Beyond Climate Change? Environmental Discourse on the Planetary Boundaries in Twitter Networks

Materials uploaded to Open Science Framework
Link to project page

File/folder name	Description
Preregistration_Green Bubbles.pdf	Study preregistration
UpdatedPreregistration.pdf	Changes and deviations from initial preregistered plan
Tweet_ids.csv	Raw data file containing IDs for tweets analysed in this study (in compliance with institutional data sharing and ethical guidelines)
Analysis script.rtf	Abridged and commented analysis script in R
Top_200_terms_per_community	Word clouds containing top 200 terms per community – resulting from topic modelling analysis on retrieved tweets
Green Communities _ Description Word Clouds	Word clouds containing the most frequently occurring terms in profile description of tweet authors
Topics.csv	Top 100 terms per topic

Supplementary materials

File name	Description
SupplementaryMaterials_1.pdf	• Appendices A-E • Detailed Methods • Determination of number of topics in topic analysis • Performance metrics of search terms used in the pilot study • Figures Accessible to people with colour-blindness • Additional figures • Additional robustness analysis – topic modelling
SupplementaryMaterials_2.xlsx	Term-frequency matrix

Supplementary Materials

Table of Contents

Beyond Climate Change? Environmental Discourse on the Planetary Boundaries in Twitter Networks

Appendices

Appendix A

Table A1

Characteristics of Green Communities

Community	Proportion of users in total network	Proportion of authors (number)
Climate change and hope	53.17%	69.3% (*n* = 396014)
Climate change and action	13.37%	18.3% (*n* = 104575)
Climate change and knowledge	8.6%	5.6% (*n* = 32001)
Climate change and science	7.45%	2.2% (*n* = 12572)
Climate change and sustainability	4.36%	1.7% (*n* = 9715)
Climate change and health	1.92%	1.2% (*n* = 6858)
Climate change and energy	1.27%	0.8% (*n* = 4572)
Climate change and solutions	1.20%	0.5% (*n* = 2857)
Climate change and society	1.05%	0.3% (*n* = 1714)

Note. Proportion of users in total network refers to the percentage of users (including authors, followers, and friends) that are within a community. Proportion of authors refers to the percentage of users within a community who posted the environment-related tweets within the network.

Appendix B

Table B1

Topic Model for Twitter Discourse About the Environment per Community

K	Topic Label	Top 5 Keywords
Climate change and hope community		
11	Justice	just, justice, public, part, history
10	Science	science, research, time, scientists, study
13	Lifestyle	organic, work, vegan, green, cost
16	Impact	impact, clean, epaa, economic, years
8	Future	future, time, back, new, years
Climate change and action community		
5	Renewables	renewable, fossil, power, zero, electricity
6	Finance	money, tax, justice, future, net
12	Nuclear	nuclear, clean, power, renewable, electricity
11	Justice	just, justice, public, part, history
14	Research	research, impact, latest, science, reports
Climate change and knowledge		
17	Industry	industry, infrastructure, tax, #esgb, production
18	Politics	biden, support, #investinourplanet, community, resources
3	Carbon	carbon, pollution, transportation, greenhouse, deforestation
19	Earth Day	#earthday, day, earth, planet, protect
14	Research	research, impact, latest, science, reports
Climate change and science community		
10	Science	science, research, time, scientists, study
18	Politics	biden, support, #investinourplanet, community, resources
11	Justice	just, justice, public, part, history
8	Future	future, time, back, new, years
3	Carbon	carbon, pollution, transportation, greenhouse, deforestation
Climate change and sustainability community		
15	Sustainability	sustainability, sustainable, #sustainability, esg, strategy
18	Politics	biden, support, #investinourplanet, community, resources
7	Energy	energy, solar, clean, green, pollution
20	Innovation	technology, future, solutions, efforts, building

11	Justice	just, justice, public, part, history
Climate change and health community		
4	Health	health, science, opportunity, support, safety
9	Nature	forest, green, plants, air, city, oxygen
13	Lifestyle	organic, work, vegan, green, cost
18	Politics	biden, support, #investinourplanet, community, resources
20	Innovation	technology, future, solutions, efforts, building
Climate change and energy community		
7	Energy	energy, solar, clean, green, pollution
20	Innovation	technology, future, solutions, efforts, building
5	Renewables	renewable, fossil, power, zero, electricity
6	Finance	money, tax, justice, future, net
12	Nuclear	nuclear, clean, power, renewable, electricity
Climate change and solutions community		
17	Industry	industry, infrastructure, tax, #esg, production
20	Innovation	technology, future, solutions, efforts, building
13	Lifestyle	organic, work, vegan, green, cost
1	Infrastructure	global, make, infrastructure, government, build
12	Nuclear	nuclear, clean, power, renewable, electricity
Climate change and society community		
18	Politics	biden, support, #investinourplanet, community, resources
9	Nature	forest, green, plants, air, city, oxygen
4	Health	health, science, opportunity, support, safety
19	Earth Day	#earthday, day, earth, planet, protect
3	Carbon	carbon, pollution, transportation, greenhouse, deforestation

Note. The topic labels are manually assigned based on the relevant terms. Since there is high overlap in the most frequently occurring terms in the corpus, relevant keywords reflecting the assigned topic labels are presented. A complete topic model results table with 100 terms per topic is available in supplementary materials.

a United States Environmental Protection Agency

b Environmental, Social, and Governance

Appendix C

Figure C1

Word clouds of environmental tweets in green communities

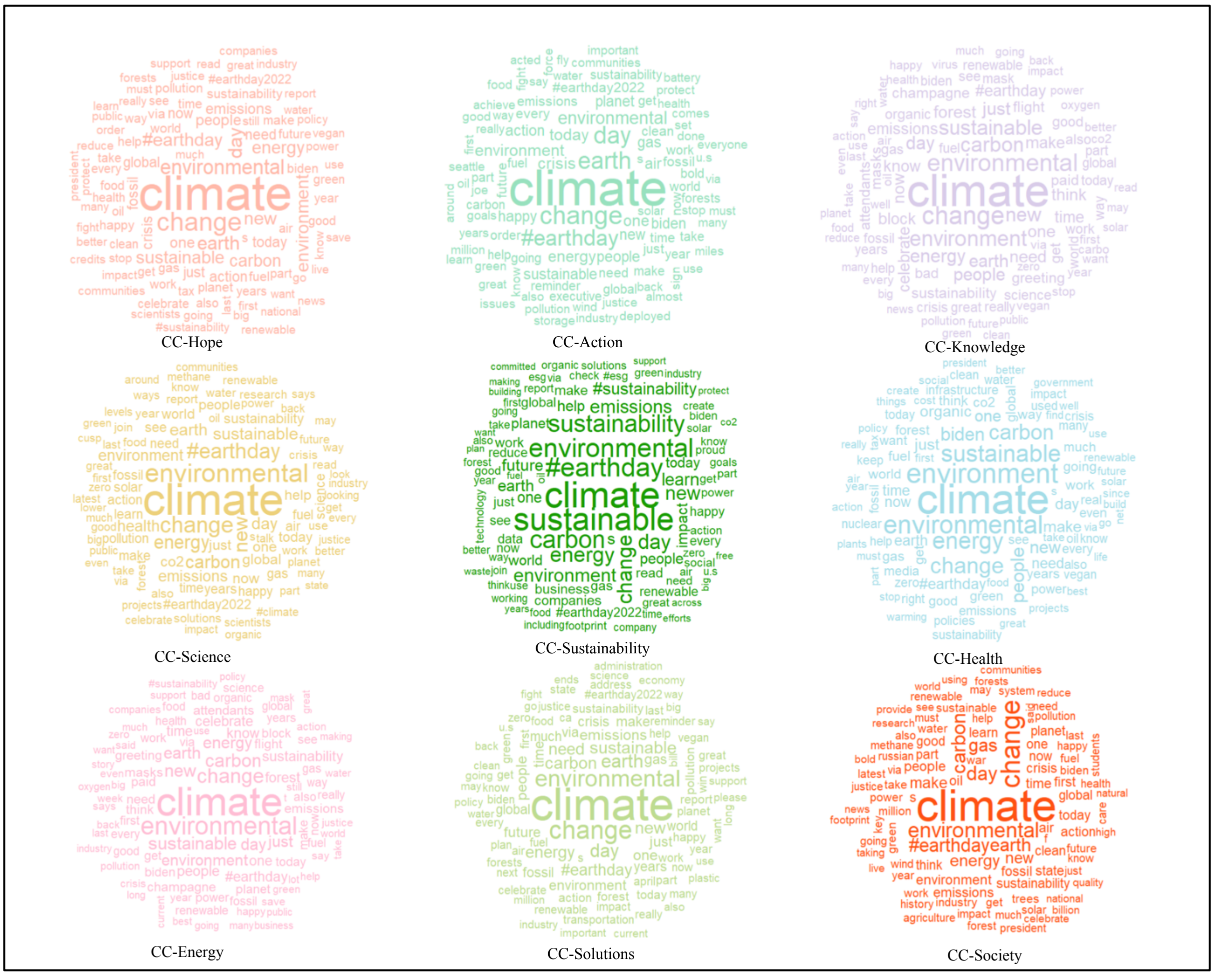


Figure C2

Word clouds of user descriptions in green communities

Appendix D

Example tweets from distinct Green Communities

Climate change and hope

“There's no sugar-coating it: The climate crisis is here and our planet is in trouble. But it's not too late to act. This Earth Day let's recommit to pulling out all the stops to protect our planet from this existential threat. Nothing short of a policy revolution will do.”

