.7395	0.9895	0.7764
Adjusted R-squared	0.6092	0.9685	0.6646
F-statistic	5.676	47.06	6.944
p-value	0.01824	0.001082	0.01026

In order to further analyse the possible effect of different socioeconomic and environmental variables on the users sentiment on sustainability we perform a multiple linear regression analysis for the six continents and most relevant countries (Table 2). Most countries and continents exhibited stronger correlations with socioeconomic variables than with environmental attributes (see their description in [76]). In those models corresponding to continents, only four variables, all being socioeconomic, were commonly present in at least two of them, which were: Net ODA received (% of GNI), Expenditure (% of GDP), Foreign direct investment, net inflows (BoP, current US$) and GDP growth per capita (annual %) (see Supplementary Material Document, Table S2 and Table S3). In the models corresponding to countries, the highest share of common characteristics in socioe-

conomic factors was 23 followed by 21, which happened in the Foreign direct investment, net inflows (BoP, current US$) and Cereal yield (kg per hectare) variables, respectively. By contrast, the highest overlap in environmental variables happened in the renewable internal freshwater resources per capita (cubic metres) attribute, which was present in the models corresponding to eleven countries (see Supplementary Material Document, Table S4).

Focusing on Europe, we observe that the explanatory variables corresponding to the model are: Gross capital formation (% of GDP), GNI per capita, Atlas method (current US$), growth of GDP per capita (% annual) and donations, excluding technical cooperation (balance of payments, current US$), which are all socioeconomic variables (see [76]). However, if we examine the most relevant countries in the region, we observe: Czech Republic with 4 socioeconomic variables, Ireland with 2 socioeconomic and 1 environmental, Italy with 4 socioeconomic and 1 environmental, France with 2 socioeconomic and 3 environmental, Great Britain with 5 socioeconomic and 1 environmental and Belgium with 1 socioeconomic and 2 environmental. None of the variables is common to the majority of the aforementioned countries or to the rest of the countries on the continent (see Supplementary Material Document Table S5).

All of the above suggests a strong relationship between the sentiment expressed on Twitter about ecological sustainability and specific socioeconomic factors in each region. The multiple linear regression models are country and continent-specific.

3.2 Keywords analysis

We have carried out an analysis of the most relevant keywords used in the tweets at the level of continents and regions. In order to do this, we have removed the initial keywords used for downloading the tweets (see Sect. 2.1). As can be seen in Table 3 the keywords that are present in the tweets posted from all the countries considered in our study are: `clean`, `make`, `power` and `project`. This result is in line with the one obtained in Sect. 3.1, which despite being concerned with actions to be taken to solve ecological sustainability issues, exhibited a positive sentiment.

The study also detected words that, while not common to all continents, were present in tweets published in several of them (see Table 3). These words also referred to the implementation of solutions to environmental problems, but focusing on more specific aspects. They were: `sun`, `reduction`, `source` or `future`. Certain words only seem important for one continent. These are those that refer to a country or city located within it

Table 3 Top 10 most tweeted keywords by continent. Only the root of the word is shown. Keywords that are common to all continents are highlighted in bold. AF: Africa, AS: Asia, EU: Europe, NA: North of America, OC: Oceania and SA: South of America

AF	AS	EU	NA	OC	SA
afric	**cle**	**cle**	**cle**	australi	chang
cle	day	futur	help	**cle**	chin
development	delhi	help	**mak**	coal	**cle**
mak	indi	london	**pow**	futur	futur
nigeri	indion	low	**project**	investment	**mak**
pow	**mak**	**mak**	sol	job	**pow**
project	**pow**	near	sourc	**mak**	**project**
reduc	**project**	news	wind	plan	sol
sol	reduc	**pow**	futur	**pow**	sourc
sourc	sol	**project**	reduc	**project**	study

(`Nigeria`, `India`, `London`, `Australia`, `China`) or to a specific problem (`day`, `low`, `wind`, `coal`). The differences and commonalities mentioned above were also observed on a smaller scale between countries on each continent (see Supplementary Material Document Table S6, Table S7, Table S8, Table S9, Table S10 and Table S11).

3.3 Interaction networks

In this section we present the main results corresponding to the constructed networks. We analyse the topology and characteristics of the global CNET and the specific CNETs for each type of interaction. For the analysis carried out in this section, loops and multiple links were removed in the raw network.

3.3.1 Network building. Community identification

We construct a network in which the nodes correspond to the 4576 communities, extracted using the Leiden method, and the links represent the interaction between users belonging to different supernodes. For this global CNET, all types of interaction between users were used, however, as users can decide to use any of the interactions mechanism of twitter, we have also studied the specific networks corresponding to each type of interaction: retweet, reply, or quote, in order to analyse if the results of our study could be influenced by the type of interaction considered [50, 110, 111]. In order to do that we carried out a dissimilarity analysis on the networks corresponding to each interaction mechanism. In CNET's analysis for each interaction typology, only that class of messages was considered. The methodology applied for the construction of the CNET retweets, replies and quotes was as follow. Firstly, each network was constructed based on an specific interaction. The users of these networks were then grouped into those supernodes to which they belonged across the entire CNET, thus maintaining consistency across various networks and allowing the interactions made by users to be compared. In this way, the CNETs of quote, reply and retweet were generated with 397 nodes and 934 links, 1089 nodes and 2196 links, and 4275 nodes and 8693 links, respectively (see Supplementary Material Table S15). By keeping the same original users in each supernode for all interaction networks, it was possible to observe the differences that the same users showed when participating in each type of interaction. The results of the examination, using the previously mentioned dissimilarity metric [71], showed strong topological analogies between the different CNETs corresponding to each interaction type, as well as between each specific interaction CNET and the global CNET (see Table S17 in the Supplementary Material). In view of the above, and given that the global CNET provides a richer repository of information than other networks corresponding to specific interactions, it is the one considered in this study.

3.3.2 Analysing the community network structure

In this section we study CNET and examine the CNET skeleton. As mentioned, for the construction of CNET, the raw network is a weighted and undirected graph. CNET is depicted in Fig. 2 and Table 4 shows its main statistical properties. The weight of the links is calculated by summing the interactions between users of the pairs of communities corresponding to the link.

Figure 2, describes the average sentiment of the supernodes, showing that the network has a mostly neutral perception, as most of the supernodes have sentiment values close

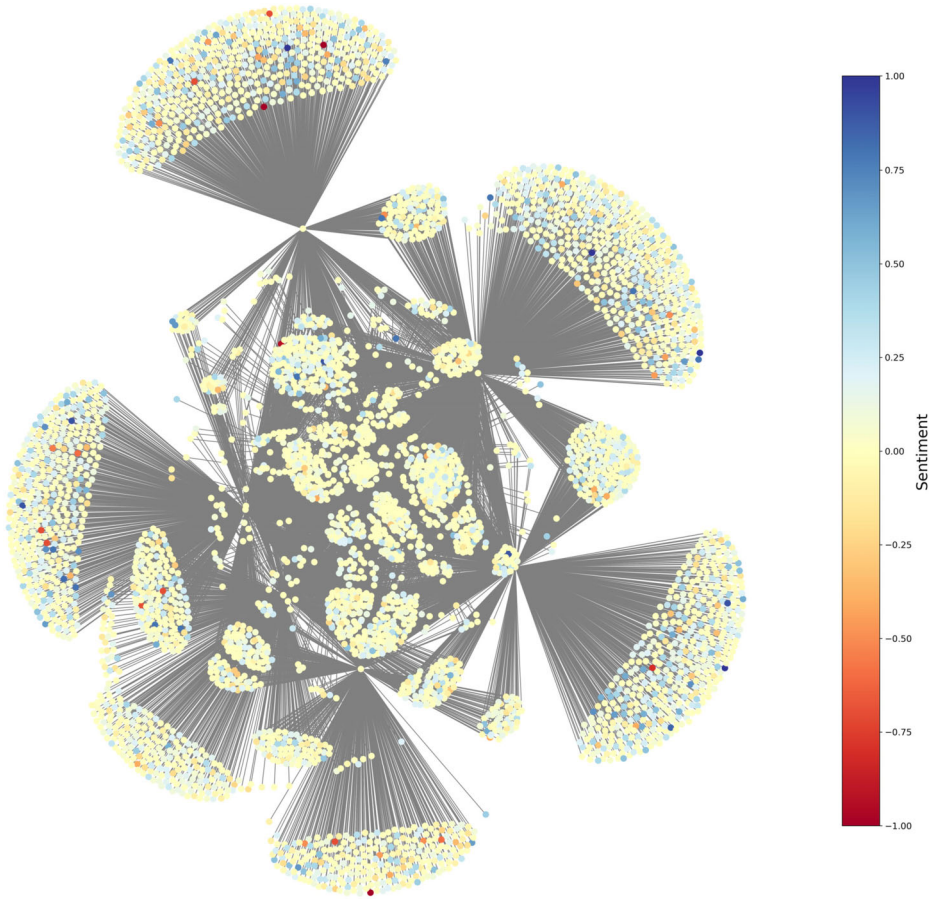

Figure 2 Visualization of CNET where each supernode represents a community. The colours in the graph represent the average sentiment of each community

Table 4 Statistical properties corresponding to CNET and CNET skeleton

Parameter	CNET	CNET Skeleton
Number of supernodes	4576	4576
Number of links	9855	8506
Maximum degree	1776	1555
Minimum degree	1	1
Average degree	4.3	3.7
Diameter	3	4
Average distance between any two nodes	2.54	3.02
Assortativity	–0.71	–0.74
Maximum degree centrality	0.39	0.34
Maximum closeness centrality	0.62	0.48
Maximum betweenness centrality	0.34	0.35
Maximum EHCC ($\delta = 0.1$)	991.97	2070.31
Maximum K-risk	46,099.29	7539.07

to 0. Therefore, although certain opinions may be considered negative or positive, they are not particularly radicalised. The supernodes with both the highest and lowest values are located in the external areas of the network, which have the lowest connectivity, so it is not possible to conclude that the debate is very extreme on the part of a group of supernodes or that there are echo chambers that divide the population according to their

opinions. This indicates that the debate is respectful and informative, despite the existence of interactions that show acceptance or rejection. Community 648 is representative of communities with a medium number of nodes and clearly positive sentiment (0.8). The most relevant hashtags in this supernode are "Earth Day" and "World Population Day", and the sentiment expressed is influenced by the perception towards these events. In contrast, community 638 also has 20 nodes and a sentiment of –0.92, with hashtags such as 'Devastating', 'Pollution' or 'Shocking', indicating a more defeatist community where the debate is more general. As communities get larger and debates cover more topics, these extreme opinions are less significant.

Regarding normalised centrality metrics in supernodes, 6 communities exhibited a degree and betweenness centrality higher than 0.1. The closeness centrality was in the range [0.33, 0.62] in all communities. Indicating high centrality values for only a few nodes in the network, while most nodes are located at an intermediate distance in the network, so that a large dispersion of nodes is not apparent.

