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 (s-ohtani@g.ecc.u-tokyo.ac.jp)
The University of Tokyo https://orcid.org/0000-0001-8414-9248

Research Article

Keywords: biodiversity, awareness, Twitter, n-gram, sentiment analysis, topic modeling

Posted Date: January 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-922908/v3

Version of Record: A version of this preprint was published at SN Computer Science on July 15th, 2022. See the published version at https://doi.org/10.1007/s42979-022-01276-w.

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, in combination with sentiment analysis and topic modeling, which are 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, the construction of topic models using Latent Dirichlet allocation (LDA), and the qualitative analysis of tweet texts based on these models, I was able to classify and analyze unstructured tweets in a meaningful way. 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 the search for people's awareness through SNS analysis still has many limitations, it is undeniable that important suggestions can be obtained. In order to further refine the research method, it will be essential to improve the skills of analysts and accumulate research examples as well as to advance data science.

Introduction

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. The Convention is one of the countless international frameworks that have been developed to address the serious deterioration of the global environment as a result of global concern. Since then, as of September 2021, 14 sessions of the Conference of the Parties (COP) have been held and two protocols have been finalized. 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 furthermore, the Aichi Biodiversity Targets, a comprehensive approach to biodiversity conservation, were adopted.

The 15th Conference of the Parties (COP15) to CBD, which was scheduled to be held in Kunming, China in October 2020, has been postponed to October 2021 due to the aftermath of the global spread of a new coronavirus (COVID-19). Furthermore, it was recently announced that the conference will be held in two parts, with the second half to be held in April 2022. Originally, 2020 was also the final year of the Aichi Biodiversity Targets and the “Mission” of The Strategic Plan for Biodiversity 2011-2020, which were 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. Various factors have been pointed out as the cause of this. Some studies have pointed to a variety of factors, such as deficiencies in global environmental governance, or individual national circumstances such as inadequate legal systems or other impediments. However, even though fragmented and localized analysis has been done, it has been a challenge to create a global and comprehensive method of analysis (Xu et al. 2021).

It is important to note that for all biodiversity conservation measures to work, they must be based on the full awareness and action of individual private citizens, not just politicians and experts. And 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, but 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”. On top of that, the Union of Ethical Bio Trade’s Biodiversity Barometer survey of the general public in 16 countries provides mainly demographic data on understanding of biodiversity (UEBT 2018), however, 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, but could also be used as a tool 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 (Cooper et al. 2019).

This study focused on the usefulness of Twitter data, and proposed a different research method. Currently, discourse on social media, including Twitter, is becoming increasingly popular as a place for many people to express their emotions and opinions, and it has been reported that these analyses are useful in various fields as a method of opinion mining (Pak & Paroubek 2010). This study attempted to explore the raising of awareness of "biodiversity" over time using the discourse space of Twitter as a subject. 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, in order to provide clues for streamlining future international governance of biodiversity conservation. At the same time, this study proposed an exploratory study aimed at exploring its usefulness and showing its potential to develop using several methods of natural language processing (NLP).

Contributions

This paper aims to make several academic contributions. First, it proposes a research methodology with a view to improving the efficiency of global environmental policy. In the field of environmental policy and other global policy issues, there are still many methods that have not been tried. It is an urgent task to introduce some of such methods into this field and to question their versatility. Second, it is necessary to show the possibility of applying the analysis methods of natural language processing (NLP) to various fields, and to demonstrate their usefulness and vulnerability. This will be an inevitable path for the future development of NLP. To this end, it is also necessary to test whether NLP can withstand several years of diachronic research.

For this reason, the novelty of this paper lies not in the presentation of a general data analysis methodology itself, but rather in its application. This is part of the challenge to create a more precise indicator of how people's posts on Twitter can be used as "globally consistent information showing trends in awareness and willingness to act on biodiversity". The methodology and results presented in this paper may serve as a touchstone for future improvements in both global environmental policy research and NLP.

The structure of this paper is as follows: Methodology provides details of the subject of analysis and the research methodology, Results details the results obtained, Discussion provides a discussion based on the results, and Conclusion presents the challenges and possible developments as well as conclusions.

Methodology

Research design

In this study, I selected tweets posted on Twitter as the object of analysis. Prior to the analysis, 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 all over the world 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 those of heads of state and dignitaries, and 330 million monthly active users, who post 500 million messages every day (statista 2021). Expressions of emotions and sentiments, such as opinions expressed on this platform, as well as 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 (Bruns & Liang, 2012). Thus, analysis of social networking sites is being applied in various fields such as market research, product reviews, traffic prediction, among which Twitter analysis is being used in many fields 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 sharp increase in the number of studies on Twitter posts about the global pandemic of COVID-19. This is because it is believed that Twitter posts can provide a lot of useful information in exploring public sentiments and concerns about an infectious disease that has rapidly spread around the world and changed people's lives (Xue et al. 2020). And this can also be applied to find out what people in 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 important to be careful about equating the discourse space on Twitter with the actual discourse space. This is 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. Therefore, when analyzing posts on Twitter and drawing certain conclusions, we should keep in mind that it is 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 National Research Council Canada (Mohammad et al. 2013), topic modeling using Latent dirichlet allocation (LDA) (Blei et al. 2003), and qualitative analysis of tweet texts. The outline of the research procedure is as follows:

1. I collected all tweets containing the keyword "biodiversity" from March 2006, when the Twitter service was launched, to December 2020, and extracted tweets that were purely in English.

2. I pre-processed all tweets from 2010 to 2020, counted n-grams (bigram and trigram), and listed the top 20.

3. In the same way, I 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, I explored LDA topic models and constructed the model that seemed to be optimal. In this paper, I discussed the models for the years 2010 and 2020 through visualization.

5. Based on the visualized information, I selected distinctive topics and created another sentiment charts to examine the contents of the texts.

6. For some distinctive topics, I provided an overview of their tweet texts.

In addition, this study aimed to grasp the whole picture of the discourse surrounding "biodiversity" posted on Twitter. Therefore, I did not take into account the nationality, title, and other attributes of the sender.

Data collection

In this study, I collected tweets containing the word “biodiversity” from March 21, 2006, when the Twitter service started, to December 31, 2020. The purpose of this study was to explore the usage of “biodiversity” in normal contexts as well as in the hashtag “#biodiversity”. For the collection, I first applied for the Academic Research product truck released by Twitter in January 2021 and obtained access to the full Twitter archive. Then, I used the open application programming interface (API) provided by Twitter and the Python programming language (ver. 3.8.8). As a result, the total number of tweets by December 31, 2020, is 2,609,834, which is outstanding compared to other major 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, I 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, I focused on “text”. This is because the purpose of this study is to understand the general speech on Twitter, and the attributes and location information of posters are out of the scope of this study, and also because there were many tweets with missing information. **Fig.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 increase in the number of tweets has clearly become larger.

Pre-processing the raw dataset

Prior to analysis, it was necessary to pre-process Twitter raw data into a form suitable for analysis. In this study, I followed a number of 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 was saved for sentiment analysis.

7. Also removed hashtagged words.

8. Tokenized the texts. Deleted tweets with less than 3 tokens.

9. n-grams (bigrams and trigrams) were counted and saved.

10.Performed lemmatization of the tokens.

Data analysis

In this study, I used both quantitative and qualitative research. First, as a quantitative study, I overviewed the data through visualization using LDA topic modeling and sentiment analysis, and second, I 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

In this study, I counted n-grams for each year to get an overview of the set of tweet texts that were narrowed down by pre-processing as described above. An n-gram is a sequence of words, where two words are called a bigram and three words are called a trigram. Here, I used Gensim, which is available in Python. By comparing the top-ranked n-grams, it was possible to understand more specific keywords that appear in each text. And 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. In this paper, the top 20 bigram and trigram terms from 2010 to 2020 were listed for comparison and discussion.

Sentiment Analysis

Sentiment analysis is defined as an automated process of mining attitudes, opinions, views, and emotions from text, speech, tweets, and database sources using natural language processing (NLP), and is said to be the process of analyzing people’s feelings, attitudes, opinions, and emotions towards elements such as products, individuals, topics, organizations, and services (Kharde & Sonawane 2016). There has been a rapid increase in the number of examples of analysis of social networking sites, the most popular of which is the categorization into Positive, Negative, and Neutral. However, in recent years, various methods such as Machine Learning Approaches, Lexicon-Based Approaches. have been devised and are showing rapid development.

