SN Computer Science (2022) 3: 371
https://doi.org/10.1007/s42979-022-01276-w

ORIGINAL RESEARCH

How is People's Awareness of "Biodiversity" Measured? Using Sentiment Analysis and LDA Topic Modeling in the Twitter Discourse Space from 2010 to 2020

Shimon Ohtani[1]

Received: 8 October 2021 / Accepted: 27 June 2022 / Published online: 15 July 2022

Abstract
The importance of biodiversity conservation is gradually being recognized worldwide, and 2020 was the final year of the Aichi Biodiversity Targets formulated at the 10th Conference of the Parties to the Convention on Biological Diversity (COP10) in 2010. Unfortunately, the majority of the targets were assessed as unachievable. While it is essential to measure public awareness of biodiversity when setting the post-2020 targets, it is also a difficult task to propose a method to do so. This study provides a diachronic exploration of the discourse on "biodiversity" from 2010 to 2020, using Twitter posts, combined with sentiment analysis and topic modeling, commonly used in data science. Through the aggregation and comparison of *n*-grams, the visualization of eight types of emotional tendencies using the NRC emotion lexicon and supplemental comparison with the machine learning model, the construction of topic models using Latent Dirichlet allocation (LDA), and the qualitative analysis of tweet texts based on these models, the analysis and classification of these unstructured tweets have been performed effectively. The results revealed the evolution of words used with "biodiversity" on Twitter over the past decade, the emotional tendencies behind the contexts in which "biodiversity" has been used, and the approximate content of tweet texts that have constituted topics with distinctive characteristics. While searching for people's awareness through SNS analysis still has many limitations, it is undeniable that essential suggestions can be obtained. To further refine the research method, it will be crucial to improve analysts' skills, accumulate research examples, and advance data science.

Keywords Biodiversity · Awareness · Twitter · *n*-gram · Sentiment analysis · Topic modeling

Introduction

Overview

The concept of "biodiversity," which began to be used officially in the 1980s, was later widely adopted by countries around the world through the implementation of the Convention on Biological Diversity (CBD), which was signed in 1992 as one of the outcomes of the United Nations Conference on Environment and Development in Rio de Janeiro, Brazil, and entered into force in 1993 [11]. The convention is one of the countless international frameworks developed to address the severe deterioration of the global environment due to global concerns. Until June 2022, 15 sessions of the Conference of the Parties (COP) have been held, and two protocols have been finalized [7]. Among them, COP10, held in Nagoya City, Aichi Prefecture, Japan, in 2010, urged the United Nations to designate the decade up to 2020 as the "United Nations Decade on Biodiversity" and the Aichi Biodiversity Targets, a comprehensive approach to biodiversity conservation, were adopted [13].

COP15 had been postponed to October 2021 due to the aftermath of the global spread of a new coronavirus (COVID-19) [6]. Furthermore, it was announced that the conference would be held in two parts, with the second half in the third quarter of 2022. Initially, 2020 was also the final year of the Aichi Biodiversity Targets and the "Mission" of The Strategic Plan for Biodiversity 2011–2020, formulated in 2010. Unfortunately, most of the goals were reported to be insufficient, according to "The Global Biodiversity Outlook 5 (GBO-5)" published in September 2020 [14]. Various

✉ Shimon Ohtani
s-ohtani@g.ecc.u-tokyo.ac.jp

[1] Graduate School of Interdisciplinary Information Studies, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

factors have been pointed out as the cause of this. Some studies have pointed to multiple factors, such as deficiencies in global environmental governance or individual national circumstances such as inadequate legal systems or other impediments [4]. However, even though fragmented and localized analysis has been done, it has been challenging to create a global and comprehensive analysis method [26].

It is important to note that for all biodiversity conservation measures to work, they must be based on individual private citizens' full awareness and action, not just politicians and experts [23]. Moreover, this has always been pointed out, as evidenced by the fact that Aichi Target 1 was "By 2020, at the latest, people are aware of the values of biodiversity and the steps they can take to conserve and use it sustainably". Unfortunately, this goal was rated as "not achieved" in GBO-5. At the same time, the reliability of that rating was considered "low." The reason cited is that "there is no globally consistent information on trends in awareness and willingness to act on biodiversity." Although UEBT (the Union of Ethical Bio Trade's Biodiversity Barometer) survey of the general public in 16 countries provided mainly demographic data on the understanding of biodiversity [25], it remains inadequate in measuring long-term and global public awareness.

Furthermore, as a focus on big data, the usefulness of a new global indicator to measure citizens' involvement in biodiversity, developed based on keyword data provided by online newspapers, Twitter, and Google Trends was simultaneously introduced in GBO-5. The indicator was described as an innovative tool that could not only be used by countries to report to the CBD Secretariat on their progress towards the Aichi Biodiversity Targets. Still, it could also be used to measure progress globally. However, due to limitations in the use of data, even with this indicator, it is not yet possible to measure long-term trends over time and to accurately measure people's awareness of the value of biodiversity and the actions they can take to conserve and sustainably use it [8].

Literature Reviews

Currently, discourse on social media, including Twitter, is becoming increasingly popular as a place for many people to express their emotions and opinions. In light of this urgent need to devise a method to measure people's awareness and willingness to act on biodiversity, this study focused on the usefulness of Twitter data and proposed a different research method. It has been reported that these analyses are helpful in various fields as a method of opinion mining [20].

Few studies have effectively attempted to analyze Twitter sentiment on environmental protection and biodiversity conservation. To cite a few that stand out, Fink et al. [10], for example, conducted a sentiment analysis of Twitter posts and online news about "rhinoceros." Here, they explored the correlation between the number of tweets about a particular event and the amount of news and analyzed the emotional tendencies of the posters. Otero et al. [19] analyzed tweets about marine plastic pollution with a detailed analysis based on topic modeling and sentiment analysis. Current state-of-the-art techniques such as bot detection and Emoji consideration were used here. These are notable examples of studies that effectively use NLP to analyze Twitter discourse on environmental issues. However, because of their focus on specific subjects, these studies are not necessarily intended to contribute to long-term global policy decisions.

This study attempted to explore the raising of awareness of "biodiversity" over time using the discourse space of Twitter as a subject from a longer-term, big-picture perspective. More specifically, this study aimed to provide a rudimentary analysis of the Twitter discourse on the concept of "biodiversity" from 2010, when the Aichi Biodiversity Targets were established, to 2020, the final year of the Targets, to provide clues for streamlining future international governance of biodiversity conservation. At the same time, this study proposed exploratory research to explore its usefulness and show its potential to develop using several natural language processing methods (NLP).

Contributions

This paper is intended to make the following scholarly contributions.

1. We proposed a method to measure people's awareness and willingness to act on biodiversity conservation, which had been considered a difficult and urgent issue.
2. We presented an example of a study to introduce NLP to a global policy issue that has rarely been tested.
3. We presented an example of a study that introduces several methods of NLP into a diachronic multi-year survey, which has rarely been done with Twitter analysis.
4. By presenting a specific analysis, especially sentiment analysis of both lexicon-based and machine learning models, we demonstrated the usefulness and vulnerability of NLP and its potential for future development.
5. In doing so, we provided a touchstone for promoting Global Governance and Computer Science integration.

For this reason, the novelty of this paper lies not in the presentation of a general data analysis methodology itself and quantitative evaluation of accuracy but rather in its application. The structure of this paper is as follows:

- *Methodology* provides details of the subject of analysis and the research methods.
- *Results* details the results obtained.

- *Discussion* provides a discussion based on the results.
- *Conclusion* presents the challenges and possible developments and findings.

Methodology

Research Design

In this study, Twitter posts have been selected as the object of the analysis. Before the investigation, it is necessary to understand the peculiarity of Twitter and its posts: since its launch in March 2006, Twitter, as a so-called microblogging site, has been expanding its users worldwide due to the ease and convenience of writing short messages of 280 characters. Today, it has grown into a leading social networking service (SNS) with 1.3 billion accounts, including heads of state and dignitaries, and 330 million monthly active users, who post 500 million messages every day [24]. Expressions of emotions and sentiments, such as opinions expressed on this platform and behavioral patterns, have become valuable targets for analysis by data and analysts. In addition, social networking sites can be linked to specific topics through hashtags, and multiple communication through likes, retweets, and replies is also possible [3]. Thus, analysis of social networking sites is being applied in various fields such as market research, product reviews, and traffic prediction [18]. The Twitter analysis is being used in many areas due to its usefulness.

Not only that, but the expression of opinions and emotions on Twitter extends to all kinds of things, including people, things, concepts, and policies. In fact, since 2020, there has been a surge of research on Twitter posts about the COVID-19 global pandemic. This fact is because it is believed that Twitter posts can provide much helpful information in exploring public sentiments and concerns about an infectious disease that has rapidly spread around the world and changed people's lives [28]. Furthermore, Twitter posts can also be applied to find out what people in the modern world are thinking about the concept of "biodiversity," which was born in the 1980s. In other words, the diachronic analysis of Twitter posts may provide us with new insights that we have not obtained before.

However, it is essential to be careful about equating the discourse space on Twitter with the actual discourse space. Because there are various obstacles to analysis, such as bias in user demographics, individual differences in tweet frequency, the existence of bots, and the inclusion of a lot of useless noise other than text [1], therefore, when analyzing posts on Twitter and drawing certain conclusions, we should consider a unique platform with the above limitations and restrictions when analyzing and interpreting them.