Histograms of communities corresponding to mean degree, maximum and minimum degree, as well as mean, maximum and minimum pagerank, and degree assortativity can be found in the Supplementary Material Document, Figure S1, Figure S2, Figure S3, Figure S4, Figure S5, Figure S6 and Figure S7. In the analysis, it could be observed that the number of nodes in communities ranged from 3 to 1,487,676 but the majority of them had a very low number of nodes, for example 3409 supernodes had a number of nodes less than 12. The number of links varied from 3 to 3,515,164; 4484 communities had a number of links smaller than 274. The average pagerank was in the range [0,0.33], and the assortativity was in the interval [–1,0.37]. The majority of communities, 4556 (99,56%), exhibited negative assortativity, and little likelihood exists that two supernodes with similar degree will tend to join together.

For $\delta = 0.1$, EHCC was in the interval [1.61, 991.97]. In 38 communities, it was higher than 50, this once again demonstrates the existence of a small number of supernodes with high network centrality, in contrast to the vast majority of nodes with low connectivity. The larger EHCC was exhibited by community 5 followed by community 2. Concerning K-risk, the supernodes that presented the highest values were labeled 0 and 6, which were at the highest level of the k-core (see Supplementary Material Document Figure S9). At the next level of the k-core were supernodes 18 and 28, which showed higher K-risk values than several communities located in the preceding core. These supernodes symbolise pivotal elements in the information flow, despite their low centrality as indicated by classical metrics (see Supplementary Material Document, Table S14).

In CNET, all higher centralities corresponded to community 2, except EHCC. The main issues in this community are presented in the section Hashtags Analysis. A few highly connected supernodes exist, and the rest of the supernodes exhibit a much lower degree. This characteristic is typical in social networks, where a few nodes have the majority of followers and number of interactions [112]. Figure 2 shows that most of the supernodes have one or two links, indicating a low information flow. The network is well connected and the presence of hubs (highly connected supernodes) indicate that information can be transmitted from one node to another efficiently, but the K-risk analysis shows the vulnerabilities of a high centralised network. As is often the case in social networks, it is the central and most well connected nodes that generate the largest amount of information,

Table 5 For CNET and CNET Skeleton, average error according to dimension, considering the adjacency and laplacian methods for the embedding

CNET				CNET skeleton			
Adjacency embedding				Adjacency embedding			
5-10	10-15	15-20	20-25	5-10	10-15	15-20	20-25
0.30717	0.09789	0.012960	0.00640	0.29217	0.09000	0.01169	0.00603
Laplacian embedding				Laplacian embedding			
5-10	10-15	15-20	20-25	5-10	10-15	15-20	20-25
0.11057	0.01071	0.00248	0.00128	0.08996	0.00857	0.00100	0.00037

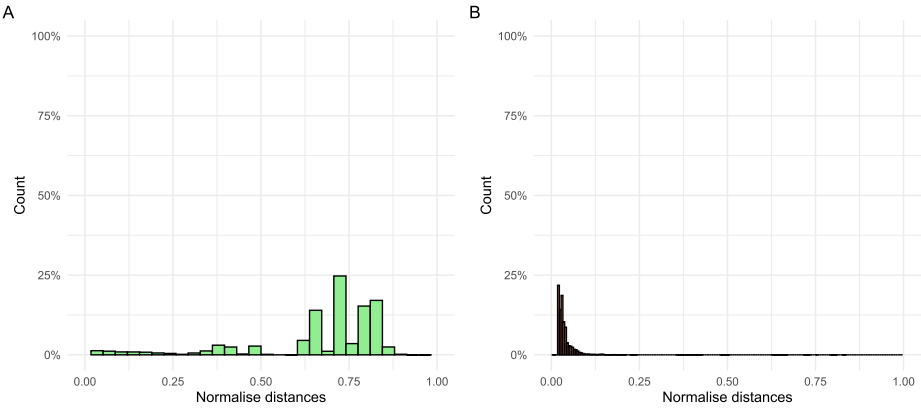

Figure 3 For CNET, normalised distances utilising the adjacency embedding with a dimension equal to 25, for both pairs of connected (A) and unconnected nodes (B)

which is in turn transmitted to the outer layers of the network. These outermost layers do not often interact with each other in discussions.

In the CNET skeleton, the maximum K-risk, degree and betweenness centralities are those of community 2, the maximum closeness centrality corresponds to community 4575.

For CNET, Table 5 depicts the obtained $\overline{E}$ in the estimation of embeddings (both adjacency and laplacian approaches). It can be observed that the optimal dimension is 25.

Figure 3 shows the normalised distances between supernode vectors, for both connected and unconnected supernode pairs, applying adjacency embedding, considering a dimension equal to 25. It can be observed that the feature vectors corresponding to pairs of unconnected supernodes look very similar to each other, indicating that they exhibit many commonalities in their connections to the rest of the network. Those vectors associated with pairs of connected supernodes show the greatest differences. As shown in the Supplementary Material Document, Figure S10, the same pattern is observed if the laplacian embedding vectors are taken into account.

For CNET, Figure S11, in the Supplementary Material Document, displays the Jaccard coefficient, resource allocation index, Adamic Adar index, and preferential attachment score metrics. The closer this value is to 1, the more likely it is that a link between supernodes will be formed. It can be observed that all these normalised similarity metrics had a value lower than 0.25.

Figure 4 depicts the CNET basic structure and the average sentiment of the supernodes after applying the `Noise Corrected` extraction method and the `Threshold filter`. Due to the particular CNET structure, only a combination of all possible meth-

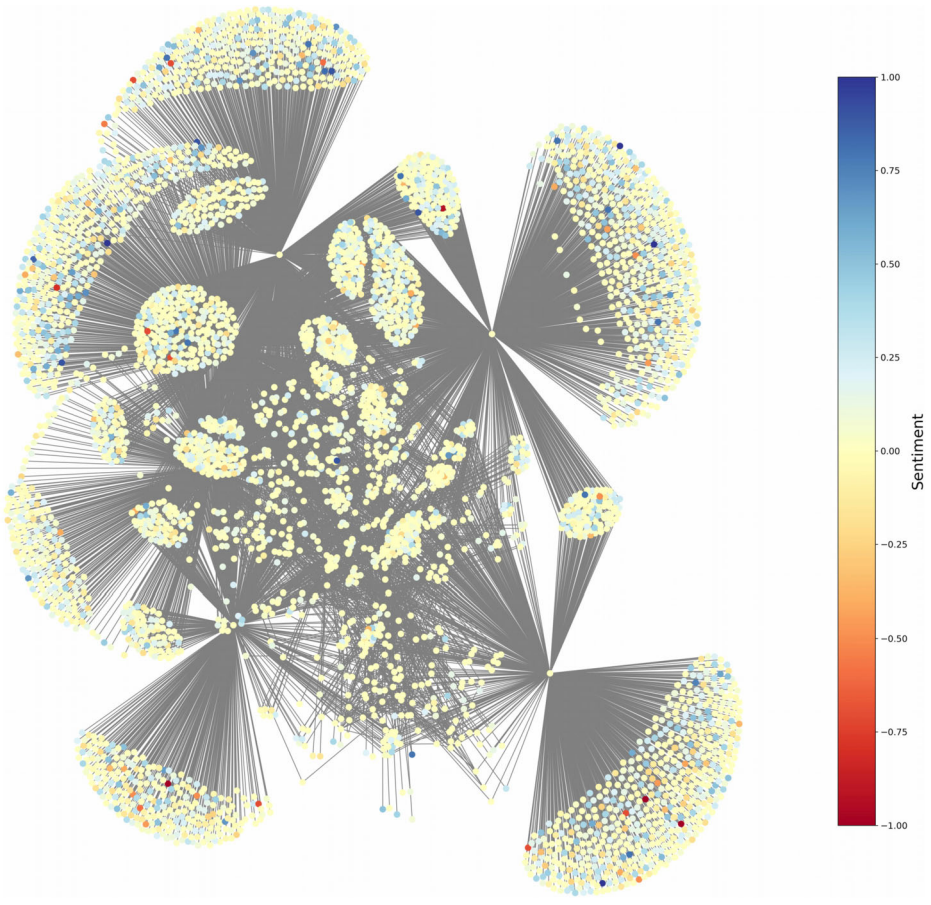

Figure 4 CNET Skeleton after applying the `Noise Corrected` extraction method and the `Threshold filter`. The colours of the supernodes represent the sentiment

ods and filters resulted in an optimal result, which was `Noise Corrected` extraction method - `Threshold Filter` [113] (see Supplementary Material Document, Table S16 and Figure S15).

The best combination of filters and methods maintained the number of nodes equal to 4576 but reduced the number of edges to 8506. Regarding normalised centrality metrics, 6 and 3 communities showed a degree and betweenness centrality higher than 0.1. The closeness centrality was in the range [0.25,0.48]. The maximum degree and betweenness centralities corresponded to community 2, the maximum closeness centrality was detected in the community 4575. Community 2 showed the highest EHCC, with community 5 coming next. With regard to K-risk, community 2 continues to exhibit the highest value, closely followed by community 1. Community 2 is possibly the most relevant community in the network and the main topics in its debate are sustainability and renewable energy, as will be explained below.

For the CNET skeleton, Table 5 displays the obtained $\overline{E}$ in the embedding computation (both adjacency and laplacian approaches). It can be seen that the best dimension is 25. Figure S12 and Figure S13, in the Supplementary Material Document, depict the normalised distances for all pairs of nodes with the optimal dimension for adjacency and laplacian embeddings, respectively. In addition, Figure S14 displays the Jaccard coefficient,

resource allocation index, Adamic Adar index, and preferential attachment score metrics. Figure S12 and Figure S13 show similar connectivity characteristics for the supernodes included in the skeleton to those observed in CNET. Figure S14 clarifies that in line with what was previously observed in CNET, all similarity metrics indicate low possibilities of new link formation between supernodes (all have a value lower than 0.25).

Regarding normalised centrality metrics in supernodes, all communities exhibited a degree in the range [0, 0.06] and 6 communities showed a betweenness centrality higher than 0.1. The closeness centrality was in the range [0.25, 0.48] in all communities. Comparing these measures with those of the original CNET shows a reduction in degree and closeness centrality due to the loss of links, although the measures do not vary much. However, the betweenness of the 6 main supernodes is not significantly reduced due to their topological position in the network and their key role in the information flow.