In this study, I used Lexicon-Based Approaches, among which is a method called Dictionary-based. I used the NRC emotion lexicon (Mohammad et al. 2013). 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 (Mohammad 2020). In my preliminary research, I found that the categorization of positive, negative, and neutral was extremely abstract and subject to wide swings, ultimately forcing me to carefully read and examine specific texts. Thus, I decided to read eight types of emotions from the tweet texts, as I needed to clarify the direction of more specific emotions.

According to the developer of the NRC emotion lexicon, the lexicon works by comparing multiple sets of data and by producing a percentage of the total number of words (Mohammad 2020). In this study, I searched for emotions expressed in tweets as a whole and in 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 word “biodiversity”. And I excluded the years from 2006 to 2009, when the total number of tweets per year was small, and calculated the total number of words by using Python, and the percentage of words constituting each emotion in the preprocessed tweet texts for each year from 2010 to 2020, and visualized them.

Latent Dirichlet allocation（LDA）

For the purpose of this study, which is to explore the discourse on “biodiversity” on Twitter, it was essential to explore the topics of each year as discussed by users. In this study, I used Latent Dirichlet allocation (LDA) as a method for this purpose (Blei 2003). LDA is a form of Unsupervised machine learning, in which 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 words, potential topics in a document and their distributions over those topics based on the data itself (Xue 2019). 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.

In this study, I 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. And the best topic models were explored, constructed, and visualized. However, for reasons of paper space, I used NRC emotion lexicon to count and calculate the percentage of emotion words in the modeled topics for 2010 and 2020 only, and the visualization results were presented for comparison and analysis.

Qualitative Analysis

After categorizing, visualizing, and interpreting the data through quantitative research, it would be beneficial to conduct specific analysis through qualitative research. In this study, I also extracted a set of keywords that constituted each topic when I 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, and the original text of the tweets in the highest range was posted as Representative Text.

In this way, it was possible to infer the dominant discourse by identifying and examining particularly important texts from a large number of 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, I focused on only a few distinctive topics and examined the tweet contents that constituted them. By using the above method, it would be possible to roughly grasp the dominant discourse of each year in the Twitter space.

Results

As a result of the pre-processing, the number of tweets to be analyzed was shown in **Fig.2**: out of 2,389,197 tweets in English 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 20 bigrams from 2010 to 2020. Observing some of them for each year, 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 2,785 in 2010, and bigram (‘biodiversity’, ‘conservation’) was 10,852 in 2020 compared to 1,358 in 2010, both of which were significant increases. As for other idioms, “climate change” was also at the top of the list, which clearly shows 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, there were 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 Environmentals Biodiversity Outreach Officer jobs. 2013 saw a noticeable increase in the number of tweets echoing the publication of “Biodiversity offsets in theory and practice”, which was published during the year. In 2015, the word “human” was used prominently along with “biodiversity. This was due to the publication of “Connecting Global Priorities: Biodiversity and Human Health” by World Health Organization (WHO) and a surge in tweets mentioning it. Also, from 2015 to 2018, “wikipedia”, “article”, “english”, and “edited” had been among the top terms. This was due to the fact that there was a major update work done on the words related to biodiversity in Wikipedia, and this was tweeted verbatim by certain account. In 2019 and 2020, “crisis,” which was not used as often in previous years, was found to be increasingly used with “biodiversity”.

Table 1 Top 20 bigrams from 2010 to 2020

	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	('climate', '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
6	('biodiversity', 'summit')	1000	('nature', 'biodiversity')	978	('biodiversity', 'outreach')	1377	('wildlife', 'biodiversity')	886
7	('iyb', 'biodiversity')	940	('eco', 'biodiversity')	791	('environmentals', 'biodiversity')	1375	('biodiversity', 'ecosystem')	886
8	('loss', 'biodiversity')	915	('wildlife', 'biodiversity')	782	('outreach', 'officer')	1375	('ecosystem', 'services')	863
9	('save', 'biodiversity')	897	('wildlife', 'conservation')	699	('global', 'biodiversity')	1317	('protect', 'biodiversity')	839
10	('global', 'biodiversity')	882	('environment', 'biodiversity')	605	('nature', 'biodiversity')	1008	('environment', 'biodiversity')	807
11	('marine', 'biodiversity')	802	('biodiversity', 'nature')	589	('loss', 'biodiversity')	973	('conservation', 'biodiversity')	800
12	('international', 'biodiversity')	782	('protect', 'biodiversity')	553	('protect', 'biodiversity')	952	('biodiversity', 'environment')	790
13	('ecosystems', 'biodiversity')	704	('decade', 'biodiversity')	550	('conservation', 'biodiversity')	871	('international', 'biodiversity')	786
14	('biodiversity', 'talks')	684	('biodiversity', 'environment')	527	('biodiversity', 'project')	845	('nature', 'biodiversity')	785
15	('biodiversity', 'conference')	639	('loss', 'biodiversity')	496	('biodiversity', 'environment')	827	('global', 'biodiversity')	741
16	('protect', 'biodiversity')	606	('ecosystem', 'services')	489	('environment', 'biodiversity')	805	('loss', 'biodiversity')	718
17	('world', 'biodiversity')	548	('global', 'biodiversity')	469	('ecosystem', 'services')	751	('world', 'biodiversity')	624
18	('biodiversity', 'iyb')	548	('biodiversity', 'strategy')	467	('wildlife', 'biodiversity')	732	('biodiversity', 'hotspot')	622
19	('nature', 'biodiversity')	530	('ecosystems', 'biodiversity')	461	('biodiversity', 'ecosystem')	704	('biodiversity', 'biodiversit')	585
20	('eco', 'biodiversity')	503	('international', 'biodiversity')	430	('biodiversity', 'nature')	683	('water', 'biodiversity')	575

	2014		2015		2016		2017	
	bigram	count	bigram	count	bigram	count	bigram	count
1	('biodiversity', 'conservation')	3345	('biodiversity', 'human')	23180	('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
6	('biodiversity', 'offsetting')	1716	('biodiversity', 'biodiversit')	1775	('biodiversity', 'human')	2878	('biodiversity', 'loss')	2959
7	('marine', 'biodiversity')	1596	('animals', 'animaux')	1588	('biodiversity', 'loss')	2562	('marine', 'biodiversity')	1989
8	('global', 'biodiversity')	1147	('protect', 'biodiversity')	1315	('marine', 'biodiversity')	1806	('conservation', 'biodiversity')	1480
9	('protect', 'biodiversity')	1092	('human', 'london')	1312	('conservation', 'biodiversity')	1569	('protect', 'biodiversity')	1459
10	('environment', 'biodiversity')	1033	('conservation', 'biodiversity')	1243	('protect', 'biodiversity')	1550	('biodiversity', 'story')	1378
11	('biodiversit', 'animals')	995	('wikipedia', 'article')	1074	('global', 'biodiversity')	1380	('nature', 'biodiversity')	1372
12	('conservation', 'biodiversity')	925	('animaux', 'biodiversity')	1069	('nature', 'biodiversity')	1355	('global', 'biodiversity')	1190
13	('biodiversity', 'ecosystem')	867	('edited', 'biodiversity')	1067	('wildlife', 'biodiversity')	1037	('biodiversity', 'ecosystem')	1134
14	('loss', 'biodiversity')	851	('english', 'wikipedia')	1065	('biodiversity', 'ecosystem')	997	('world', 'biodiversity')	1070
15	('world', 'biodiversity')	851	('biodiversity', 'hotspot')	957	('soil', 'biodiversity')	994	('wildlife', 'biodiversity')	1033
16	('biodiversity', 'climate')	783	('nature', 'biodiversity')	955	('environment', 'biodiversity')	985	('loss', 'biodiversity')	1032
17	('biodiversity', 'story')	782	('world', 'biodiversity')	950	('biodiversity', 'nature')	961	('environment', 'biodiversity')	1020
18	('wildlife', 'biodiversity')	759	('biodiversity', 'ecosystem')	913	('loss', 'biodiversity')	959	('biodiversity', 'hotspot')	968
19	('biodiversity', 'hotspot')	756	('environment', 'biodiversity')	903	('biodiversity', 'hotspot')	943	('ecosystem', 'services')	917
20	('ecosystem', 'services')	752	('biodiversity', 'story')	894	('biodiversity', 'climate')	923	('rich', 'biodiversity')	863