Climate change and action

“As Australia battles wild weather and coastal erosion, we should learn from our mistakes - a clear call for action and direction on coastal management and climate change from @username who is seeing policy failure first hand at Collaroy Beach <external link>”

Climate change and knowledge

“Celebrating #EarthDay tomorrow? Find climate change focused resources from @username in this collection on PBS LearningMedia. Find video resources, interactives, teaching guides and more! <external link>”

Climate change and science

as shown by the following tweet – “A letter signed by 135 scientists, including founding dean and professor emeritus Norman Christensen, called on Biden to protect mature and old-growth forests as a critical climate solution.”

Climate change and sustainability

“This Earth Week, investors must consider climate, and beyond that, ecological impacts of companies, to truly capture all the risks and upside present in their holdings. #ESG #EarthWeek #climate <external link>”.

Climate change and health

“Health is more than just our genetics. Environmental factors, like water quality, can affect our overall health too. #JoinAllofUs on 4/21 to learn more about how factors like this may impact health and how @username is helping find answers. More at <external link>”.

Climate change and energy

“@username The biggest issue is the people who continue to block nuclear plants which is by far the most efficient energy generator by land use and for the environment.”

Climate change and solutions

“This #EarthWeek, our new analysis takes a deep dive into an important environmental issue: plastic. It may surprise people how little plastic is recycled, and how much of it ultimately goes into the landfill, said Daniel Coffee, LCI project manager. <external link>.”

Climate change and society

“The U.S gives the fossil fuel industry $20,000,000,000 in subsidies every year and nobody blinks an eye. We can't sit back and accept this corporate welfare as business as usual. We have the power to change the system.”

Appendix E

Interactions among green communities

We found that the users in CC-Hope community followed those in CC-Health community and vice-versa. The CC-Health community also mutually followed users in CC-Solutions community. Other communities that mutually followed each other to a large extent were CC-Action and CC-Science, CC-Science and CC-Knowledge, and CC-Knowledge and CC-Sustainability. The CC-Science community was generally followed by all other communities and followed the CC-Hope and CC-Society communities back.

The CC-Hope community was followed by users in CC-Energy community, but the interaction was largely one-sided as users in CC-Energy community did not follow back the users in CC-Hope community. Users in CC-Solutions community followed those in CC-Knowledge community and were widely followed by users in CC-Action community. Users in CC-Society community were followed by those in CC-Knowledge and followed back users in CC-Solutions community. Finally, we found little evidence for interaction of CC-Sustainability community with CC-Solutions community, CC-Society community with CC-Solutions, CC-Action, and CC-Health communities.

1. Detailed methods

Data Collection

We retrieved global English language tweets with the full archive search on Twitter Application Programming Interface (API v2)[1] implemented using the academictwitteR package in R (Barrie et al., 2022). The package allows for collecting tweets based on a full archive search and makes it possible to omit retrieval of tweets based on certain criteria such as whether a tweet is promoted (i.e., advertised) or not. We sampled tweets posted between January 2019 and April 2022 based on a predetermined list of search terms. Advertisements and tweets containing media were excluded.

Selecting the tweets

As the selection of search terms can impact further textual analysis of the retrieved data (Mahl et al., 2022; Stryker et al., 2006), we measured the efficiency of the search terms as recommended in previous studies (Lacy et al., 2015; Stryker et al., 2006). This was done in two steps.

First, we created and preregistered 58 open search terms based on the nine planetary boundaries framework (Rockström et al., 2009), previous literature (Cameletti et al., 2022.; Gaytan Camarillo et al., 2021; Pilař et al., 2019), and researcher consensus (all authors). As the Twitter API does not allow conventional regex, several but not all expansions of the pre-registered search terms were utilised (for example, "sustainable" and "sustainability"). A pilot dataset (*n* = 5000) containing tweets, retweets, and quote tweets retrieved with this list was used to evaluate the efficiency search terms by measuring their recall (Stryker et al., 2006). Since it is not possible to estimate the total number of relevant documents (i.e., number of tweets containing the search terms), we calculated recall as the proportion of tweets retrieved

[1] Since the sale of Twitter in October 2022, the access to Twitter API for academic research is no longer available free of cost.

by each search term in our sample. This means that search terms that retrieved more tweets in the sample had a greater recall whereas search terms that retrieved relatively less number of tweets had a lower recall.

Additionally, for a subset of the data (*n* = 500), we manually calculated the precision of search terms. A relevant tweet referred to a tweet related to the environment, i.e., it mentioned an environment-related topic associated with the natural world. In doing so, we observed that some search terms with a high recall showed little precision, retrieving many irrelevant tweets. For instance, the search term "toxic" retrieved the second highest number of tweets (14.6%, 73 of 500 tweets) but only 3 of these tweets were relevant (precision = 4%). Hence, we deviated from our preregistered selection criteria for the search terms. To ensure a relevant collection of tweets, we removed search terms with precision values less than .5 instead of removing those with recall values lower than .6. The open search terms and their performance metrics are provided in supplementary materials.

Moreover, 11 search terms retrieved no tweets but we decided to keep them in our list of closed search terms because absence of tweets relating to these terms (e.g., "planetary boundaries", "circular economy", etc.) may indicate that although relevant in academic literature, some topics are yet to become mainstream. Finally, we shortlisted 50 closed search terms for collecting tweets for the main dataset (*n* = 2,000,000). These included original tweets, retweets, replies, and comments, but not promotional tweets and advertisements.

Retrieving the Twitter network

We were interested in examining the environmental discourse within networks of users as the content of a user's Twitter feed was largely based on who they followed (at the time of data collection; it is arguably different now). Given this objective of analysing environmental discourse of networks of people who do not exclusively post about the environment (i.e., the general public), we decided to create an egocentric network comprising of followers and friends of the authors of the environment-related tweets. This also allowed us to avoid the risk of including only those users who mostly post about environmental issues, which can happen when networks are created based on user replies and/or comments to environmental tweets.

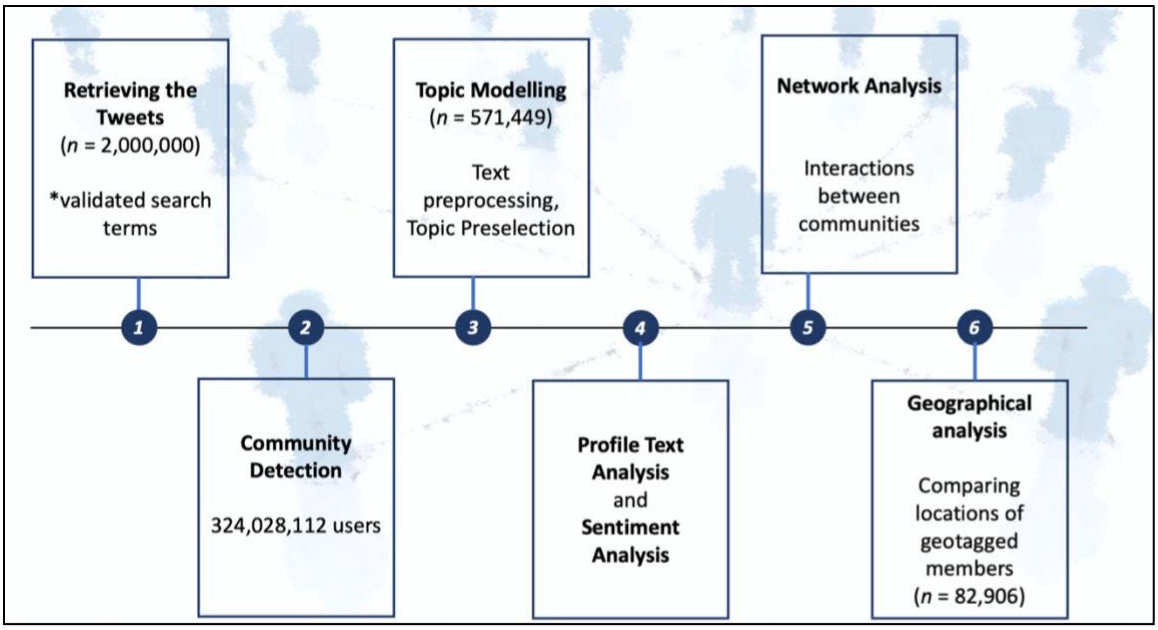
To construct the egocentric user network for community detection, we retrieved the connections of the authors of the collected tweets. A total of 1,133,748 unique users were identified as authors in our sample of 2 million tweets. We excluded private users because while their profile information and tweets cannot be accessed, user IDs for all replies and comments were provided in the main dataset comprising original tweets, replies, and comments by default. We compiled all the authors (author IDs) from the main dataset and used the relevant endpoint to retrieve information for these author IDs. It is at this stage that we excluded private users because the API returned no results for these author IDs. So, private profiles were not included in our analyses. After removing protected users with private profiles (n = 229,705), the number of authors in the network was 904,043. The total number of followers (users who follow the authors) was 6,053,998,317 whereas the total number of friends (users followed by the authors) was 1,181,490,733. The average number of followers was 6,723 whereas the average number of friends was 1,312. As per API limitations, we retrieved a maximum of 1 million followers for users with more than a million followers (n = 187) and restricted the maximum number of friends to 5000. Thus, for 904,043 authors, we retrieved 532,297,774 followers and 131,073,118 friends. Following the approach by Lutkenhaus et al. (2019), we examined the distribution of the number of connections for the authors to determine a cut-off point to exclude users (friends and followers) who were not connected to a minimum of 5 authors, so that the retrieved network would reflect shared interests and audiences. The overall network contained of 324,028,112 users comprising total number of authors, followers, friends in an egocentric network with 367,465,523 connections.