3.3.3 Analysis of supernodes

In order to better understand the different characteristics of the supernodes we carry out an analysis of their inner structure, using the aforementioned metric described in [71]. As previously mentioned, the aforementioned metric described in [71] was applied for studying the structural dissimilarity between supernodes. After running 200 experiments, the means in each of the three analysed percentiles were 0.18308 ($\pm$ 0.09428), 0.24632 ($\pm$ 0.13342), and 0.14951 ($\pm$ 0.07278). The median in each took the values 0.18264, 0.23934, and 0.14476. The minimum and maximum magnitudes were also analysed. The lowest values corresponded to 0, 0.01097 and 0, while the highest ones were 0.44583, 0.50828 and 0.36844. The topological structure of supernodes was also examined at global level. 1000 experiments were carried out, which provided mean and median dissimilarity values equal to 0.18898 ($\pm$ 0.09357) and 0.19718. 0 and 0.61172 were the minimum and maximum magnitudes, respectively.

It can be observed that a low topological dissimilarity exists both globally and in all percentiles. This suggests that the variability in the internal structure of the supernodes is small and the networks have a relatively uniform organisation in terms of the topological structure of the nodes.

3.3.4 Hashtags analysis

Hashtags allow tweets to be categorised by topic, making it easier to identify contextualised conversations, aiding marketing activities and measuring trends in discussions. Table 6 shows the 10 most used hashtags by the users of the 5 most relevant communities. Unlike keyword analysis, the purpose of hashtags is not to identify important information in discussions, but instead to narrow down the topic to which the hashtags refer and to disseminate them effectively across the network. We detected propagandistic or sensationalist hashtags such as `EarthDay`, `CleanAirDay`, `WeMeanToClean` or `ActOnClimate`. If we focus on the hashtags of community 2, due to their relevance in terms of centrality, we observe that these hashtags are related to the concepts `renewable`, `pollution`, `sustainable` and `energy`, as shown in Table 6. Only one hashtag was common to all six communities in the top 10 most prominent hashtags. A comprehensive examination of the hashtags associated with each of these communities makes it possible to identify a primary and distinctive theme for each. For community 0, this is the pollution of the oceans. Regarding community 1, it is the air pollution in India. Community 2 focuses on various

Table 6 Main hashtags of the 6 largest communities. Hashtags that are common to all communities are highlighted in bold

0	1	2	3	4	5
pollution	**pollution**	RenewableEnergy	**pollution**	ActOnClimate	**pollution**
GoodOmens	Delhi	renewable	airpollution	**pollution**	renewable
illustration	AirPollution	AirPollution	London	Trump	energy
NSTnation	indianarmy	**pollution**	ukair	EarthDay	green
sghaze	WeMeanToClean	energy	renewable	climate	climate
EarthDay	India	plastic	plastic	RenewableEnergy	solar
WorldOceansDay	OddEven	sustainable	sustainable	GreenNewDeal	ActOnClimate
RunForTheOceans	DELHI	solar	CleanAirDay	ClimateChange	RenewableEnergy
GreenNewDeal	WEnvironmentDay	renewables	energy	EPA	plastic
lockdowneffect	Diwali	climatechange	Fife	ClimateCrisis	sustainable

clean energies. For community 3, it is air pollution and plastics in Europe. In relation to community 4, it is global warming. Finally, Community 5 focuses on renewable energies and environmental sustainability, similar to community 2. These two communities show similarities not only in terms of the topics of debate but also in terms of being the ones with the highest network centrality, which is consistent with dealing with issues centred on environmental sustainability, which is the main focus of this study.

4 Discussion

Today, ecological sustainability is a global concern that requires action in a variety of contexts. This study analyses the perception of this issue in Twitter conversations. Using data from Twitter 2007 to 2022, the application of Network Theory allowed us to unravel the complex interactions occurring on the network and the keywords used. Data analytics, in particular, sentiment analysis and modelling techniques also allowed us to gain insight into existing perceptions and their relationships. All of this enabled us to obtain an excellent temporal and geographical perspective on the analysed topic.

The number of messages downloaded from Twitter exceeded 70 million, representing more than 13 million unique users, corresponding to 6 continents and 178 countries. According to the analysis carried out, messages showed a mostly positive but very close to neutral sentiment, in all studied regions. The huge number of tweets and the temporal extension (17 years) contemplated in the analysis minimise the possibility of biases. The sentiment displayed contrasts with sentiments detected in other works. These works studied debates focusing on environmental catastrophes or climate change, where a extreme sentiment and polarisation was found [49]. Since this research focuses on environmental sustainability, as well as on human-life-environment relationship, it uses a different set of terms to the issues mentioned above. The subject of ecological sustainability seems to elicit a less enthusiastic or passionate response among users.

The analysis identified the existence of common terms in tweets from different geographic locations, suggesting a common understanding of ecological sustainability in different locations. It was also unexpectedly detected that the strongest linear relationships occurred with socioeconomic variables, rather than environmental factors. While the characteristic sentiment was found to be common to the locations examined, this was not the case for all studied socioeconomic and environmental attributes (see Supplementary Material Document, Table S3 and Table S4). As result, the actions required to change individuals' perceptions had to be undertaken at a regional level. Notwithstanding the above, in a few places certain similitudes between variables were detected, suggesting the

possibility of carrying out some work on them collectively. Keyword and hashtag studies showed a strong user commitment to environmental sustainability.

Regarding the raw interaction network, it was composed of 8,840,532 nodes and 28,403,555 links. The computation of the communities (supernodes) of this network to create CNET (4576 supernodes) allowed a comprehensive topological analysis of it. This study revealed a significant disparity between the communities. The structure of connections throughout the entire CNET was analysed using a variety of metrics, including adjacency and laplacian embeddings, as well as Chebyshev, Euclidean, Minkowski, and Manhattan distances. This study found similarities in the connections established by supernodes that are not connected to each other, while the biggest differences were observed in the connections of supernode pairs that were connected, due to the structure of CNET, where large numbers of similar peripheral nodes are linked to few nodes with high centrality. The overwhelming majority of supernodes (4427) correspond to communities of less than 100 nodes, and are largely external with low centrality. Conversely, a small number of supernodes (21) have a high connectivity, occupying central positions within CNET and facilitating communication between disparate supernodes (see Supplementary Material Document, Figure S8). Nonetheless, the application of new centrality measures, such as K-risk, allowed the identification of communities that might have seemed peripheral but played a key role in the dissemination of information.

One such community is 18, where the K-risk value was even superior even to that of supernodes with a higher K-core. It seems to play an important role in disseminating information on a branch of the CNET. This community had 73,028 users and its primary hashtags were 'MSGTreePlantationDrive', 'MSGwelfare4All' and 'MSGmissionOfWelfare', as well as some related to pollution. A detailed analysis of this community revealed that it was primarily made up of a religious core led by MSG "the saint." Discourses common to Eastern religions, such as spirituality and respect for nature, were central to this community. The most relevant user within this community is @Gurmeetramrahim. Although it is not typical to consider this community as a significant contributor to the discourse on environmental sustainability, it is in fact, a supernode worth mentioning. This is due to the fact that it spreads positive and neutral messages concerning the protection of nature and the interrelationship that human beings have with it. Another community worth to mention is 28 (mainly related to India), comprising 17,788 nodes, which utilised the hashtags 'PollutionKiAsliJadbhagwa', 'EcoFriendlyChristmas', and 'VedicHoliHealthyHoli'. This community was populated primarily with users from India, where environmental sustainability was discussed at select events and festivals, including the Holi festival. Globally, the sentiment of the messages exchanged in this community leans towards neutrality. However, it should be noted that, if the negativity of the sentiment, exhibited in some messages, related to mass events criticising air and noise pollution, it was mitigated by others with positive sentiment, which were disseminated around these festivities.

In relation to the classic centrality measures, these revealed that the central communities of the CNET were significantly larger than the rest, and their discussions, as well as the hashtags associated with them, were more focused on pollution, renewable energies and the environment. Nevertheless, certain particularities are observed, as is the case of community 6, whose most prominent users are @BillGates and @BBCWorld. The main focus of their discussions was China, India, covid19 and technology, with special emphasis on environmental sustainability. Community 5, which was identified as a notable case

in the EHCC analysis, showed a prevalence of hashtags related to environmental sustainability, similar to those observed in community 2. Therefore, it is not surprising that one of the main users of this community was @EcoWatch, an environmental news organisation. In Community 4 the predominant hashtag was Trump. A closer look at the data reveals that notable users include @HuffPost and @TIME, two prominent newspapers, @CoriBush and @laurenboebert, two American politicians, and finally @WhiteHouse, the official White House account. This community approaches environmental sustainability with a political vision, influenced by the American political situation, and, consequently, with references to Trump, identifying the community 4 as a predominantly North American hub. It is worth highlighting the existence of a neutral sentiment despite highly polarised topics such as politics and the environment being discussed in the debates. The above seems to suggest that the terms used in obtaining the data set influenced the exclusion of the most radical tweets, showing the debates as a focused and respectful dialogue.

The debate in Community 2, which is the most notable according to the centrality metrics with the exception of EHCC, is about global sustainability in the world. Its users include @ClimateGroup, a non-governmental organisation dedicated to addressing solutions to climate change; @GerdaVerburg, former Dutch Minister of Agriculture, Nature and Food Quality; and @fbirol, executive director of the International Energy Agency (IEA). The participation of people who have qualifications, an influence and a strong commitment to improving environmental sustainability is important within this community.

Despite the differences in the approaches and issues that the examined communities show in discussions concerning environmental sustainability, the network maintains a coherent and cohesive debate, with a largely neutral sentiment in the messages in the main communities (see Table S18 in the Supplementary Material). Notice that the nature of the conversations analysed in this research is different from others ones with a higher ideological content [114, 115] and [116]. In the network we studied, although certain communities might have different opinions, there is no clear confrontation between ideologies in relation to environmental sustainability. This could be the reason why we did not find the same level of radicalization and confrontation as in [114, 115] and [116]. It should be noted that we found mostly neutral and slightly positive sentiment, which makes our study not comparable to the toxicity analyses conducted by [116] and [117]. In addition, the analysis methodology of [116] and [117] is based on deep learning models of toxic language, which is different from the sentiment analysis we performed. As in [117], our study finds size differences between the retweet network and the reply/quote networks. Additionally, we also detected different sizes in CNETs. The analysis in [117] shows high toxicity, which may be expected in climate protest network conversations, both in discussions occurring internally and between different ideological communities. It is worth mentioning that the communities in the network examined in the research that we described in this document, as mentioned, were established by analysing the topological structures of the network (using Leiden's method) and not by ideology. In contrast, sustainability debates tend to be less controversial because they are less burdened by ideological and political divisions than the climate change debate.

Focusing on the period spanning the last 5 years, we observed some events related to environmental sustainability that were relevant on the network. These included 'Sustainable Food Development in the EU' and the 'Banning of Palm Oil in EU products' in 2020, with a majority of positive tweets for them. Globally, both events generated a slightly pos-

itive but very close to neutral sentiment. In addition, other events took place such as the 'Industrial Emissions Reduction Plan in China' in 2021 registering a majority of positive tweets but also a high number of neutral and negative ones. Another was the 'Tesla Renewable Energy Project in Australia' in 2017 with a vast majority of neutral tweets. The specificity of the search that was carried out regarding environmental sustainability and not climate change excluded some events such as climate summits and shows other events more related to environmental policies, seeming to generate sentiments close to neutrality (see Supplementary Material Document, Table S19).