The methods used in this study include *n*-gram counting and comparison, sentiment analysis using the NRC emotion lexicon developed by the National Research Council Canada [16], topic modeling using Latent Dirichlet allocation (LDA) [2], and qualitative analysis of tweet texts. The outline of the research procedure is as follows:

1. We had collected all tweets containing the keyword "biodiversity" from March 2006, when the Twitter service was launched, to December 2020 and extracted purely in English tweets.
2. We have pre-processed all tweets from 2010 to 2020, counted *n*-grams (bigram and trigram), and listed the top 20.
3. In the same way, we have counted eight types of emotion words using the NRC emotion lexicon for pre-processed tweets from 2010 to 2020, calculated the percentage of each type used in the total number of words, and visualized the results on charts.
4. For each year, we have explored LDA topic models and constructed the model that seemed to be optimal. In this paper, we have discussed the models for the years 2010 and 2020 through visualization.
5. Based on the visualized information, we have selected specific topics and created sentiment charts to examine the texts' contents.
6. We have provided an overview of their tweet texts for specific topics.

Furthermore, for step 6 above, we also included the analysis results by DistilBERT [22], one of the machine learning models. This text classification method was conducted on a pilot basis. By presenting the results of DistilBERT, for which few practical examples of analysis have been given so far, we sought to promote the introduction of NLP into global governance research as soon as possible. In addition, this study aimed to grasp the whole picture of the discourse surrounding "biodiversity" posted on Twitter. Therefore, we did not consider the sender's nationality, title, and other attributes.

Data Collection

In this study, we have collected tweets containing the word "biodiversity" posted from March 21, 2006, when the Twitter service started, to December 31, 2020. This study aimed to explore the usage of "biodiversity" in ordinary contexts and the hashtag "#biodiversity." For the collection, we first applied for the Academic Research product truck released by Twitter in January 2021 and obtained access to the entire Twitter archive. Then, we used the open application programming interface (API) provided by Twitter and the Python programming language (ver. 3.8.8). As a result, the

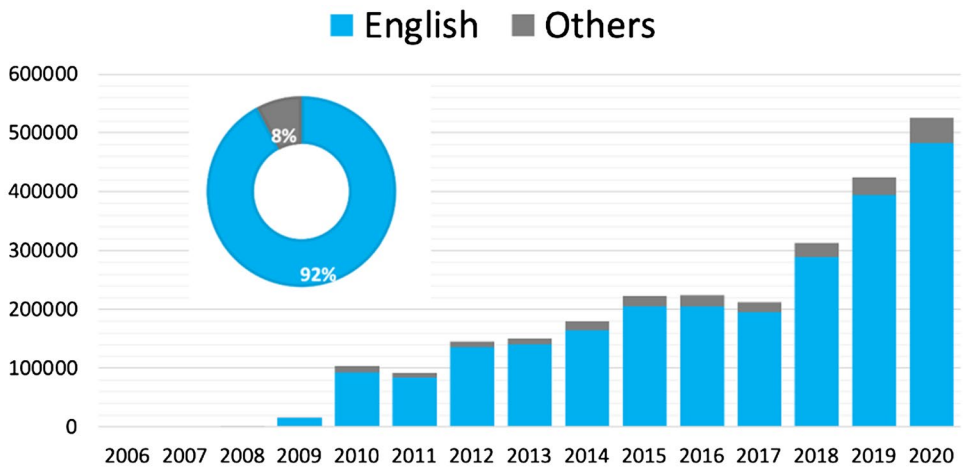

Fig. 1 Total number of tweets containing the word "biodiversity" by year from 2006 to 2020 (distinguishing between English text and other language text)

total number of tweets by December 31, 2020, is 2,609,834, which is outstanding compared to other primary language expressions of "biodiversity" (biodiversité, biodiversidad, biodiversität). Of these tweets, 2,405,937 tweets were purely in English text, accounting for 92% of the total (Fig. 1). Therefore, we have decided to focus on tweets in English in this study. The collected tweet information includes "text," "author_id," "created_at," "lang," "entities," "geo," "public_metrics," and "text." In this study, we have focused on "text." Because this study aims to understand the general speech on Twitter, the attributes and location information of posters are out of the scope of this study because many tweets lacked information. Figure 1 shows the total number of tweets and the number of tweets in English for the period covered. According to this figure, the number of relevant tweets increased by about two times from 2010 to 2017, but the increase was relatively slow. However, since 2018, the rise in tweets has become more extensive.

Pre-processing the Raw Dataset

Before analysis, it was necessary to pre-process Twitter raw data into a form suitable for analysis. In this study, we have followed several related studies on Twitter analysis and performed two types of pre-processing in Python, one for sentiment analysis and the other for topic model building. The specific pre-processing is as follows:

1. Extracted only English tweets from the collected tweets.
2. Removed tweets with duplicate text.
3. Removed @usernames and links (pasted URLs such as HTTP and www) in the text.
4. Removed special characters and punctuations from the text.
5. Other strings that did not have any particular meaning were excluded by designating them as "stop words."
6. The texts in the above state were saved for sentiment analysis.
7. Also removed hashtagged words.
8. Tokenized the texts—deleted tweets with less than three tokens.s.
9. *n*-grams (bigrams and trigrams) were counted and saved.
10. Performed lemmatization of the tokens.

Data Analysis

In this study, we have used both quantitative and qualitative research. First, as a quantitative study, we overviewed the data through visualization using LDA topic modeling and sentiment analysis, and second, we qualitatively examined the content of the specific tweet texts that were categorized. The following sections describe each of the analysis methods.

Counting and Comparing *n*-grams

An n-gram is a sequence of words, where two words are called a bigram and three are called a trigram. In this study, we have counted *n*-grams for each year to get an overview of the set of tweet texts narrowed down by pre-processing as described above. Here, we have used Gensim, which is available in Python. By comparing the top-ranked *n*-grams, it was possible to understand more specific keywords in each text. Moreover, it was also possible to see the characteristics of the words used with "biodiversity" throughout the entire period and to identify the words that were characteristic of each year. The top 20 bigram and trigram terms from 2010 to 2020 were listed for comparison and discussion in this paper.

Sentiment Analysis

Sentiment analysis is an automated process of mining attitudes, opinions, views, and emotions from text, speech, tweets, and database sources using natural language processing (NLP). It is said to be analyzing people's feelings, attitudes, opinions, and emotions towards elements such as products, individuals, topics, organizations, and services [15]. There has been a rapid increase in the number of examples of analysis of social networking sites, the most popular of which is categorizing them into Positive, Negative, and Neutral. However, in recent years, various methods such as Machine Learning Approaches and Lexicon-Based Approaches have been devised and are showing rapid development.

In this study, we have used one of the Lexicon-Based Approaches, the NRC emotion lexicon [16]. It is a crowdsourced task for tens of thousands of English words, manually curated and encoded with emotions (positive or negative) and discrete models of emotions covering anger, expectation, disgust, fear, joy, sadness, surprise, and trust via binary variables for each emotion [17]. Our preliminary research found that positive, negative, and neutral categorization was highly abstract and subject to wide swings, ultimately forcing us to read and scrutinize specific texts. Thus, we have decided to read eight types of emotions from the tweet texts, as we needed to clarify the direction of more specific emotions.

According to the NRC emotion lexicon developer, the lexicon works by comparing multiple data sets and producing a percentage of the total number of words [17]. In this study, we have searched for emotions expressed in tweets as a whole and individual tweets by finding the total number of words belonging to each of the eight types of emotions in the NRC emotion lexicon and their percentage of the total number of words in the text of tweets containing the expression "biodiversity." Furthermore, we excluded the years from 2006 to 2009, when the total number of tweets per year was small. We calculated the total number of words by using Python and the percentage of terms constituting each emotion in the pre-processed tweet texts for each year from 2010 to 2020 and visualized them.

Latent Dirichlet Allocation (LDA)

For this study, which explores the discourse on "biodiversity" on Twitter, it was essential to explore each year's topics discussed by users. In this study, we have used Latent Dirichlet allocation (LDA) [2]. LDA is a form of Unsupervised machine learning. To date, it has been applied to all kinds of sociological research, including the analysis of news articles, and is considered to be an efficient method for identifying patterns, themes, and structures in large, unstructured groups of text, such as tweets in Twitter, and classifying them by topic based on these patterns. The model assumes that each document consists of a mixture of various potential topics and that each topic is characterized using a distribution of linguistic units. Furthermore, the algorithm generates pairs of frequently mentioned words, pairs of co-occurring terms, potential topics in a document, and their distributions over those topics based on the data itself [27].

In this study, we have used the Python library "Gensim" and the java open-source software "MALLET" to run multiple trials on the tweet texts of each year from 2010 to 2020 with different numbers of topics. Furthermore, the best topic models were explored, constructed, and visualized. However, for reasons of paper space, we have used the NRC emotion lexicon to count and calculate the percentage of emotion words in the modeled topics for 2010 and 2020 only.

In addition, emotion classification results from DistilBERT, a distilled version of BERT (short for Bidirectional Encoder Representations from Transformers, Devilin et al. [9]), were also presented simultaneously in this step using the Python library Transformers. DistilBERT is created by distilling knowledge in the pre-training phase to reduce the size of the BERT model by 40% while retaining 97% of the language comprehension and is said to be smaller and faster than BERT and other BERT-based models [22]. In this study, we used the fine-tuned model "Distilbert-based-uncased-emotion" from the Hugging Face model hub (https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion?text=I+feel+a+bit+let+down), to calculate six emotion scores for each sadness, joy, love, anger, fear, and surprise for all tweet texts on the selected topic. The emotion with the highest score was labeled for each tweet, and the percentages were tabulated.