Data Analysis

We examined the first research question in four steps. First, to answer which green communities exist on Twitter (**RQ1a**) we identified network clusters using the Louvain algorithm. Second, we examined the content of the tweets (**RQ1b**) with topic modelling. We compared the extracted topics across communities to distinguish environmental discourse in distinct green communities and created word clouds to describe the environmental content in these communities. Third, we analysed the profile texts of the users to describe the users in these communities (**RQ1c**). Fourth, we conducted a manual analysis on a subset of one hundred tweets per community to explore the underlying sentiments within each green community (**RQ1d**). To examine if distinct green communities interact with each other (**RQ2**), we conducted a network analysis to visualise the communication flows between green communities. Finally, to investigate where in the world these tweets originated **(RQ3)**, we compared the proportions of geo-tagged authors belonging to the Global North and Global South. Figure A depicts the data analysis steps.

Figure A

Analysis steps



Community Detection

Community detection was used to identify green communities (**RQ1a**), i.e., groups of users based on common connections in their Twitter network. We used the igraph package in R (Csardi & Nepusz, 2005) which applies the Louvain algorithm for community detection (Blondel et al., 2008). The Louvain algorithm tries to optimise modularity for each identified cluster by first assigning each node to its own cluster and then shuffling the nodes until modularity is maximised. The modularity measure indicates the relative density of edges within a cluster compared to a random network. The Louvain method is known as a relatively accurate and fast method for community detection in large networks (Mukerjee, 2021). In our retrieved Twitter network, users (authors, followers, and friends) are represented as nodes, while the connections between them are represented as edges. We retained communities that comprised at least 1% of users in the network.

Topic Modelling

We employed topic modelling to examine the environmental themes (topics) of tweets in green communities (**RQ1b**). For this we extracted the original tweets (n = 571, 449) written by authors, i.e., tweets that were not quote tweets and retweets. The tweets from each community were denoted as a separate document and the distribution of topics and terms was determined using the latent Dirichlet allocation (LDA) approach (Blei et al., 2003). For analysing the tweets, we used the topicmodels package (Grün & Hornik, 2011) in R to create a document feature matrix (DFM). To create the matrix, we pre-processed the tweets according to the preregistered steps as recommended in previous research (Maier et al., 2018). First, we removed the irrelevant tweet components such as URLs, mentions (@), and retweets (RT). Second, we divided the tweets into word units (tokenisation) and converted all letters to lower case. Third, special characters and punctuation marks were deleted. Fourth, we removed English stop words such as ‘the’, ‘is’, ‘are’, etc using the list included in the topicmodels package. Finally, in a deviation from our preregistered criteria, we applied only the lower limit of pruning to remove all the terms that occurred in less than 1% of all the tweets. This was decided upon observing that removing terms that occurred in more than 95% of the tweets resulted in elimination of highly relevant terms such as “climate”, “environment”, “sustainability”, etc. The resulting DFM comprised of 9 documents and 209,532 features.

To identify the suitable number of topics (K), we calculated the evaluation metrics for different topics by varying the different combinations for number of topics, i.e., K (20, 25, 30, 35, 40, and 45) and document-topic densities, i.e., α (.01, .05, .1, .2, .5, 1) while keeping the value of β fixed at $1/K$, using the ldatuning package (Nikita & Chaney, 2020). To ensure qualitative interpretability of our selected number of topics, we manually examined the top words and top tweets for each model for each model. Following this approach, we decided on $K = 20$ with $\alpha = .01$ for our topic model.

To understand the environmental discourse in each community, we examined the per-document-per-topic probabilities (gamma), since each document comprised of tweets from a green community identified via community detection. We created word clouds to show the most frequently occurring terms for each community with the wordcloud package in R (Fellows, 2018). SD and MM discussed the resulting topics and labelled them. All authors agreed on the descriptions of green communities. Specifically, the topics were labelled based on the most frequently occurring terms and the green communities were described based on the most prominent topics discussed by the users in these communities.

Profile text analysis

Profile description texts for each user in the network (authors, followers, and friends) were examined to understand which types of users constitute the green communities (**RQ1c**). The text pre-processing was conducted using the same pre-processing steps as for the topic modelling (Maier et al., 2018). The profile descriptions of user in each community were concatenated into one document per community and the most frequent terms in user profile texts were obtained by assigning TF-IDF weights using the tm package (Feinerer et al., 2008). Word clouds of most frequently occurring terms were also visualised for the profile descriptions of users in each green community using the wordcloud package in R (Fellows, 2018).

Sentiment analysis

To identify the underlying sentiment in the tweets of the different green communities (**RQ1d**), we manually examined 900 tweets (100 tweets per green community) containing the most likely terms and topics. That is, we identified the topic most strongly associated with a community and retrieved one hundred tweets that contained the most likely terms in that topic. Then these tweets were coded by the first author as positive, negative, or neutral based on their underlying sentiment. Specifically, tweets were coded as positive if they pertained to pro-environmental issues and action, negative if they dismissed pro-environmental issues and action, and neutral if they did not contain an explicit pro- or anti- environmental sentiment.

For example, the following tweet contains a positive sentiment - *"As Australia battles wild weather and coastal erosion, we should learn from our mistakes - a clear call for action and direction on coastal management and climate change from @username who is seeing policy failure first hand at Collaroy Beach <external link>* [sic]" whereas the tweet *"The $2.1 billion natural gas liquefaction plant will be built on roughly 250 acres of state-owned land. @UserName reports*." was considered neutral. Finally, the tweet – *""@UserName The climate naturally changes get over it already <external link>"* was coded to contain a negative sentiment.

Network analysis

To answer **RQ2**, we examined the mutual connections between distinct green communities to visualise interactions. Specifically, we quantified the number of users within each community that followed or friended users from other communities to show the extent to which users in a particular green community are likely to be exposed to users from another community, as described in Lutkenhaus et al. (2019).

Comparing geotagged members

Lastly, to investigate to what extent tweets in green communities originate from the Global North vs the Global South answer (**RQ3**), we looked at geo-tagged tweets. Specifically, we extracted the geographical metadata of authors using the Twitter API. In line with previous estimates (Huang & Carley, 2019; Karami et al., 2021), location of only 14.51% of all authors was tagged in our sample (*n* = 82,906). We found that 7,433 unique locations were retrieved for these authors in identified green communities, which were then labelled by the first author as countries belonging Global South and Global North using the Human Development Index categorisation[2].

The authors from these unique locations (countries, cities, towns, etc.) were categorised as belonging to either the Global North or Global South. We compared the proportion of authors from these two regions for each green community to understand which region contributes more to the online environmental discourse.

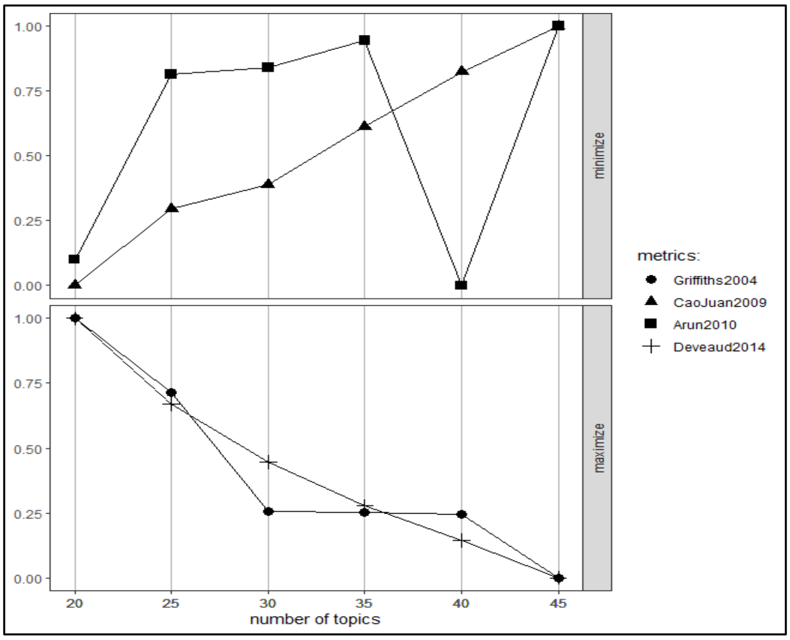
[2] https://worldpopulationreview.com/country-rankings/global-north-countries

2. Determining the number of topics in topic analysis

The plot in figure A shows that the suitable number of topics for our model is K = 20 as shown by the maximisation of Griffiths and Deveaud metrics and minimisation of Arun and Cao Juan metrics.