In the CNET, as observed in Fig. 2, few supernodes perform most of the interactions. The large number of small supernodes demonstrates a lack of multi-level debate. Based on both Fig. 2 and Twitter's own features, it seems that information is disseminated to the larger supernodes. By contrast, the smaller supernodes receive information but do not participate very actively in the conversations. It is interesting to highlight that the metrics used to study the possibilities of link formation between supernodes also showed small values.

The low connectivity of some supernodes shown on CNET might be a difficulty for communication between certain communities. Communities that are relatively small in relation to large central nodes need central nodes to exchange information, so communities with air pollution dialogues such as 269 and 349 could exchange information in a biassed way, as they would be forced to communicate with intermediary central communities. Greater connectivity between lower level supernodes could help solve issues at the local level. According to CNET structure, dialogue between smaller regions with similar problems may be reduced because they require a higher entity to centralise the dialogue. If horizontal communication occurred more frequently between supernodes, local conditions that prevent solutions to global issues involving ecological sustainability could be shared. This would prove to be a huge aid in tackling such problems.

In conclusion, the results provide a comprehensive view on the perception on ecological sustainability on Twitter. The differences and commonalities found between regions (keywords, hashtags and multi-linear regression models) prove both the enormous complexity of the problem and the need to find solutions for each location. It is necessary to promote participation as well as achieve more diverse debates, and consequently, increase connectivity between supernodes. The quicker people become aware of current environmental problems, the sooner it will be possible to take steps to preserve the natural environment, protect human health and reverse the negative effects of human activity on the planet.

In future research, based on data already downloaded from Twitter, we will analyse the CNET dynamics, studying its evolution by time ranges. This will make it possible to know if new structures are formed as a consequence of the modification of both nodes and links. Some competitive strategies can also be investigated as models for exchanging messages between supernodes in order to increase their connectivity and influence users' perceptions.

Abbreviations

CNET, Community Network; HCC, Hybrid Characteristic Centrality; SDGs, Sustainable Development Goals.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00545-x.

Additional file 1. Supplementary materials (PDF 5.0 MB)

Acknowledgements
This work was supported by a Predoctoral Research Grant which was obtained in the internal call in 2021 at the Universidad Francisco de Vitoria. This work was also partially supported by Telefónica Chair at the Universidad Francisco de Vitoria and by the Spanish Ministry of Science and Innovation (Contract No. PID2021-122711NB-C21). The authors would like to express gratitude to Mari Luz Congosto Martínez for her instruction in the use of the T-Hoarder software. They also thank Marta Subirán for her assistance with the software programs for preparing tweets for analysis.

Author contributions
JGS-S contributed in the literature review, in downloading and processing tweets, in sentiment, keywords and hashtags study. In addition to analysing the raw network, the community network and its skeleton. He also revised the paper. MLM-L contributed in the analysis of communities, in connectivity examination using embedding adjacency and laplacian embeddings, as well as in the link formation analysis. She also contributed to defining the fundamental of the research, writing and reviewed the manuscript. RMB participated in approach, establishment of methods, writing - examination and guidance supervision.

Funding information
Not applicable.

Data Availability
We only utilised public tweets. Additionally, we utilise the repository WorldBank|Data|Indicators . 2014. Data retrieved from World Development Indicators. http://data.worldbank.org/indicator Terms of Use, Reproduction, and Citation of WorldBank|Data|Indicators repository are detailed in: https://www.worldbank.org/en/archive/using-the-archives/terms-of-use-reproduction-and-citation. World Bank Group makes data available according open data standards and licensed under the Creative Commons Attribution license (CC-BY 4.0). We also indicate this information in the Section "Ethical approval". The data utilised to support the findings of this research are also available from the corresponding author upon reasonable request.

Code availability
The software code implemented in this research is available from the corresponding author upon reasonable request.

Declarations

Software programs
Several programs were developed using the R language, in order to: (1) To detect relationships between sentiments and both socioeconomic and environmental variables. The following packages were utilised *caret* [118, 119] and *stats* [120] packages. (2) to carry out the Keywords analysis, where the packages *udpipe* [121, 122] and *igraph* [123–125] were used. (3) Plotting of graphs, the package *ggplot2* [126, 127] was applied.
(4) To study the networks, to calculate embeddings and related metrics. The following packages were utilised: *dplyr* [128], *data.table* [129], *parallel* [130], *Rfast* [131–136], *igraph* [123–125], *iGraphMatch* [137], *fastnet* [138, 139], *multivariance* [140–145], *proxy* [146], *abdiv* [147], *plyr* [148], *matrix* [149], *stats* [120], *stringr* [150] (5) To plot histograms. The following packages were used: *ggplot2* [126, 127], *patchwork* [151], *scales* [152]. (6) Other utilised packages: *zoo* [153, 154], *ggpubr* [155], *purrr* [156], *tidyverse* [157], *rstatix* [158], *quanteda* [159], *car* [160, 161], *caTools* [162], *GGally* [163], *glmnet* [164–167], *MASS* [168, 169], *sqldf* [170], *sjmisc* [171, 172], *reshape2* [173], *FactoMineR* [174, 175], *Factoextra* [176], *splus2R* [177], *astsa* [178–180], *magrittr* [181], *lubridate* [182, 183], *tseries* [184], *openair* [185, 186], *jsonlite*, [187, 188], *tm* [189–191], *tmap* [192], *NLP*[193], *OpenNLP* [194], *gtools* [195], *tnet* [196, 197], *xml2* [198], *SnowballC* [199], *base* [200].
Several functionalities were developed using Python language, they were: (1) To prepare the tweets for analysis, where the following packages were utilised: *TextBlob* [77], *emoji* [201], *demoji* [202], *stop-words* [203], *nltk* [204, 205]. (2) To build and study the interaction networks, where *networkx* [206, 207], *pickle* [208, 209] and *copy* [210] were used. (3) Identification of the CNET skeleton, comparison of filters for obtaining it, and metrics analysis where *netbone* [68] was used. (4) General uses, where *os* [211], *numpy* [212], matplotlib [213], itertools [209, 214], csv [209, 215] and *pandas* [216–218] were applied. (5) Other utilised packages: warnings [219], *operator* [220], *sys* [221], *regex* [222], *glob* [223], *random* [224], *dask* [225], *cdlib* [226, 227], *collections* [228], *seaborn* [229, 230], *subprocess* [231], *re* [232], *pyvis* [233].
In addition to the above we use the software provided in [69–71].

Ethics approval
The research complied with all the relevant national regulations.
We only used public tweets. Terms of Use, Reproduction, and Citation of WorldBank|Data|Indicators repository [76] can be found here: https://www.worldbank.org/en/archive/using-the-archives/terms-of-use-reproduction-and-citation. World Bank Group makes data available according open data standards and licensed under the Creative Commons Attribution license (CC-BY 4.0).

Consent to participate
There is consent from the authors to participate in the manuscript.

Consent for publication
There is consent to the publication of the manuscript.

Competing interests
The authors declare no competing interests.

Author details
[1]Higher Polytechnic School, Universidad Francisco de Vitoria, carretera Pozuelo a, Av de Majadahonda, Km 1.800, 28223 Madrid, Spain. [2]Grupo de Sistemas Complejos, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Avda. Puerta de Hierro 2-4, 28040 Madrid, Spain.