	2018		2019		2020	
	bigram	count	bigram	count	bigram	count
1	('biodiversity', 'conservation')	9701	('climate', 'change')	18253	('biodiversity', 'loss')	20274
2	('climate', 'change')	7717	('biodiversity', 'loss')	14396	('climate', 'change')	18945
3	('biodiversity', 'loss')	6951	('biodiversity', 'conservation')	8234	('biodiversity', 'conservation')	10852
4	('marine', 'biodiversity')	2767	('climate', 'biodiversity')	6034	('climate', 'biodiversity')	7412
5	('funds', 'biodiversity')	2762	('loss', 'biodiversity')	6024	('protect', 'biodiversity')	6584
6	('english', 'wikipedia')	2730	('biodiversity', 'crisis')	4356	('loss', 'biodiversity')	6153
7	('wikipedia', 'article')	2730	('nature', 'biodiversity')	4024	('nature', 'biodiversity')	6085
8	('edited', 'biodiversity')	2729	('protect', 'biodiversity')	3919	('global', 'biodiversity')	5630
9	('nature', 'biodiversity')	2676	('global', 'biodiversity')	3349	('biodiversity', 'crisis')	5021
10	('conservation', 'biodiversity')	2581	('change', 'biodiversity')	3334	('environment', 'biodiversity')	4629
11	('loss', 'biodiversity')	2512	('biodiversity', 'ecosystem')	3092	('biodiversity', 'climate')	4555
12	('protect', 'biodiversity')	2427	('biodiversity', 'climate')	2998	('change', 'biodiversity')	4300
13	('conservation', 'views')	2406	('marine', 'biodiversity')	2937	('world', 'environment')	4062
14	('views', 'contribute')	2371	('ecosystem', 'services')	2832	('biodiversity', 'nature')	3930
15	('contribute', 'clicks')	2286	('wildlife', 'biodiversity')	2786	('rich', 'biodiversity')	3736
16	('biodiversity', 'nature')	2208	('conservation', 'biodiversity')	2715	('wildlife', 'biodiversity')	3527
17	('wildlife', 'biodiversity')	2198	('climatechange', 'biodiversity')	2379	('marine', 'biodiversity')	3421
18	('global', 'biodiversity')	2191	('environment', 'biodiversity')	2332	('conservation', 'biodiversity')	3294
19	('turn', 'pics')	2188	('million', 'species')	2298	('protecting', 'biodiversity')	3116
20	('pics', 'microstocka')	2153	('biodiversity', 'nature')	2281	('world', 'biodiversity')	2993

Also shown in **Table 2** are the top 20 trigrams, 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 due to the fact that many English tweets were tagged with French hashtags. And it was found that in 2015 and 2016, a number of 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, but this was due to the simple fact that there were many tweets warning that various species on the planet are on the risk of extinction.

Looking at the increase of trigrams during the past decade, trigram ('climate', 'change', 'biodiversity') was 362 in 2010 and 4,016 in 2020. The number of trigram ('biodiversity', 'ecosystem', 'services') is 1,069 in 2020 compared to 185 in 2010, both of which are significant increases. Thus, it is possible to obtain certain suggestions even by observing only the number of bigrams and trigrams.

Table 2 Top 20 trigrams from 2010 to 2020

	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
6	('biodiversity', 'climate', 'change')	225	('biodiversity', 'heritage', 'library')	258	('biodiversity', 'climate', 'change')	304
7	('biodiversity', 'media', 'alliance')	223	('biodiversity', 'ecosystem', 'services')	227	('forests', 'biodiversity', 'declining')	302
8	('help', 'save', 'biodiversity')	205	('pollution', 'ecosystem', 'biodiversity')	195	('global', 'biodiversity', 'percent')	293
9	('ecosystems', 'biodiversity', 'climatechange')	199	('button', 'biodiversity', 'project')	183	('biodiversity', 'percent', 'years')	290
10	('beaty', 'biodiversity', 'museum')	185	('food', 'security', 'biodiversity')	159	('biodiversity', 'heritage', 'library')	279
11	('biodiversity', 'ecosystem', 'services')	185	('economics', 'ecosystems', 'biodiversity')	149	('international', 'biological', 'diversity')	264
12	('brands', 'help', 'save')	177	('halt', 'biodiversity', 'loss')	144	('climate', 'change', 'pollution')	255
13	('jason', 'clay', 'brands')	165	('nature', 'wildlife', 'biodiversity')	139	('beaty', 'biodiversity', 'museum')	247
14	('nagoya', 'biodiversity', 'summit')	164	('protect', 'palawan', 'biodiversity')	137	('biodiversity', 'could', 'casualty')	204
15	('environment', 'ecosystems', 'biodiversity')	161	('global', 'biodiversity', 'loss')	136	('effects', 'biodiversity', 'loss')	201
16	('halt', 'biodiversity', 'loss')	152	('biodiversity', 'unesco', 'sites')	134	('report', 'global', 'biodiversity')	197
17	('declared', 'international', 'year')	146	('unesco', 'sites', 'mining')	134	('ecosystem', 'effects', 'biodiversity')	189
18	('loss', 'biodiversity', 'ecosystems')	134	('palawan', 'biodiversity', 'unesco')	134	('nature', 'wildlife', 'biodiversity')	183
19	('biodiversity', 'daily', 'read')	132	('help', 'protect', 'palawan')	132	('worse', 'climate', 'change')	182
20	('daily', 'read', 'newspaper')	131	('plummets', 'despite', 'growth')	132	('biodiversity', 'crisis', 'worse')	179

	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
6	('biodiversity', 'outreach', 'officer')	293	('biodiversity', 'ecosystem', 'services')	445	('biodiversity', 'biodiversit', 'animals')	490
7	('outreach', 'officer', 'jobs')	293	('biodiversity', 'climate', 'change')	413	('nabhan', 'food', 'biodiversity')	483
8	('climate', 'water', 'land')	270	('birds', 'oiseaux', 'biodiversity')	378	('biodiversity', 'ecosystem', 'services')	463
9	('water', 'land', 'biodiversity')	269	('biodiversity', 'heritage', 'library')	366	('food', 'biodiversity', 'prudent')	455
10	('international', 'biological', 'diversity')	267	('oiseaux', 'biodiversity', 'biodiversit')	361	('biodiversity', 'climate', 'change')	453
11	('harms', 'climate', 'water')	237	('global', 'biodiversity', 'targets')	237	('biodiversity', 'prudent', 'hedging')	450
12	('food', 'waste', 'harms')	214	('international', 'biological', 'diversity')	224	('biodiversity', 'heritage', 'library')	449
13	('waste', 'harms', 'climate')	213	('beaty', 'biodiversity', 'museum')	215	('gary', 'nabhan', 'food')	430
14	('beaty', 'biodiversity', 'museum')	195	('manu', 'national', 'park')	186	('climate', 'change', 'biodiversity')	409
15	('biodiversity', 'biodiversit', 'animals')	172	('biodiversity', 'policy', 'practice')	184	('heianaturen', 'heianaturen', 'heianaturen')	401
16	('biodiversit', 'animals', 'animaux')	164	('national', 'biodiversity', 'teach')	178	('biodiversit', 'animals', 'animaux')	389
17	('nature', 'wildlife', 'biodiversity')	161	('meet', 'conservation', 'targets')	146	('farming', 'benefits', 'biodiversity')	389
18	('happy', 'international', 'biodiversity')	155	('fight', 'climate', 'change')	145	('organic', 'farming', 'benefits')	385
19	('spark', 'broad', 'biodiversity')	138	('biodiversity', 'report', 'highlights')	143	('prudent', 'hedging', 'strategy')	379
20	('broad', 'biodiversity', 'loss')	138	('iisd', 'biodiversity', 'policy')	141	('hedging', 'strategy', 'dealing')	377