Qualitative Analysis

After categorizing, visualizing, and interpreting the data through quantitative research, it would be beneficial to conduct specific analysis through qualitative research. This study also extracted keywords that constituted each topic when we built the topic model and identified representative tweets for each topic. All tweet texts were assigned a score (Topic_Perc_Contrib) for their weight within each topic. Furthermore, the original text of the tweets in the highest range was posted as Representative Text.

It was possible to infer the dominant discourse by identifying and examining significant texts from many tweets. At the same time, this was an attempt to minimize the drawbacks of quantitative analysis of tweets. While it would have been possible to examine and define all the categorized topics in detail, we have focused on only a few specific topics and examined the tweet contents that constituted them. Using the above method, it would be possible to roughly

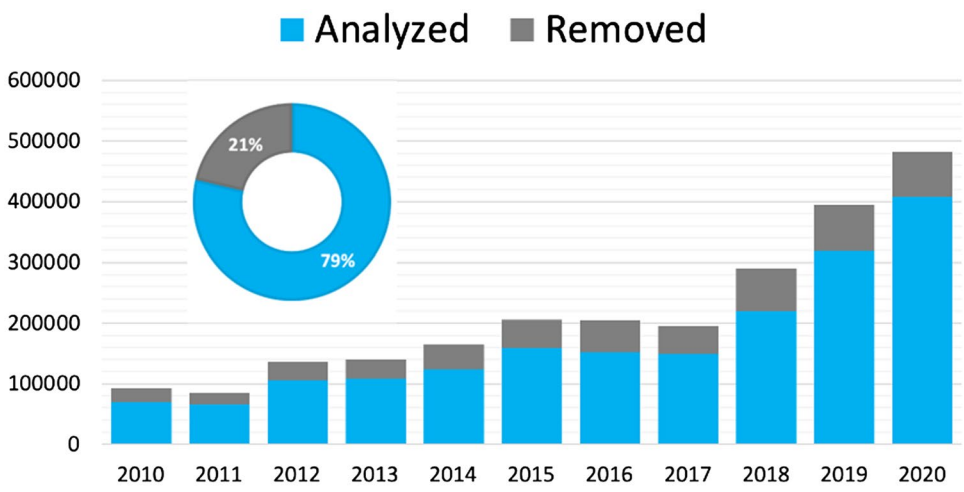

Fig. 2 Total number of tweets in English text containing the word "biodiversity" by year from 2010 to 2020 (distinguishing between analyzed and removed tweets)

grasp the dominant discourse of each year in the Twitter space.

Result

As a result of the pre-processing, the number of tweets to be analyzed, as shown in Fig. 2: out of 2,389,197 tweets from January 1, 2010, to December 31, 2020, about 21% were removed, and the final number was narrowed down to 1,879,221 tweets.

Counting and Comparing *n*-grams

Table 1 shows the top 5 bigrams from 2010 to 2020, the examples from the top 20. According to these, it could be seen that "conservation," "loss," "marine," "protect," "nature," and "wildlife" were consistently used with "biodiversity." Bigram ('biodiversity,' 'loss') was 20,274 in 2020 compared to 2785 in 2010, and bigram ('biodiversity,' 'conservation') was 10,852 in 2020 compared to 1358 in 2010, both of which were significant increases. As for other idioms, "climate change" was also at the top of the list, clearly showing that "biodiversity" was often used in conjunction with climate change issues.

In addition, by observing the other bigrams, some of the characteristics of each year could be identified. Since 2010 was the United Nations' Year of Biodiversity, many words related to the "International Year of Biodiversity (IYB)" and related topics, or COP10, held in Nagoya, Aichi Prefecture, Japan. The year 2012 saw the impact of increased tweets about the Environmental Biodiversity Outreach Officer jobs. 2013 saw a noticeable increase in tweets echoing "Biodiversity offsets in theory and practice" published. In 2015, the word "human" was used prominently along with "biodiversity" due to the publication of "Connecting Global Priorities: Biodiversity and Human Health" by the World Health Organization (WHO) and a surge in tweets mentioning it. Also, from 2015 to 2018, "Wikipedia," "article," "English," and "edited" were among the top terms because there was a significant update work done on the words related to biodiversity in Wikipedia, and this was tweeted verbatim by specific accounts. In 2019 and 2020, "crisis," which was not used as often in previous years, was increasingly used with "biodiversity."

Also shown in Table 2 are the top 5 trigrams, the examples from the top 20, and observing these revealed further details of the trends seen in the observations of the bigrams. In 2011, it could be seen that tweets about "The Belly Button Biodiversity Project" came out on top. In 2014, French words such as "biodiversit," "animaux," and "oiseaux" topped the list because many English tweets were tagged with French hashtags. Furthermore, it was found that in 2015 and 2016, several tweets were made regarding writings by Gary Paul Nabhan and, in 2017, by Pankaj Oudhia. 2019 is the only year in which the word "extinction" was found at the top of the list. This result is because many tweets warned that various species on the planet are at the risk of extinction.

The increase of trigrams during the past decade shows that trigram ('climate,' 'change,' 'biodiversity') was 362 in 2010 and 4016 in 2020. The number of trigrams ('biodiversity,' 'ecosystem,' 'services') is 1069 in 2020 compared to 185 in 2010, which are significant increases. Thus, it is possible to obtain particular suggestions even by observing only the number of bigrams and trigrams.

Sentiment Analysis

Figure 3 shows the number of words corresponding to the eight types of emotions using the NRC emotion lexicon for the pre-processed tweet texts, calculated as a percentage of the total number of words from 2010 to 2020 and visualized on an area chart. Even though the 11 years of

Table 1 Top 5 bigrams from 2010 to 2020 (examples from top 20)

	2010		2011		2012		2013	
	Bigram	Count	Bigram	Count	Bigram	Count	Bigram	Count
1	(‘year’, ‘biodiversity’)	3550	(‘biodiversity’, ‘conservation’)	1920	(‘biodiversity’, ‘conservation’)	3295	(‘biodiversity’, ‘conservation’)	2835
2	(‘biodiversity’, ‘loss’)	2785	(‘climate’, ‘change’)	1755	(‘climate’, ‘change’)	3251	(‘clima te’, ‘change’)	2228
3	(‘international’, ‘year’)	2731	(‘biodiversity’, ‘loss’)	1460	(‘biodiversity’, ‘loss’)	2438	(‘biodiversity’, ‘loss’)	1853
4	(‘climate’, ‘change’)	1705	(‘marine’, ‘biodiversity’)	1347	(‘marine’, ‘biodiversity’)	2046	(‘biodiversity’, ‘offsetting’)	1806
5	(‘biodiversity’, ‘conservation’)	1358	(‘conservation’, ‘biodiversity’)	1215	(‘officer’, ‘jobs’)	1379	(‘marine’, ‘biodiversity’)	1301
	2014		2015		2016		2017	
	Bigram	Count	Bigram	Count	Bigram	Count	Bigram	Count
1	(‘biodiversity’, ‘conservation’)	3345	(‘biodiversity’, ‘human’)	23,180	(‘biodiversity’, ‘conservation’)	4311	(‘biodiversity’, ‘conservation’)	4740
2	(‘climate’, ‘change’)	2587	(‘biodiversity’, ‘conservation’)	3923	(‘climate’, ‘change’)	3148	(‘climate’, ‘change’)	3670
3	(‘biodiversity’, ‘biodiversit’)	2455	(‘climate’, ‘change’)	2808	(‘wikipedia’, ‘article’)	2997	(‘wikipedia’, ‘article’)	3316
4	(‘biodiversity’, ‘loss’)	1797	(‘biodiversity’, ‘loss’)	2169	(‘edited’, ‘biodiversity’)	2994	(‘edited’, ‘biodiversity’)	3312
5	(‘animals’, ‘animaux’)	1722	(‘marine’, ‘biodiversity’)	1852	(‘english’, ‘wikipedia’)	2993	(‘english’, ‘wikipedia’)	3310

	2018		2019		2020	
	Bigram	Count	Bigram	Count	Bigram	Count
1	(‘biodiversity’, ‘conservation’)	9701	(‘climate’, ‘change’)	18,253	(‘biodiversity’, ‘loss’)	20,274
2	(‘climate’, ‘change’)	7717	(‘biodiversity’, ‘loss’)	14,396	(‘climate’, ‘change’)	18,945
3	(‘biodiversity’, ‘loss’)	6951	(‘biodiversity’, ‘conservation’)	8234	(‘biodiversity’, ‘conservation’)	10,852
4	(‘marine’, ‘biodiversity’)	2767	(‘climate’, ‘biodiversity’)	6034	(‘climate’, ‘biodiversity’)	7412
5	(‘funds’, ‘biodiversity’)	2762	(‘loss’, ‘biodiversity’)	6024	(‘protect’, ‘biodiversity’)	6584

data were represented on a single chart, the shape of the data was almost uniform, resulting in good visibility. This result was because, in all years, the use of words corresponding to “trust” and “anticipation” was high. At the same time, “joy” and “fear” were slightly elevated, showing almost the same tendency to use emotional words.

There are several possible interpretations for this uniformity in the distribution of words of emotion use, despite the more than five-fold increase in the number of tweets over the past 10 years. The most straightforward interpretation is that “biodiversity”-related discourse has remained constant in this way on Twitter. However, even if sentiment analysis results show approximate trends in the use of words of emotion on the chart, it is possible to infer that there were subtle differences in the individual tweets each year by comparing the *n*-grams. Furthermore, one of the limitations of using the NRC emotion lexicon as-is may be that it needs to be strictly customized for each analysis target. It is also possible that changes in the analysis procedure may yield different results.