Received: 4 June 2024 Accepted: 22 March 2025 Published online: 17 April 2025

References

1. Jiang H, Sun Z, Guo H, Weng Q (2021) An assessment of urbanization sustainability in China between 1990 and 2015 using land use efficiency indicators. npj Urban Sustain 1:34. https://doi.org/10.1038/s42949-021-00032-y
2. Bologna M, Aquino G (2020) Deforestation and world population sustainability: a quantitative analysis. Sci Rep 10:7631. https://doi.org/10.1038/s41598-020-63657-6
3. Chen Y, Bai Y, Liu H, Alatalo JM, Jiang B (2020) Temporal variations in ambient air quality indicators in Shanghai municipality, China. Sci Rep 10:11350. https://doi.org/10.1038/s41598-020-68201-0
4. Juginović A, Vuković M, Aranza I, Biloš V (2021) Health impacts of air pollution exposure from 1990 to 2019 in 43 European countries. Sci Rep 11:22516. https://doi.org/10.1038/s41598-021-01802-5
5. Mahmood H, Irshad URA, Tanveer M (2024) Do innovation and renewable energy transition play their role in environmental sustainability in western Europe? J Zhejiang Univ (Humanit Soc Sci) 11(1):22. https://doi.org/10.1057/s41599-023-02539-4
6. Riera JA, Lima RM, Hoteit I, Knio O (2022) Simulated co-optimization of renewable energy and desalination systems in neom, Saudi Arabia. Nat Commun 13(1):3514. https://doi.org/10.1038/s41467-022-31233-3
7. Stokes LC, Warshaw C (2017) Renewable energy policy design and framing influence public support in the United States. Nat Energy 2(8):17107. https://doi.org/10.1038/nenergy.2017.107
8. Wang J, Chen L, Tan Z, Du E, Liu N, Ma J, Sun M, Li C, Song J, Lu X, Tan C-W, He G (2023) Inherent spatiotemporal uncertainty of renewable power in China. Nat Commun 14(1):5379. https://doi.org/10.1038/s41467-023-40670-7
9. Sonter LJ, Dade MC, Watson JEM, Valenta RK (2020) Renewable energy production will exacerbate mining threats to biodiversity. Nat Commun 11(1):4174. https://doi.org/10.1038/s41467-020-17928-5
10. Li S, Zhuang Y, Liu H, Wang Z, Zhang F, Lv M, Zhai L, Fan X, Niu S, Chen J, Xu C, Wang N, Ruan S, Shen W, Mi M, Wu S, Du Y, Zhang L (2023) Enhancing rice production sustainability and resilience via reactivating small water bodies for irrigation and drainage. Nat Commun 14(1):3794. https://doi.org/10.1038/s41467-023-39454-w
11. Chakraborti R, Davis KF, DeFries R, Rao ND, Joseph J, Ghosh S (2023) Crop switching for water sustainability in India's food bowl yields co-benefits for food security and farmers' profits. Nat Water 1(10):864–878. https://doi.org/10.1038/s44221-023-00135-z
12. United Nations (w.d.) Academic Impact. https://www.un.org/en/academic-impact/sustainability
13. Zhiming S, Xianglong C, Hanfa X, Hongtao M, Yuan M (2020) Regional differences in socioeconomic trends: the spatiotemporal evolution from individual cities to a megacity region over a long time series. PLoS ONE 15(10):e0244084. https://doi.org/10.1371/journal.pone.0244084
14. Nuralina K, Baizholova R, Aleksandrova N, Konstantinov V, Biryukov A (2023) Socio-economic development of countries based on the Composite Country Development Index (CCDI). Reg Sustain 4:115–128. https://doi.org/10.1016/j.regsus.2023.03.005
15. Tsiotas D, Dialesiotis S, Christopoulou O (2023) Examining the relationship between regional economic resilience and epidemiologic spread of covid-19: evidence from Greece. Environ Dev Sustain. https://doi.org/10.1007/s10668-023-04240-7
16. Mouronte-López ML, Savall J (2024) Exploring socioeconomic similarity-inequality: a regional perspective. J Zhejiang Univ (Humanit Soc Sci) 11:1–16. https://doi.org/10.1057/s41599-024-02730-1
17. Sener B, Akpinar E, Ataman MB (2023) Unveiling the dynamics of emotions in society through an analysis of online social network conversations. Sci Rep 13(1):14997. https://doi.org/10.1038/s41598-023-41573-9
18. Cebral-Loureda M, Hernández-Baqueiro A, Tamés-Muñoz E (2023) A text mining analysis of human flourishing on Twitter. Sci Rep 13(1):3403. https://doi.org/10.1038/s41598-023-30209-7
19. Pröllochs N, Bär D, Feuerriegel S (2021) Emotions explain differences in the diffusion of true vs. false social media rumors. Sci Rep 11(1):22721. https://doi.org/10.1038/s41598-021-01813-2
20. Mouronte-López ML, Savall J, Mora A (2023) Analysing the sentiments about the education system trough Twitter. Educ Inf Technol 28:10965–10994. https://doi.org/10.1007/s10639-022-11493-8
21. Mouronte-López ML, Subirán M (2023) Analysis of worldwide greenhouse and carbon monoxide gas emissions: which countries exhibit a special pattern? A closer look via Twitter. Int J Environ Res 17(19):1–20. https://doi.org/10.1007/s41742-023-00510-4
22. Mouronte-López ML, Subirán M (2022) What do Twitter users think about climate change? Characterization of Twitter interactions considering geographical, gender, and account typologies perspectives. Weather Clim Soc 14(4):1039–1064. https://doi.org/10.1175/WCAS-D-21-0163.1
23. Doshi D, Garschagen M (2023) Assessing social contracts for urban adaptation through social listening on Twitter. npj Urban Sustain 3(1):30. https://doi.org/10.1038/s42949-023-00108-x
24. Merz E, Saberski E, Gilarranz LJ, Isles PDF, Sugihara G, Berger C, Pomati F (2023) Disruption of ecological networks in lakes by climate change and nutrient fluctuations. Nat Clim Change 13(4):389–396. https://doi.org/10.1038/s41558-023-01615-6
25. Rocha LEC, Ryckebusch J, Schoors K, Smith M (2021) The scaling of social interactions across animal species. Sci Rep 11(1):12584. https://doi.org/10.1038/s41598-021-92025-1
26. O'Garra T, Mangubhai S, Jagadish A, Tabunakawai-Vakalalabure M, Tawake A, Govan H, Mills M (2023) National-level evaluation of a community-based marine management initiative. Nat Sustain 6(8):908–918. https://doi.org/10.1038/s41893-023-01123-7
27. French RK, Anderson SH, Cain KE, Greene TC, Minor M, Miskelly CM, Montoya JM, Wille M, Muller CG, Taylor MW, Digby A, Crane J, Davitt G, Eason D, Hedman P, Jeynes B, Latimer S, Little S, Mitchell M, Osborne J, Philp B, Salton A, Uddstrom L, Vercoe D, Webster A, Holmes EC, Kākāpō Recovery Team (2023) Host phylogeny shapes viral

transmission networks in an island ecosystem. Nat Ecol Evol 7(11):1834–1843. https://doi.org/10.1038/s41559-023-02192-9
28. Talaga S, Stella M, Swanson TJ, Teixeira AS (2023) Polarization and multiscale structural balance in signed networks. Commun Phys 6(1):349. https://doi.org/10.1038/s42005-023-01467-8
29. Ting-Ting G, Gang Y (2022) Autonomous inference of complex network dynamics from incomplete and noisy data. Nat Comput Sci 2(3):160–168. https://doi.org/10.1038/s43588-022-00217-0
30. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. In: The web conference. https://api.semanticscholar.org/CorpusID:1508503
31. Morales AJ, Borondo J, Losada JC, Benito RM (2014) Efficiency of human activity on information spreading on Twitter. Soc Netw 39:1–11. https://doi.org/10.1016/j.socnet.2014.03.007
32. Weng L, Menczer F, Ahn Y-Y (2013) Virality prediction and community structure in social networks. Sci Rep 3(1):2522. https://doi.org/10.1038/srep02522
33. Kawamoto T (2023) Single-trajectory map equation. Sci Rep 13(1):6597. https://doi.org/10.1038/s41598-023-33880-y
34. Radicioni T, Saracco F, Pavan E, Squartini T, Garlaschelli D (2021) Analysing Twitter semantic networks: the case of 2018 Italian elections. Sci Rep 11(1):13207. https://doi.org/10.1038/s41598-021-92337-2
35. Borondo J, Morales AJ, Benito RM, Losada JC (2015) Multiple leaders on a multilayer social media. Chaos Solitons Fractals 72:90–98. https://doi.org/10.1016/j.chaos.2014.12.023
36. Fu J, Zhang W, Wu J (2017) Identification of leader and self-organizing communities in complex networks. Sci Rep 7(1):704. https://doi.org/10.1038/s41598-017-00718-3
37. Lyu H, Kureh YH, Vendrow J, Porter MA (2024) Learning low-rank latent mesoscale structures in networks. Nat Commun 15(1):224. https://doi.org/10.1038/s41467-023-42859-2
38. Caldarelli G, De Nicola R, Del Vigna F, Petri G, Righetti L, Ricchiuti F, Schiavone A, Scala A, Porcelli F (2020) The role of bot squads in the political propaganda on Twitter. Commun Phys 3:81. https://doi.org/10.1038/s42005-020-0340-4
39. Yang C, Harkreader R, Gu G (2013) Empirical evaluation and new design for fighting evolving Twitter spammers. IEEE Trans Inf Forensics Secur 8(8):1280–1293. https://doi.org/10.1109/tifs.2013.2267732
40. Shevtsov A, Tzagkarakis C, Antonakaki D, Ioannidis S (2022) Identification of Twitter bots based on an explainable machine learning framework: the US 2020 elections case study, vol 16. Association for the Advancement of Artificial Intelligence (AAAI), Washington D.C., pp 956–967. https://doi.org/10.1609/icwsm.v16i1.19349
41. Shao C, Hui P-M, Wang L, Jiang X, Flammini A, Menczer F, Ciampaglia GL (2018) Anatomy of an online misinformation network. PLoS ONE 13(4):1–23. https://doi.org/10.1371/journal.pone.0196087
42. Yoshida M, Sakaki T, Kobayashi T, Toriumi F (2021) Japanese conservative messages propagate to moderate users better than their liberal counterparts on Twitter. Sci Rep 11:19224. https://doi.org/10.1038/s41598-021-98349-2
43. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559
44. Ma S, Feng L, Lai CH (2018) Mechanistic modelling of viral spreading on empirical social network and popularity prediction. Sci Rep 8:13126. https://doi.org/10.1038/s41598-018-31346-0
45. Mønsted B, Sapieżyński P, Ferrara E, Lehmann S (2017) Evidence of complex contagion of information in social media: an experiment using Twitter bots. PLoS ONE 12(9):1–12. https://doi.org/10.1371/journal.pone.0184148
46. Carballosa A, Mussa-Juane M, Muñuzuri AP (2021) Incorporating social opinion in the evolution of an epidemic spread. Sci Rep 11:1772. https://doi.org/10.1038/s41598-021-81149-z
47. Daume S, Bjersér P, Galaz V (2023) Mapping the automation of Twitter communications on climate change, sustainability, and environmental crises — a review of current research. Curr Opini Environ Sustain 65:101384. https://doi.org/10.1016/j.cosust.2023.101384
48. Veltri GA, Atanasova D (2015) Climate change on Twitter: content, media ecology and information sharing behaviour. Public Underst Sci 26(6):721–737. https://doi.org/10.1177/0963662515613702
49. Falkenberg M, Galeazzi A, Torricelli M, Di Marco N, Larosa F, Sas M, Mekacher A, Pearce W, Zollo F, Quattrociocchi W, Baronchelli A (2022) Growing polarization around climate change on social media. Nat Clim Change 12(12):1114–1121 https://doi.org/10.1038/s41558-022-01527-x
50. Meyer H, Peach AK, Guenther L, Kedar HE, Brüggemann M (2023) Between calls for action and narratives of denial: climate change attention structures on Twitter. Media Commun. https://doi.org/10.17645/mac.v11i1.6111
51. Caravaggi A, Olin AB, Franklin KA, Dudley SP (2021) Twitter conferences as a low-carbon, far-reaching and inclusive way of communicating research in ornithology and ecology. Ibis 163(4):1481–1491. https://doi.org/10.1111/ibi.12959
52. Vaibhav DS, Annalise BF, Finnegan B, Kaitlin C (2024) Green consumerism: a cross-cultural linguistic and sentiment analysis of sustainable consumption discourse on Twitter (X). J Curr Issues Res Advert 45:1–28. https://doi.org/10.1080/10641734.2024.2318705
53. Segerberg A, Bennett WL (2011) Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests. Commun Rev 14(3):197–215. https://doi.org/10.1080/10714421.2011.597250
54. Perrotti D, Hyde K, Otero D (2020) Can water systems Foster commoning practices? Analysing leverages for self-organization in urban water commons as social–ecological systems. Sustain Sci 15(3):781–795. https://doi.org/10.1007/s11625-020-00782-1
55. Morales AJ, Borondo J, Losada JC, Benito RM (2015) Measuring political polarization: Twitter shows the two sides of Venezuela. Chaos, Interdiscip J Nonlinear Sci 25(3):033114. https://doi.org/10.1063/1.4913758
56. Martin-Gutierrez S, Losada JC, Benito RM (2023) Multipolar social systems: measuring polarization beyond dichotomous contexts. Chaos Solitons Fractals 169:113244. https://doi.org/10.1016/j.chaos.2023.113244
57. Mouronte-López ML, Subirán M (2022) Modeling the interaction networks about the climate change on Twitter: a characterization of its network structure. Complexity 2022:8924468. https://doi.org/10.1155/2022/8924468
58. Mouronte-López ML, Gómez J, Benito RM (2024) Patterns of human and bots behaviour on Twitter conversations about sustainability. Sci Rep 14:3223. https://doi.org/10.1038/s41598-024-52471-z

