Twitter data analysis to assess the interest of citizens on the impact of marine plastic pollution

Otero[1*], P., J. Gago[1], and P. Quintas[1].

[1]Centro Oceanográfico de Vigo (IEO, CSIC), Subida a Radio Faro, 50, 36390, Vigo, Spain, Phone Number: 0034986492111, Fax Number: 0034986498626

*Corresponding author: pablo.otero@ieo.es

CRediT author statement

Pablo Otero: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Visualization, Writing- Original draft preparation. **Jesús Gago**: Conceptualization, Resources, Writing- Reviewing and Editing, Funding acquisition. **Patricia Quintas**: Conceptualization, Validation, Writing- Reviewing and Editing.

Abstract

Few studies have mined social media platforms to assess environmental concerns. In this study, Twitter was scraped to obtain a ~140,000 tweet dataset related specifically to marine plastic pollution. The goal is to understand what kind of users profiles are tweeting and how and when they do it. In addition, topic modelling and graph theory techniques have allowed us to identify main concerns on this topic: i) impact on wildlife, ii) microplastics/water pollution, iii) estimates/reports, iv) legislation/protection, and v) recycling/cleaning initiatives. Results reveal a scarce influence of organizations involved in research and marine environmental awareness, so some guidelines are depicted that could help to adjust their communication plans. This is relevant to engage society through reliable information, change habits and reinforce sustainable behaviour. A visualization tool has been created to analyze the results over time.

Keywords: Twitter, social media, marine litter, plastic pollution, topic modelling, COVID-19

Highlights:

- Twitter is a valuable tool to analyze the social aspects of marine pollution
- Topic modelling helped to identify 5 main relevant subtopics
- COVID-19 pandemic impacted the marine plastic pollution topic on Twitter
- Low presence of academic or environmental bodies compared to personal opinions
- An interactive app is released to facilitate further analysis

Funding: This work was co-financed by both the European Regional Development Fund through the CleanAtlantic INTERREG (Atlantic Area) project [grant number EAPA_046/2016] and the Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación, Government of Spain) through ANDROMEDA project [grant number PCI2020-112047].

1. Introduction

Marine litter is a planetary threat, affecting nearly every marine ecosystem globally (GESAMP, 2015). In particular, plastic constitutes more than 80 per cent of marine litter (European Commission, 2018), and it is estimated that quantities from 4.8 to 12.7 million metric tons per year are entering our seas and oceans (Jambeck et al., 2015). Despite the ambitious commitments currently set by several governments to reduce marine litter, Borrelle et al., (2020) estimated that the annual input may reach up to 53 million metric tons by 2030.

The impacts of plastics on marine ecosystems are broad, including habitat degradation and a wide range of negative effects on marine organisms. The impacts include from wildlife dead due to ingestion, starvation or entanglement in marine litter (Gall and Thompson, 2015), to attaching and drifting invasive species and pathogens (i.e., hitchhiking), among others. The socioeconomic effects are also evident in sectors such as fisheries (e.g., damaged gear during trawling activities or reduction of catches), on tourism due to the presence of beach litter or the economy of coastal areas due to clean up actions (GESAMP, 2015). Indirect effects on human health are still being discussed, including the sources and transport dynamics of antibiotic resistance (Bank et al., 2020) and the still unknown effects of microplastics (<5 mm) along the food chain (GESAMP, 2020). Those microplastics can enter the marine environment as primary microplastics (e.g. manufactured pellets and microbeads) or secondary microplastics after the fragmentation and degradation of larger plastics. The presence of microplastics in all marine environments, including marine biota, has been reported in several scientific studies (Filgueiras et al., 2020; Gago et al., 2020; GESAMP, 2020, 2015).

Nowadays, it is impossible even to try imagining our world without plastics, given the extreme importance and the number of functions it has in a broad range of aspects of the industry and everyone’s daily life. There are no other manufactured materials whose production has grown as plastic has over the last 70 years (Geyer et al., 2017). From 1950 until 2015, 8,300 million metric tons of plastics have been produced: 30% of products are in use, 10% has been incinerated, only 7% has been recycled and 55% has been discarded (Geyer et al., 2017). It is clear that there is an excess of consumption and that many single use articles could be substituted by other materials. Areas of high population density, poor waste management or lack of environmental education, become factors that favour littering of the aquatic environment by plastics (Duckett et al., 2015; Napper and Thompson, 2020).

As a consequence, the potential solutions to mitigate the problem are widespread, and the governance solutions become complex. Government and legislative initiatives, changes in the industry and a greater environmental awareness of citizens are factors that can help reduce the arrival of plastics into the sea (Vince and Hardesty, 2017). As part of this strategy, understanding public perceptions, opinions and knowledge about marine plastic litter issue is a critical step in effectively engaging society and changing human behaviour (Forleo and Romagnoli, 2021).

In the last decade, the information about marine litter has circulated from the scientific community to the public, through reports, awareness campaigns, events and informative material of all kinds. The disclosure has contributed to raising a critical conscience in society (Heidbreder et al., 2019; Mitrano and Wohlleben, 2020; Vince and Hardesty, 2017). This information has been increasingly echoed in part by generalist media. However, another part of the success must be attributed to Social Media (SM), which has begun to open the eyes of many people regarding some of these environmental threats. SM users have passed the 3.8 billion mark and were estimated that more than half of the world's total population was using SM by mid-2020 (Kemp, 2020). Viral messages, photos and videos can reach audiences of millions (Parton et al., 2019), so data created and shared by users on SM platforms have emerged as a potentially useful source of information in marine environmental research, management and conservation (see e.g, Abreo et al., 2019; Becken et al., 2017; Ghermandi et al., 2020; Parton et al., 2019; Retka et al., 2019; Ruiz-Frau et al., 2020).

Twitter is one of the most popular SM and microblogging sites with more than 330 million monthly active users worldwide (Kemp, 2020), who post ~500 million comments (the so-called *tweets*) per day with up to 280-characters containing their thoughts and opinions. Despite the limited number of characters, the possibility of including links allows to increase the information and reach a higher impact. These tweets are affected by both real-world events and the trends of other messages posted in SM (Zubiaga and Ji, 2014). Nowadays, Twitter is the most important SM on science dissemination where journalists, science dissemination professionals, scientists and many research institutions interact, talk and share science with other colleagues and the public (Collins et al., 2016; Letierce et al., 2010; Mohammadi et al., 2018; Van Noorden, 2014). About 50% of scientists use Twitter to follow conversations or debates about their discipline and about 40% consider this network as a tool to talk about their progress or that of other colleagues (Van Noorden, 2014). The most common perceived benefits of Twitter were the size and diversity of the audience, the ability to network with other scientists and the ability to engage with the public (Collins et al., 2016; Smith, 2015).

Twitter has increasingly become a world-wide choice to raise awareness and disseminate information on a variety of topics, as the promotion of cancer screening and early diagnosis through specific campaigns (Plackett et al., 2020; Teoh et al., 2018; Vraga et al., 2018; Yoosefi Nejad et al., 2019), identify cancer barriers and policy solutions (Shimkhada et al., 2021), identify mental health discourses (Budenz et al., 2020; Makita et al., 2021), detect and predict the epidemic of diseases (Dang et al., 2018), analyze pro- and anti-vaccination discourses (Milani et al., 2020), aware about emerging technologies (Li et al., 2017), emergencies (Barker and Macleod, 2019; Martínez-Rojas et al., 2018; Zhou et al., 2021), and so on. Karami et al. (2020) found 38 different topics in more than 18,000 Twitter-related papers published between 2006 and 2019, using analysis techniques like sentiment analysis, topic modelling or graph mining, among others. These techniques have been applied here to a dataset of more than 140,000 tweets to analyze the interests of citizens about marine litter. Text mining and natural language processing were used to learn about what people commented on this particular SM and how they did it, identifying dominant topics and analysing the word and hashtag frequencies. Sentiment analysis was also employed to explore which were their feelings, whereas geo-tagged tweets and information from the user profile were combined to know the main hot spots of discussion. Additionally, manual and automatic identification of image content in tweets was conducted.

This study aims to describe the interest and awareness of marine pollution by plastics and microplastics in Twitter, to understand the spatio-temporal trends and sentiment of tweets, to distinguish different subtopics from the general discourse. Additionally, those images associated with tweets have been explored to assess, among other things, their suitability for discerning amounts and types of litter, particularly in coastal areas.

Taking into account that people engage with information posted by people they trust (Huber et al., 2019; Media Insight Project, 2017), this study will provide new insights to governmental, academia and NGOs involved in marine environmental protection to reanalyze their communication strategy on Twitter. Understanding who tweets about the marine litter issue and how they do it will help institutions to design effective communication on this channel to reinforce the commitment of users who are already engaged, facilitate greater public understanding of solutions and enable action.

2. Methodology

This study aims to perform an exploratory analysis of a collection of tweets that cover frequency analysis, sentiment analysis, graph theory and topic modelling, among others. Scraping and data mining techniques involve different steps from data acquisition and data cleaning to data analysis (see Figure 1). In addition, an automatic classification image analysis technique has been tested with the aim of characterizing litter in coastal areas.

2.1. Dataset creation

Figure 1 shows a flow chart of the methodology applied during dataset creation and data analysis phases. The first step consists of collect data from Twitter database. The set of streaming APIs offered by Twitter gives developers low latency access to Twitter's global data, which include the tweet text along with the associated metadata (post time, geographical coordinates if geolocation is enabled, information about the user profile, etc.). In this study, free Twitter's standard search API v1.1 (search/tweets) was used for simple queries against the indices of recent or popular tweets and behaves similarly to, but not exactly like the search UI feature available in Twitter mobile or web clients. The Twitter Search API works as a keyword search method against a sampling of recent tweets published in the past 7 days (further details available in https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets); words and not hashtags were used to perform a query to capture a higher number of tweets. To analyze a longer period, a Python script scrapped the Twitter service every week.

Several queries were done to retrieve original tweets (not retweets) by a combination of the keywords '*plastic*' or '*microplastic*' with at least one keyword related to the marine environment: '*ocean[s]*', '*sea*', '*beach*', '*coast*' and/or '*marine*'. To ensure representative results of the global use made of this platform, main languages used on Twitter were exposed to queries: English, Japanese, Spanish, Portuguese, French, Italian, Malaysian, German, Turkish, Thai, Korean and Indi (sorted in descending order in our dataset). Tweets were retrieved weekly during ~8 months (Mar 19 – Nov 16, 2020). Table 1 displays a few tweets from the final dataset of 147,552 tweets. The dataset needs to be cleaned up before analysis. Tweets usually contain colloquial and informal sentences, URLs, emojis and emoticons. Therefore, some cleaning is needed to facilitate its understanding and analysis. At the same time that the original text is preserved for each tweet for future reference, a "*sanitized*" text is added to the dataset. Hashtags, user mentions, URLs, media and symbols are stripped out from the full text. Although only original tweets were retrieved, preventive cleaning of "RT" (retweet) at the beginning of the text and symbols such as *at signs* were stripped out from the full text. The remaining text was lower cased and automatic translation of those non-English tweets was done using the TextBlob library in Python (https://textblob.readthedocs.io/). Internally, TextBlob relies on Google Translate's API. To skip the rate limit, a delay was done between consecutive queries to the API. Finally, English text was cleaned of grammatical contractions (e.g., "*ain't*" to "*is not*", "*I'll*" to "*I will*"); with this purpose, a set of 125 contractions was used (https://github.com/PabloOtero/twitter-python). While some abbreviations and acronyms may be common across all SM sites, others are unique to the microblogging platform. Some terms of this Twitter lingo were also fixed (e.g., "*u*" to "*you*", "*ya*" to "*yeah*").