	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
6	('biodiversity', 'heritage', 'library')	313	('oudhia', 'cancer', 'drug')	400	('clickasnap', 'getting', 'funds')	1406
7	('nabhan', 'food', 'biodiversity')	298	('cancer', 'drug', 'interactions')	400	('getting', 'funds', 'biodiversity')	1392
8	('soil', 'biodiversity', 'atlas')	298	('biodiversity', 'climate', 'change')	353	('biodiversity', 'conservation', 'views')	1384
9	('food', 'biodiversity', 'prudent')	296	('global', 'biodiversity', 'loss')	297	('viagetting', 'funds', 'biodiversity')	1318
10	('biodiversity', 'safe', 'levels')	290	('nabhan', 'food', 'biodiversity')	293	('climate', 'change', 'biodiversity')	1094
11	('biodiversity', 'prudent', 'hedging')	279	('wildlife', 'nature', 'biodiversity')	274	('biodiversity', 'flora', 'nature')	1012
12	('gary', 'nabhan', 'food')	275	('food', 'biodiversity', 'prudent')	270	('biodiversity', 'ecosystem', 'services')	999
13	('wildlife', 'nature', 'biodiversity')	273	('biodiversity', 'sustainable', 'tourism')	269	('funds', 'biodiversity', 'flora')	961
14	('safe', 'levels', 'across')	243	('gary', 'nabhan', 'food')	269	('nature', 'conservation', 'views')	944
15	('strategy', 'dealing', 'food')	238	('biodiversity', 'prudent', 'hedging')	245	('flora', 'nature', 'conservation')	936
16	('hedging', 'strategy', 'dealing')	236	('international', 'biological', 'diversity')	243	('views', 'pxrtg', 'botany')	866
17	('across', 'half', 'world')	234	('daily', 'thanks', 'biodiversity')	230	('click', 'link', 'contribute')	852
18	('prudent', 'hedging', 'strategy')	232	('dealing', 'food', 'insecurity')	224	('link', 'contribute', 'projects')	836
19	('levels', 'across', 'half')	231	('food', 'insecurity', 'climate')	224	('stop', 'biodiversity', 'loss')	759
20	('dealing', 'food', 'insecurity')	226	('strategy', 'dealing', 'food')	224	('biodiversity', 'loss', 'could')	708

	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
6	('species', 'risk', 'extinction')	1057	('nature', 'based', 'solutions')	1085
7	('biodiversity', 'climate', 'change')	1056	('biodiversity', 'ecosystem', 'services')	1069
8	('wastelaying', 'polluting', 'contaminating')	1039	('climate', 'biodiversity', 'crises')	997
9	('altering', 'natural', 'world')	969	('biodiversity', 'loss', 'climate')	911
10	('extinction', 'altering', 'natural')	939	('climate', 'biodiversity', 'crisis')	906
11	('million', 'species', 'risk')	783	('loss', 'climate', 'change')	811
12	('fuels', 'climate', 'change')	749	('fight', 'climate', 'change')	773
13	('climate', 'biodiversity', 'crises')	744	('reverse', 'biodiversity', 'loss')	738
14	('destruction', 'wastelaying', 'polluting')	643	('international', 'biological', 'diversity')	708
15	('plant', 'animal', 'species')	626	('keep', 'soil', 'alive')	680
16	('economic', 'growth', 'fuels')	625	('protect', 'soil', 'biodiversity')	679
17	('biodiversity', 'destruction', 'wastelaying')	622	('soil', 'alive', 'protect')	674
18	('halt', 'biodiversity', 'loss')	618	('beinnature', 'justbreathe', 'environment')	673
19	('global', 'biodiversity', 'framework')	611	('justbreathe', 'environment', 'biodiversity')	673
20	('planet', 'biodiversity', 'destruction')	610	('climatechange', 'biodiversity', 'loss')	657

Sentiment Analysis

Fig.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 in the texts for each year from 2010 to 2020, and visualized on an area chart. Even though the 11 years of data were represented on a single chart, the shape of the data was almost uniform, resulting in good visibility. This was due to the fact that in all years, the use of words corresponding to “trust” and “anticipation” was high, while “joy” and “fear” were slightly high, showing almost exactly 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 ten years. The most straightforward interpretation is that “biodiversity”-

related discourse has remained constant in this way on Twitter. 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. However, even if the results of sentiment analysis 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 in each year by comparing the n-grams.

Topic Modeling

Using the Python library Gensim, I explored the LDA topic models for each year and built the model that was considered best for each. In each year, I built the Gensim model and the MALLET model, respectively, and derived the best model from the Coherence Scores (Röder 2015) and the topic distribution of the visualization results using the Python library pyLDAvis, while varying the number of topics. In this paper, for reasons of paper space, I included models for 2010 and 2020 for comparison, and **Fig.4** and **Fig.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, I adopted the Gensim model for 2010, which consists 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 (**Fig.8** & **Fig.9**). 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. **Fig.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, I was able to classify indiscriminately posted tweets of each year into a meaningful form. This enabled me to perform efficient qualitative analysis. Of course, it was possible to define all the topics from keywords and representative tweets, but in this case, I only examined the characteristic topics.

Qualitative Analysis

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

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 distribution, many of the 60 topics

overlapped and did not spread out as a whole. 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 the one where “joy” was prominent. **Fig.11** shows this on a chart, with “trust” and “joy” being prominent, and a little “anticipation” standing out. Looking at the textual content of the tweets that made up this topic, it was dominated by introductions 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”. As a result of visualizing this on the chart, it was found that the shape of the chart was clearly different from that of Topic 3 and the chart showing the overall trend for 2010. 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 how to infer from keywords, when looking for something containing “economy,” Topic 20 was applicable, and the chart shows a tendency for “trust” to be somewhat common. The main content of the tweets that made up this topic was related to the conference in Nagoya, Japan, although there were also tweets that associated the loss of biodiversity with economy.

Next, from the key words 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. With this in mind, an examination of the content of the tweet texts that make up this topic show that there were references to the 2010 Biodiversity Target, which had not been met, as well as 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. This may be related to the limited popularity of Twitter as a social networking service in 2010, as well as the limited user base and usage patterns. And most importantly, it should be noted that until 2017, there was a limit of 140 words in a tweet.

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

The year 2020 shows an almost identical shape 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”, and “sadness” are prominent (**Fig.12**). Looking at the keywords and the actual content of the tweet texts, 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, 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 explosive spread of infection 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. On the chart, it was confirmed that words such as “trust” and “anticipation” were frequently used. Looking at the details of the tweet texts, the majority of the tweets were positive in nature, 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).

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 a lot of 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 may be attributed to the fact that the tweet limit has been raised to 280 words, and 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.

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, in order 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 about the topics and contexts in which the word and concept “biodiversity” were 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 changes in 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, in order to grasp the details, this study attempted to facilitate intuitive understanding through topic modeling and sentiment analysis and its 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, due to the limited space, only the years 2010 and 2020 were mentioned. Nevertheless, it is an example of how the debate on “biodiversity” has converged in specific contexts over the past decade and how the debate 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, with regard to some of the goals established in 2010, it was indicated that discourse has already begun to look beyond 2020, while reflecting on the status of achievement in the final year, 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 me to classify the large group of tweets produced daily by the algorithm into topics with meaning, and then to 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, or conversely, to examine keywords and text content beforehand and then use heat maps and bar charts 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, and not many of them allocate them to specific emotions. 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 normal to eliminate duplicates such as bots and remove as many superfluous items as possible (handle names, link URLs, images). However, of course, it is impossible to completely remove those unnecessary elements. For example, it frequently happens that a tweet that is actually 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 as well, as a result of 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 the use of NRC emotion lexicon, the developer also pointed out that while it is a simple and powerful tool for analyzing text, the use of lexicon also poses the risk of inappropriate bias (Mohammad 2020). As an example, among the 2020 topics, Topic 33, which stands out in the chart for its high number of “disgusts” compared to the others, required special confirmation (Fig. 12). A closer examination of the tweet texts on this topic revealed that it was dominated by tweets emphasizing the importance of soil conservation, along with 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 lexicon is needed for individual analysis to minimize such cases.

The implication of this study is that it may be possible to narrow down and classify large scale text data to some extent using developing quantitative methods such as NLP, and then provide deep insights through qualitative analysis based on the knowledge and insights 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 special research by institutions and organizations, and even as a tool for propaganda (Guarino et al. 2020). On the other hand, it cannot be denied that it is also a subject of analysis that can provide many useful suggestions. Therefore, in order to present more precise research results, it is especially important for the analysts themselves to acquire a high level of literacy in the field of social networking.

In the case of research aimed at exploring public awareness toward a particular concept, 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 are awaited. Furthermore, in an unsupervised topic model such as LDA, each topic needs to be scrutinized by the analyst from the results obtained, in which case 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, it is recommended to explore other possibilities, such as using another method of semi-supervised model.