59. Piñeiro V, Arias J, Dürr J, Elverdin P, Ibáñez AM, Kinengyere A, Morales Opazo C, Owoo N, Page JR, Prager SD, Torero M (2020) A scoping review on incentives for adoption of sustainable agricultural practices and their outcomes. Nat Sustain 3(10):809–820
60. Rashid A, Salwa H, Evans S, Longhurst P (2008) A comparison of four sustainable manufacturing strategies. Int J Sustain Eng 1(3):214–229. https://doi.org/10.1080/19397030802513836
61. Winkler L, Pearce D, Nelson J, Babacan O (2023) The effect of sustainable mobility transition policies on cumulative urban transport emissions and energy demand. Nat Commun 14(2357):1–14
62. European Comission Business (w.d.) European Comission. Business, Economy, Euro. Internal Market, Industry, Entrepreneurship and SMEs. https://single-market-economy.ec.europa.eu/industry/sustainability_en
63. EPA United States Environmental Protection Agency (w.d.) https://www.epa.gov/sustainability/sustainable-manufacturing
64. U. S. Department of Transportation (w.d.) Sustainability. https://www.transportation.gov/tags/sustainability
65. United Nations (w.d.) Objetivos de desarrollo sostenible. La Asamblea General adopta la Agenda 2030 para el Desarrollo Sostenible. https://www.un.org/sustainabledevelopment/es/2015/09/la-asamblea-general-adopta-la-agenda-2030-para-el-desarrollo-sostenible/
66. Stanley N, Kwitt R, Niethammer M, Mucha PJ (2018) Compressing networks with super nodes. Sci Rep 8(10892):10892. https://doi.org/10.1038/s41598-018-29174-3
67. Borondo J, Morales AJ, Benito RM, Losada JC (2014) Mapping the online communication patterns of political conversations. Phys A, Stat Mech Appl 414:403–413. https://doi.org/10.1016/j.physa.2014.06.089
68. Yassin A, Haidar A, Cherifi H, Seba H, Togni O (2023) An evaluation tool for backbone extraction techniques in weighted complex networks. Sci Rep 13(1):17000. https://doi.org/10.1038/s41598-023-42076-3
69. García-Algarra J, Pastor JM, Iriondo JM, Galeano J (2017) Ranking of critical species to preserve the functionality of mutualistic networks using the k-core decomposition. PeerJ 5:1–17. https://doi.org/10.7717/peerj.3321
70. Liu J, Zheng J (2023) Identifying important nodes in complex networks based on extended degree and E-shell hierarchy decomposition. Sci Rep 13:1–10. https://doi.org/10.1038/s41598-023-30308-5
71. Schieber T, Carpi L, Díaz-Guilera A, Pardalos PM, Masoller C, Ravetti MG (2017) Quantification of network structural dissimilarities. Nat Commun 8:1–10. https://doi.org/10.1038/ncomms13928
72. Congosto ML, Basanta-Val P, Sanchez-Fernandez L (2017) T-hoarder: a framework to process Twitter data streams. J Netw Comput Appl 83:28–39. https://doi.org/10.1016/j.jnca.2017.01.029
73. Esmukov K (2023) geopy. Version 2.4.1. https://geopy.readthedocs.io/en/stable/
74. Nominatim developer community (w.d.) Open-source geocoding with OpenStreetMap data. https://nominatim.org/
75. OpenStreetMap Foundation (w.d.) OpenStreetMap. https://osmfoundation.org/wiki/Main_Page
76. The World Bank (2014) Indicator. Data retrieved from World Development Indicators. https://www.worldbank.org/en/archive/using-the-archives/terms-of-use-reproduction-and-citation. https://creativecommons.org/licenses/by/4.0/. http://data.worldbank.org/indicator
77. Loria S (w.d.) TextBlob: simplified Text Processing. https://textblob.readthedocs.io/en/dev/
78. Zeileis A, Hothorn T (2002) Diagnostic Checking in Regression Relationships. https://CRAN.R-project.org/doc/Rnews/
79. Lai TL, Robbins H, Wei C (1979) Strong consistency of least squares estimates in multiple regression ii. J Multivar Anal 9(3):343–361. https://doi.org/10.1016/0047-259X(79)90093-9
80. Abilhoa WD, De Castro LN (2014) A keyword extraction method from Twitter messages represented as graphs. Appl Math Comput 240:308–325. https://doi.org/10.1016/j.amc.2014.04.090
81. Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9(5233):5233. https://doi.org/10.1038/s41598-019-41695-z
82. Anuar SHH, Abas ZA, Yunos NM, Zaki NHM, Hashim NA, Mokhtar MF, Asmai SA, Abidin ZZ, Nizam AF (2021) Comparison between Louvain and Leiden algorithm for network structure: a review. J Phys Conf Ser 2129(1):012028. https://doi.org/10.1088/1742-6596/2129/1/012028
83. Wang C, Wang F, Onega T (2020) Network optimization approach to delineating health care service areas: spatially constrained Louvain and Leiden algorithms. Trans GIS 25(2):1065–1081. https://doi.org/10.1111/tgis.12722
84. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1):107–117. https://doi.org/10.1016/S0169-7552(98)00110-X. Proceedings of the Seventh International World Wide Web Conference
85. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701. https://doi.org/10.1103/physrevlett.89.208701
86. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67(2):026126. https://doi.org/10.1103/physreve.67.026126
87. Dessì D, Cirrone J, Recupero DR, Shasha D (2018) SuperNoder: a tool to discover over-represented modular structures in networks. BMC Bioinform 19:318. https://doi.org/10.1186/s12859-018-2350-8
88. Chan SY, Morgan K, Parsons N, Ugon J (2022) Supernodes: a generalization of the rich-club. J Complex Netw 10(1):052. https://doi.org/10.1093/comnet/cnab052. https://academic.oup.com/comnet/article-pdf/10/1/cnab052/42192854/cnab052.pdf
89. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67(2):026126
90. Foster JG, Foster DV, Grassberger P, Paczuski M (2010) Edge direction and the structure of networks. Proc Natl Acad Sci 107(24):10815–10820
91. Nieminen J (1974) On the centrality in a graph. Scand J Psychol 15(1):332–336. https://doi.org/10.1111/j.1467-9450.1974.tb00598.x
92. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239. https://doi.org/10.1016/0378-8733(78)90021-7
93. Wasserman S, Faust K (1994) Social network analysis: methods and applications. The Press Syndicate of the University of Cambridge.
94. Brandes U (2001) A faster algorithm for betweenness centrality*. J Math Sociol 25(2):163–177. https://doi.org/10.1080/0022250x.2001.9990249

95. Brandes U (2008) On variants of shortest-path betweenness centrality and their generic computation. Soc Netw 30(2):136–145. https://doi.org/10.1016/j.socnet.2007.11.001
96. Brandes U, Pich C (2007) Centrality estimation in large network. Int J Bifurc Chaos 17(07):2303–2318. https://doi.org/10.1142/s0218127407018403
97. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35. https://doi.org/10.2307/3033543
98. Batagelj V, Zaversnik M (2003) An O(m) Algorithm for Cores Decomposition of Networks. https://arxiv.org/abs/cs/0310049
99. Sussman D, Tang M, Fishkind D, Priebe C (2011) A consistent adjacency spectral embedding for stochastic blockmodel graphs. J Am Stat Assoc 107:1119–1128. https://doi.org/10.1080/01621459.2012.699795
100. Gu W, Tandon A, Ahn Y-Y, Radicchi F (2021) Principled approach to the selection of the embedding dimension of networks. Nat Commun 12:3772. https://doi.org/10.1038/s41467-021-23795-5
101. Coghetto R (2016) Chebyshev distance. Formaliz Math 24:121–141. https://doi.org/10.1515/forma-2016-0010
102. Liberti L, Lavor C, Maculan N, Mucherino A (2014) Euclidean distance geometry and applications. SIAM Rev 56(1):3–69. https://doi.org/10.1137/120875909
103. Thant A, Aye S (2020) Euclidean, Manhattan and Minkowski distance methods for clustering algorithms. Int J Sci Res Sci Eng Technol 7:553–559. https://doi.org/10.32628/IJSRSET2073118
104. Falcón R Practicing machine Learning interview questions in R
105. Jaccard P (1901) Distribution de la flore Alpine dans le bassin des dranses et dans quelques régions voisines. Bull Soc Vaud Sci Nat 37:241–272
106. Tao Z, Linyuan L, Yi-Cheng Z (2009) Predicting missing links via local information. Eur Phys J B 71(4):623–630. https://doi.org/10.1140/epjb/e2009-00335-8
107. Adamic LA, Adar E (2003) Friends and neighbors on the web. Soc Netw 25(3):211–230. https://doi.org/10.1016/s0378-8733(03)00009-1
108. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512. https://doi.org/10.1126/science.286.5439.509
109. Yuxin Z, Dazuo T, Feng Y (2020) Effectiveness of entropy weight method in decision-making. Math Probl Eng 2020:1–5. https://doi.org/10.1155/2020/3564835
110. Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: is online political communication more than an echo chamber? Psychol Sci 26(10):1531–1542. https://doi.org/10.1177/0956797615594620
111. Barberá P (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Polit Anal 23(1):76–91. https://doi.org/10.1093/pan/mpu011
112. Wang W, Chen X, Jiang S, Wang H, Yin M, Wang P (2020) Exploring the construction and infiltration strategies of social bots in sina microblog. Sci Rep 10(1):19821. https://doi.org/10.1038/s41598-020-76814-8
113. Coscia M, Neffke F (2017) Network Backboning with Noisy Data. arXiv:1701.07336
114. Borondo J, Morales AJ, Losada JC, Benito RM (2012) Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish presidential election as a case study. Chaos 22:1–6. https://doi.org/10.1063/1.4729139
115. Martin-Gutierrez S, Morales J, Torcal M, Losada JC, Benito RM (2024) In-party love spreads more efficiently than out-party hate in online communities. Sci Rep 14:1–12. https://doi.org/10.1038/s41598-024-65688-9
116. Falkenberg M, Zollo F, Quattrociocchi W, Pfeffer J, Baronchelli A (2024) Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. Nat Commun 15:1–12. https://doi.org/10.1038/s41467-024-53868-0
117. Meyer H, Pröschel L, Brüggemann M From Disruptive Protests to Disrupted Networks? Analyzing Levels of Polarization in the German Twitter Discourses Around "Fridays for Future" and "The Last Generation". https://doi.org/10.31219/osf.io/nd68z
118. Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):1–26. https://doi.org/10.18637/jss.v028.i05
119. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C, Hunt T (2023) Classification and Regression Training. https://doi.org/10.32614/CRAN.package.caret. https://cran.r-project.org/web/packages/caret/index.html
120. R Core Team (2024) R: a Language and Environment for Statistical Computing, Vienna, Austria. https://www.R-project.org/
121. Wijffels J BNOSAC, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic, Straka M, Straková J (2023) udpipe: tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. https://doi.org/10.32614/CRAN.package.udpipe. https://cran.r-project.org/web/packages/udpipe/index.html
122. Straka M, Straková J (2017) Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Hajič J, Zeman D (eds) Proceedings of the CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. Association for Computational Linguistics, Vancouver, pp 88–99. https://doi.org/10.18653/v1/K17-3009. https://aclanthology.org/K17-3009/
123. Csardi G, Nepusz T (2006) The igraph software package for complex network research. Int J Res Complex Syst 1695
124. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, Müller K (2024) igraph: network Analysis and Visualization in R. R package version 2.0.3. https://doi.org/10.5281/zenodo.7682609. https://CRAN.R-project.org/package=igraph
125. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, Müller K, Salmon M, Antonov M, Chan Zuckerberg Initiative (2024) igraph. Network Analysis and Visualization. https://doi.org/10.32614/CRAN.package.igraph. https://cran.r-project.org/web/packages/igraph/index.html
126. Wickham H (2016) Ggplot2: elegant Graphics for Data Analysis. Springer, New York. https://ggplot2.tidyverse.org
127. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke KO, Woo K, Yutani H, Dunnington D, van den Brand T, Posit PBC (2024) ggplot2: create Elegant Data Visualisations Using the Grammar of Graphics. https://doi.org/10.32614/CRAN.package.ggplot2. https://cran.r-project.org/web/packages/ggplot2/index.html