Topic Modeling

Using the Python library Gensim, we have explored the LDA topic models for each year and built the model that was considered best for each. We created the Gensim model and the MALLET model each year, respectively. We derived the best model from the Coherence Scores [17] and the topic distribution of the visualization results using the Python library pyLDAvis while varying the number of topics. For paper space, we included models for 2010 and 2020 for comparison, and Figs. 4 and 5 show the results of the search for the Coherence Scores when the numbers of topics are determined, respectively. As a result of the trials, we adopted the Gensim model for 2010, consisting of 60 topics (Fig. 6), and the MALLET model for 2020, which consists of 40 topics (Fig. 7).

In addition, the topics were arranged in order from the major ones, and the number of the eight types of emotion words used by the NRC emotion lexicon was counted for each topic and represented on a color scale (Figs. 8 and 9).

Table 2 Top 5 trigrams from 2010 to 2020 (examples from top 20)

	2010		2011		2012	
	Trigram	Count	Trigram	Count	Trigram	Count
1	('international', 'year', 'biodiversity')	2618	('wildlife', 'conservation', 'biodiversity')	620	('environmentals', 'biodiversity', 'outreach')	1375
2	('climate', 'change', 'biodiversity')	362	('belly', 'button', 'biodiversity')	296	('biodiversity', 'outreach', 'officer')	1375
3	('economics', 'ecosystems', 'biodiversity')	324	('climate', 'change', 'biodiversity')	294	('outreach', 'officer', 'jobs')	1375
4	('species', 'iyb', 'biodiversity')	238	('beaty', 'biodiversity', 'museum')	280	('biodiversity', 'ecosystem', 'services')	429
5	('iucn', 'species', 'iyb')	235	('biodiversity', 'climate', 'change')	270	('climate', 'change', 'biodiversity')	423
	2013		2014		2015	
	Trigram	Count	Trigram	Count	Trigram	Count
1	('biodiversity', 'ecosystem', 'services')	542	('biodiversity', 'biodiversit', 'animals')	983	('biodiversity', 'human', 'london')	1312
2	('climate', 'change', 'biodiversity')	374	('biodiversit', 'animals', 'animaux')	920	('english', 'wikipedia', 'article')	1065
3	('biodiversity', 'heritage', 'library')	336	('animals', 'animaux', 'biodiversity')	587	('animals', 'animaux', 'biodiversity')	1060
4	('biodiversity', 'climate', 'change')	294	('animaux', 'biodiversity', 'biodiversit')	543	('animaux', 'biodiversity', 'biodiversit')	747
5	('environmentals', 'biodiversity', 'outreach')	293	('climate', 'change', 'biodiversity')	509	('twii', 'sittelle', 'twii')	611
	2016		2017		2018	
	Trigram	Count	Trigram	Count	Trigram	Count
1	('english', 'wikipedia', 'article')	2993	('english', 'wikipedia', 'article')	3310	('english', 'wikipedia', 'article')	2729
2	('biodiversity', 'climate', 'change')	528	('biodiversity', 'ecosystem', 'services')	554	('views', 'contribute', 'clicks')	2272
3	('biodiversity', 'ecosystem', 'services')	490	('climate', 'change', 'biodiversity')	514	('conservation', 'views', 'contribute')	2182
4	('climate', 'change', 'biodiversity')	423	('biodiversity', 'heritage', 'library')	461	('turn', 'pics', 'microstocka')	2153
5	('global', 'soil', 'biodiversity')	353	('pankaj', 'oudhia', 'cancer')	400	('funds', 'biodiversity', 'conservation')	1606

	2019		2020	
	Trigram	Count	Trigram	Count
1	('climate', 'change', 'biodiversity')	3104	('climate', 'change', 'biodiversity')	4016
2	('biodiversity', 'ecosystem', 'services')	1757	('change', 'biodiversity', 'loss')	2186
3	('change', 'biodiversity', 'loss')	1501	('global', 'biodiversity', 'framework')	1511
4	('climate', 'biodiversity', 'emergency')	1103	('post', 'global', 'biodiversity')	1222
5	('english', 'wikipedia', 'article')	1096	('biodiversity', 'climate', 'change')	1183

In this way, a color gradation was created on the heat map. However, in 2020, the weight of the topics became almost uniform due to the adoption of the MALLET model. Figure 10 shows a bar chart of the trend in the use of words of emotion in both years, and as mentioned earlier, the shape of the chart was similar in each year. At the same time, it is also possible to output and store the keywords and most representative tweet texts that make up each topic for both years.

Through the above series of processes, we were able to classify indiscriminately posted tweets of each year into a meaningful form. These processes enabled us to perform efficient qualitative analysis. Of course, it was possible to define all the topics from keywords and representative tweets, but we only examined the characteristic topics in this case.

Qualitative Analysis

Looking at the text of the tweets representing each topic, 2010, as mentioned earlier, was the International Year of Biodiversity, so naturally, the topic related to that was at the top of the list. However, as Fig. 6 shows the

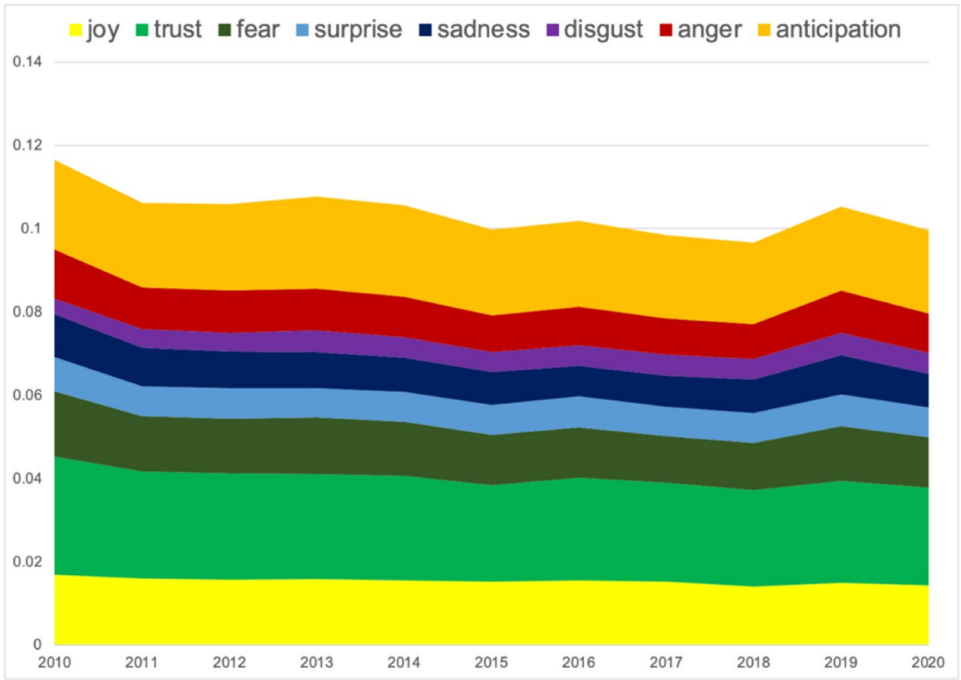

Fig. 3 Chart showing the usage rate of emotion words to total tweet text words from 2010 to 2020 (8 types of classification by NRC emotion lexicon)

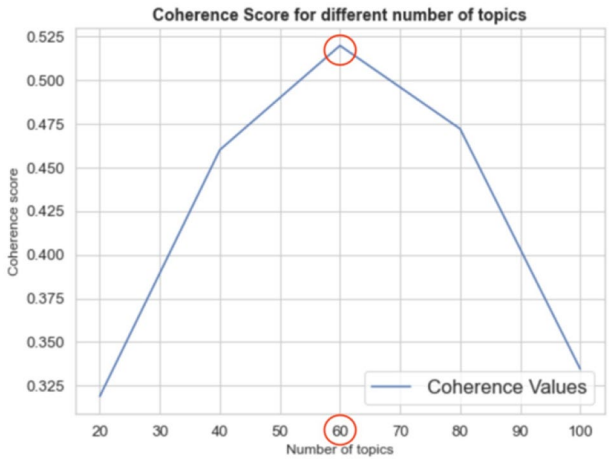

Fig. 4 Graph showing the optimal number of topics for 2010

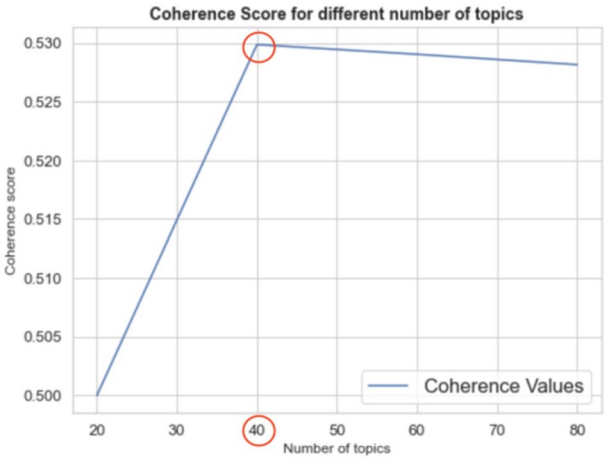

Fig. 5 Graph showing the optimal number of topics for 2020

distribution, many of the 60 topics overlapped and did not spread out. In addition, by overviewing the heat map (Fig. 8), in which the usage of eight types of emotion words can be recognized with a single glance, it became possible to refer to individual texts based on this. Table 3 shows a selection of topics that show significant characteristics in 2010, with Keywords and Representative Text displayed. For example, among the 60 topics, Topic 3 was where "joy" was prominent. Figure 11 shows this on a chart, with "trust" and "joy" being superior and a little "anticipation" standing out. Introductions dominated the textual content of the tweets that made up this topic to videos about biodiversity or observations of greening and biodiversity in private and public gardens.