Figure A

Topic Fit Metrics for Varying Number of Topics (K) with α = .01



3. Performance metrics of the search terms used in the pilot study

Table SM 1.1

Performance metrics of the search terms used in the pilot study

Search term[a]	Number of tweets	Proportion	Precision
climate*	959	19.2%	99%
toxic	610	12.2%	4%
garbage	383	7.7%	3%
nature	363	7.3%	22%
environment*	359	7.2%	60%
fossil fuel	168	3.4%	100%
forest	161	3.2%	65%
vegan	148	3.0%	93%
emissions	146	2.9%	100%
carbon	146	2.9%	94%
innovation	102	2.0%	29%
sustainable*	93	1.9%	50%
organic	92	1.8%	70%
renewable*	85	1.7%	57%
global warming	53	1.1%	100%
pollution	49	1.0%	63%
renewable energy	46	0.9%	100%
vegetarian	44	0.9%	100%
drought	41	0.8%	33%
agriculture*	40	0.8%	100%

CO2	28	0.6%	0%
ecological*	22	0.4%	100%
plant based	21	0.4%	0%
greenhouse gas	21	0.4%	100%
methane	18	0.4%	0%
biodegredable	18	0.4%	0%
recycle*	15	0.3%	100%
net zero	15	0.3%	0%
reuse	14	0.3%	50%
biodiversity*	10	0.2%	100%
rainforest	8	0.2%	0%
pollutants	6	0.1%	0%
zero waste	5	0.1%	0%
sdg	4	0.08%	0%
smog	4	0.08%	100%
deforestation	4	0.08%	100%
palm oil	4	0.08%	0%
agw	3	0.06%	100%
food waste	3	0.06%	0%
solar energy	3	0.06%	0%
greenwashing	2	0.04%	0%
oil spill*	2	0.04%	0%
ozone layer	2	0.04%	0%
coral bleaching	2	0.04%	100%

ozone depletion	1	0.02%	0%
algal blooms	1	0.02%	0%
biosphere*	1	0.02%	0%
nitrogen fertiliser	1	0.02%	0%
planetary boundaries	0	0%	0%
circular economy	0	0%	0%
donut economy	0	0%	0%
plastic soup	0	0%	0%
ocean acidification	0	0%	0%
water stress	0	0%	0%
overfish*	0	0%	0%
plastic free	0	0%	0%
nitrogen fertilizer	0	0%	0%
freshwater use	0	0%	0%
water conservation	0	0%	0%

Note.
a. For search terms marked with an asterisk (*), alternative (but not all) word forms such as plurals and verb forms were used to retrieve tweets. This was done because the Twitter API does not allow for boolean expressions. For search terms with a recall of 0, alternative word forms also did not yield any results.
b. Since some search terms may occur simultaneously in a single tweet, there is some overlap in the calculation of percent occurrence values. Precision was calculated manually for a subset of 500 tweets, environmentally relevant tweets were determined to be accurately retrieved while those not related to the environment were determined to be inaccurately retrieved.

4. Figures Accessible to People with Colour-blindness

Figure 1

Network Graph Showing the Identified Green Communities

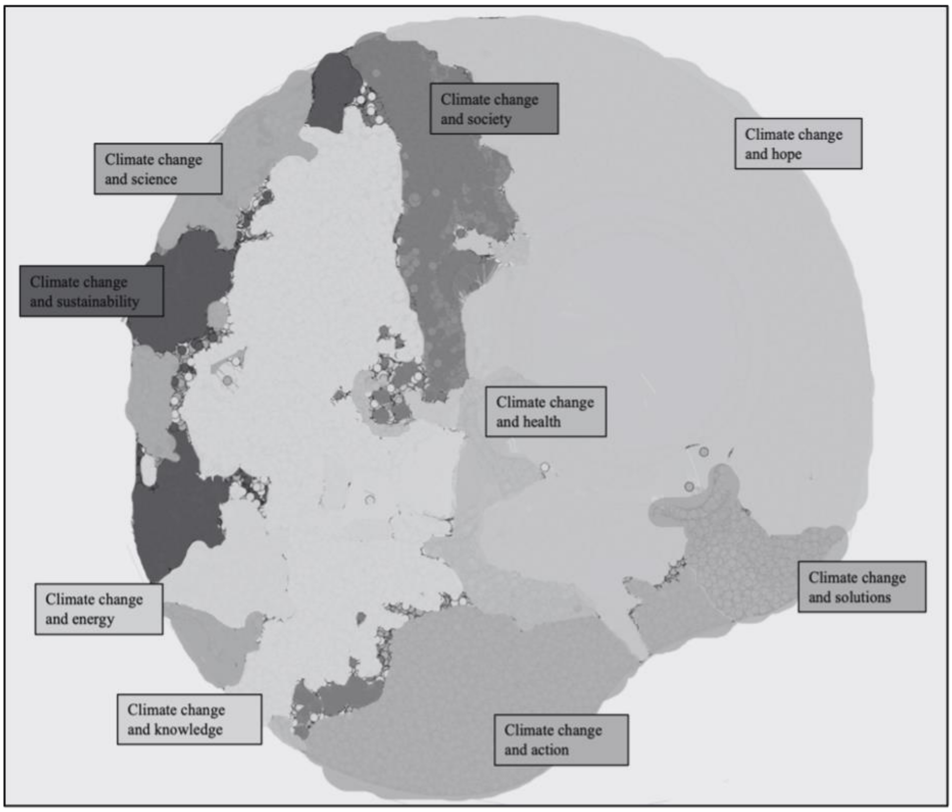


Figure 2

Topic Distribution for All Green Communities

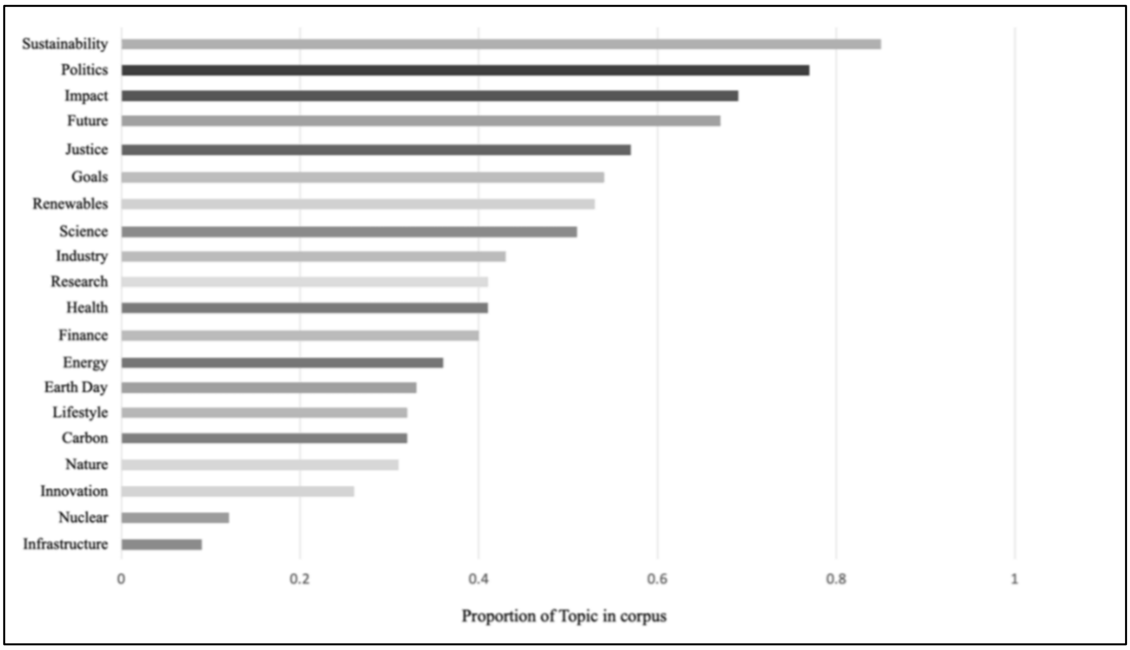
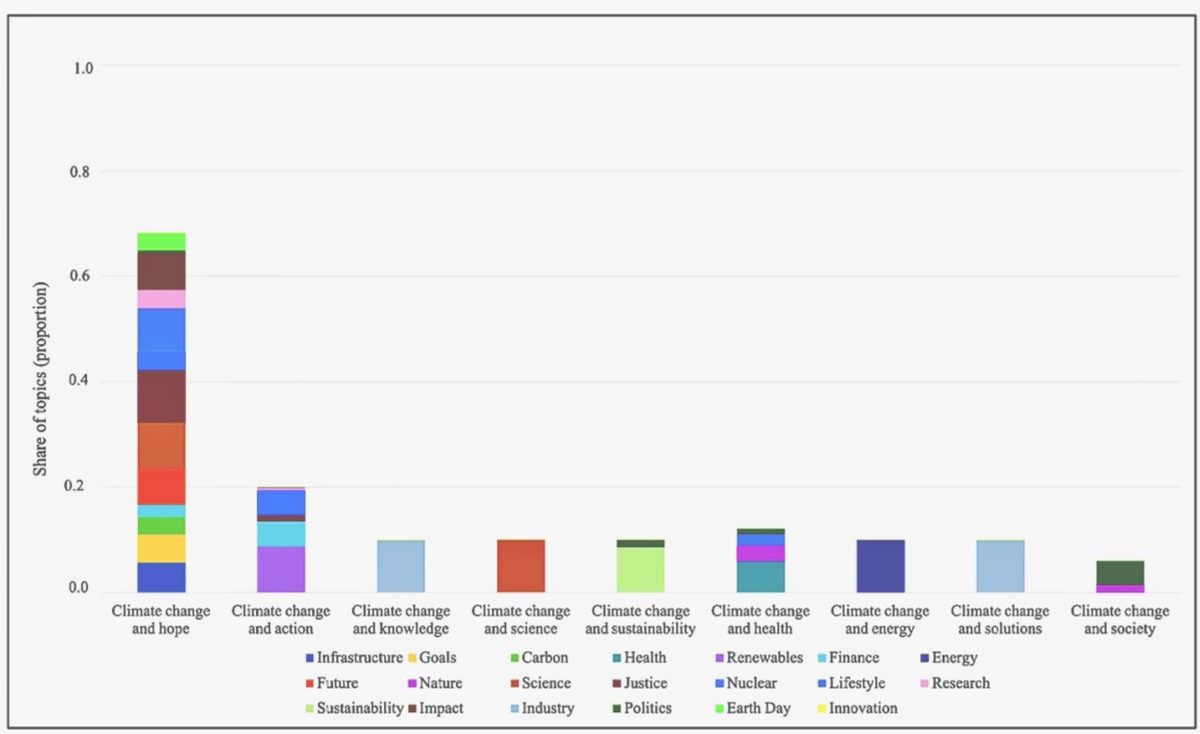