With clean text, it is easier to apply sentiment analysis and store the results along with our dataset. The algorithms of sentiment analysis mostly focus on defining opinions, attitudes, and even emoticons in a corpus of texts. In this study, the approach of the TextBlob package based on a pre-defined set of categorized words was used. The sentiment property returns both *polarity* and *subjectivity*. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a floating value within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. Two sentiment scores were computed for each tweet: one based only on words and the other one on words plus emojis and emoticons; the meaning of an emoji will depend on the context of the current text. Whereas emoticons are handled well by TextBlob, emojis were searched in the message based on a Unicode list (http://www.unicode.org/emoji/charts/full-emoji-list.html) and converted to official name or any known short name (https://github.com/alexmick/emoji-data-python).

Whenever possible, the spatial information associated with the tweet was also obtained and added to the dataset. Geographical coordinates were extracted from tweet metadata (geotagged tweets) in those cases in which the user had the geolocation enabled in its device, something that only happened in 1.98% of the cases. Coordinates were exposed to Nominatim API (https://nominatim.org/) to perform reverse geocoding on OpenStreetMap (https://www.openstreetmap.org/) and obtain city, state —when possible— and country data fields. From those non geotagged tweets, location information from the user profile was obtained and forward and reverse geocoding performed through the API. Here, it is necessary to understand that the user does not necessarily have to indicate an existing or recognizable place. In fact, although 70.6% of the tweets presented a location in the user's profile, only 56.7% had a valid position on the globe.

As summarized in Figure 1, further cleaning of the dataset may be necessary to facilitate some analysis and the interpretation of results. This is the case of frequency analyses where the social bots influence the volume data. Although there is no universally agreed-upon definition of a bot, they can be considered malicious actors that create inauthentic social media accounts partially controlled by algorithms. Most automated bots reply or post tweets or simply follow other users based on triggers or according to some scripted patterns (e.g., retweeting all messages from certain accounts). To detect social bots, the popular API Botometer® v4 (Sayyadiharikandeh et al., 2020) was used. Botometer is a machine learning algorithm trained to calculate a score where low scores indicate likely human accounts and high scores indicate likely bot accounts. From several types of scores provided by Botometer, we used the "*overall score*" based on a comparison of several models trained on different kinds of bots and human accounts and language-independent (“*universal*”). In the [0-1] range, here we considered 0.6 as a limit to consider an account as a bot.

In the same way that the detection and elimination of bots are important, it is also necessary to discard as much as possible those tweets that are not directly related to marine plastic pollution. As a notable example, are those tweets that refer to the British virtual band “*Gorillaz*” —with about 1 million followers on Twitter— and their studio album “*Plastic Beach*”. Thanks to the Latent Dirichlet Allocation (LDA) technique (Blei et al., 2003) that will be described later, we could determine the most frequent words used when talking about this subtopic (e.g., ‘*song*’, ‘*album[s]*’, ‘*music*’, ‘*demon*’, ‘*humanz*’, ‘*listening*’) and consequently, eliminate the related tweets.

2.2. Data analysis

2.2.1. Topic modelling

Topic modelling is an efficient and systematic approach to analyze thousands of documents in a few minutes (Karami et al., 2020). Among topic models, LDA is a valid and widely used generative probabilistic model (see e.g., Blei et al., 2003). LDA identifies semantically related words, which occur together in multiple documents (i.e., tweets) of a corpus (i.e., our preprocessed dataset as shown in Section 2.1). As a result, several groups of multinomial distributions over the terms in the vocabulary of the corpus represent the topics. To interpret a topic by human intuition as a meaningful "*theme*", one typically examines the top terms in a ranked list of the most probable terms in that topic. The problem with this method is that common terms in the corpus often appear near the top of several lists, making interpretation difficult. For this reason, these lists were sorted by relevance according to Sievert and Shirley (2014). They defined the relevance r of word ω to topic k given a weigh parameter λ [0-1] as:

$$r(\omega, k|\lambda) = \lambda \log(\phi_{k\omega}) + (1-\lambda)\log(\frac{\phi_{k\omega}}{p_{\omega}}) \text{ [Equation 1]}$$

where $\phi_{k\omega}$ denotes the probability of term ω for topic k and p_{ω} the marginal probability of term ω in the corpus. Whereas λ = 1 shows the classical ranking of terms by their probability in the topic, a lower λ also weights by the probability of appearing in the corpus. At the other extreme, λ = 0 classifies terms by the ratio in a logarithmic scale of their probability within a topic to its marginal probability across the corpus, also known as *lift* (Taddy, 2011); this is, ranks words that appear exclusively in that topic but not in the others. Thus, playing to vary λ can help to better define the associated topic.

In this study, the technique was only applied to English words after removing from the corpus those keywords used in data acquisition; in this way, the subtopics produced by this technique are meaningful and not dominated by the same keywords. Stop words, which are the most common without significant contextual meaning in a sentence (e.g., "*a*", "*the*", "*and*", "*but*", and so on) were also filtered out from the corpus. Those ampersands (&) written in tweets via a mobile device appear in the document as "*amp*" with no inherent meaning, so they were removed as well. Apostrophes were also deleted and words in plural were converted to the singular as far as possible.

The number of topics must be chosen before LDA is run; however, it is unclear how many topics the dataset should be divided into. A low number of topics can cause the loss of a detailed view of the text when merging topics. Alternatively, a high number of topics can lead to too many top words being shared and make interpretation difficult. In this study, different tests were run with some topics varying from 5 to 15 and finally, 10 words in 6 topics were used to train the model. To better understand the underlying fitted LDA model, the LDAvis tool (Sievert and Shirley, 2014) over Python was used (https://github.com/bmabey/pyLDAvis). This tool allows flexibility in exploring topic-term relationships using *relevance*.

2.2.2. Graph theory analysis

To complement topic modelling, analysis of networks using graph theory was performed. A graph (network) is a collection of vertices (nodes) with a collection of edges that are connections between the different vertices in a network. In this study, nodes are represented by words, while edges illustrate the connections between words in the same tweet and the frequency of those connections. Within the network, it is possible to distinguish communities. A community is defined as a group of nodes where the density of the edges between the nodes inside the group is greater than the connections with the rest of the network. To find communities in our network, a semi-synchronous label propagation method was used (Cordasco and Gargano, 2011). This method combines the advantages of both synchronous and asynchronous models. If a score is given to the number of links between two nodes and the process is repeated for the complete network landscape, the *modularity* —a measure of the strength of the division of a network into communities— can be computed. Networks with high modularity have dense connections between the nodes within communities but sparse connections between nodes in different communities. According to Clauset et al. (2004), modularity can be computed as:

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right] \text{ [Equation 2]}$$

where the sum iterates over all community c, m is the number of edges, L_c is the number of intra-community links for community c and k_c is the sum of degrees of the nodes in the community c. Modularity ranges from -1 to 1, and the higher the value, the better the community structure.

On the other hand, different types of centrality measures can be used to identify which nodes are the biggest influencers on the network. Here, an eigenvector centrality, which is based on the centrality of its neighbours was used (Newman, 2010). A node with a high score will influence multiple nodes, which in turn are highly connected. The advantage of this method is that it can highlight nodes that exercise control behind the scenes. For the creation, manipulation, and study of the structure, dynamics, and function of our network the library NetworkX was used (https://github.com/networkx/networkx; Hagberg et al., 2008). By default, the layout of the nodes and edges is automatically determined by the Fruchterman-Reingold force-directed algorithm (Fruchterman and Reingold, 1991). Frequencies of word pairs, also called bigrams, were analyzed.

2.2.3. Image analysis

Photos associated with tweets were also downloaded and stored for image content analysis (6,172 images only in English tweets). Duplicated images were removed after comparing their associated Message Digest Algorithm 5 (MD5) hash values. The Computer Vision API (v3.1) from Azure Cognitive Services (https://azure.microsoft.com/es-es/services/cognitive-services/) was used to process and obtain information from images. This free tool allows, among other features, to estimate the dominant and accent colours, categorize the content of the images, tag and create a short description. In the present study, the goal was to use this tool to filter images based on categories and tags to assess the type of media content uploaded by users, as a previous step to object detection techniques in future studies.

3. Results and Discussion

We first present results from a social perspective to know, among other aspects, which languages were used the most, from where people were tweeting, when they were active, the most frequent words and hashtags, the main subtopics, the positive or negative feelings and who capitalized on the conversation. Second, image analysis was performed to determine the type of content most used and to assess whether this information could be used to monitor coastal areas directly from images present on Twitter.

3.1. Regional analysis

Figure 2 displays a heatmap for the dataset where positions of each tweet were obtained either from the metadata in the case of geotagged tweets or estimated from the user's profile. The spatial information obtained in this way accounted for 56.7% of the total volume of tweets. The map shows a greater concentration of tweets on the coasts of the USA, Japan, Western Europe, the west coast of South America, Indonesia and the west coast of Australia. Hot spots in the map can be compared with the number of tweets per country shown in Figure 3a. The top four countries were the USA with 16,111 tweets, the UK with 9,908, Japan with 5,909 and Canada with 3,548 tweets. These countries accounted for 52.4% of the total tweets with associated spatial information in their metadata.

To achieve a vision as global as possible we have made a multilingual approach, in contrast with the majority of published studies that queried Twitter with hashtags or only with keywords in English. The frequency of languages in our dataset was English (63.1%), Japanese (16.5%), Spanish (8.5%), Portuguese (2.9%), French (2.8%), Italian (1.9%), Malaysian (1.8%), German (0.8%), Turkish (0.6%), Thai (0.5%), Korean (0.4%) and Indi (0.2%), a list that does not coincide in order with the classification by most widespread native languages (Eberhard et al., 2019) nor the usage statistics of content languages for websites (W3Techs, 2020). If we narrow the list in Figure 3a to only English-speaking countries, the top 4 countries by volume of tweets become the USA, UK, Canada and Australia, the same countries with a greater number of comments in Twitter on the climate change issue (Dahal et al., 2019).