On the contrary, Topic 6 is the one that shows the most negative trends in the heat map, namely "anger," "fear," and "sadness." Visualizing this on the chart, the shape of the chart was different from that of Topic 3. In addition, looking at the content of the tweet text that constituted this topic, the top posts expressed concern about human health and the negative effects of biodiversity loss on the human body, as shown in the Representative Text.

As an example of inference from the keywords, the chart shows that when looking for anything with "economy" in

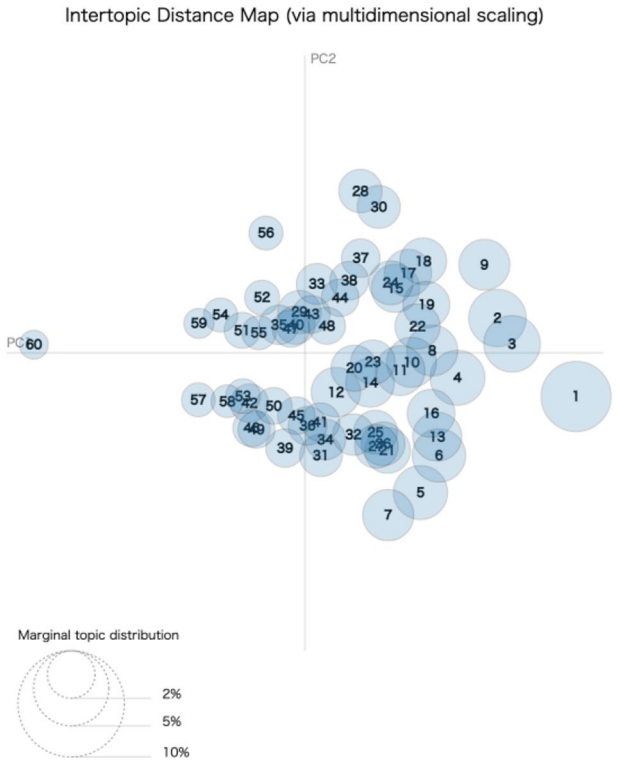

Fig. 6 Topic distribution in 2010 (output of 60 topics by pyLDAvis)

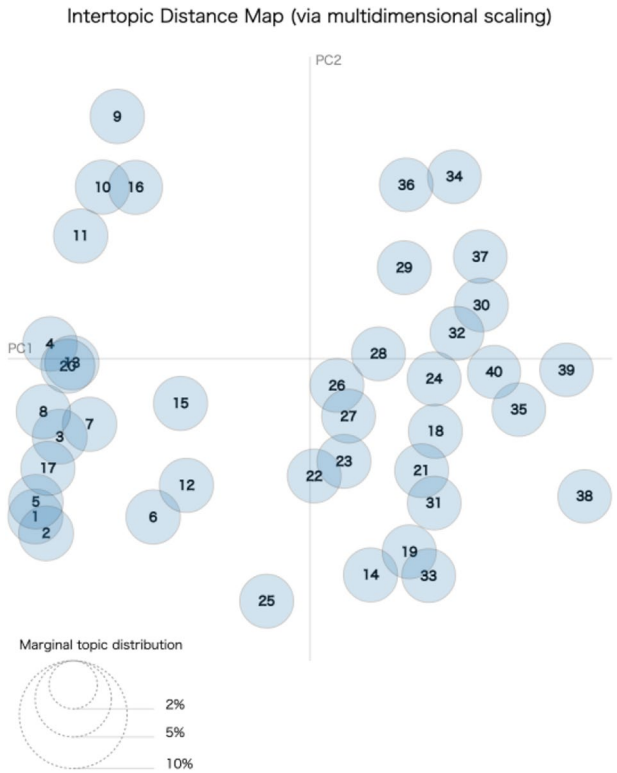

Fig. 7 Topic distribution in 2020 (output of 40 topics by pyLDAvis)

it, Topic 20 is applicable, with "trust" tending to stand out somewhat. The main content of the tweets that made up this topic was related to the conference in Nagoya, Japan. However, some tweets associated the loss of biodiversity with the economy.

Next, from the keywords of each topic, it could be inferred that Topic 25, which contains the word "target," is related to the international goals of biodiversity conservation. In the chart, "anticipation" was particularly prominent. An examination of the content of the tweet texts that make up this topic shows that there were references to the 2010 Biodiversity Target, which had not been met, and references to the awareness of "biodiversity" and news quotes and announcements about the newly developed targets. The original texts of some of the tweets are posted below.

World governments fail to halt biodiversity loss on 2010 targets. #Unreport

Shockingly, EU admits it has failed to reach the 2010 target to halt biodiversity loss:

How much do you know about biodiversity? Test yourself!

Press Release—Bold New Targets Needed to Halt Biodiversity Loss

UN biodiversity targets now need to be implemented say campaigners

UW prof: trade-offs necessary to reach biodiversity targets

The results of examining text content based on the 2010 topic model show that overall, tweets tended to be relatively short texts that announced events or awards, introduced brochures or videos, or tweeted links to news articles. These facts may be related to the limited popularity of Twitter as a social networking service in 2010 and the narrow user base and usage patterns. Furthermore, it should be noted that until 2017, there was a limit of 140 words in a tweet.

Figure 12 shows the percentage of tweets on the same topic group that was experimentally labeled with six types of emotion using the machine learning model "Distilbert-based-uncased-emotion." Note that this is an aggregate of labeling the highest-scoring emotion per tweet, and the visualization method is slightly different. However, even after considering this, the disproportionate number of tweets labeled "joy" for each topic. Even Topic 6, which shows the most negative sentiment trend in 2010, offers such a tendency; however, how to interpret this requires examination of more analysis.

2020 shows an almost identical to 2010 on the overall tweet sentiment chart. However, when looking at the details of individual topics, the content turns out to be completely different. Table 4 shows a selection of topics that show significant characteristics in 2020, with Keywords and Representative Text displayed. For example, out of the 40 topics, the chart for Topic 5, where the use of "fear" was relatively high on the heat map (Fig. 9), shows that "anger," "fear,"

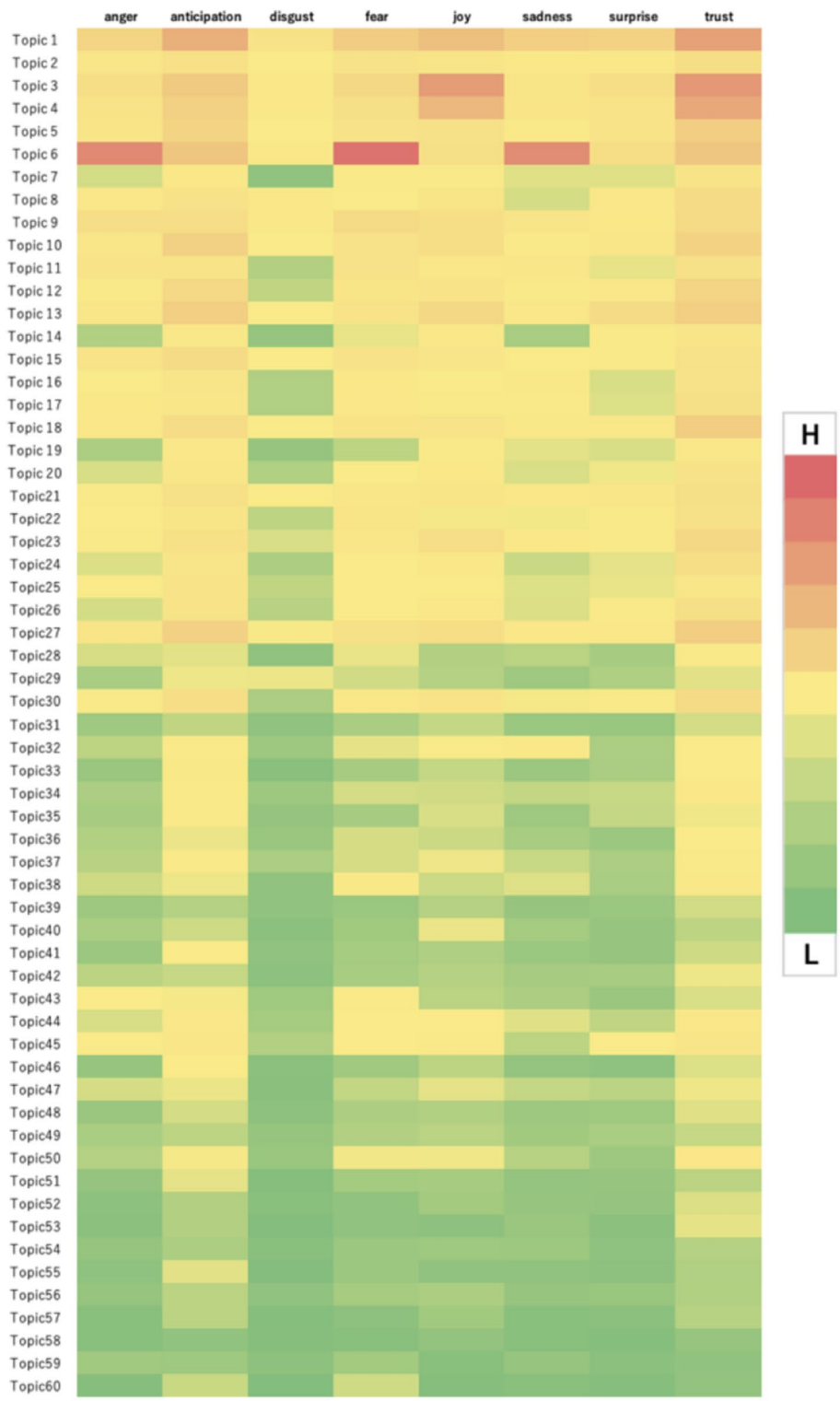

Fig. 8 Heatmap showing the total number of sentiment words per 60 topics in 2010

and "sadness" are prominent (Fig. 13). According to the keywords and the actual content of the tweet texts on this topic, there was a tendency for many tweets to warn of risks to the earth's resources, mainly due to population growth and harmful events (such as extreme weather and starvation) brought about by inaction.