Figure 3

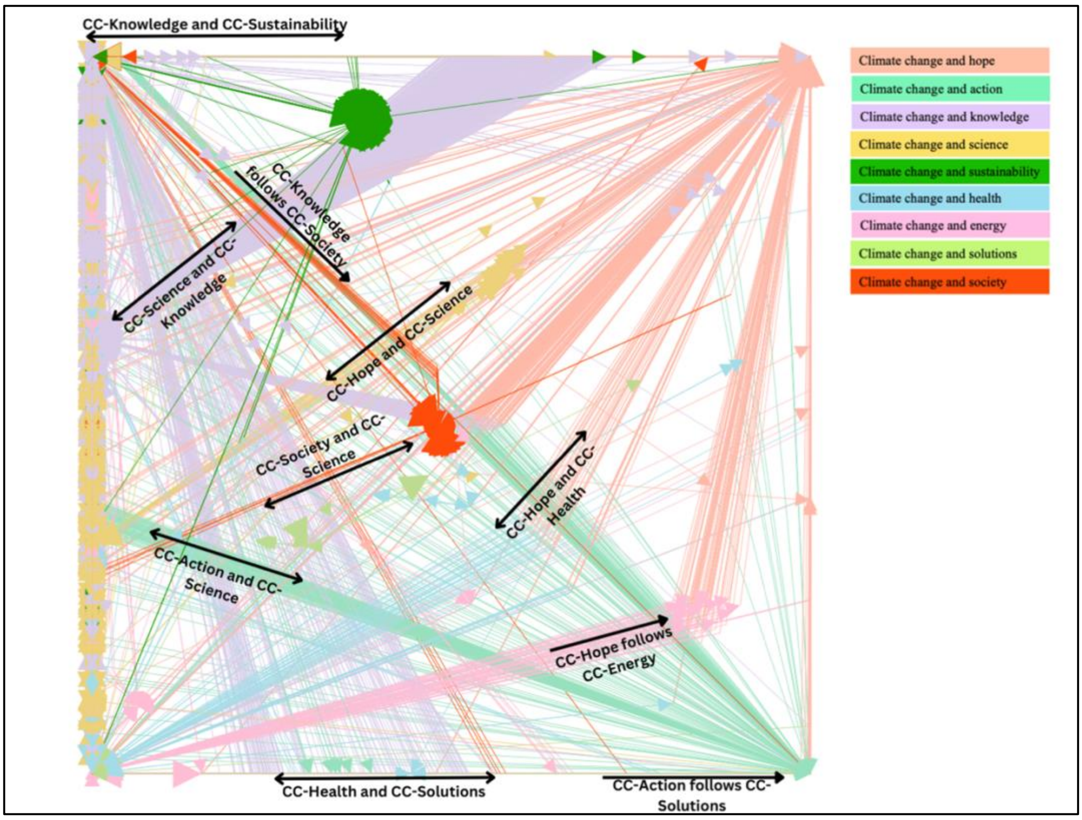
Topic Distribution in Each Green Community



Note. Since each community was represented by a document consisting of the tweets of its authors, this proportion of topics is obtained by calculating the per-document-per-topic probabilities (gamma) of the corpus.

Figure 4

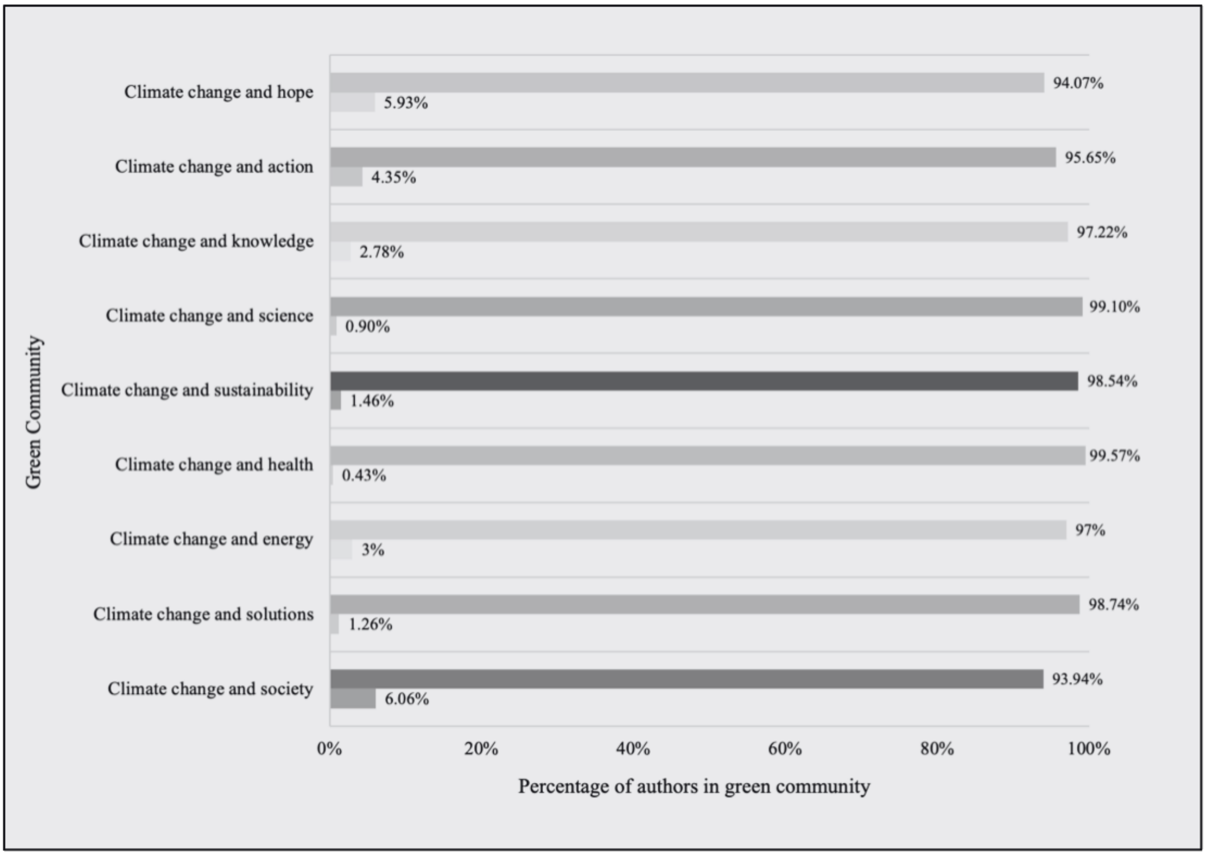
Interactions Between Green Communities – Following Users from Other Communities



Note. The direction of the arrows indicates the target of a node, i.e., the users being followed.

Figure 5

Representation of the Global North and Global South in Twitter Environmental Discourse

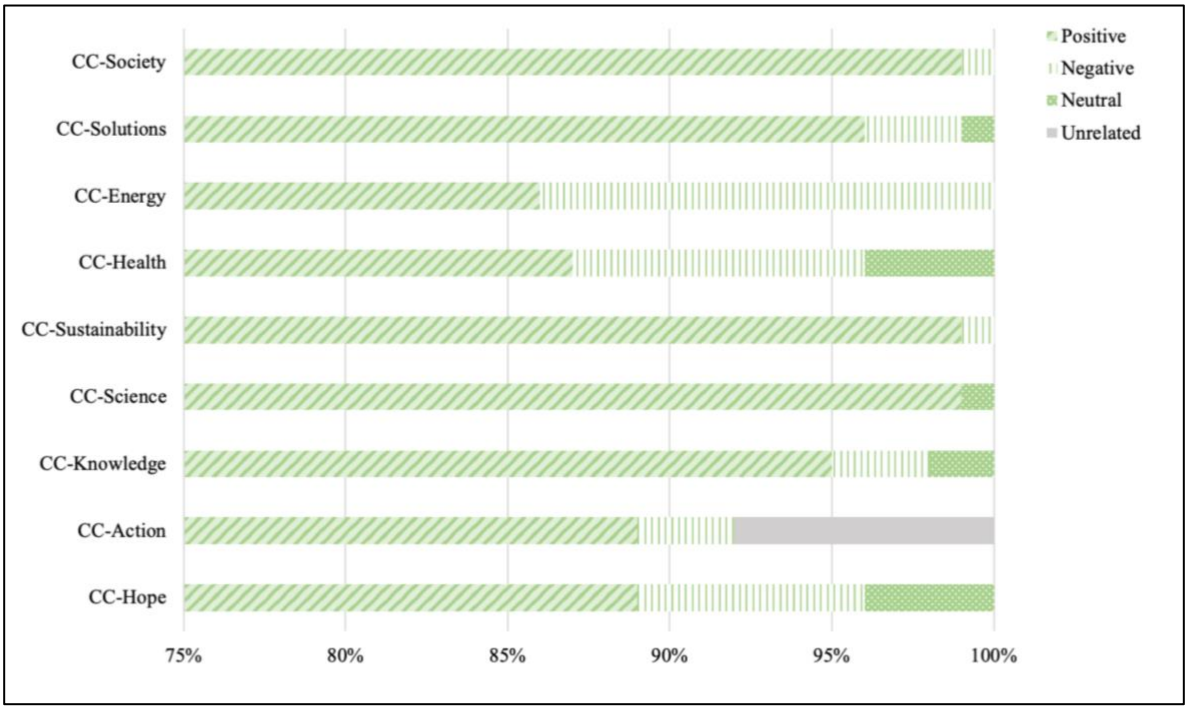


Note. The top bar represents authors from the Global North whereas the bottom bar indicates authors from the Global South.