Conclusion

This study is just one example of a wide variety of methods, which can be applied not only to "biodiversity" but also to other concepts and keywords. And 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 postss in other languages and of other major 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.

It is known that new goals have been developed for the COP16 of CBD, which has been 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 effective 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 whether it will end up being a temporary fad, depends largely on the training and improvement of researchers' skills, as well as the progress of data science. In any case, however, the increase and accumulation of exploratory research such as this study is essential for the efficiency of global environmental governance.

Statements And Declarations

The corresponding author states that there is no conflict of interest. This article does not contain any studies involving human participants performed by any of the authors.

References

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 September 7, 202

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

Buchanan, Graeme M., et al. "Assessment of national-level progress towards elements of the Aichi Biodiversity Targets." Ecological Indicators 116 (2020): 106497.

Convention on Biological Diversity (https://www.cbd.int/) Accessed August 30, 2021

Cooper, Matthew W., et al. "Developing a global indicator for Aichi Target 1 by merging online data sources to measure biodiversity awareness and engagement." *Biological Conservation* 230 (2019): 29-36.

Guarino, Stefano, et al. "Characterizing networks of propaganda on twitter: a case study." *Applied Network Science* 5.1 (2020): 1-22.

Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." *arXiv preprint arXiv*:1601.06971 (2016).

Mohammad, Saif M., and Peter D. Turney. "Nrc emotion lexicon." *National Research Council, Canada* 2 (2013).

Mohammad, Saif M. "Practical and ethical considerations in the effective use of emotion and sentiment lexicons." *arXiv preprint arXiv:2011.03492* (2020).

Morshed, Syed Ahnaf, et al. "Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis." *Journal of Urban Management* (2021).

Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." *LREc.* Vol. 10. No. 2010. 2010.

Röder, Michael, Andreas Both, and Alexander Hinneburg. "Exploring the space of topic coherence measures." *Proceedings of the eighth ACM international conference on Web search and data mining.* (2015).

statista (https://www.statista.com/topics/737/twitter/) Accessed August 29, 2021

Union for Ethical BioTrade UEBT Biodiversity Barometer 2018 https://static1.squarespace.com/static/577e0feae4fcb502316dc547/t/5b51dbaaaa4a99f62d26454d/1532091316690/UEBT+-+Baro+2018+Web.pdf Accessed August 30, 2021

Xu, Haigen, et al. "Ensuring effective implementation of the post-2020 global biodiversity targets." *Nature Ecology & Evolution* 5.4 (2021): 411-418.

Xue, Jia, Junxiang Chen, and Richard Gelles. "Using data mining techniques to examine domestic violence topics on Twitter." *Violence and gender* 6.2 (2019): 105-114.

Xue, Jia, et al. "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter." *PloS one* 15.9 (2020): e0239441.

Figures

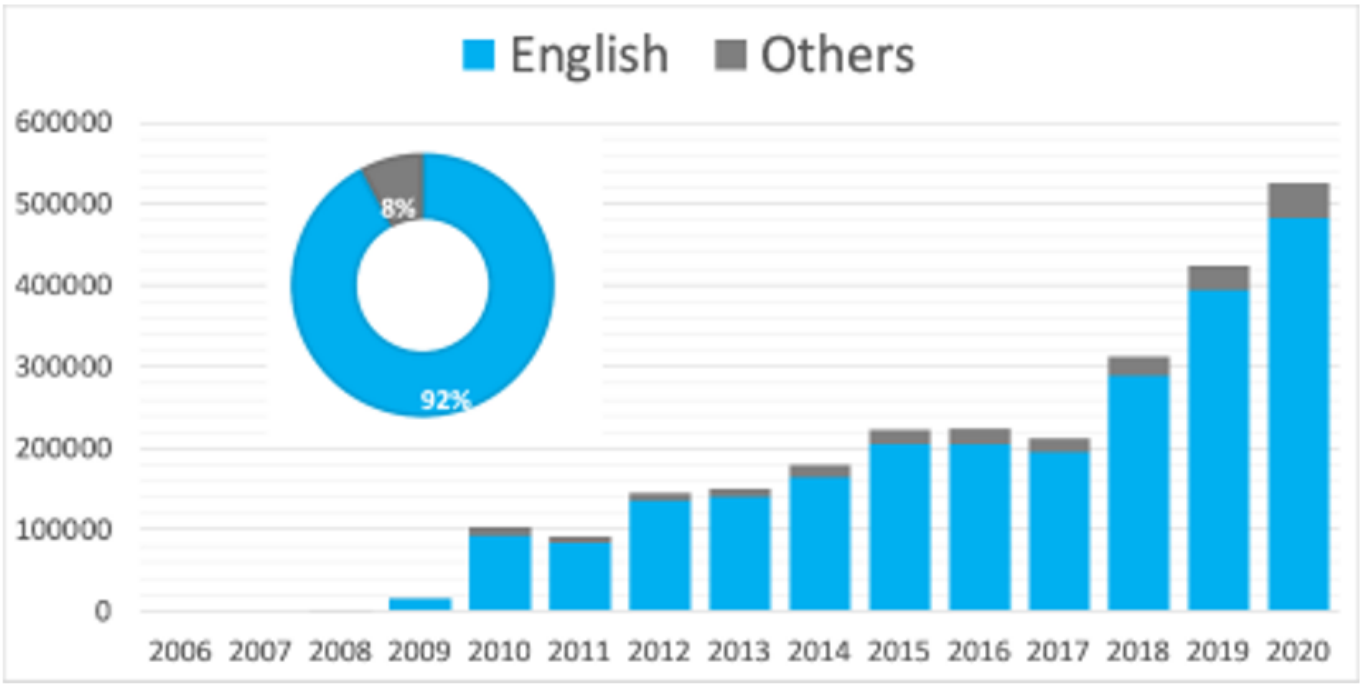

Figure 1

Total number of tweets containing the word “biodiversity” by year from 2006 to 2020 (distinguishing between English text and other language text)