To compare with the population size, the tweet volume was normalized by the population of each country and by the corresponding maximum ratio from these top-25 countries (Figure 3b); the list is restricted to prevent a high bias by little populated countries with a large relative volume of tweets (e.g., small island states). This figure reveals that the UK was the country with the highest number of tweets per capita followed by Ireland, Canada, New Zealand and Spain. Although Japanese was the second language in our dataset, Spain appeared as the first non-English-speaking country with the highest relative weight. A third approach could be taken to weigh based on the digitization of the country as well as the engagement of this social network. Here, the tweet volume per country was divided by the number of active users on Twitter. Although we have only had access to data from countries with the highest number of active users (STATISTA, 2020), it was enough to verify that the order of the previous lists would be modified, becoming now the top 4 countries: UK (with 16.6 Million Active Users, MAU), Spain (7.5 MAU), France (7.9 MAU) and Germany (5.4 MAU). The USA occupied the fifth position as the interest in this topic dissolved among its high number of active users (68.7 MAU). Saudi Arabia occupied the eighth position in the number of active users however it was not represented in our dataset, so Arabian should be also considered in the queries to Twitter in future studies.

The difference in data volume between countries is a combination of population size, the degree of digitization, the number of active users on this social network and, finally, interest in this specific topic. For example, India is the third country in the world in terms of active users (18.9 MAU) on Twitter surpassing slightly UK. If we focus on a highly topical issue during the data acquisition period such as COVID-19, India was ranked third globally in line with its number of active users (Banda et al., 2020). However, India occupied the ninth position in data volume and was the thirtieth in terms of tweets per capita in our study, contrasting with the UK that demonstrated to be the country with the greatest interest in the marine plastic pollution topic on Twitter. This lower relative interest in comparison with its number of active users was also observed in countries such as Japan or Brazil, among others.

Over a quarter of tweets with spatial reference in the dataset came from the USA which invited to deepen the analysis in this country. California was the state with the largest volume of domestic tweets (15.7%) followed by New York (9.7%), Kansas (8.1%) and Florida (7.8%). Dahal et al., (2019) found that the northeast region had a relatively high amount of climate change discussion and this could be caused, among other hypotheses, by the cultural and political differences, since climate change is treated as a political issue by many Twitter users.

A high amount of tweets in densely populated coastal states was one of the expected results in this study, yet the inland state of Kansas was surprising. In this state, there was not high activity of any particular user or group of users. Nor had a higher immersion been observed in environmental campaigns held online. Therefore, it will be interesting to analyze the domestic behaviour of USA in future studies that include a longer period.

3.2. Temporal analysis

The temporal evolution is impacted by different events as depicted in Figure 4. The largest peak corresponds to the celebration of the World Environment Day on 5th June 2020 with the lesser impact of other world-wide celebrations and campaigns with hashtags such as #EarthDay (22nd June), #WorldSeaTurtleDay (16th June) or #PlasticFreeJuly (1st-3rd July). The second and third largest peaks on the series are related to comments on environmental reports published by The Pew Charitable Trusts and SYSTEMIQ (2020) and OCEANA (Warner et al., 2020) on 14th July and 19th November, respectively. Scientific publications in high impact journals also impacted the volume of tweets, like the study of Pabortsava and Lampitt (2020) on 18th August, Borrelle et al. (2020) on 6th October and Law et al. (2020) on 5th November. The comments on Twitter about the scientific studies echo news in high-audience media (e.g. *The New York Times*, *FOX News*, *The Guardian*, *The Economist*, etc.), which explains a certain delay between the publication date on scientific journals and their peak in the time series on Twitter.

The acquisition data began with an exceptional situation due to the COVID-19 pandemic, with national lockdown measures, particularly in most of Europe, Asia and South America, followed by a period of greater freedom over the summer months. Twitter is what´s happening and what people are talking about right now. In this sense, the temporal analyses were affected by the confinement period with people at home and face to face activities focused on marine litter cancelled (e.g. beach clean-ups, events, etc.).

Nevertheless, the conversation around this topic remained active on Twitter, probably due to the concern about the increase of single-use plastics (especially masks and gloves) during the COVID-19. Thus, before 1st June, the average number of daily tweets was 351.5+/-104.6, whereas after this date the average was significantly higher 472.7+/-172.69 (p<0.05). The COVID-19 situation also pulls the comments, with a peak after the article in "*The Economist*" entitled "*COVID-19 has led to a pandemic of plastic pollution*" and published on 22nd June. (https://www.economist.com/international/2020/06/22/covid-19-has-led-to-a-pandemic-of-plastic-pollution).

If tweets are grouped in time hour slots and days of the week (see Figure 5), results show the highest number of posts on business days, between 12 and 18 UTC, and increases as the workweek progress. As expected, there are differences if the analysis is by time zone, mainly due to differences in social habits and work schedules. Peaks often coincide with catching up on the coffee break, lunchtime, and the time people are on their way home. For example, in Japan (figures not shown), the activity is high from 7 am to midnight and during all days of the weeks, with some peaks at noon and 6 pm. In Spain, Twitter activity in this topic is high from 9 am to 9 pm, with peaks at noon and 2 pm; the increase in activity on Sunday is noteworthy. In the USA, the activity is high between 7 am and 6 pm, mainly during business days and particularly on Thursdays. In this country, there are also important differences between the Eastern and Central Time Zones. Habits and work breaks determine the time of use of this SM and being aware of this reality is relevant to increase the engagement, particularly taking into account the "*short lifespan*" of a tweet (Wilson, 2019).

3.3. Sentiment analysis

Figure 4 also shows the volume of positive and negative tweets per day. Positive tweets (n=67,470) always exceeded the negative ones (n=33,612). An increase of positive tweets was noticeable during the celebration of both the Earth Day and World Environment Day. In contrast, the volume of negative tweets in social networks increased during the celebration of Sea Turtle Day with many references to the entanglement of turtles. Negative tweets were also accompanied by references to Pew´s report (The Pew Charitable Trusts and SYSTEMIQ, 2020), which warn about poor environmental conditions.

Figure 6a shows a histogram of the polarity of tweets at 0.25 bin intervals. On average, tweets exhibit significantly greater sentiment when emoticons are included (0.085±0.264) than when these are not taken into account (0.065±0.256) (paired t-test; $p<0.01$). Negative tweets are significantly more objective (0.435±0.265) than positive tweets (0.516±0.202) (2-sample t-test; $p<0.01$), although the frequency distribution encourages us to interpret this result with caution (see Figure 6b and 6c).

The comparison of average sentiment between the 10-top countries (those with more than 2,000 tweets) also shows significant differences (ANOVA; $p<0.01$), with Spain the most positive country (0.123±0.262) and Japan the less positive (0.038±0.219). Internally in the USA, no significant differences were found on average sentiment between the west and east coasts of the USA and neither between the coastal and inland states.

3.4. User´s activity and engagement

Just as important as knowing where and when people tweet about marine plastic pollution, it to know who does it, how often, and the success of their message. From 81,664 users that composed our dataset, we have sorted the top 100 users by the number of tweets, by engagement and by the number of followers. To avoid ethical and privacy problems, the data showed here (Figure 7) has been aggregated in different categories; explicit mentions in the text are only related to large organizations or companies and not individuals.

The first way to categorize relies on a binary classification between bots and human-like user profiles. Based on results from the Botometer API, those user profiles with an overall bot score greater than 0.6 and probability above 0.8 were directly classified as bots and the rest were supervised based on the number of published tweets, number of followers, the type of content, etc.; this binary classification can sometimes be subjective, as bots are becoming more and more refined. In turn, human-like profiles were classified as companies, individuals, NGOs/foundations/nonprofit organizations, academic institutions, official organisms or initiatives/projects.

As expected, bots are the main group when users are classified by tweet volume (32%), followed by individuals (28%) that double the rest of the categories. Generally, these individual users are not celebrities or influencers, in contrast to the list of users with more engaging tweets like actors, musicians, soccer players, writers, politicians or even astronauts. Companies are the second category in terms of successful tweets, particularly those coming from big media (BBC, CNN, New York Times, The Economist, ABC, Globo News, Huff Post, Le Monde, The Guardian, etc.); nothing surprising considering that they have the highest number of followers. The Ocean Clean Up, Oceana, Earth Day Network, WWF Japan, No Plastic Waste or Greenpeace are some examples of NGOs/Foundations/Nonprofit organizations with tweets with large engagement. Within the initiatives/projects category, Lost at sea, Aplastic Planet and Blue Planet Society are some examples. The presence of both academic institutions and official bodies responsible for the care and protection of the environment is scarce.

3.5. Hashtags and topic modeling

The use of hashtags has the advantage of classifying tweets within a certain topic with some independence of the language. Figure 8 shows the most used hashtags in the complete dataset. Results show a classification of events like the World Ocean Day or Plastic Free July, being the most used hashtag *#plastic* followed by *#plasticpollution*. The positive sentiment outweighs the negative in all hashtags, although neutrality predominates. Although hashtags allow immediate thematic classification of the tweet, they may not be sufficient to determine subtopics. Here, the corpus of the dataset must be used. To simply analyze the occurrence of words, only English tweets were processed to avoid problems related to automatic translation, which could alter the meaning of the word or use different synonyms in the translation. The words *pollution* and *waste* appear in similar frequency (9,121 and 9,115 occurrences, respectively), followed by *use* (8,514), *help* (7,477) and *people* (6,781).

Table 2 shows the top-10 most probable words in the 6 topics generated by LDA, with a weigh parameter λ of 1 and 0.4. Whereas $\lambda = 1$ shows the ranking of terms by their probability in the topic, a lower λ also weights by the probability of appearing in the corpus. At the other extreme, $\lambda = 0$ ranks words that appear exclusively in that topic but not in the others. Thus, decreasing the value of λ means that less frequent and more exclusive words of that topic will rise in the ranking, although that does not necessarily imply that it helps to better define the topic by a human. From model results, the following six topics ordered from larger to lower marginal topic distribution were defined:

i) "*Impact on wildlife*" (20.4% of tokens) that defines concerns about the impact of marine litter, especially plastic bags and straws on marine biota, with particular mention to turtles;
ii) "*Microplastics/Water pollution*" (20.2%) referring to the pollution produced by plastics and microplastics derived from items such as plastic bottles and bags;
iii) "*Estimate of quantities/Reports*" (17.5%) with comments on the amounts of marine litter that impact the

environment based on news from recent studies and reports;

iv) "*Legislation/Protection*" (14.6%) concerning problems of legislation and the need for global treaties to tackle this problem;

v) "*Recycling/Cleaning initiatives*" (11.1%) with comments on citizen initiatives, private companies and NGOs related to the collection and cleaning of marine litter and the reduction of single-use plastics, among others;

vi) tweets with comments about the album "*Plastic beach*" of the British virtual music band "*Gorillaz*" (16.2%), completely unrelated to the subject of this study.