Similarly, the chart of Topic 17, where the use of "fear" was relatively high, shows a certain number of "trust" and "anticipation" as well. Looking at the keywords and the contents of tweet texts, tweets that relate biodiversity loss to public health risks caused by global pandemics such as COVID-19, which showed the explosive spread of infection

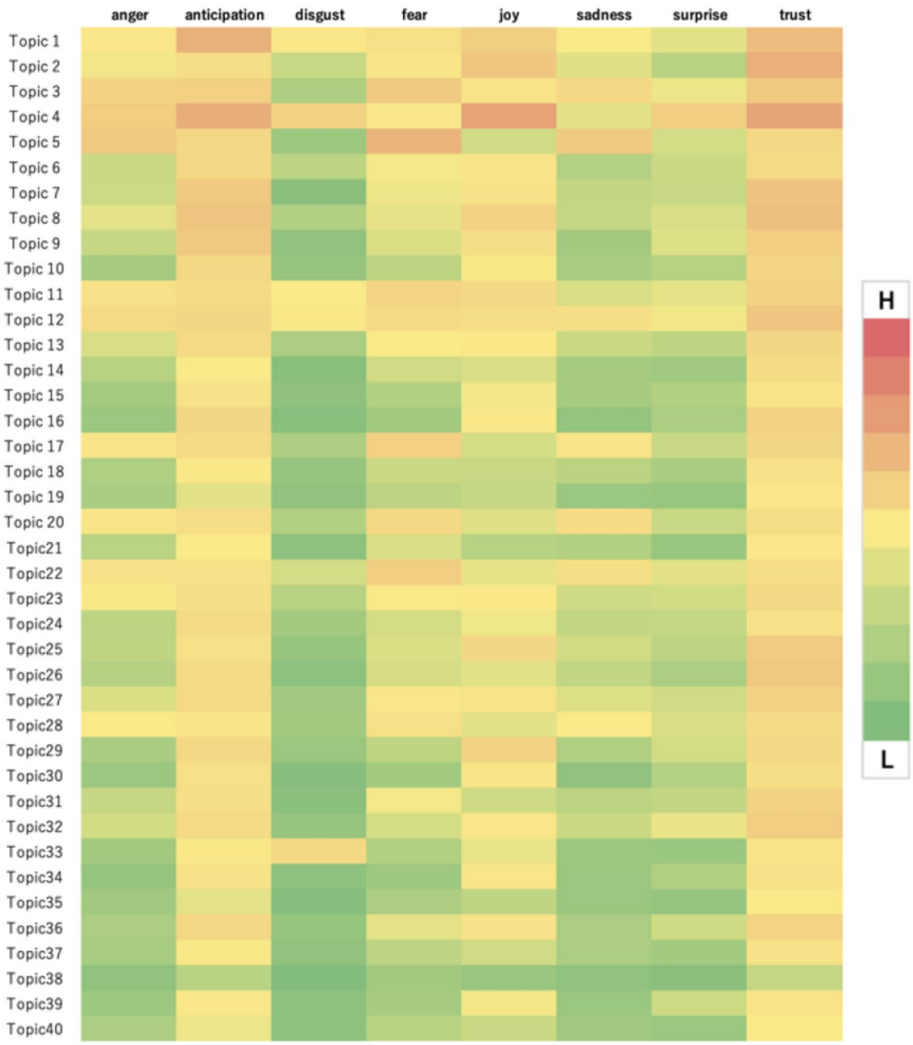

Fig. 9 Heatmap showing the total number of sentiment words per 40 topics in 2020

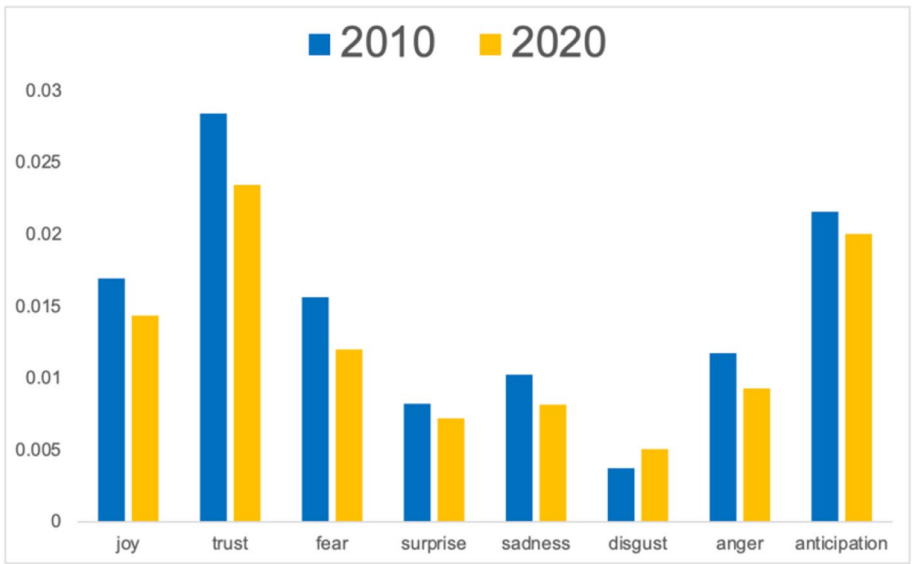

Fig. 10 Chart showing the usage rate of emotion words to total tweet text words in 2010 and 2020 (8 types of classification by NRC emotion lexicon)

Table 3 Characteristic topics for 2010 (Keywords, Representative Text)

Topic	Keywords	Representative text
3	Green, biodiversity, support, video, garden, come, eco, Africa, nature, challenge	rt SCB_SSWG Pythons in Florida Stalked by Hunters and Tourists Alike (NYT) #green #eco #nature #biodiversity #fb
6	Loss, biodiversity, human, intl, disappear, continue, provide, follower, film, term	Loss Of Biodiversity = End Of Human Race: -humans-are-rapidly-destroying-the-biodiversity-ne/
20	Biodiversity, thank, city, economy, wetland, Nagoya, lecture, get, cite, healthy	Brilliant! Permaculture in the City—#biodiversity #permaculture #growyourown
25	Biodiversity, target, know, plan, policy, source, halt, mean, damage, winner	What do u mean by biodiversity. What are d -do-u-mean-by-biodiversity-what-are-demerits-and-merits-of-biodiversity

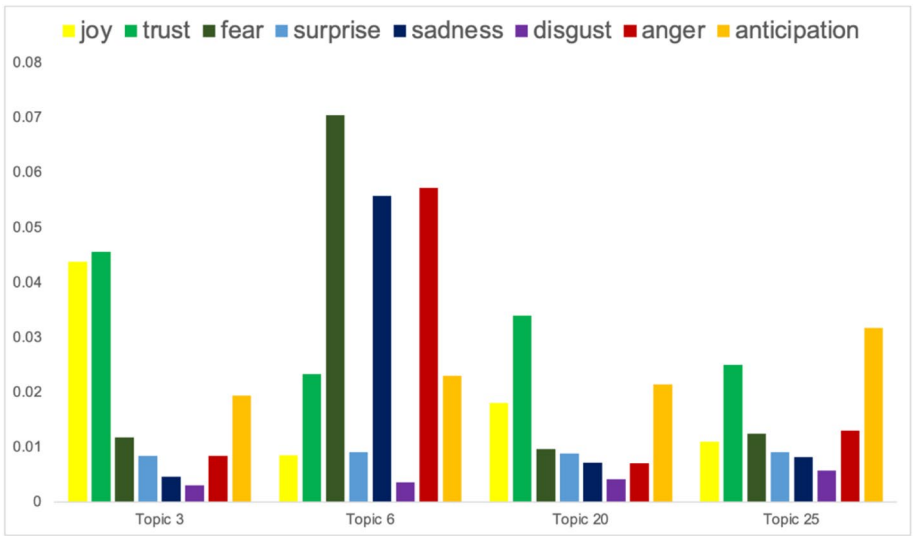

Fig. 11 Chart showing the usage rate of emotion words to total tweet text words for the characteristic topics in 2010 (8 types of classification by NRC emotion lexicon)

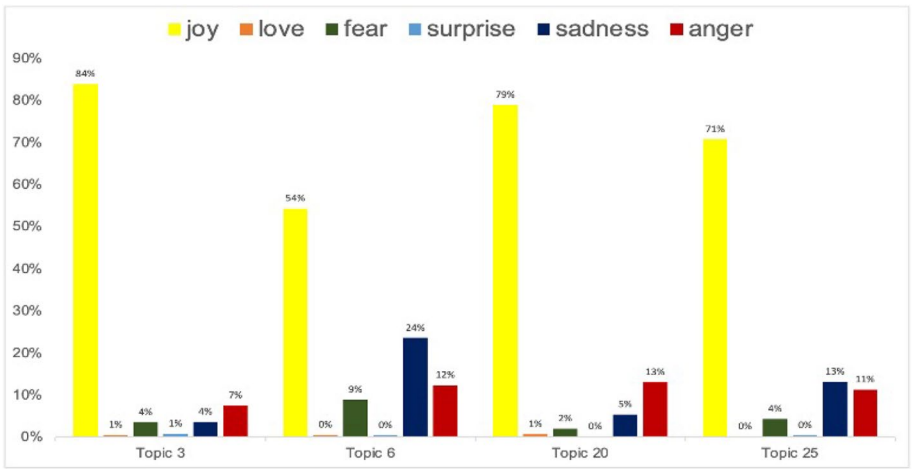

Fig. 12 Chart showing the percentage of each emotion-labeled tweet to the total number of tweets for the characteristic topics in 2010 (6 types of classification by Distilbert-based-uncased-emotion)

that year, stand out. In addition, the tweet text of Topic 22, which also had a large number of "fear," was about forest fires in Australia and other parts of the world.