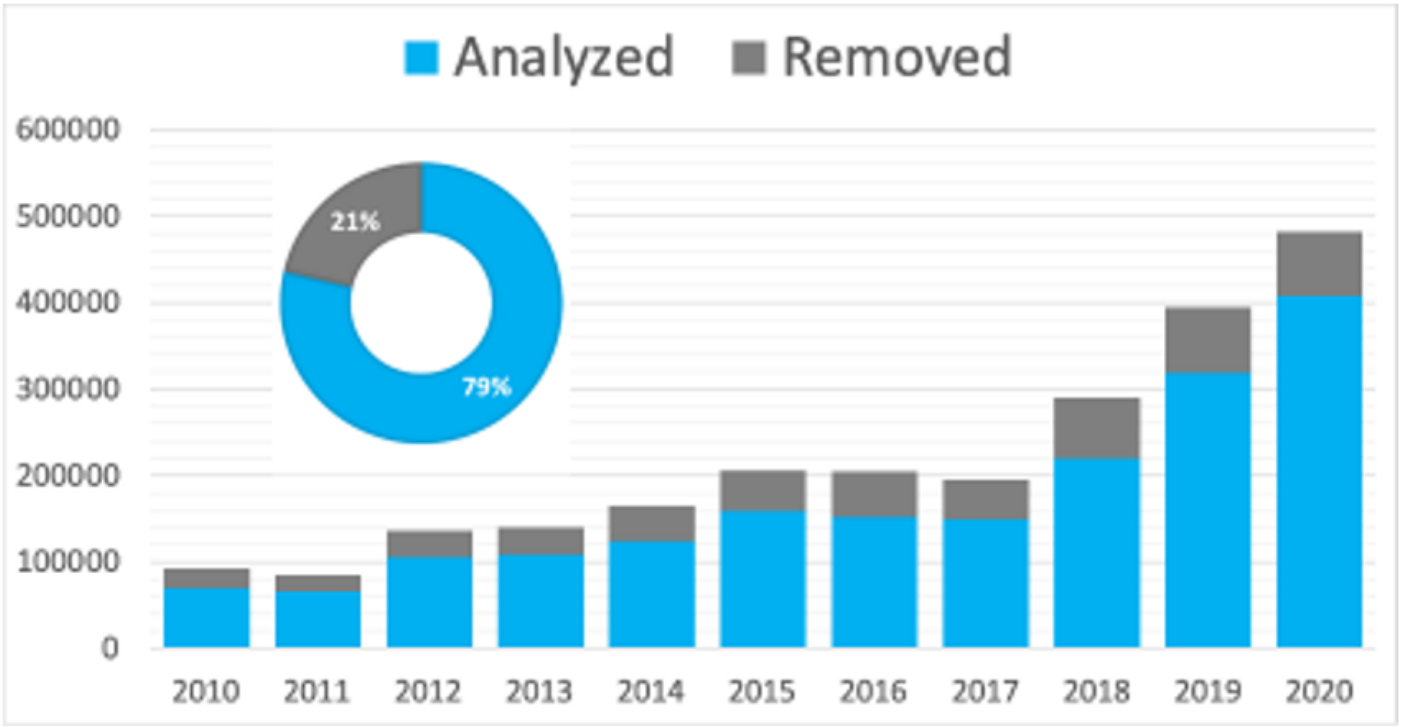

Figure 2

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

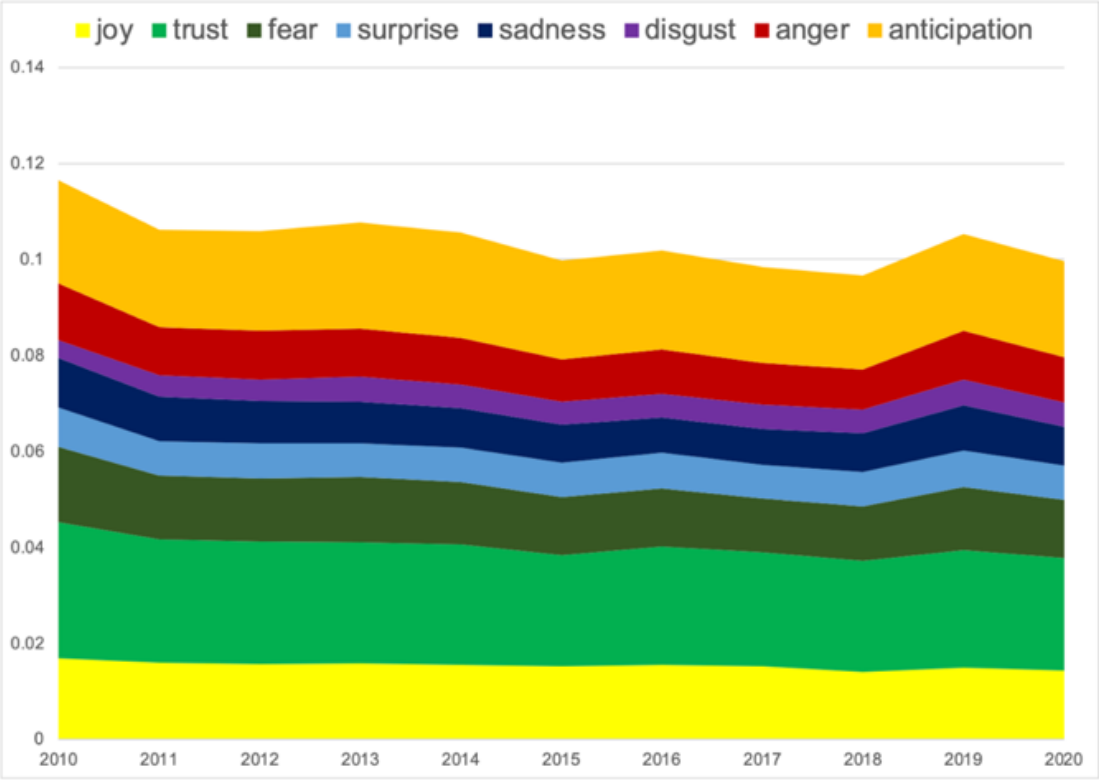

Figure 3

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)

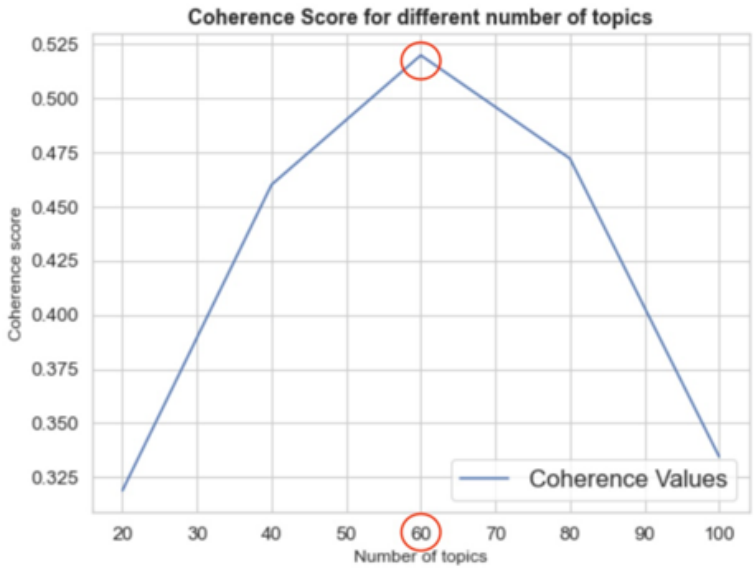

Figure 4

Graph showing the optimal number of topics for 2010

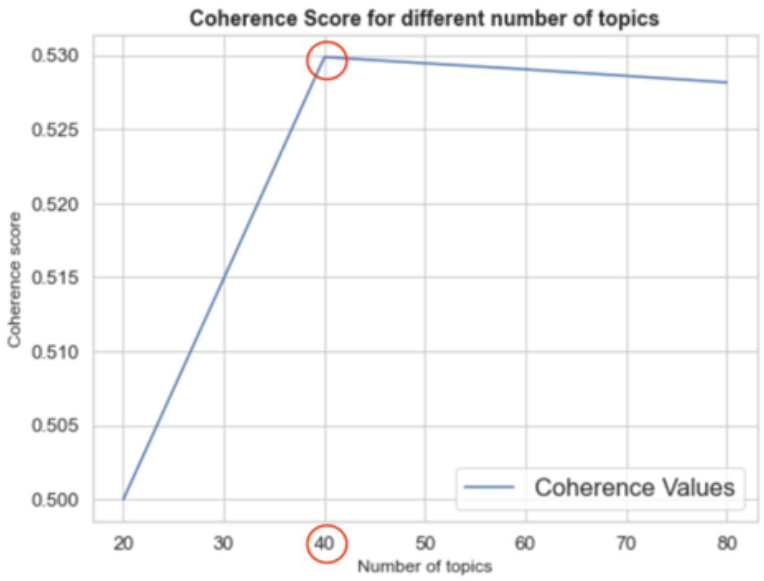

Figure 5

Graph showing the optimal number of topics for 2020

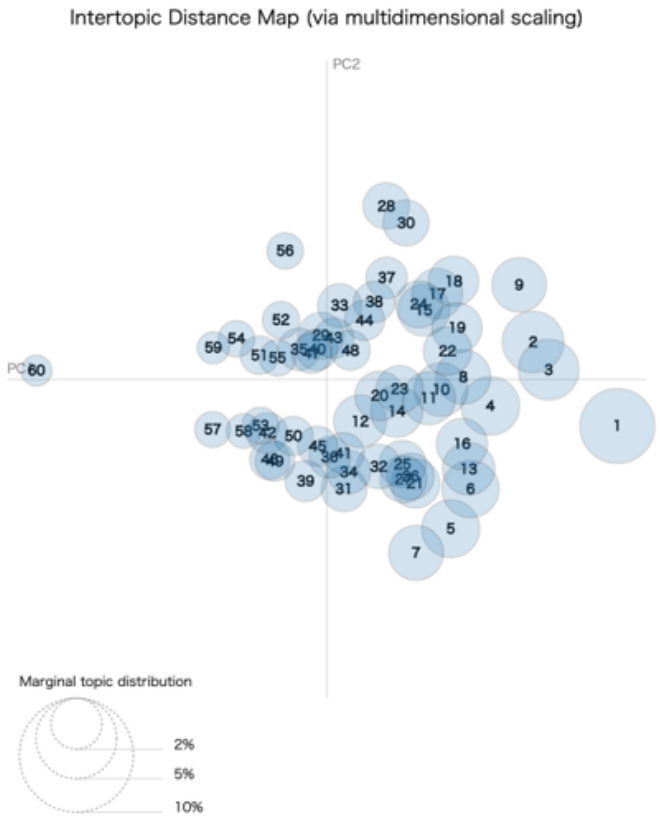

Figure 6

Topic distribution in 2010

(output of 60 topics by pyLDAvis)