Table 2 also includes a selection of words that, without being in the highest positions in the ranking, may be relevant. Thus, for example, words like *birds, mammals* or *entangled* also define the topic "Impact on wildlife", whereas words like *butts*, *cigarettes* or *fibers* complement the definition of the topic "*Microplastics/Water pollution*". Words about the COVID-19 pandemic situation were mainly related to the topic "*Estimate of quantities/Reports*".

According to Mehrotra et al. (2013), the performance of the topic models produced by LDA on Twitter data is significantly improved when tweets are aggregated by some common factor to produce pseudo-documents for the corpus. Thus, we have also merged documents (tweets) from the same users before performing LDA, similar to a recent study by Dahal et al. (2019). Same hyperparameters and corpus cleaning methods were applied to both methods. To compare the quality by topic between both models, the metric of UMass coherence (Röder et al., 2015) was examined.

The UMass coherence measure assesses topic quality by looking at how frequently words within a topic co-occur in the corpus. The average UMass coherence of the author-pooled LDA was -2.96 and the classical LDA was -4.63. Despite the author-pooled LDA performed better in statistical terms, we found easier to interpret topics determined by classical LDA. For example, The LDAvis tool (Sievert and Shirley, 2014) used to interpret results showed overlap of token clusters from three topics in the author-pooled LDA and some of the words assigned to the topics had relatively little meaning. LDA is fundamentally a statistically trained model and its performance does not always directly translate to better human interpretability. In fact, topics in Table 2 are indeed meaningful, which was the desired result. As Twitter users add hashtags to align their tweets with a specific topic, it is expected that pooling tweets that share the same hashtag would produce better topic models (e.g., Dahal et al., 2019; Steinskog et al., 2017). However, only a fraction of tweets contain hashtags (28%) and the selection of a particular one or group of them would imply abandonee the most of our dataset. Pooling by hashtags is however straightforward in the analysis of other global and well identified issues, as could be the #MeToo movement (e.g., Goel and Sharma, 2020; Manikonda et al., 2018) or #COVID-19 (e.g., Xue et al., 2020).

In conjunction with topic modelling, network visualization has been used to make sense and explore our dataset. Figure 9 shows networks of the 100 most frequent pairs of co-occurring words (bigrams) in tweets on marine pollution by plastics from March 19th to June 1st, 2020, when a large part of countries world-wide were in strict confinement measures due to COVID-19 pandemic. This period is particularly interesting because without differing too much from the network graph for the entire study period (not shown), it highlights relationships with words typical of the pandemic status: *lockdown*, *COVID-19* and *pandemic*. The label propagation method detects 14 communities in the subset, with a modularity value of 0.39. This poor value can be partially explained by the shortlist of pairs used to build the network. Many of these communities are composed of only a pair of words and the largest by 25 nodes. The word *plastic*, our main keyword in this study, is crucial to the network and its centrality value is used to normalize the rest of the nodes. Its community is composed of words such as *bag*, *bottle*, *debris*, *pollution* or *straws*, but also the word *pandemic* belongs to this community. This last word links to a different community formed by the words *led* and *COVID-19.* The *ocean* is the second word in the network with the highest eigenvector value, meaning that it is highly influencing other strongly-connected words in the net. As an example of other useful information that can be extracted from the graph, *plastic* links through *pollution* with a community related to *legislation* and the need of *treaties* at a *global* scale; the low centrality values of this community suggest that it is a well-defined community separated from the rest.

Finally, filtering our dataset by the presence of the word *stomach* in the corpus could help to list the marine species in which plastic ingestion has been observed; alerting even before there is a scientific publication with the observation. Although this word was not among the most frequent ones, 506 tweets were found related to this subtopic.

3.6. Image analysis

A total of 33,285 tweets (24%) contained associated media resources. It is well known that including an image in a tweet increases engagement (Wadhwa et al., 2017). Publishing a tweet with visual content makes the publication more attractive and suggestive for the user. The image catches the interest of the user, and surprises and acts as a claim. The message takes up more physical space on the user's screen and helps the message to be better understood (Polinario, 2016). Tweets with pictures generate greater engagement independently of their content (Carrasco-Polaino et al., 2019).

To know which tweets with associated images aroused the most interest among users, the tweets were classified by the total number of retweets and favorites. In our dataset only the original tweets were kept and not the retweets, therefore, only direct interactions with the original tweet were taken into account. A top-100 list was done and engagement examined. Focusing on the top 10 tweets on this list, 2 of them belonged to accounts that were blocked, 6 to influencers (>180,000 followers), 1 to an NGO with less than 2,000 followers and the last one to an individual (not influencer) with environmental concerns. As some images could not be recovered as the accounts were blocked, the ranking was redone with those images that could be recovered and eliminating those repeated after comparing their hash values. Most of the images (53) belonged to various topics such as various objects, images from awareness campaigns, groups of people and cartoons. Twenty of these images unequivocally captured trash in beach areas and the open sea. The rest of 27 images contained animals for raising awareness purposes and in fact, some of them were damaged or entangled; the most common animal was the turtle (10), followed by fish (5), marine mammals (5), octopus (4), birds (2) and crustaceans (1).

Another objective of this study was to automatically classify tags and images using available artificial intelligence tools and explore their utility to inspect in a second phase the type and quantity of accumulated marine litter in coastal areas. Unfortunately, the use of the Computer Vision API (v3.1; Azure Cognitive Services) was deemed unsatisfactory in this study. From a random selection of 100 images, only 19 were properly described by the tool. The tags attached by the software have not been helpful either. For example, the image of a bucket full of cigarette butts on a sandy floor was described as “a bowl of nuts” and the tags of this image were: ‘bowl’, ‘ground’, ‘floor’, ‘plate’ and ‘tea’. As stated in the API documentation, objects are generally not detected if they are small (less than 5% of the image) and they are not detected if they are arranged closely together, as in the case of hot spots of marine litter. This prospective study prevents us from developing greater efforts, such as the training of a model for object detection, at least until there is a major advance in this field of technology.

4. Conclusions

Twitter promotes the theory of public engagement, allowing users to have conversations, form communities, share content and build relationships (Kietzmann et al., 2011). This paper advises of the potential of this platform to create and spread environmental awareness, in this case, to combat the problem of marine litter in the world connecting leaders, actors, companies, students and the public. This is the first time —at least to the knowledge of the authors— that a scientific article explores the social network Twitter to analyze public awareness about this issue.

The study describes a snapshot of an extremely dynamic social network spanning a period with an exceptional pandemic situation world-wide. Most of the previous studies of various kinds that analyze Twitter do so by only focusing on hashtags and keywords in English. The largest volume of tweets in our dataset is in English, however and thanks to our multilingual approach, it is possible to analyze the differences between countries from a broader point of view. The results show that countries such as the USA, UK, Japan and Canada with a high population and digitization are those with the highest volume of tweets, but when weighted by the number of active users then the topic is led by the UK and European countries like Spain, France or Germany. In contrast, other high populated and digitized countries have relatively low interest in this specific topic. The results also show the high engagement that occurs during the celebration of "World Days" (e.g., #EarthDay, #WorldSeaTurtleDay or #PlasticFreeJuly) and related to the dissemination of reports and scientific studies in traditional media such as newspapers or television which are echoed in this SM. If we put the focus on a higher temporal resolution, we have found that the activity is mainly concentrated on weekdays with differences between countries according to habits and work breaks.

Tweets are often informal, unstructured, making it challenging for deciphering a general discourse when read individually. To know what the part of society that uses this social network is talking about, it is necessary to analyze the tweets in an aggregate way. In this regard, the LDA technique has been effective in distinguishing between different subtopics: i) "*Impact on wildlife*", ii) "*Microplastics/Water pollution*", iii) "*Estimate of quantities/Reports*", iv) "*Legislation/Protection*" and v) "*Recycling/Cleaning initiatives*". Besides, this technique has been useful to distinguish topics that were not directly associated with our objective, allowing us to improve the cleaning technique of the original dataset. The impact of the COVID-19 pandemic has also been evident in the messages with many of these messages referencing to mask and glove waste and its impact on the environment. Within the topic "*Impact on wildlife*" the high number of comments referring to entangled turtles is noteworthy. These tweets presented a slightly negative sentiment as opposed to positive sentiment on the most general topic.

Our results show that NGOs, international organizations and academic institutions do not lead the conversation on marine litter issues, in spite of their high research and environmental awareness efforts on this topic. Bearing in mind the characteristics of the described snapshot may help to adjust the communication plan in Twitter of those institutions that wish to play a relevant role to fight against marine pollution by plastics, environmental awareness and scientific dissemination. This is relevant to offer citizens reliable and certified information, as well as to change habits and to reinforce sustainable behaviour aimed at protecting our seas. By identifying where it is tweeted from and in what language, institutions can focus their efforts in those areas by combining, for example, several institutional Twitter accounts with a regional perspective. Major events may impact how users discuss socio-scientific issues in online media. Thus, to increase recruitment is useful to follow environmental "World Days" to identify and approach people in the marine plastic pollution field with a more general public discourse, as well as invite influencers to join the cause with an objective and truthful discourse. By identifying events and their participants, institutions can increase and diversify their network and reach, useful to identify the type of audience they usually communicate to.

To know the terminology used, the different sub-topics, feelings and reactions are useful clues/guidelines to design an efficient communication strategy. The communication should be bidirectional considering that Twitter is conversation and users choose who to follow. In addition, knowing favourite hours and days for publication is relevant taking into account the “short lifespan” of a tweet (Wilson, 2019). Positive messages are expected to reinforce recruitment and promote activism, for this reason, sentiment analysis is an interesting approach to analyze the before and after of campaigns launched by an institution and that could be followed under the same hashtag.

The images associated with a tweet contribute to increasing its impact; therefore, it is relevant to understand what attracts the attention of the users of the platform. At this point, the use of images containing animals for raising awareness purposes is a popular resource. Our results have also shown an increase in activity after the dissemination of news about relevant or impacting scientific advances. Generally, it is believed that the more individuals use social media, even if just to communicate and connect, the more likely they are to encounter news (Huber et al., 2019). For this reason, institutions should be a source of scientific news that helps to spread a truthful and contrasted discourse.

Another initial objective of this study was to verify if images that corresponded to waste in coastal areas could be automatically filtered and used to create a map of coastal pollution by marine litter. However, the artificial intelligence tool tested could not create correct descriptions of these images, among other reasons, because the objects present were too small and appeared distorted. However, this social network —as well as other popular ones in the use of images such as Instagram— have the potential to support local or regional programs for coastal monitoring of marine litter, through the use of a specific hashtag or user mention. That is why future studies should find the optimal way to use this social network to photograph coastal areas with waste.