On the contrary, judging from the keywords, it could be inferred that Topic 26 was related to global targets such as the Aichi Biodiversity Targets. The chart confirmed that words such as "trust" and "anticipation" were frequently used. Looking at the details of the tweet texts, most of the tweets were positive, pointing out the lack of achievement of the goals but explaining the need to set more ambitious goals in the future. The followings are some examples (original tweets).

Table 4 Characteristic topics for 2020 (Keywords, Representative Text)

Topic	Keywords	Representative text
5	Loss, global, threat, decline, risk, population, collapse, deforestation, big, lead	"Capping global warming at 2.7 degrees Fahrenheit would decrease the risk of ecosystem failures significantly, but allowing global warming to continue unchecked would lead to widespread biodiversity decline quickly"
17	Human, covid, pandemic, health, risk, future, prevent, disease, loss, link	How biodiversity loss is hurting our ability to combat pandemics via,#pandemics #covid #coronavirus #pandemic #staysafe #virus #healthcare #outbreak #quarantine #who #corona #lockdown #viruses #pandemicsurvival #cov #mask #cdc #stayhome #z
22	Biodiversity, stop, destroy, Australia, fire, destruction, damage, lose, continue, burn	Unfortunate: Huge Wildfire At Dzuko Valley At Manipur-Nagaland BorderThe massive fire is likely to have caused huge damage to biodiversity in Dzuko, also known as "the valley of the flowers".#wildfire #fire #firefighter #wildfires #firefighters #firefighting #fireseason
26	global, biodiversity, report, post, target, goal, framework, achieve, meet, decade	Last day of the thematic consultation on transparent implementation, monitoring, reporting and review for the post2020 Global Biodiversity Framework. Delays in NBSAP updating should not delay implementation of the post-2020 global biodiversity framework
33	Biodiversity, soil, healthy, diversity, life, ecosystem, protect, biodiversityday, health, matter	Keep soil alive, Protect soil Biodiversity Soil is essential to sustain all forms of life on Earth. Healthy soil can ensure a healthy & sustainable life. Let us aims to raise awareness of the importance of sustaining healthy ecosystems by protecting Soil Health. #WorldSoilDay

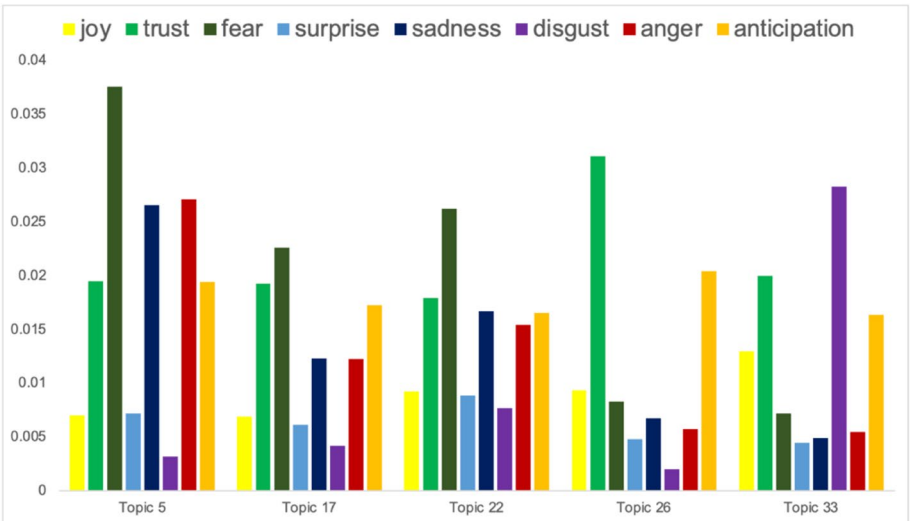

Fig. 13 Chart showing the usage rate of emotion words to total tweet text words for the characteristic topics in 2020 (8 types of classification by NRC emotion lexicon)

In 2010, country leaders gathered to set the Aichi Biodiversity Targets: a series of 10-year goals designed to preserve the world's biodiversity. At a global level, not a single target has been met, according to the UN Global Biodiversity Outlook report

A decade later, the world failed in meeting the ambitious Aichi 2020 Biodiversity targets. Some achievements reported, which is progress. But overall we maintained the bad situation and moved backwards in meeting some targets. We have 10 years left to meet the SDGs targets

The failure of the CBD 2010 Aichi biodiversity targets has shown just having a "vision" does not guarantee its fulfilment. The first draft for the post-2020 biodiversity framework looks bare when compared with the landmark Paris Agreement on climate change. Needs actions as well

The CBD Acting Executive Secretary now closing the OEWG 2 on a new global biodiversity framework. Interesting meeting, great opportunity to exchange views. Now these have to be narrowed down to ambitious, coherent set of goals and targets. Still much work ahead!

Compared to 2010, the tweet texts in 2020 tended to use "biodiversity" in diverse and specific contexts, and the texts tended to be somewhat longer. This phenomenon may be attributed to the tweet limit being raised to 280 words moreover the expansion of the user base with the spread of the Twitter platform over the past decade. However, it also suggests that the concept of "biodiversity" may be spreading, at least slightly in the Twitter space.

Figure 14 shows the percentage of tweets on the same topic group that was experimentally labeled with six types of emotion using the machine learning model "Distilbert-based-uncased-emotion." The visualization method was the same as in 2010, but the percentage of tweets labeled "joy" was significant for each topic. In contrast, negative sentiments such as "fear," "sadness," and "anger" were also labeled, suggesting that tweets with neutral sentiments could be labeled as "joy" tweets. Given the above pilot results, it is recommended that further learning and refinement are needed before practical analysis can be performed with this method.

Discussion

As described above, the purpose of this study was to present a rudimentary analysis of the Twitter discourse on the concept of "biodiversity" from 2010 to 2020, to explore the awareness of a large number of people, and to obtain clues to improve the efficiency of future international governance of biodiversity conservation. In this regard, although limited to one of the SNS spaces, the analysis method used in this study provides some insight into the topics and contexts in which the word and concept "biodiversity" was used and what other words occurred in association with them, and what emotional tendencies could be observed.

First of all, even by simply comparing the top *n*-grams from 2010 to 2020, it was possible to roughly estimate the changes in the discourse surrounding "biodiversity" on Twitter over the past 10 years. Next, to grasp the details, this study attempted to facilitate intuitive understanding through topic modeling, sentiment analysis, and visualization. Finally, it was shown that the dominant discourse of each year on Twitter could be inferred by identifying the representative tweets within each topic. In this paper, only the years 2010 and 2020 were mentioned due to the limited space. Nevertheless, it is an example of how the debate on "biodiversity" has converged in specific contexts over the past decade and how the discussion has become more sophisticated. In particular, this study also found that the number of tweets has shown a clear upward trend since 2018 and that the content of the tweets has always been more positive rather than pessimistic. In addition, it was suggested that discourse has already begun to look beyond 2020.

The other objective was to introduce several methods of NLP and show examples of analysis to gain new insights and present the possibility of developing the research. The analysis method used in this study has enabled us to classify the large group of tweets produced daily by the algorithm into topics with meaning and then interpret individual texts in more depth. Furthermore, it was confirmed that the visualization of the eight types of emotional tendencies of the NRC emotion lexicon, one of the Lexicon-Based Approaches, on multiple charts facilitates interpretation. In this way, it became possible to identify topics from heat maps and bar charts and qualitatively analyze keywords and specific text. Conversely, to examine keywords and text content beforehand and then use heat maps and bar graphs to infer emotional trends. The combination of sentiment analysis and topic modeling has been proposed in many studies, and there are various ways to visualize them. However, most of them still allocate them to positive, negative, or neutral polarity,

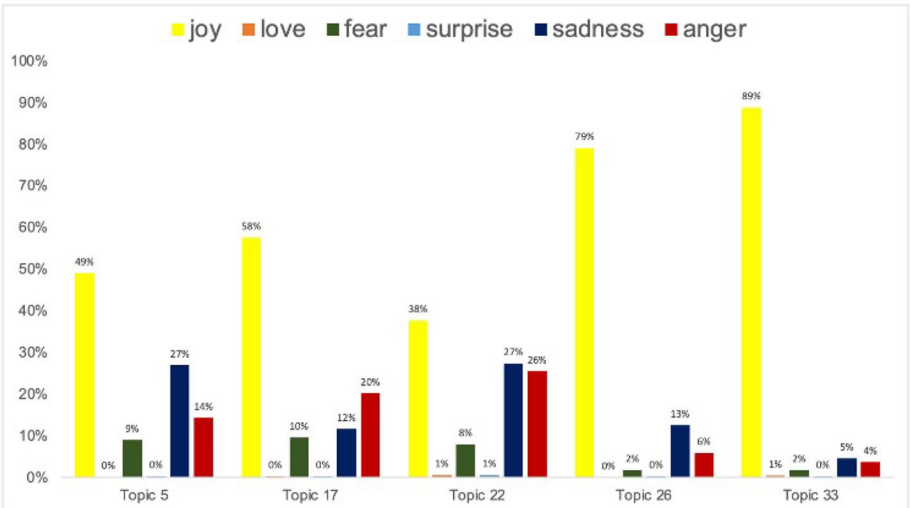

Fig. 14 Chart showing the percentage of each emotion-labeled tweet to the total number of tweets for the characteristic topics in 2020 (6 types of classification by Distilbert-based-uncased-emotion)

and not many give them to specific emotions. In addition, the results of the sentiment analysis of the machine learning model were also presented supplementally for a brief comparison with the lexicon-based model. The accumulation of research examples in this way would be meaningful for the future progress of analytical methods.