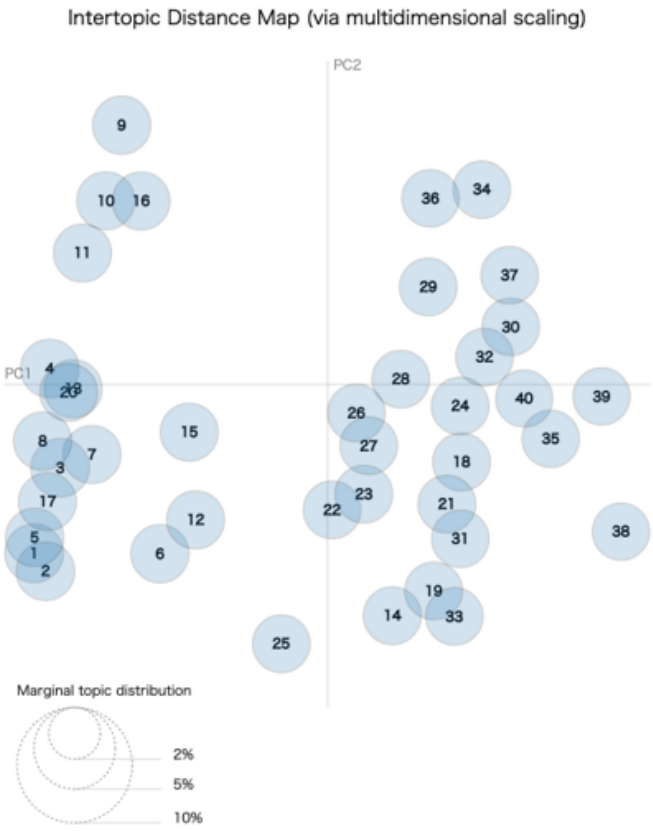

Figure 7

Topic distribution in 2020

(output of 40 topics by pyLDAvis)