5. Additional Figures

Figure 6

Representation of sentiment of environment related tweets in Green Communities



Figures 7-15: Topic distribution in each green community

Figure 7

Topic distribution in the climate change and hope (CC-Hope) green community

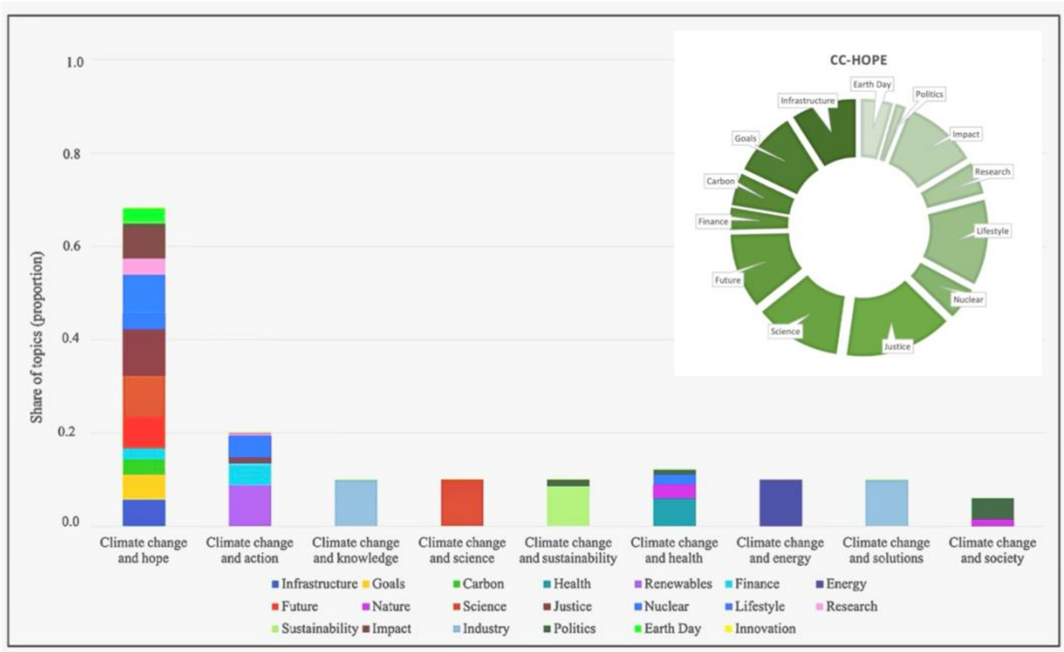


Figure 8

Topic distribution in the climate change and action (CC-Action) green community

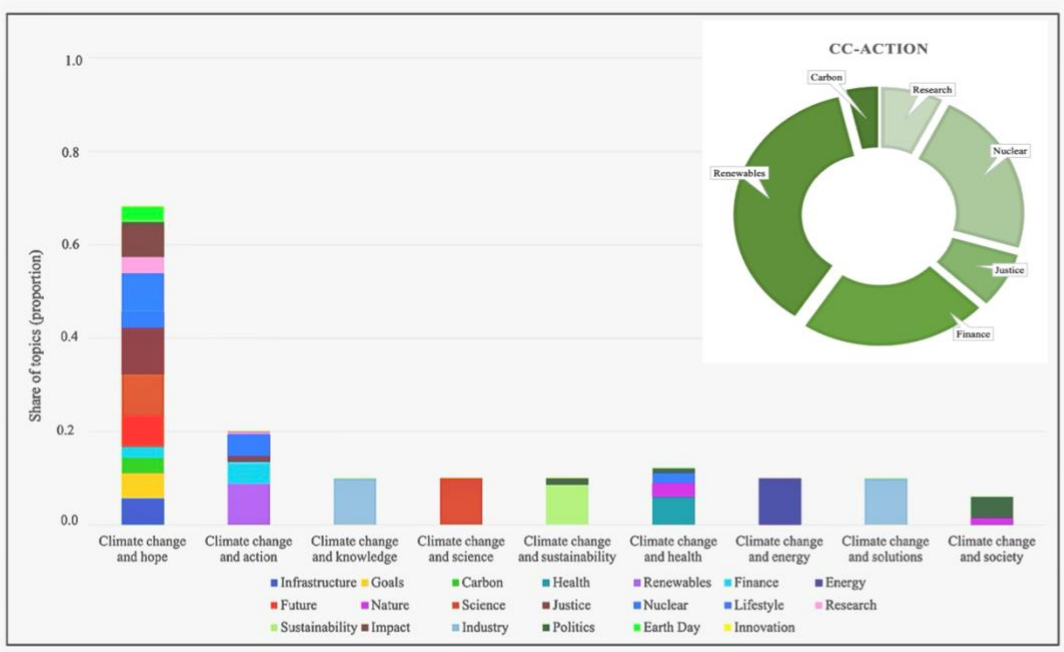


Figure 9

Topic distribution in the climate change and knowledge (CC-Knowledge) green community

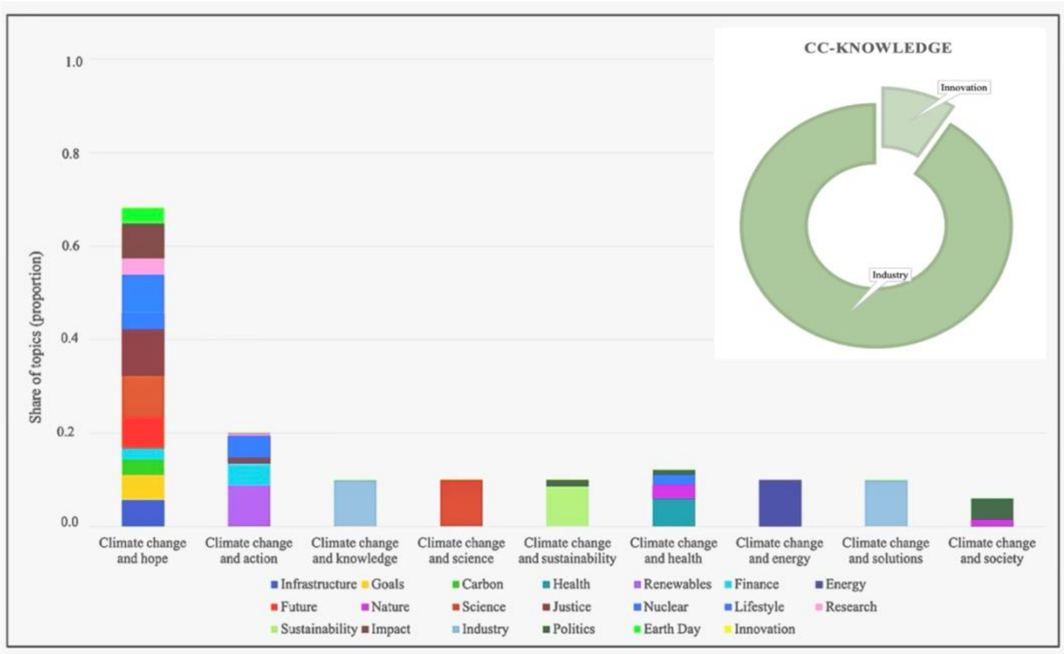


Figure 10

Topic distribution in the climate change and science (CC-Science) green community

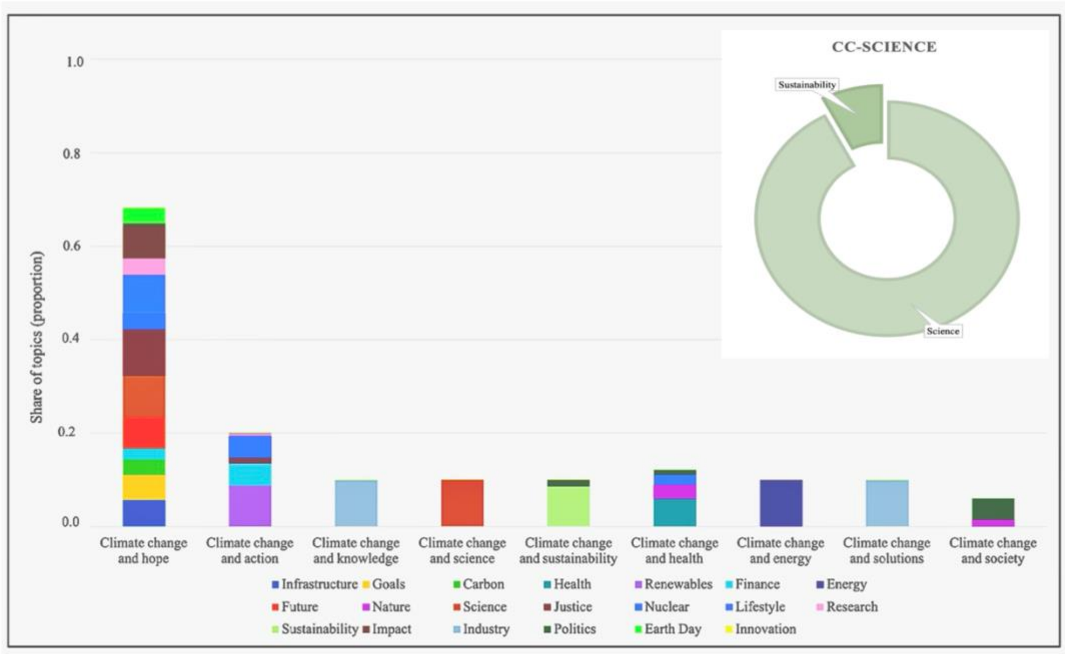


Figure 11

Topic distribution in the climate change and sustainability (CC-Sustainability) green community