128. Wickham H, François R, Henry L, Müller K, Vaughan D, Posit Software, PBC (2023) dplyr: a Grammar of Data Manipulation. https://doi.org/10.32614/CRAN.package.dplyr. https://cran.r-project.org/web/packages/dplyr/index.html
129. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B, Krylov I, Stetsenko P, Short T, Lianoglou S, Antonyan E, Bonsch M, Parsonage H, Ritchie S, Ren K, Tan X, Saporta R, Seiskari O, Dong X, Lang M, Iwasaki W, Wenchel S, Broman K, Schmidt T, Arenburg D, Smith E, Cocquemas F, Gomez M, Chataignon P, Blaser N, Selivanov D, Riabushenko A, Lee C, Groves D, Possenriede D, Parages F, Toth D, Yaramaz-David M, Perumal A, Sams J, Morgan M, Quinn M, @javrucebo, @marc-outins, Storey R, Saraswat M, Jacob M, Schubmehl M, Vaughan D, Silvestri L, Hester J, Damico A, Freundt S, Simons D, Sales de Andrade E, Miller C, Meldgaard JP, Tlapak V, Ushey K, Eddelbuettel D, Fischetti T, Shilon O, Khotilovich V, Wickham H, Becker B, Haynes K, Kamgang BC, Delmarcell O, O'Brien J, de Mezquita D, Czekanski M, Shemetov D, Jha N, Wu J, Giné-Vázquez I, Chetia A, Amoakohene D, Feliz A, Young M, Seeto M, Grosjean P, Runge V, Wia C, Maigné E, Rocher V, Lulla V, Sluga A, Evans B (2024) data.table: extension of 'data.frame' https://doi.org/10.32614/CRAN.package.data.table. https://cran.r-project.org/web/packages/data.table/index.html
130. R Core Team (2024) Package 'parallel'. https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf
131. Tsagris M, Papadakis M (2018) Taking R to its limits: 70+ tips. https://doi.org/10.7287/peerj.preprints.26605v1
132. Tsagris M, Papadakis M (2021) Forward regression in R: from the extreme slow to the extreme fast. Data Sci 16:771–780. https://doi.org/10.6339/JDS.201810_16(4).00006
133. Chatzipantsiou C, Dimitriadis M, Papadakis M, Tsagris M (2018) Extremely efficient permutation and bootstrap hypothesis tests using R
134. Papadakis M, Tsagris M, Dimitriadis M, Fafalios S, Fasiolo M, Jacob M, Borboudakis G, Burkardt J, Zou C (2023) Rfast: a Collection of Efficient and Extremely Fast R Functions. https://doi.org/10.32614/CRAN.package.Rfast. https://cran.r-project.org/web/packages/Rfast/index.html
135. Tsagris M, Papadakis M, Alenazi A, Alzeley O (2024) Computationally efficient outlier detection for high-dimensional data using the mdp algorithm. Computation 12(9):1–10. https://doi.org/10.3390/computation12090185
136. Tsagris M, Papadakis M (2025) Fast and light-weight energy statistics using the R package Rfast. https://doi.org/10.48550/arXiv.2501.02849
137. Sussman D, Qiao Z, Agterberg J, Wang L, Lyzinski V (2024) iGraphMatch: tools for Graph Matching. https://doi.org/10.32614/CRAN.package.iGraphMatch. https://cran.r-project.org/web/packages/iGraphMatch/index.html
138. Dong X, Castro L, Shaikh N (2020) fastnet: an R package for fast simulation and analysis of large-scale social networks. J Stat Softw 96(7):1–23. https://doi.org/10.18637/jss.v096.i07
139. Shaikh N, Dong X, Castro L, Llano C (2020) fastnet: large-Scale Social Network Analysis. https://doi.org/10.32614/CRAN.package.fastnet. https://cran.r-project.org/web/packages/fastnet/index.html
140. Böttcher B (2019) Dependence and dependence structures: estimation and visualization using the unifying concept of distance multivariance. Open J Stat 1:1–46. https://doi.org/10.1515/stat-2020-0001
141. Böttcher B (2020) Dependence and dependence structures: estimation and visualization using the unifying concept of distance multivariance. Open J Stat 1(1):1–48. https://doi.org/10.1515/stat-2020-0001
142. Berschneider G, Böttcher B (2019) On complex Gaussian random fields, Gaussian quadratic forms and sample distance multivariance. https://doi.org/10.48550/arXiv.1808.07280. arXiv:1808.07280
143. Böttcher B, Keller-Ressel M, Schilling R (2019) Distance multivariance: new dependence measures for random vectors. Ann Stat 47:2757–2789. https://doi.org/10.1214/18-AOS1764
144. Böttcher B (2020) Copula versions of distance multivariance and dhsic via the distributional transform – a general approach to construct invariant dependence measures. Statistics 54(3):577–594. https://doi.org/10.1080/02331888.2020.1748029
145. Böttcher B, Keller-Ressel M (2021) multivariance: measuring Multivariate Dependence Using Distance. https://doi.org/10.32614/CRAN.package.multivariance. https://cran.r-project.org/web/packages/multivariance/index.html
146. Meyer D, Buchta C (2022) proxy: distance and Similarity Measures. https://doi.org/10.32614/CRAN.package.proxy. https://cran.r-project.org/web/packages/proxy/index.html
147. Bittinger K (2020) abdiv: alpha and Beta Diversity Measures. https://doi.org/10.32614/CRAN.package.abdiv. https://cran.r-project.org/web/packages/abdiv/index.html
148. Wickham H (2023) plyr: tools for Splitting, Applying and Combining Data. https://doi.org/10.32614/CRAN.package.plyr. https://cran.r-project.org/web/packages/plyr/index.html
149. Bates D, Maechler M, Jagan M Davis TA, Karypis G, Riedy J, Oehlschlägel J, R Core Team ROR ID (2023) Matrix: sparse and Dense Matrix Classes and Methods and Methods. https://doi.org/10.32614/CRAN.package.Matrix. https://cran.r-project.org/web/packages/Matrix/index.html
150. Wickham H, Posit Software, PBC (2023) stringr: simple, Consistent Wrappers for Common String Operations. https://doi.org/10.32614/CRAN.package.stringr. https://cran.r-project.org/web/packages/stringr/index.html
151. Pedersen TL (2024) patchwork: the Composer of Plots. https://doi.org/10.32614/CRAN.package.patchwork. https://cran.r-project.org/web/packages/patchwork/index.html
152. Wickham H, Pedersen TL, Seidel D, Posit, PBC (2023) scales: scale Functions for Visualization. https://doi.org/10.32614/CRAN.package.scales. https://cran.r-project.org/web/packages/scales/index.html
153. Zeileis A, Grothendieck G (2005) zoo: S3 infrastructure for regular and irregular time series. J Stat Softw 14(6):1–27. https://doi.org/10.18637/jss.v014.i06
154. Zeilei A, Grothendieck G, Ryan JA, Ulrich JM, Andrews F (2025) zoo: s3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations). https://doi.org/10.32614/CRAN.package.zoo. https://cran.r-project.org/web/packages/zoo/index.html
155. Kassambara A (2023) ggpubr: 'ggplot2' Based Publication Ready Plots. https://doi.org/10.32614/CRAN.package.ggpubr. https://cran.r-project.org/web/packages/ggpubr/index.html
156. Wickham H, Henry L, Posit Software, PBC ROR ID (2025) purrr: functional Programming Tools. https://doi.org/10.32614/CRAN.package.purrr. https://cran.r-project.org/web/packages/purrr/index.html
157. Wickham H, RStudio (2023) tidyverse: easily Install and Load the 'Tidyverse'. https://doi.org/10.32614/CRAN.package.tidyverse. https://cran.r-project.org/web/packages/tidyverse/index.html