Finally, to contribute to this analysis over time, an interactive web application has been made available at http://twilitter.herokuapp.com/. The tool allows the user to follow the temporal evolution, examine areas with the highest volume of tweets, analyze sentiments or check the highest frequency of hashtags, among others.

References

Abreo, N.A.S., Thompson, K.F., Arabejo, G.F.P., Superio, M.D.A., 2019. Social media as a novel source of data on the impact of marine litter on megafauna: The Philippines as a case study. Marine Pollution Bulletin 140, 51–59. https://doi.org/10.1016/j.marpolbul.2019.01.030

Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Chowell, G., 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research -- an international collaboration. arXiv:2004.03688.

Bank, M.S., Ok, Y.S., Swarzenski, P.W., 2020. Microplastic's role in antibiotic resistance. Science 369, 1315–1315. https://doi.org/10.1126/science.abd9937

Barker, J.L.P., Macleod, C.J.A., 2019. Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. Environmental Modelling & Software 115, 213–227. https://doi.org/10.1016/j.envsoft.2018.11.013

Becken, S., Stantic, B., Chen, J., Alaei, A.R., Connolly, R.M., 2017. Monitoring the environment and human sentiment on the Great Barrier Reef: Assessing the potential of collective sensing. Journal of Environmental Management 203, 87–97. https://doi.org/10.1016/j.jenvman.2017.07.007

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Borrelle, S.B., Ringma, J., Law, K.L., Monnahan, C.C., Lebreton, L., McGivern, A., Murphy, E., Jambeck, J., Leonard, G.H., Hilleary, M.A., Eriksen, M., Possingham, H.P., De Frond, H., Gerber, L.R., Polidoro, B., Tahir, A., Bernard, M., Mallos, N., Barnes, M., Rochman, C.M., 2020. Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution. Science 369, 1515–1518. https://doi.org/10.1126/science.aba3656

Budenz, A., Klassen, A., Purtle, J., Yom Tov, E., Yudell, M., Massey, P., 2020. Mental illness and bipolar disorder on Twitter: implications for stigma and social support. J Ment Health 29, 191–199. https://doi.org/10.1080/09638237.2019.1677878

Carrasco-Polaino, R., Villar-Cirujano, E., Martín-Cárdaba, M.-Á., 2019. Redes, tweets y engagement: análisis de las bibliotecas universitarias españolas en Twitter. Profesional de la Información 28. https://doi.org/10.3145/epi.2019.jul.15

Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70, 066111. https://doi.org/10.1103/PhysRevE.70.066111

Collins, K., Shiffman, D., Rock, J., 2016. How Are Scientists Using Social Media in the Workplace? PLOS ONE 11, e0162680. https://doi.org/10.1371/journal.pone.0162680

Cordasco, G., Gargano, L., 2011. Community Detection via Semi-Synchronous Label Propagation Algorithms. arXiv:1103.4550 [physics]. https://doi.org/10.1504/..045103

Dahal, B., Kumar, S.A.P., Li, Z., 2019. Topic modeling and sentiment analysis of global climate change tweets. Soc. Netw. Anal. Min. 9, 24. https://doi.org/10.1007/s13278-019-0568-8

Dang, T., Nguyen, N.V.T., Pham, V., 2018. HealthTvizer: Exploring Health Awareness in Twitter Data through Coordinated Multiple Views, in: 2018 IEEE International Conference on Big Data (Big Data). Presented at the 2018 IEEE International Conference on Big Data (Big Data), pp. 3647–3655. https://doi.org/10.1109/BigData.2018.8622445

Duckett, P.E., Repaci, V., Duckett, P.E., Repaci, V., 2015. Marine plastic pollution: using community science to address a global problem. Mar. Freshwater Res. 66, 665–673. https://doi.org/10.1071/MF14087

Eberhard, D., Simons, G.F., Fenning, C., 2019. Ethnologue: Languages of the World. SIL International.

European Commission, 2018. Single-use plastics: New EU rules to reduce marine litter [WWW Document]. Single-use plastics: New EU rules to reduce marine litter. URL https://ec.europa.eu/commission/presscorner/detail/en/MEMO_18_3909 (accessed 12.28.20).

Filgueiras, A.V., Preciado, I., Cartón, A., Gago, J., 2020. Microplastic ingestion by pelagic and benthic fish and diet composition: A case study in the NW Iberian shelf. Marine Pollution Bulletin 160, 111623. https://doi.org/10.1016/j.marpolbul.2020.111623

Forleo, M.B., Romagnoli, L., 2021. Marine plastic litter: public perceptions and opinions in Italy. Marine Pollution Bulletin 165, 112160. https://doi.org/10.1016/j.marpolbul.2021.112160

Fruchterman, T.M.J., Reingold, E.M., 1991. Graph drawing by force-directed placement. Softw: Pract. Exper. 21, 1129–1164. https://doi.org/10.1002/spe.4380211102

Gago, J., Portela, S., Filgueiras, A.V., Salinas, M.P., Macías, D., 2020. Ingestion of plastic debris (macro and micro) by longnose lancetfish (Alepisaurus ferox) in the North Atlantic Ocean. Regional Studies in Marine Science 33, 100977. https://doi.org/10.1016/j.rsma.2019.100977

Gall, S.C., Thompson, R.C., 2015. The impact of debris on marine life. Marine Pollution Bulletin 92, 170–179. https://doi.org/10.1016/j.marpolbul.2014.12.041

GESAMP, 2020. Proceedings of the GESAMP International Workshop on assessing the risks associated with plastics and microplastics in the marine environment, Journal Series GESAMP Reports and Studies. GESAMP Office, London.

GESAMP, 2015. Sources, fate and effects of microplastics in the marine environment: a global assessment (No. 90), Rep. Stud. GESAMP. IMO, London.

Geyer, R., Jambeck, J.R., Law, K.L., 2017. Production, use, and fate of all plastics ever made. Science Advances 3, e1700782. https://doi.org/10.1126/sciadv.1700782

Ghermandi, A., Camacho-Valdez, V., Trejo-Espinosa, H., 2020. Social media-based analysis of cultural ecosystem services and heritage tourism in a coastal region of Mexico. Tourism Management 77, 104002. https://doi.org/10.1016/j.tourman.2019.104002

Goel, R., Sharma, R., 2020. Understanding the MeToo Movement Through the Lens of the Twitter, in: Aref, S., Bontcheva, K., Braghieri, M., Dignum, F., Giannotti, F., Grisolia, F., Pedreschi, D. (Eds.), Social Informatics, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 67–

80. https://doi.org/10.1007/978-3-030-60975-7_6

Hagberg, A., Swart, P., S Chult, D., 2008. Exploring network structure, dynamics, and function using networkx (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab. (LANL), Los Alamos, NM (United States).

Heidbreder, L.M., Bablok, I., Drews, S., Menzel, C., 2019. Tackling the plastic problem: A review on perceptions, behaviors, and interventions. Science of The Total Environment 668, 1077–1093. https://doi.org/10.1016/j.scitotenv.2019.02.437

Huber, B., Barnidge, M., Gil de Zúñiga, H., Liu, J., 2019. Fostering public trust in science: The role of social media. Public Underst Sci 28, 759–777. https://doi.org/10.1177/0963662519869097

Jambeck, J.R., Geyer, R., Wilcox, C., Siegler, T.R., Perryman, M., Andrady, A., Narayan, R., Law, K.L., 2015. Plastic waste inputs from land into the ocean. Science 347, 768–771. https://doi.org/10.1126/science.1260352

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

Kemp, S., 2020. Digital 2020: Global Digital Overview. Essential insights into how people around the world use the internet, mobile devices, social media and ecommerce. We Are Social and Hotsuite®.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S., 2011. Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons, SPECIAL ISSUE: SOCIAL MEDIA 54, 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Law, K.L., Starr, N., Siegler, T.R., Jambeck, J.R., Mallos, N.J., Leonard, G.H., 2020. The United States' contribution of plastic waste to land and ocean. Science Advances 6, eabd0288. https://doi.org/10.1126/sciadv.abd0288

Letierce, J., Passant, S., Brelisn, J.G., 2010. Understanding how Twitter is used to spread scientific message. Presented at the Proceedings of the WebSci10: Extending the Frontiers of Society On-Line.

Li, X., Xie, Q., Huang, L., Yuan, Z., 2017. Twitter Data Mining for the Social Awareness of Emerging Technologies, in: 2017 Portland International Conference on Management of Engineering and Technology (PICMET). Presented at the 2017 Portland International Conference on Management of Engineering and Technology (PICMET), pp. 1–10. https://doi.org/10.23919/PICMET.2017.8125279

Makita, M., Mas-Bleda, A., Morris, S., Thelwall, M., 2021. Mental Health Discourses on Twitter during Mental Health Awareness Week. Issues in Mental Health Nursing 42, 437–450. https://doi.org/10.1080/01612840.2020.1814914

Manikonda, L., Beigi, G., Liu, H., Kambhampati, S., 2018. Twitter for Sparking a Movement, Reddit for Sharing the Moment: #metoo through the Lens of Social Media. arXiv:1803.08022 [cs].

Martínez-Rojas, M., Pardo-Ferreira, M. del C., Rubio-Romero, J.C., 2018. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. International Journal of Information Management 43, 196–208. https://doi.org/10.1016/j.ijinfomgt.2018.07.008

Media Insight Project, 2017. ‘Who shared it?’ How Americans decide what news to trust on social media. American Press Institute. URL https://www.americanpressinstitute.org/publications/reports/survey-research/trust-social-media/ (accessed 5.3.21).

Mehrotra, R., Sanner, S., Buntine, W., Xie, L., 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’13. Association for Computing Machinery, New York, NY, USA, pp. 889–892. https://doi.org/10.1145/2484028.2484166

Milani, E., Weitkamp, E., Webb, P., 2020. The Visual Vaccine Debate on Twitter: A Social Network Analysis. Media and Communication 8, 364–375. https://doi.org/10.17645/mac.v8i2.2847

Mitrano, D.M., Wohlleben, W., 2020. Microplastic regulation should be more precise to incentivize both innovation and environmental safety. Nature Communications 11, 5324. https://doi.org/10.1038/s41467-020-19069-1

Mohammadi, E., Thelwall, M., Kwasny, M., Holmes, K.L., 2018. Academic information on Twitter: A user survey. PLOS ONE 13, e0197265. https://doi.org/10.1371/journal.pone.0197265

Napper, I.E., Thompson, R.C., 2020. Plastic Debris in the Marine Environment: History and Future Challenges. Global Challenges 4, 1900081. https://doi.org/10.1002/gch2.201900081

Newman, M.E.J., 2010. Networks: an introduction. Oxford University Press, Oxford ; New York.