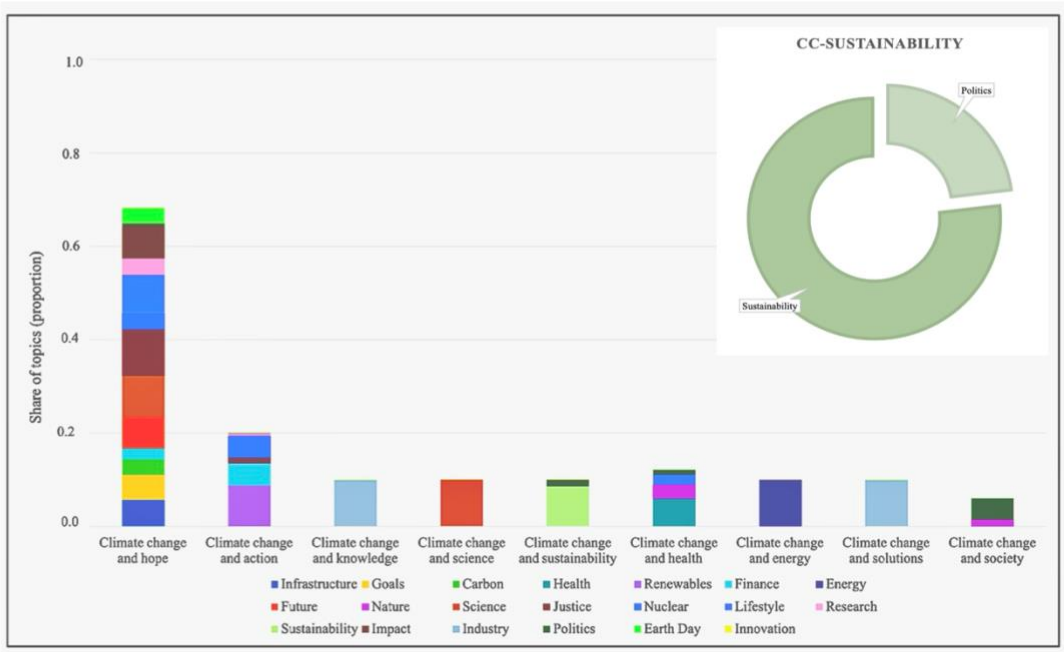


Figure 12

Topic distribution in the climate change and health (CC-Health) green community

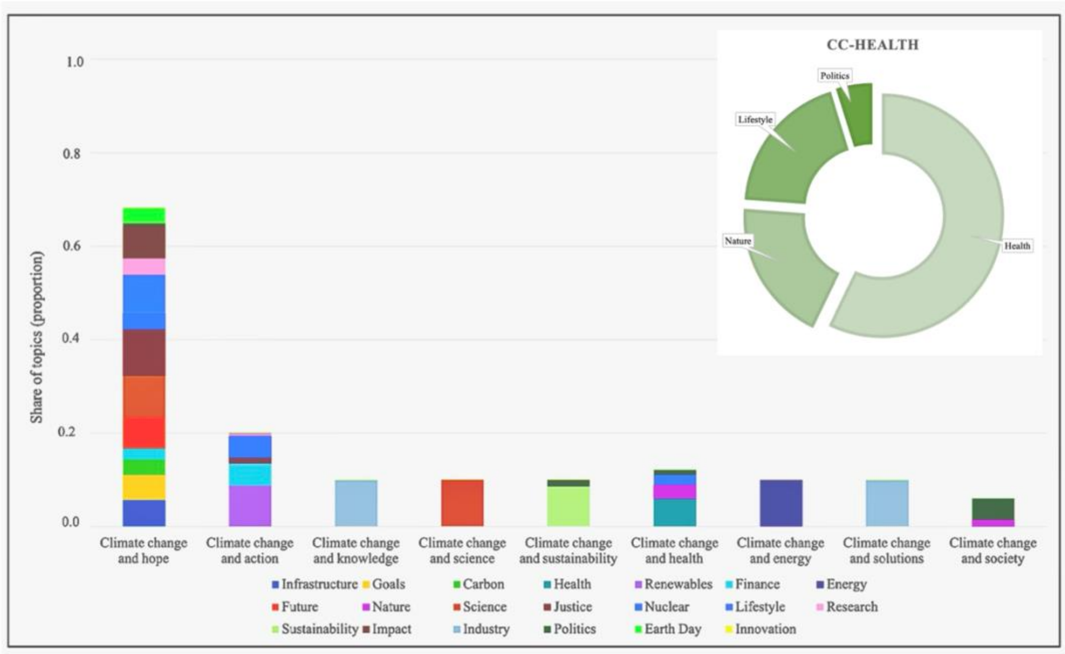


Figure 13

Topic distribution in the climate change and energy (CC-Energy) green community

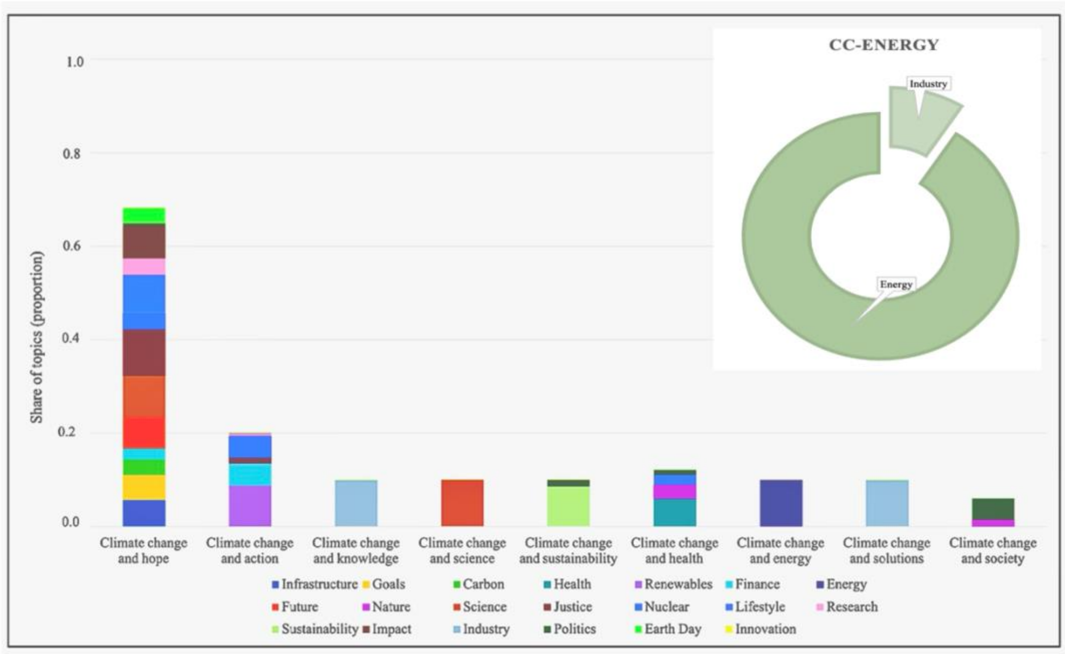


Figure 14

Topic distribution in the climate change and solutions (CC-Solutions) green community