However, there are also some limitations to this study. First of all, as is generally the case with Twitter analysis, when pre-processing a large group of tweets by programming, it is customary to eliminate duplicates such as bots and remove as many superfluous items as possible (handle names, link URLs, images). However, it is impossible to remove those unnecessary elements altogether. For example, it frequently happens that a tweet that is a duplicate is not judged as a "duplicate" in programming terms but remains due to slight modifications by inserting spaces or adding other elements to the text. Several such examples were identified in this study due to visual confirmation. The impact of these cases on the final results needs to be examined separately. The improvement of analysis accuracy will depend on the future progress of data science and the skillfulness of individual analysts.

Secondly, regarding using the NRC emotion lexicon, the developer also pointed out that while it is a simple and powerful tool for analyzing text, the lexicon also poses the risk of inappropriate bias [21]. For example, among the 2020 topics, Topic 33, which stands out in the chart for its high number of "disgusts" among the others, required unique confirmation (Fig. 13). A closer examination of the tweet texts on this topic revealed that it was dominated by tweets emphasizing the importance of soil conservation and the theme of "WORLD SOIL DAY 2020" on December 5, 2020, "Keep soil alive, Protect soil Biodiversity." However, it turned out that the word "soil," which was frequently mentioned in this topic, and words such as "bacteria" and "fungi," which were used at the same time, were often classified as "disgust" in the NRC emotion lexicon. As seen in this case, the fact that the results may be contradictory to the specific contents of the tweet texts should also be kept in mind when using the NRC emotion lexicon. Although it was not modified in this study, careful customization of the lexicon is needed for individual analysis to minimize such cases. In addition, pilot results from DistilBERT's sentiment analysis suggested that the method requires more training and fine-tuning depending on the case in which it is used. Although BERT has attracted attention as a next-generation NLP method, and there have been many reports on its high accuracy (e.g., Chiorrini et al. [5]), only a few practical examples of analysis have yet to be presented. Careful research design and fine-tuning would be required when using this method for practical analysis.

This study implies that it may be possible to narrow down and classify large-scale text data using developing quantitative methods such as NLP and then provide deep insights through a qualitative analysis based on the knowledge and understanding of the analyst. Social networking sites are used not only for everyday purposes, such as expressing personal opinions and feelings, communication, and announcements, but also for planning and developing fictional stories, advertising and unique research by institutions and organizations, and even as a tool for propaganda [12]. On the other hand, it cannot be denied that it is also a subject of analysis that can provide many valuable suggestions. Therefore, the analysts themselves need to acquire a high literacy level in social networking to present more precise research results.

In the case of research exploring public awareness toward a particular concept, or policy, we should be cautious about placing too much faith in the results, no matter what method is used. However, since the analysis of SNS by NLP is a field that is constantly developing despite the many limitations, it is expected that the accuracy will be improved by using multiple analysis methods together and repeating trial and error. For example, when it comes to sentiment analysis, more accurate results may be obtained by comparing results using multiple different dictionaries or by introducing the analysis of Emoji, which was not included in this study. In addition, in this study, hashtags were deleted in the pre-processing stage of the topic model search, but there is room for future analysis that focuses on the hashtags used. Furthermore, research focusing on "like" and "reply" and examples of research focusing on geo-information is awaited. Again, in an unsupervised topic model, each topic needs to be scrutinized by the analyst from the results obtained. The predicted topic may not necessarily appear, and depending on that, the research purpose may not be fulfilled. Therefore, if the goal is to collect and analyze tweets related to a specific keyword or theme, exploring other possibilities, such as using another semi-supervised model, is recommended.

Conclusion

This study is one example of various methods applied to "biodiversity" and other concepts and keywords. More to the point, although this study only analyzed posts in purely English text, there is room to consider country- and region-specific research designs, such as conducting surveys of positions in other languages and of other primary language expressions of "biodiversity" (biodiversité, biodiversidad, biodiversität). Therefore, if the method is further refined through additional tests and applications by skilled researchers in the future, unexpected results may appear, and more knowledge may be obtained through more sophisticated and detailed analysis. Furthermore, synergistic effects of research in multiple fields can be obtained.

New goals have been developed for the COP15 of CBD, postponed from May 2020. In addition, The Vision 2050 on

biodiversity (living in harmony with nature) is still in progress. In the future, we need to formulate more practical and realistic goals and clarify the measures we should take on a national, regional, and individual basis. Whether the analysis of social networking sites will become more sophisticated in the future or will be a temporary fad depends mainly on the training and improvement of researchers' skills and the progress of data science. In any case, however, the increase and accumulation of exploratory research such as this study are essential for the efficiency of global environmental governance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42979-022-01276-w.

Acknowledgements The author would like to thank two anonymous reviewers for their valuable suggestions and comments on this paper. This work was supported by JST SPRING, Grant number JPMJSP2108.

Declarations

Conflict of interest The corresponding author states that there is no conflict of interest.

Human and animal rights This article does not contain any studies involving human participants performed by any authors.

References

1. Alizadeh, K. Limitations of Twitter Data Issues to be aware of when using Twitter text data In: towards data science. https://towardsdatascience.com/limitations-of-twitter-data-94954850cacf. Accessed 7 Sept 2021.
2. Blei DM, Andrew YN, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
3. Bruns A, Liang YE. Tools and methods for capturing Twitter data during natural disasters. First Monday 2012;17(4). https://doi.org/10.5210/fm.v17i4.3937.
4. Buchanan GM, et al. Assessment of national-level progress towards elements of the Aichi Biodiversity Targets. Ecol Ind. 2020;116:106497.
5. Chiorrini A, Diamantini C, Mircoli A, Potena D. Emotion and sentiment analysis of tweets using BERT. In: EDBT/ICDT Workshops. Dallas: Texas; 2021.
6. Ciotti M, Ciccozzi M, Terrinoni A, Jiang WC, Wang CB, Bernardini S. The COVID-19 pandemic. Crit Rev Clin Lab Sci. 2020;57(6):365–88.
7. Convention on Biological Diversity. https://www.cbd.int/. Accessed 30 Aug 2021.
8. Cooper MW, et al. Developing a global indicator for Aichi Target 1 by merging online data sources to measure biodiversity awareness and engagement. Biol Conserv. 2019;230:29–36.
9. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint arXiv:1810.04805.
10. Fink C, Hausmann A, Di Minin E. Online sentiment towards iconic species. Biol Conserv. 2020;241: 108289.
11. Glowka L, Burhenne-Guilmin F, Synge H, McNeely JA, Gündling L. A guide to the convention on biological diversity. International Union for the Conservation of Nature (IUCN). Gland, Switzerland; 1994.
12. Guarino S, et al. Characterizing networks of propaganda on twitter: a case study. Appl Netw Sci. 2020;5(1):1–22.
13. Herkenrath P, Harrison J. The 10th meeting of the Conference of the Parties to the Convention on Biological Diversity—a breakthrough for biodiversity? Oryx. 2011;45(1):1–2.
14. Hirsch T, Mooney K, Cooper D. Global biodiversity outlook 5. Secretariat of the Convention on Biological Diversity; 2020.
15. Kharde V, Sonawane P. Sentiment analysis of twitter data: a survey of techniques. 2016. arXiv preprint arXiv:1601.06971.
16. Mohammad SM, Turney PD. Nrc emotion lexicon. National Research Council; 2013.
17. Mohammad SM. Practical and ethical considerations in the effective use of emotion and sentiment lexicons. 2020. arXiv preprint arXiv:2011.03492.
18. Morshed SA, et al. Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis. J Urban Manage. 2021;10(2):155–65. https://doi.org/10.1016/j.jum.2021.03.002
19. Otero P, Gago J, Quintas P. Twitter data analysis to assess the interest of citizens on the impact of marine plastic pollution. Mar Pollut Bull. 2021;170:11262.
20. Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. LREc. 2010;10:2010.
21. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on web search and data mining. 2015. pp. 399–408. https://doi.org/10.1145/2684822.2685324.
22. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. arXiv preprint arXiv:1910.01108.
23. Sarkar S. Biodiversity and environmental philosophy: an introduction. Cambridge University Press; 2005.
24. statista. https://www.statista.com/topics/737/twitter/. Accessed 29 Aug 2021.
25. Union for Ethical BioTrade UEBT Biodiversity Barometer 2018. https://static1.squarespace.com/static/577e0feae4fcb502316dc547/t/5b51dbaaaa4a99f62d26454d/1532091316690/UEBT+-+Baro+2018+Web.pdf. Accessed 30 Aug 2021.
26. Xu H, et al. Ensuring effective implementation of the post-2020 global biodiversity targets. Nat Ecol Evol. 2021;5(4):411–8.
27. Xue J, Chen J, Gelles R. Using data mining techniques to examine domestic violence topics on Twitter. Violence Gend. 2019;6(2):105–14.
28. Xue J, et al. Public discourse and sentiment during the COVID 19 pandemic: using latent Dirichlet Allocation for topic modeling on Twitter. PLoS ONE. 2020;15(9):e0239441.