Pabortsava, K., Lampitt, R.S., 2020. High concentrations of plastic hidden beneath the surface of the Atlantic Ocean. Nature Communications 11, 4073. https://doi.org/10.1038/s41467-020-17932-9

Parton, K., Galloway, T., Godley, B., 2019. Global review of shark and ray entanglement in anthropogenic marine debris. Endang. Species. Res. 39, 173–190. https://doi.org/10.3354/esr00964

Plackett, R., Kaushal, A., Kassianos, A.P., Cross, A., Lewins, D., Sheringham, J., Waller, J., von Wagner, C., 2020. Use of Social Media to Promote Cancer Screening and Early Diagnosis: Scoping Review. J Med Internet Res 22, e21582. https://doi.org/10.2196/21582

Polinario, J., 2016. Cómo divulgar ciencia a través de las redes sociales, Investigación. Circulo Rojo.

Retka, J., Jepson, P., Ladle, R.J., Malhado, A.C.M., Vieira, F.A.S., Normande, I.C., Souza, C.N., Bragagnolo, C., Correia, R.A., 2019. Assessing cultural ecosystem services of a large marine protected area through social media photographs. Ocean & Coastal Management 176, 40–48. https://doi.org/10.1016/j.ocecoaman.2019.04.018

Röder, M., Both, A., Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM ’15. Association for Computing Machinery, New York, NY, USA, pp. 399–408. https://doi.org/10.1145/2684822.2685324

Ruiz-Frau, A., Ospina-Alvarez, A., Villasante, S., Pita, P., Maya-Jariego, I., de Juan, S., 2020. Using graph theory and social media data to assess cultural ecosystem services in coastal areas: Method development and application. Ecosystem Services 45, 101176. https://doi.org/10.1016/j.ecoser.2020.101176

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., Menczer, F., 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. Proceedings of the 29th ACM International Conference on Information & Knowledge Management 2725–2732. https://doi.org/10.1145/3340531.3412698

Shimkhada, R., Attai, D., Scheitler, A.J., Babey, S., Glenn, B., Ponce, N., 2021. Using a Twitter Chat to Rapidly Identify Barriers and Policy Solutions for Metastatic Breast Cancer Care: Qualitative Study. JMIR Public Health and Surveillance 7, e23178. https://doi.org/10.2196/23178

Sievert, C., Shirley, K., 2014. LDAvis: A method for visualizing and interpreting topics, in: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 63–70. https://doi.org/10.3115/v1/W14-3110

Smith, A., 2015. “Wow, I didn’t know that before; thank you”: How scientists use Twitter for public engagement. Journal of Promotional Communications 3.

STATISTA, 2020. Leading countries based on number of Twitter users as of October 2020. URL https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed 12.28.20).

Steinskog, A., Therkelsen, J., Gambäck, B., 2017. Twitter Topic Modeling by Tweet Aggregation, in: Proceedings of the 21st Nordic Conference on Computational Linguistics. Association for Computational Linguistics, Gothenburg, Sweden, pp. 77–86.

Taddy, M.A., 2011. On Estimation and Selection for Topic Models. arXiv:1109.4518 [stat].

Teoh, D., Shaikh, R., Vogel, R.I., Zoellner, T., Carson, L., Kulasingam, S., Lou, E., 2018. A Cross-Sectional Review of Cervical Cancer Messages on Twitter during Cervical Cancer Awareness Month. J Low Genit Tract Dis 22, 8–12. https://doi.org/10.1097/LGT.0000000000000363

The Pew Charitable Trusts and SYSTEMIQ, 2020. Breaking the Plastic Wave: A Comprehensive Assessment of Pathways Towards Stopping Ocean Plastic Pollution. The Pew Charitable Trusts and SYSTEMIQ.

Van Noorden, R., 2014. Online collaboration: Scientists and the social network. Nature News 512, 126. https://doi.org/10.1038/512126a

Vince, J., Hardesty, B.D., 2017. Plastic pollution challenges in marine and coastal environments: from local to global governance. Restoration Ecology 25, 123–128. https://doi.org/10.1111/rec.12388

Vraga, E.K., Stefanidis, A., Lamprianidis, G., Croitoru, A., Crooks, A.T., Delamater, P.L., PFOSER, D., Radzikowski, J.R., Jacobsen, K.H., 2018. Cancer and Social Media: A Comparison of Traffic about

Breast Cancer, Prostate Cancer, and Other Reproductive Cancers on Twitter and Instagram. Journal of Health Communication 23, 181–189. https://doi.org/10.1080/10810730.2017.1421730

W3Techs, 2020. Usage statistics of content languages for websites. URL https://w3techs.com/technologies/overview/content_language (accessed 12.28.20).

Wadhwa, V., Latimer, E., Chatterjee, K., McCarty, J., Fitzgerald, R.T., 2017. Maximizing the Tweet Engagement Rate in Academia: Analysis of the AJNR Twitter Feed. American Journal of Neuroradiology 38, 1866–1868. https://doi.org/10.3174/ajnr.A5283

Warner, K., Linske, E., Mustain, P., Valliant, M., Leavitt, C., 2020. Choked, Strangled, Drowned: The Plastics Crisis Unfolding In Our Oceans. OCEANA.

Wilson, C., 2019. Updated: Lifespan of a Social Media Post. Max Influence. URL https://mtomconsulting.com/updated-lifespan-of-a-social-media-post/ (accessed 12.28.20).

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

Yoosefi Nejad, M., Delghandi, M.S., Bali, A.O., Hosseinzadeh, M., 2019. Using Twitter to raise the profile of childhood cancer awareness month. Netw Model Anal Health Inform Bioinforma 9, 3. https://doi.org/10.1007/s13721-019-0206-4

Zhou, S., Kan, P., Huang, Q., Silbernagel, J., 2021. A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura. Journal of Information Science 01655515211007724. https://doi.org/10.1177/01655515211007724

Zubiaga, A., Ji, H., 2014. Tweet, but verify: epistemic study of information verification on Twitter. Soc. Netw. Anal. Min. 4, 163. https://doi.org/10.1007/s13278-014-0163-y

List of tables

Table 1. Example of some fields from tweets in the dataset. Time and original message were directly obtained from Twitter API. Clean text after processing the original message is also shown in the table. The city, state, and country fields were calculated from user profile and added to the dataset. Polarity [-1, 1] and subjectivity [0, 1] are also shown. User identifier is not shown to protect the identities of the Twitter users.

Time (UTC)	Original message	Clean text	City, State, Country	(Polarity, subjectivity)
Sat Aug 08 19:52:52 2020	A pocos metros de una playa del sur.. recogimos en 10 minutos 7 bolsas de latas y botellas de plástico abandonadas...falta concienciación de limpieza y civismo en la isla. @GranCanariaCab https://t.co/A4pCuWyP2G	meters southern beach collected minutes bags abandoned cans plastic bottles lack awareness cleanliness civility island	Las Palmas de Gran Canaria, Islas Canarias, Spain	(-0.1, 0.05)
Tue Jul 28 22:34:55 2020	Last week, we did a beach cleanup with @LagunaOceanFdn at Aliso Beach, the end of the Aliso Creek watershed! We picked up 10 pounds of trash in ONE hour, prevented gulls from eating plastic cups, saw native &… https://t.co/hUpG2Vu03B	last week beach cleanup aliso beach end aliso creek watershed picked pounds trash one hour prevented gulls eating plastic cups saw native	Costa Mesa, California, USA	(0, 0.07)
Thu Aug 13 10:45:04 2020	Nearly half of the plastic found in the ocean comes from fishing nets. People are reducing the consumption of plastics, but given that the scientific community warned that by 2050 there will be more plastic in the ocean than fish, it is not enough. https://t.co/CmDTOyhBoL	nearly half plastic found ocean comes fishing nets people reducing consumption plastics given scientific community warned plastic ocean fish enough	Allerdale, England, UK	(0.11, 0.38)

Table 2. The 10-top most salient words for 6 topics generated from LDA analysis. Words are in descending order of relevance following the definition of Sievert et al. (2014), computed with a weight parameter λ = 1 and λ = 0.4.

Human interpretation	**Top-10 salient terms (λ = 1)**	**Top-10 salient terms (λ = 0.4)**	**Selection of terms (λ < 0.4)**
Impact on wild life	bag, waste, turtle, use, life, straws, stop, animals, trash, single	turtle, bag, animals, straws, waste, stop, use, life, killing, fishing	birds, creatures, jellyfish, dolphins, killing, nets, harming, mammals, entangled, whales, stomach
Microplastics / Water pollution	water, bottle, new, pollution, clean, bag, micro, study, food, litter	water, bottle, micro, atlantic, new, particles, tiny, times, study, surface	butts, cigarettes, cans, technology, discovered, clothes, fragments, fibers
Estimate of quantities / Reports	pollution, million, year, waste, tons, world, use, fish, people, end	million, year, tons, pollution, estimated, metric, likely, lets, report, waste	lockdown, coronavirus, consequence, biodiversity, mediterranean
Legislation / Protection	global, pollution, help, hi, states, climate, protecting, members, legislation, treaty	global, hi, states, protecting, climate, member, legislation, treaty, requires, encouraging	unenvironment, truth, warming, loveplanet, protectdepends, spain, fuels, fosil, oil, forest, melting
Recycling / Cleaning initiatives	free, pollution, help, single, use, people, clean, save, july, make	free, july, solution, act, minute, challenge, cleaner, helps, communities, signed	Trump, refuse, congress, fund, movement
Gorillaz*	gorillaz, album, days, like, masks, good, best, demon, song, love	gorillaz, album, days, masks, demon, best, song, face, covid, good	-

*Gorillaz topic (British virtual band), completely unrelated to marine pollution by plastics and microplastics. LDA revealed this topic and helped to improve cleaning in the dataset.

List of figures

Figure 1: Flowchart of the dataset preparation and data analysis.