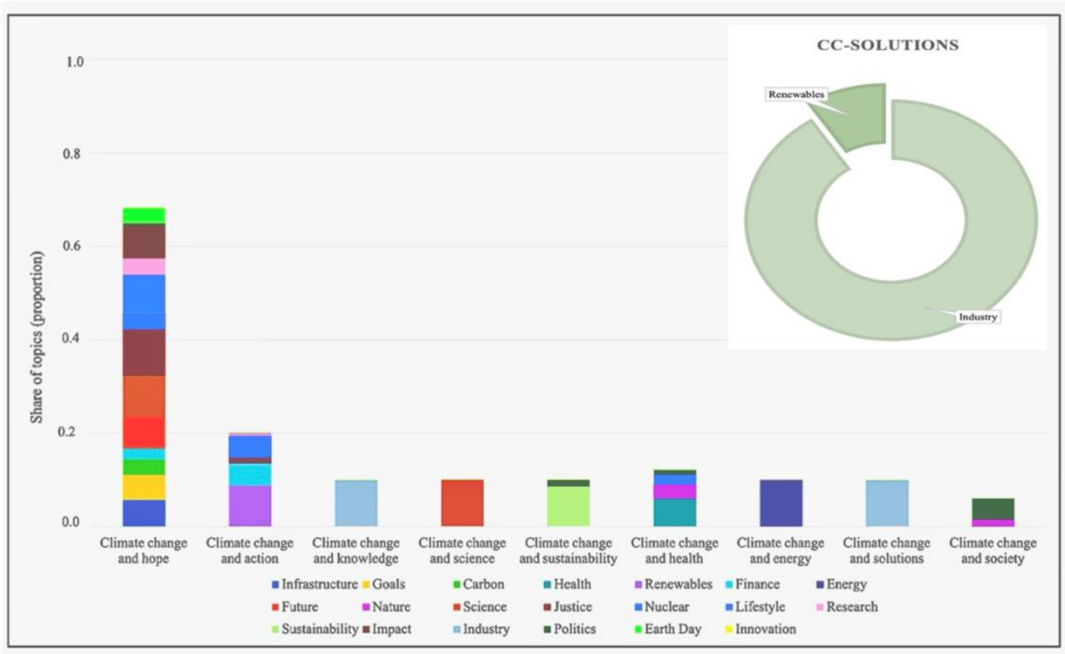
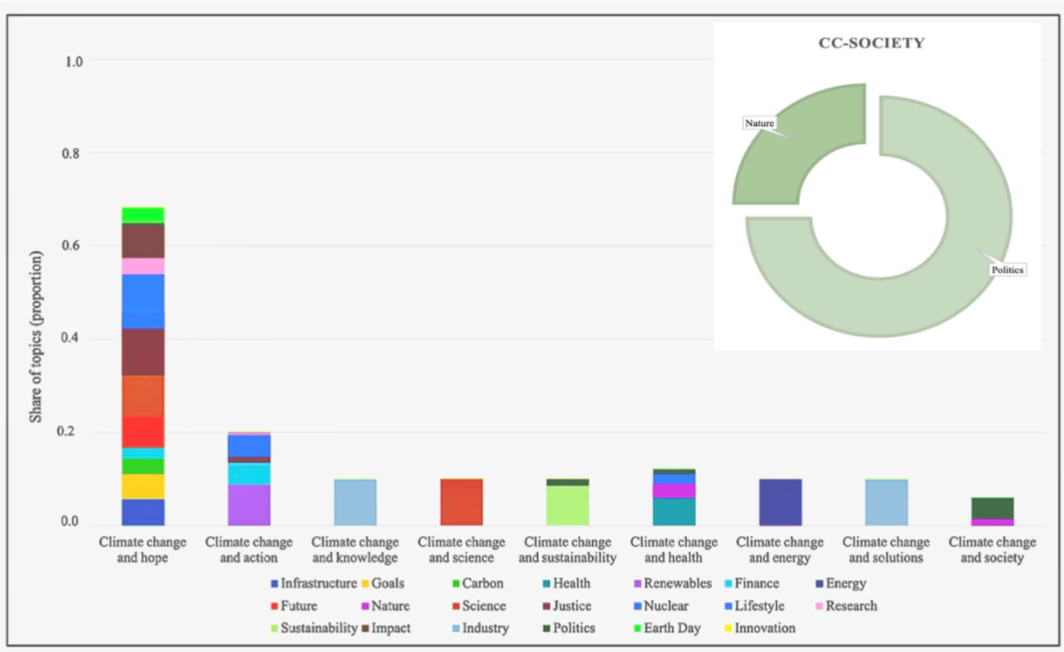


Figure 15

Topic distribution in the climate change and society (CC-Society) green community



Appendix

Additional robustness analysis – Topic modelling

Rationale for additional analysis

During the submission process, a reviewer pointed out that our choice of pooling the tweets by community may have resulted in 9 quite long documents, i.e., one document per green community which contains all the tweets written by the authors in that community. This could have resulted in a dilution of inferred topics if the tweets in a document covered a wide range of topics. Based on the reviewer's suggestion, we conducted additional analysis of a subsample of tweets (n = 100000) with varying document lengths (d = 100 and d = 1000) to examine the extent to which the length of documents may have impacted the topics inferred in this study.

Method

We randomly sampled a subset of 100,000 tweets from the complete dataset of 571, 449 tweets. These 100,000 tweets were pooled in two ways to reflect varying document lengths: (a) 1,000 tweets each pooled by author into 100 documents, and (b) 100 tweets each pooled by author in 1,000 documents.

Then, we repeated the steps of topic modelling as in the initial analysis. We used the topicmodels package in R (Grün & Hornik, 2011) to create a document feature matrix (DFM) by preprocessing the data according to the steps recommended in previous research (Maier et al., 2018). Finally, we estimated respective topic models for K = 20, and compared the top 10 terms per topic in each of these models (d = 100 and d = 1000 respectively) with the results of the topic modelling from our main analysis (d = 9).

Results

Table 1 shows the number of features and perplexity scores for the 3 models. The perplexity scores suggest that pooling the data into 100 documents was led to a better fitting model than pooling them into 1000 documents. The perplexity score is lowest for the topic model reported in the main study, for which we optimised the number of topics.

Table 1.

Perplexity scores of the three topic models comparing the model fit

No. of tweets (*n*)	No. of documents (*d*)	No. of terms in the DFM	Perplexity
571449	$d = 9$	209532	2920.86
100000	$d = 100$	19979	3706.43
100000	$d = 1000$	20259	3878.55

We excluded the topic model with 1000 documents from subsequent analysis as it was deemed to be a poor fit based on relatively high perplexity scores and low interpretability of topics. To infer the extent to which topics may have shifted, we examined the top 10 terms within the topics for the model with 100 documents and compared them to the top 10 terms per topic for the original model with 9 documents. We found that for both models, top words included those related to climate change, sustainability, environment, and earthday (Table 2).

Finally, we also examined the mean probabilites of topics in the documents, finding that one topic (top terms: climate, change, environmental, carbon, earthday) was the most likely to occur in the corpus than others (mean probability = 0.42), confirming our finding that climate change and sustainability were indeed the most discussed environmental issue. The top three topics amounted to 81% of the topic proportions, indicating that these topics were way more likely to occur than the others.

Table 2.

Comparing top terms for robustness analysis of inferred topics

Topic	Top terms *(d = 9)*[a]	Top terms *(d = 100)*[b]
1	climate, environment, environmental, think, flight, energy, day, know, sustainable, attendants	climate, environmental, day, think, flight, energy, know, sustainable, attendants, masked
2	climate, sustainable, change, environmental, energy, emissions, sustainability, learn, future, #sustainability	climate, hope, energy, emissions, sustainable, environment, world, act, clean, great

3	climate, day, earth, new, change, carbon, energy, emissions, people, sustainable	climate, change, environmental, carbon, earthday, energy, sustainable, day, environment, emissions
4	climate, #earthday, change, environmental, new, people, sustainable, carbon, health, energy	sustainability, earth, power, organic, health, year, clean, day, project, carbon
5	climate, environmental, energy, new day, sustainable, environment, carbon, time, think	sustainable, green, earth, learn, climate, sky, time, emissions, #earthday, carbon
6	climate, carbon, one, change, environmental, day, block, earth, #earthday, people	earth, sustainable, climate, learn, hope, carbon, media, green, long, reduce
7	climate, change, carbon, environmental, day, energy, one, make, fuel, environment	flight, energy, day, emissions, fuel, environmental, crisis, forest, ukraine, climate
8	change, energy, future, global, good, health, make, time, back, crisis	health, year, clean, great, reduce, back, time, climate, long, join
9	climate, sustainable, new, environment, change, people, just, need, forest, make	climate, sustainability, earth, masked, sky, qatar, change, environmental, earthday, energy
10	climate, environmental, energy, new, now, co2, day, earth, world, today	global, ecoguy, climate, environment, energy, earthday, time, back, long, act
11	climate, change, just, one, sustainability, carbon, energy, need, make, greeting	climate, environmental, energy, learn, important, year, agriculture, join, just, health
12	climate, change, environmental, carbon, environment, energy, make, earth, going, sustainable	environment, #earthday, report, join, co2, planet, impact, action, gas, health
13	climate, environment, people, carbon, environmental, organic, sustainable, just, fuel, energy	environment, impact, organic, today, goal, #sustainability, solutions, says, food, join
14	climate, change, carbon, #earthday, gas, energy, earth, day, make, new	emissions, make, year, gas, forest, one, now, help, see, need

15	sustainability, carbon, new, day, one, environmental, sustainable, energy, climate, #earthday	green, sustainable, polar, organic, advertisement, day, earthday, future, forest, gas
16	sustainable, energy, environmental, change, biden, going, even, want, new, environment	sustainable, environmental, carbon, clean, carbonnegative, paid, defence, billgates, end, hope
17	biden, climate, sustainable, environment, energy, people, #earthday, improve, industry, earth	justice, biden, government, change, people, sustainability, climate, agriculture, permafrost
18	environmental, change, happy, carbon, one, energy, day, climate, sustainable, sustainability	day, environment, world, emissions, people, global, gas, future, fossil, forest
19	#earthday, climate, sustainability, carbon, environment, environmental, new, energy, sustainable, green	earth, etsy, sustainable, green, rediscova, learn, carbon, sign, food, government
20	#earthday, climate, day, read, change, carbon, sustainable, environmental, global, world	sustainability, earth, year, clean, health, reduce, great, justice, biden, time

Note. The topics are listed in decreasing order of probability of occurrence, i.e., topic 1 indicates the most probable topic and topic 20 indicates the least probable topic. The terms within each topic differ more as the probability of topics decreases because of the different corpora used in the analyses.

[a]Intial analysis reported on 571,449 tweets

[b]Robustness analysis reported on a random subsample of 100,000 tweets

References

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*, 1–30. https://doi.org/10.18637/jss.v040.i13

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754