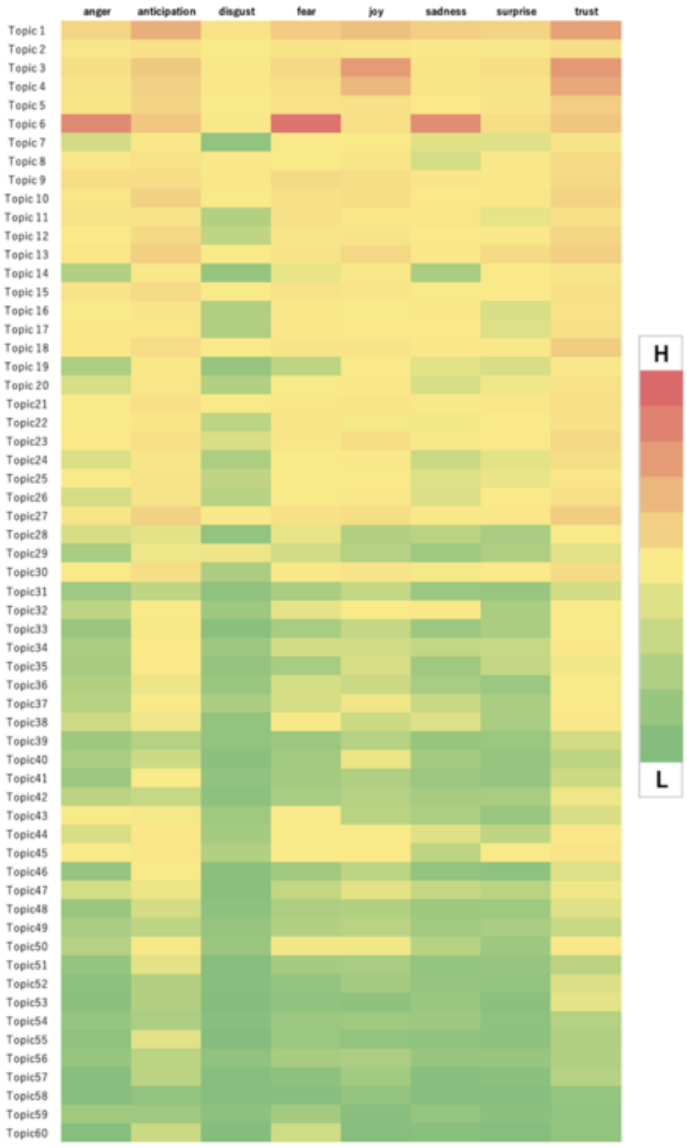

Figure 8

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

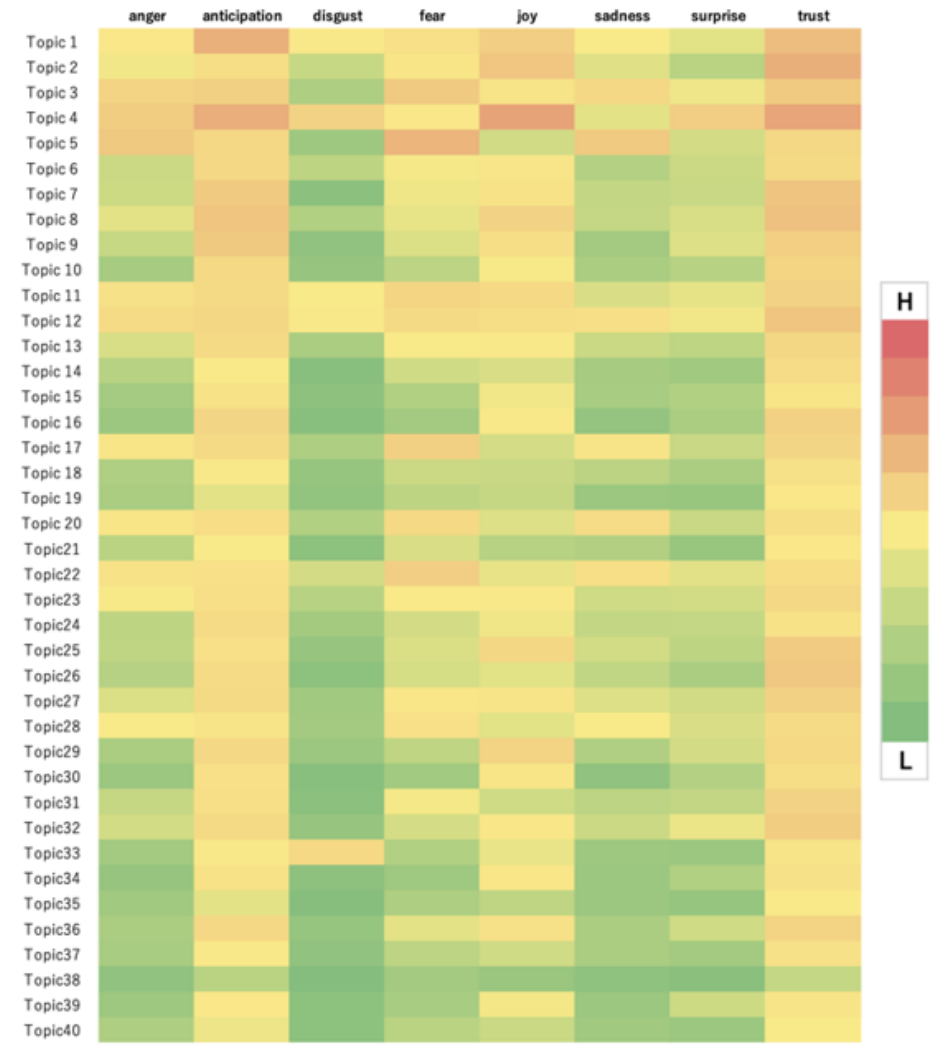

Figure 9

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

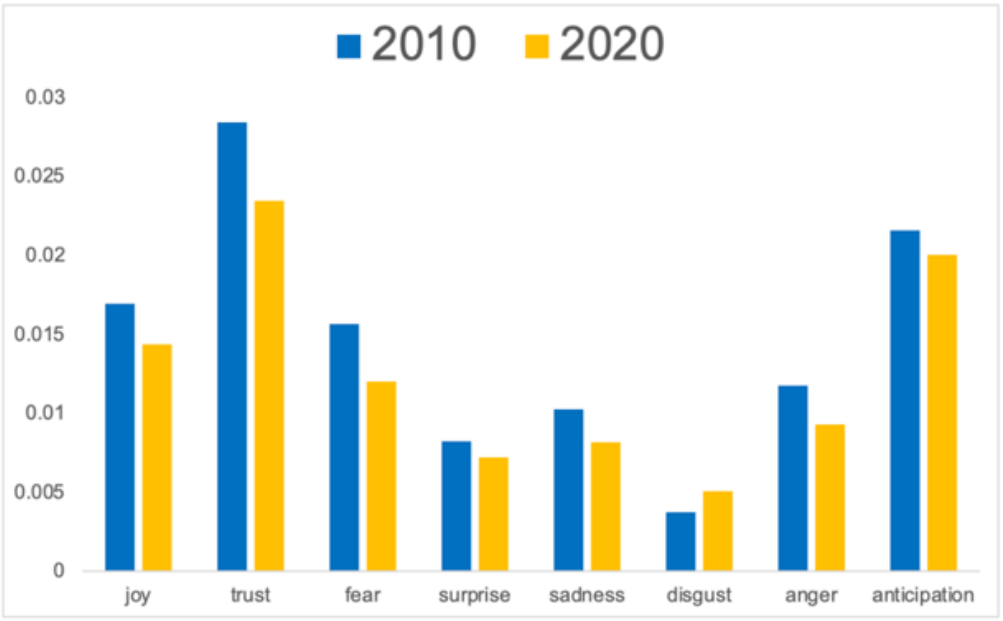

Figure 10

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)

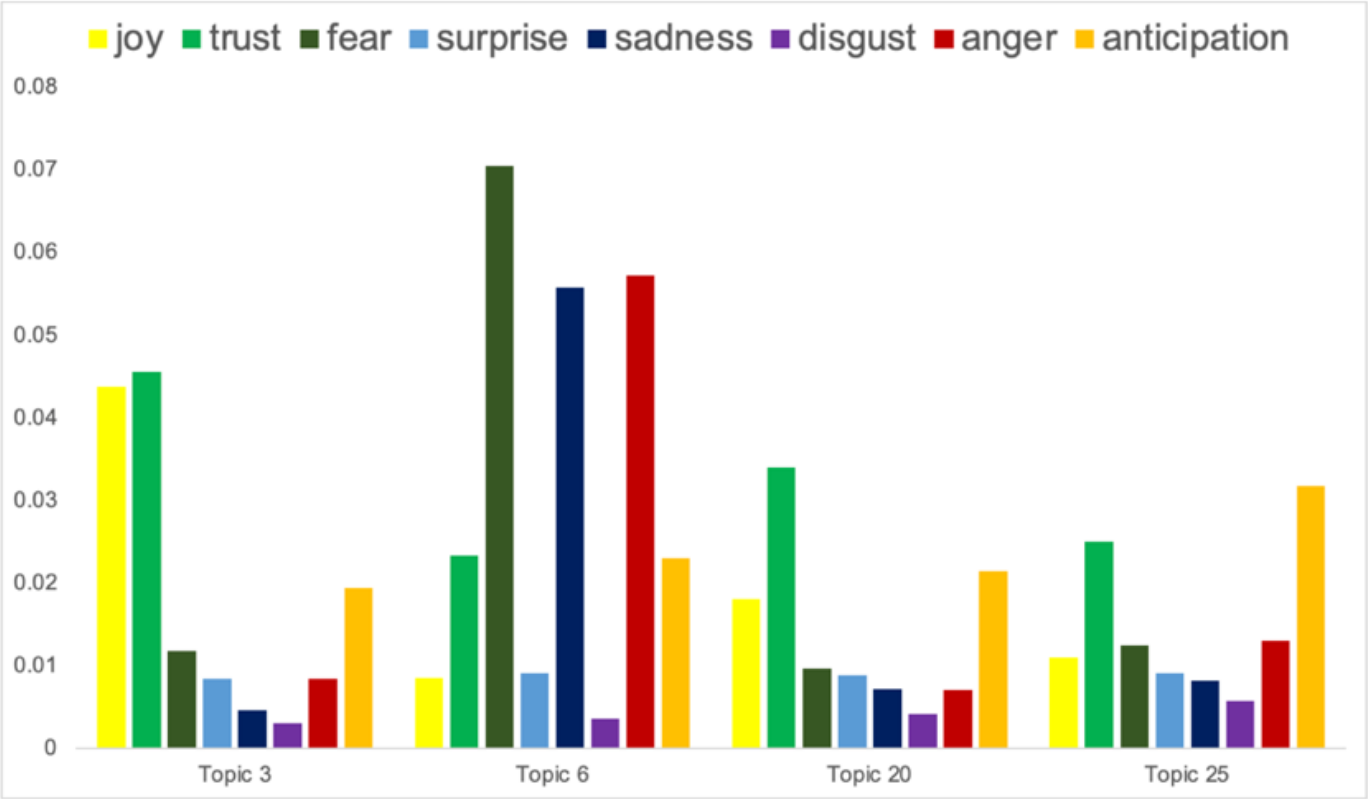

Figure 11

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)

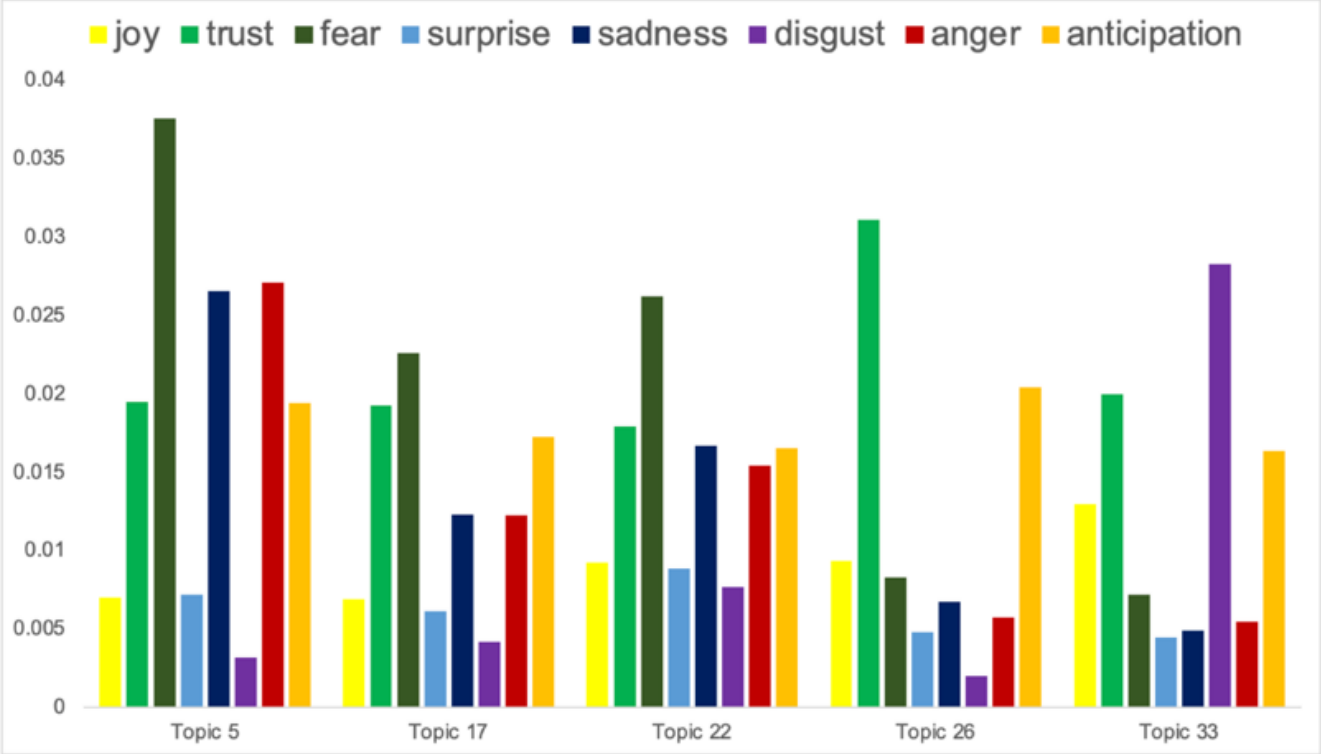

Figure 12

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)