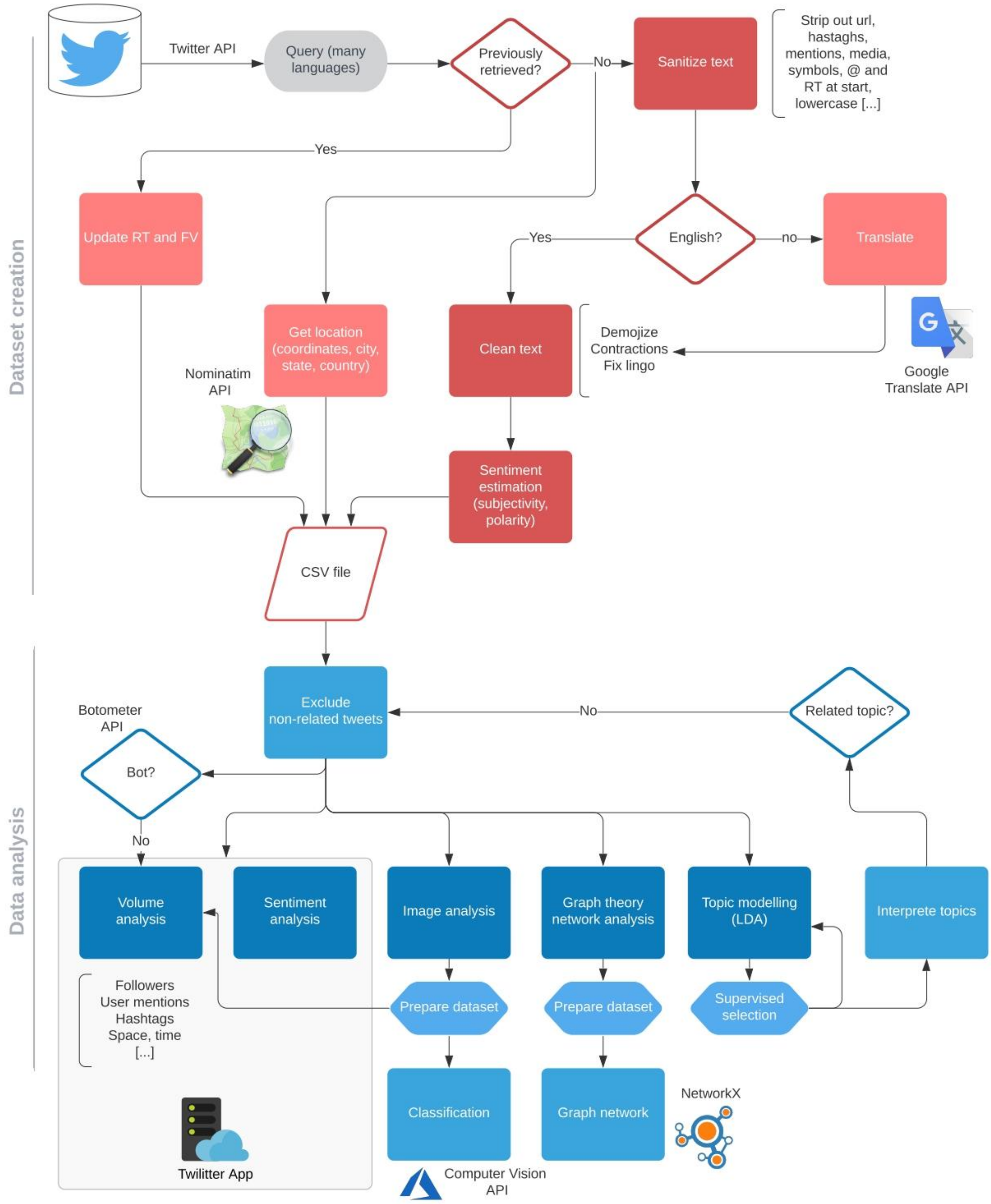

Figure 2: Heatmap of tweets from 19th March to 26th November 2020. Locations were retrieved from both geotagged tweets and from Twitter user's profile.

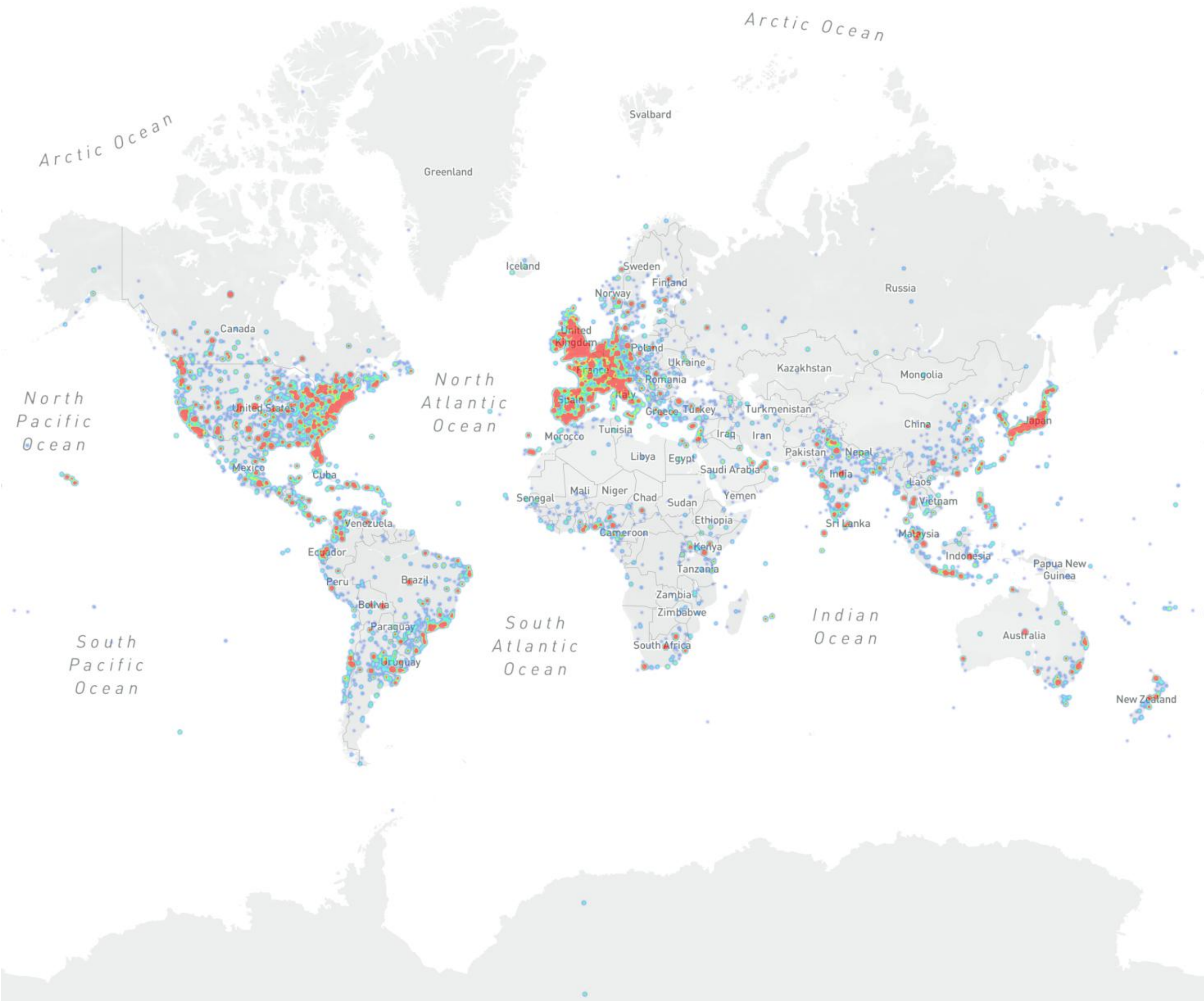

Figure 3: Top-25 countries sorted by a) number of tweets and b) number of tweets per capita normalized by the maximum ratio among countries (UK).

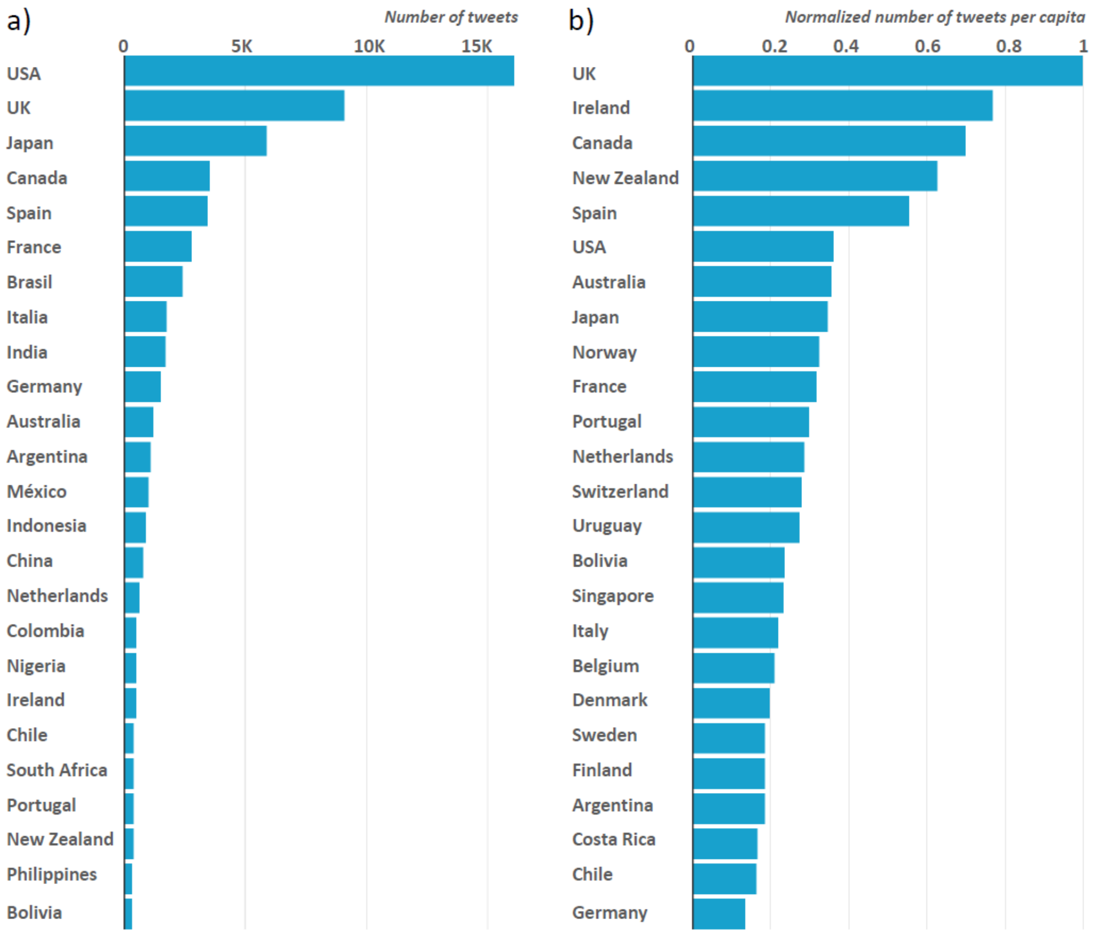

Figure 4: Total number of tweets per day (dark grey), with positive sentiment (green) and negative (red). Relevant events are annotated on the figure.