158. Kassambara A (2023) rstatix: pipe-Friendly Framework for Basic Statistical Tests. https://doi.org/10.32614/CRAN.package.rstatix. https://cran.r-project.org/web/packages/rstatix/index.html
159. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A, Lowe W Müller C, Delmarcelle O, European Research Council (2025) quanteda: quantitative Analysis of Textual Data. https://doi.org/10.32614/CRAN.package.quanteda. https://cran.r-project.org/web/packages/quanteda/index.html
160. Fox J, Weisberg S, Price B, Adler D, Bates D, Baud-Bovy G, Bolker B, Ellison S, Firth D, Friendly M, Gorjanc G, Graves S, Heiberger R, Krivitsky P, Laboissiere R, Maechler M, Monette G, Murdoch D, Nilsson H, Ogle D, Ripley B, Short T, Venables W, Walker S, Winsemius D, Zeileis A, R-Core (2024). car: companion to Applied Regression. https://doi.org/10.32614/CRAN.package.car. https://cran.r-project.org/web/packages/car/index.html
161. Fox J, Weisberg S (2019) An R companion to applied regression, 3rd edn. Sage, Thousand Oaks. https://www.john-fox.ca/Companion/
162. Tuszynski J, Dietze M (2024) caTools: tools: moving Window Statistics, GIF, Base64, ROC AUC, etc. https://doi.org/10.32614/CRAN.package.caTools. https://cran.r-project.org/web/packages/caTools/index.html
163. Schloerke B, Cook LJ, Briatte F, Marbach M, Thoen E, Elberg A, Toomet O, Crowley J, Hofmann H, Wickham H (2024) GGally: extension to 'ggplot2'. https://doi.org/10.32614/CRAN.package.GGally. https://cran.r-project.org/web/packages/GGally/index.html
164. Friedman J, Hastie T, Tibshirani B, Narasimhan B, Tay K, Simon N, Qian J, Yang J (2023) glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. https://doi.org/10.32614/CRAN.package.glmnet. https://cran.r-project.org/web/packages/glmnet/index.html
165. Friedman J, Tibshirani R, Hastie T (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22. https://doi.org/10.18637/jss.v033.i01
166. Simon N, Friedman J, Tibshirani R, Hastie T (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 39(5):1–13. https://doi.org/10.18637/jss.v039.i05
167. Tay JK, Narasimhan B, Hastie T (2023) Elastic net regularization paths for all generalized linear models. J Stat Softw 106(1):1–31. https://doi.org/10.18637/jss.v106.i01
168. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D (2025) MASS: support Functions and Datasets for Venables and Ripley's MASS. https://doi.org/10.32614/CRAN.package.MASS. https://cran.r-project.org/web/packages/MASS/index.html
169. Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. ISBN 0-387-95457-0. https://www.stats.ox.ac.uk/pub/MASS4/
170. Grothendieck G (2017) sqldf: manipulate R Data Frames Using SQL. https://doi.org/10.32614/CRAN.package.sqldf. https://cran.r-project.org/web/packages/sqldf/index.html
171. Lüdecke D, Giné-Vázquez I, Bartel A (2024) sjmisc: data and Variable Transformation Functions. https://doi.org/10.32614/CRAN.package.sjmisc. https://cran.r-project.org/web/packages/sjmisc/index.html
172. Lüdecke D (2018) sjmisc: data and variable transformation functions. J Soc Struct 3(26):754. https://doi.org/10.21105/joss.00754
173. Wickham H (2022) reshape2: flexibly Reshape Data: a Reboot of the Reshape Package. https://doi.org/10.32614/CRAN.package.reshape2. https://cran.r-project.org/web/packages/reshape2/index.html
174. Husson F, Josse J, Le S, Mazet J (2024) FactoMineR: multivariate Exploratory Data Analysis and Data Mining. https://doi.org/10.32614/CRAN.package.FactoMineR. https://cran.r-project.org/web/packages/FactoMineR/index.html
175. Lê S, Josse J, Husson F (2008) FactoMineR: a package for multivariate analysis. J Stat Softw 25(1):1–18. https://doi.org/10.18637/jss.v025.i01
176. Kassambara A, Mundt F (2020) factoextra: extract and Visualize the Results of Multivariate Data Analyses. https://doi.org/10.32614/CRAN.package.factoextra. https://cran.r-project.org/web/packages/factoextra/index.html
177. Constantine W, Hesterberg T Wittkowski K, Tingting S, Dunlap B, Kaluzny S (2024) splus2R: supplemental S-PLUS Functionality in R. https://doi.org/10.32614/CRAN.package.splus2R. https://cran.r-project.org/web/packages/splus2R/index.html
178. Stoffer D, Poison N (2025) astsa: applied Statistical Time Series Analysis. https://doi.org/10.32614/CRAN.package.astsa. https://cran.r-project.org/web/packages/astsa/index.html
179. Shumway RH, Stoffer DS (2005) Time series analysis and its applications (Springer texts in statistics). Springer, Berlin
180. Shumway R, Stoffer D (2019) Time series: a data analysis approach using R. CRC Press. Taylor & Francis Group, Boca Raton
181. Bache SM, Wickham H, Henry L, RStudio (2022) magrittr: a Forward-Pipe Operator for R. https://doi.org/10.32614/CRAN.package.magrittr. https://cran.r-project.org/web/packages/magrittr/index.html
182. Spinu V, Grolemund G, Wickham H, Vaughan D, Lyttle I, Costigan I, Law J, Mitarotonda D, Larmarange J, Boiser J, Lee CH (2024) lubridate: make Dealing with Dates a Little Easier. https://doi.org/10.32614/CRAN.package.lubridate. https://cran.r-project.org/web/packages/lubridate/index.html
183. Grolemund G, Wickham H (2011) Dates and times made easy with lubridate. J Stat Softw 40(3):1–25
184. Trapletti A, Hornik K, LeBaron B (2024) tseries: time Series Analysis and Computational Finance. https://doi.org/10.32614/CRAN.package.tseries. https://cran.r-project.org/web/packages/tseries/index.html
185. Carslaw D, Davison J, Ropkins K (2024) openair: tools for the Analysis of Air Pollution Data. https://doi.org/10.32614/CRAN.package.openair. https://cran.r-project.org/web/packages/openair/index.html
186. Carslaw CD, Ropkins K (2012) openair — an R package for air quality data analysis. Environ Model Softw 27–28:52–61. https://doi.org/10.1016/j.envsoft.2011.09.008
187. Ooms J (2014) The jsonlite Package: a Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805
188. Ooms J, Duncan Temple Lang, Hilaiel L (2025) jsonlite: a Simple and Robust JSON Parser and Generator for R. https://doi.org/10.32614/CRAN.package.jsonlite. https://cran.r-project.org/web/packages/jsonlite/index.html
189. Feinerer I, Hornik K, Artifex Software, Inc. (2025) tm: text Mining Package. https://doi.org/10.32614/CRAN.package.tm. https://cran.r-project.org/web/packages/tm/index.html
190. Feinerer I, Hornik K (2025) Tm: text Mining Package. R package version 0.7-16. https://CRAN.R-project.org/package=tm

191. Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in R. J Stat Softw 25(5):1–54. https://doi.org/10.18637/jss.v025.i05
192. Tennekes M, Nowosad J, Gombin J, Jeworutzki S, Russell K, Zijdeman R, Clouse J, Lovelace R, Muenchow J, Roy O, Pebesma E, Graham H, Sumner MD, Appelhans T, Bearman N (2025) tmap: thematic Maps. https://doi.org/10.32614/CRAN.package.tmap. https://cran.r-project.org/web/packages/tmap/index.html
193. Hornik K (2024) NLP: natural Language Processing Infrastructure. https://doi.org/10.32614/CRAN.package.NLP. https://cran.r-project.org/web/packages/NLP/index.html
194. Hornik K (2019) openNLP: apache OpenNLP Tools Interface. https://doi.org/10.32614/CRAN.package.openNLP. https://cran.r-project.org/web/packages/openNLP/index.html
195. Warnes GR, Bolker B, Lumley T, Magnusson A, Venables B, Ryodan G, Moeller S, Wilson I, Davis M, Jain N, Chamberlain S (2023) gtools: various R Programming Tools. https://doi.org/10.32614/CRAN.package.gtools. https://cran.r-project.org/web/packages/gtools/index.html
196. Opsahl T (2020) tnet: weighted, Two-Mode, and Longitudinal Networks Analysis. https://doi.org/10.32614/CRAN.package.tnet. https://cran.r-project.org/web/packages/tnet/index.html
197. Opsahl T (2009) Structure and evolution of weighted networks. University of London (Queen Mary College), London, pp 104–122. http://toreopsahl.com/publications/thesis/
198. Wickham H, Hester J, Ooms J, Posit Software, PBC, R Foundation (2025) xml2: parse XML. https://doi.org/10.32614/CRAN.package.xml2. https://cran.r-project.org/web/packages/xml2/index.html
199. Bouchet-Valat M (2023) nowballC: snowball Stemmers Based on the C 'libstemmer' UTF-8 Library. https://doi.org/10.32614/CRAN.package.SnowballC. https://cran.r-project.org/web/packages/SnowballC/index.html
200. R Core Team and contributors (2024) The R Base Package
201. Kim T, Wurster K (w.d.) emoji. https://pypi.org/project/emoji/
202. Solomon B (w.d.) demoji. https://pypi.org/project/demoji/
203. Savand A (w.d.) stop-words. https://pypi.org/project/stop-words/
204. Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., Sebastopol
205. NLTK Project (w.d.) https://www.nltk.org/
206. Hagberg A, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using network. Technical report, Los Alamos National Lab. (LANL), Los Alamos, NM (United States)
207. Aric A, Hagberg DA, Schult PJ (2008) Exploring network structure, dynamics, and function using NetworkX". In: Vaught T, Millman J (eds) Proceedings of the 7th Python in science conference (SciPy2008), Gäel Varoquaux. SciPy, Pasadena, pp 11–15
208. Python Software Foundation pickle — Python object serialization. https://docs.python.org/3/library/pickle.html
209. Guido VR (2020) The Python Library Reference, Release 3.8.2. Python Software Foundation, Beaverton, OR USA
210. Python Software Foundation (w.d.) copy — Shallow and deep copy operations¶. https://docs.python.org/3/library/copy.html
211. Python Software Foundation (w.d.) os — Miscellaneous operating system interfaces. https://docs.python.org/3.9/library/os.html
212. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Fernández del Río J, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2
213. Hunter JD (2007) Matplotlib: a 2d graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55
214. Python Software Foundation itertools — Functions creating iterators for efficient looping. https://docs.python.org/3/library/itertools.html
215. Python Software Foundation csv — CSV File Reading and Writing. https://docs.python.org/3/library/csv.html
216. Pandas Development Team (w.d.) pandas: powerful Python data analysis toolkit. https://pypi.org/project/pandas/
217. Pandas Development Team (2020) pandas-dev/pandas: pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134
218. Wes MK (2010) Data structures for statistical computing in Python. In: van der Walt S, Millman J (eds) Proceedings of the 9th Python in science conference, pp 56–61. https://doi.org/10.25080/Majora-92bf1922-00a
219. Python Software Foundation (w.d.) warnings — Warning control. https://docs.python.org/3/library/warnings.html
220. Python Software Foundation (w.d.) operator — Standard operators as functions. https://docs.python.org/3/library/operator.html
221. Python Software Foundation (w.d.) sys — System-specific parameters and functions. https://docs.python.org/3/library/sys.html
222. Barnett M (w.d.) regex. https://pypi.org/project/regex/
223. Python Software Foundation (w.d.) glob — Unix style pathname pattern expansion. https://docs.python.org/3/library/glob.html
224. Python Software Foundation (w.d.) random — Generate pseudo-random numbers. https://docs.python.org/3/library/random.html
225. Dask Core Developers (w.d.) dask. https://pypi.org/project/dask/
226. Rossetti G, Team¶ CD (w.d.) CDlib - Community Detection Library . https://cdlib.readthedocs.io/en/latest/
227. Rossetti G, Milli L, Cazabet R (2019) Cdlib: a Python library to extract, compare and evaluate communities from complex networks. Appl Netw Sci 4:1–27. https://doi.org/10.1007/s41109-019-0165-9
228. Python Software Foundation (w.d.) collections — Container datatypes. https://docs.python.org/3/library/collections.html
229. Waskom ML (2021) seaborn: statistical data visualization. J Soc Struct 6(60):3021. https://doi.org/10.21105/joss.03021
230. Waskom ML (w.d.) seaborn: statistical data visualization. https://seaborn.pydata.org/
231. Python Software Foundation (w.d.) subprocess — Subprocess management. https://docs.python.org/3/library/subprocess.html
232. Python Software Foundation (w.d.) re — Regular expression operations. https://docs.python.org/3/library/re.html
233. West Health Institute (w.d.) Interactive network visualizations. https://pyvis.readthedocs.io/en/latest/