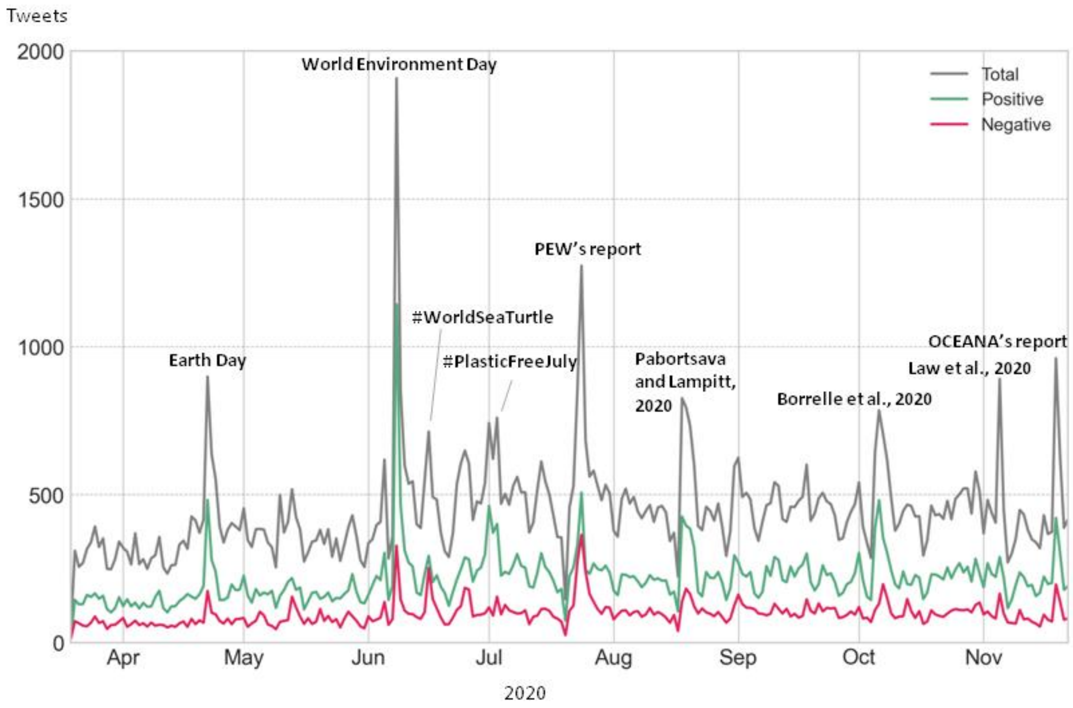

Figure 5: Average weekly tweets within each time slot (UTC hours)

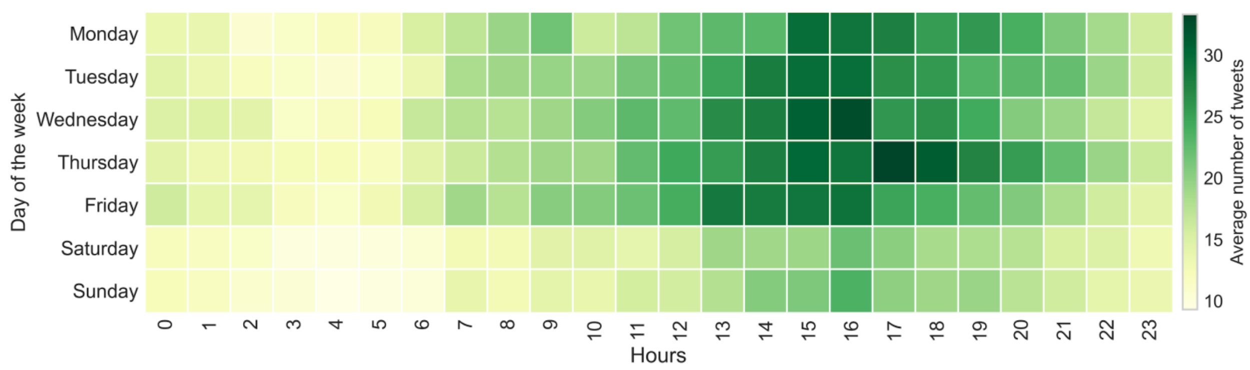

Figure 6: Sentiment analysis distribution for the total of tweets: a) polarity of positive and negative tweets, b) subjectivity of positive tweets and c) subjectivity of negative tweets. Subjectivity ranges from 0 (very objective) to 1 (very subjective).

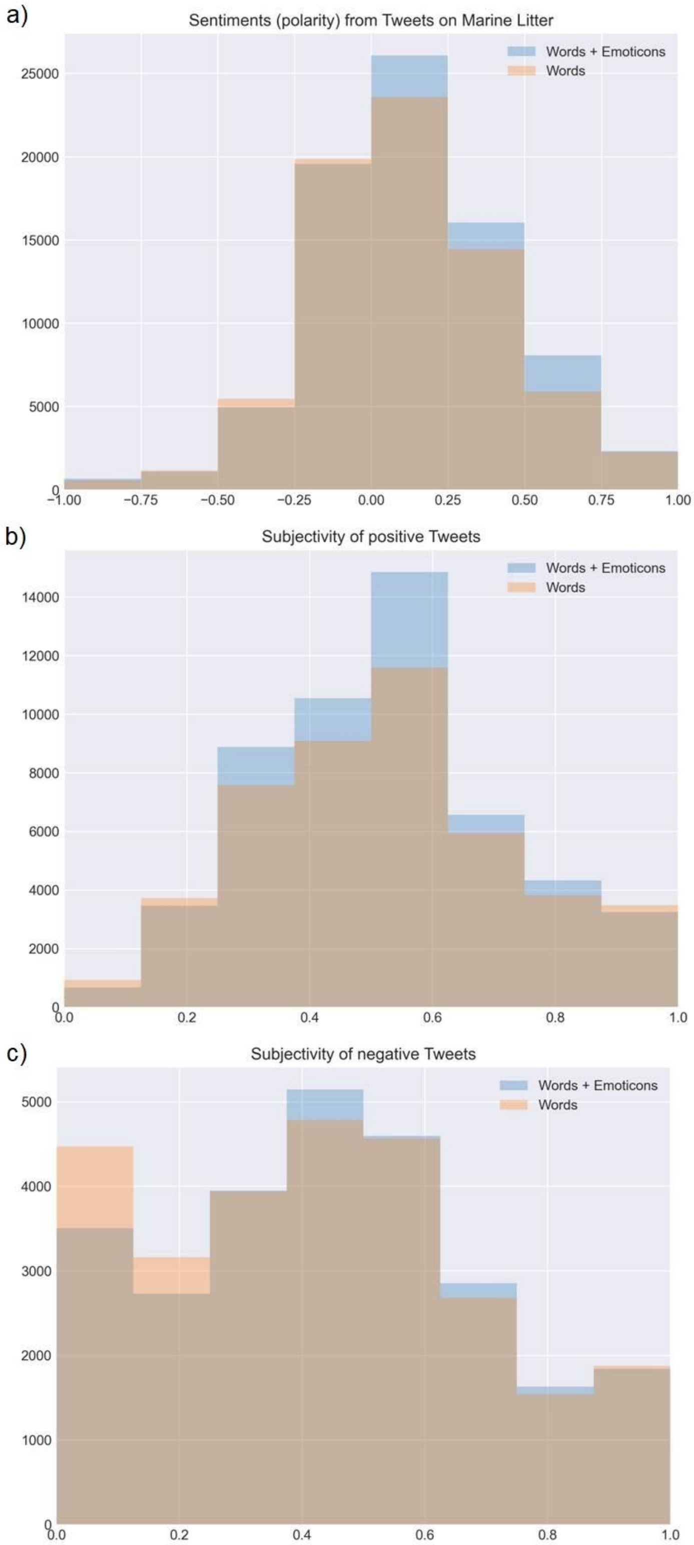

Figure 7: Top-100 users by tweet volume, engagement (likes plus retweets) and the number of followers categorized by bots, companies, official institutions, initiatives/projects and NGOs/foundation/Nonprofit organizations. In the analysis by followers, the category of companies is made up of 90% by mass media.

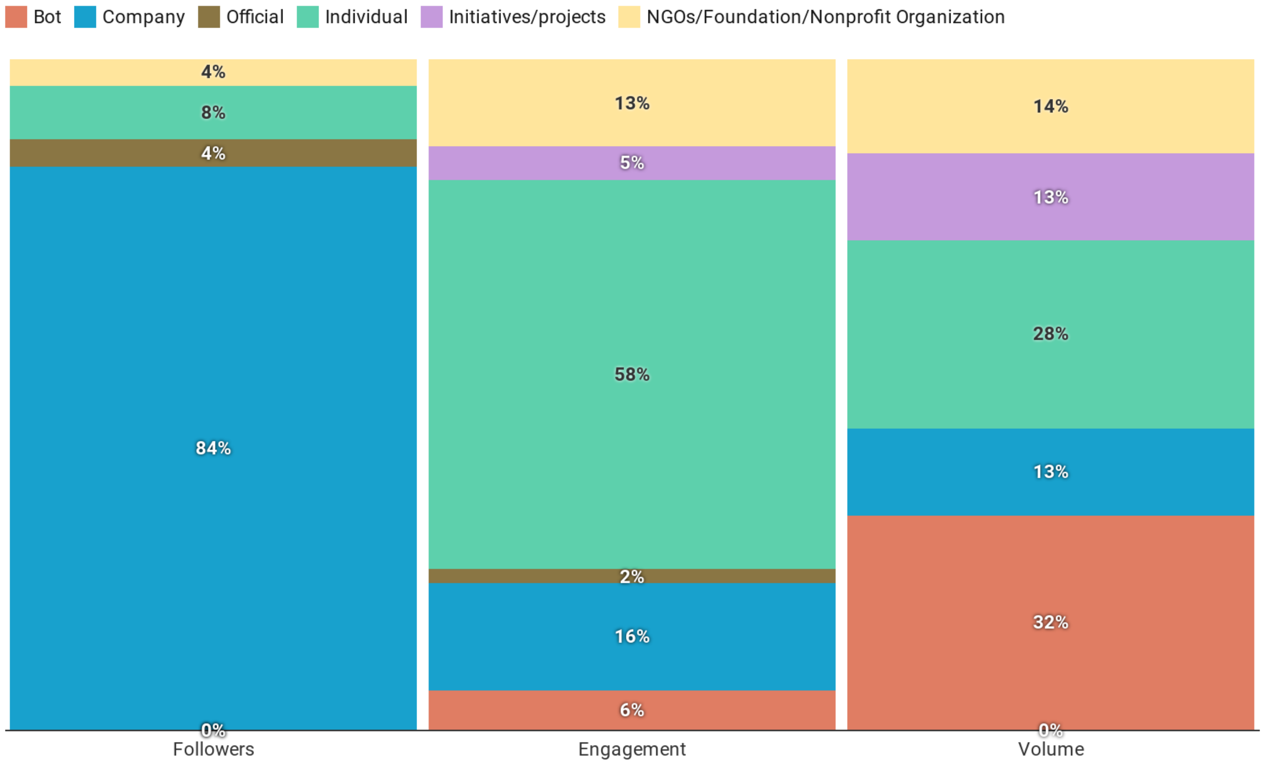

Figure 8: Top-25 most used hashtags sorted by the number of appearances in the dataset. The proportion of negative (<-0.3), positive (>0.3) and neutral sentiment is also shown.

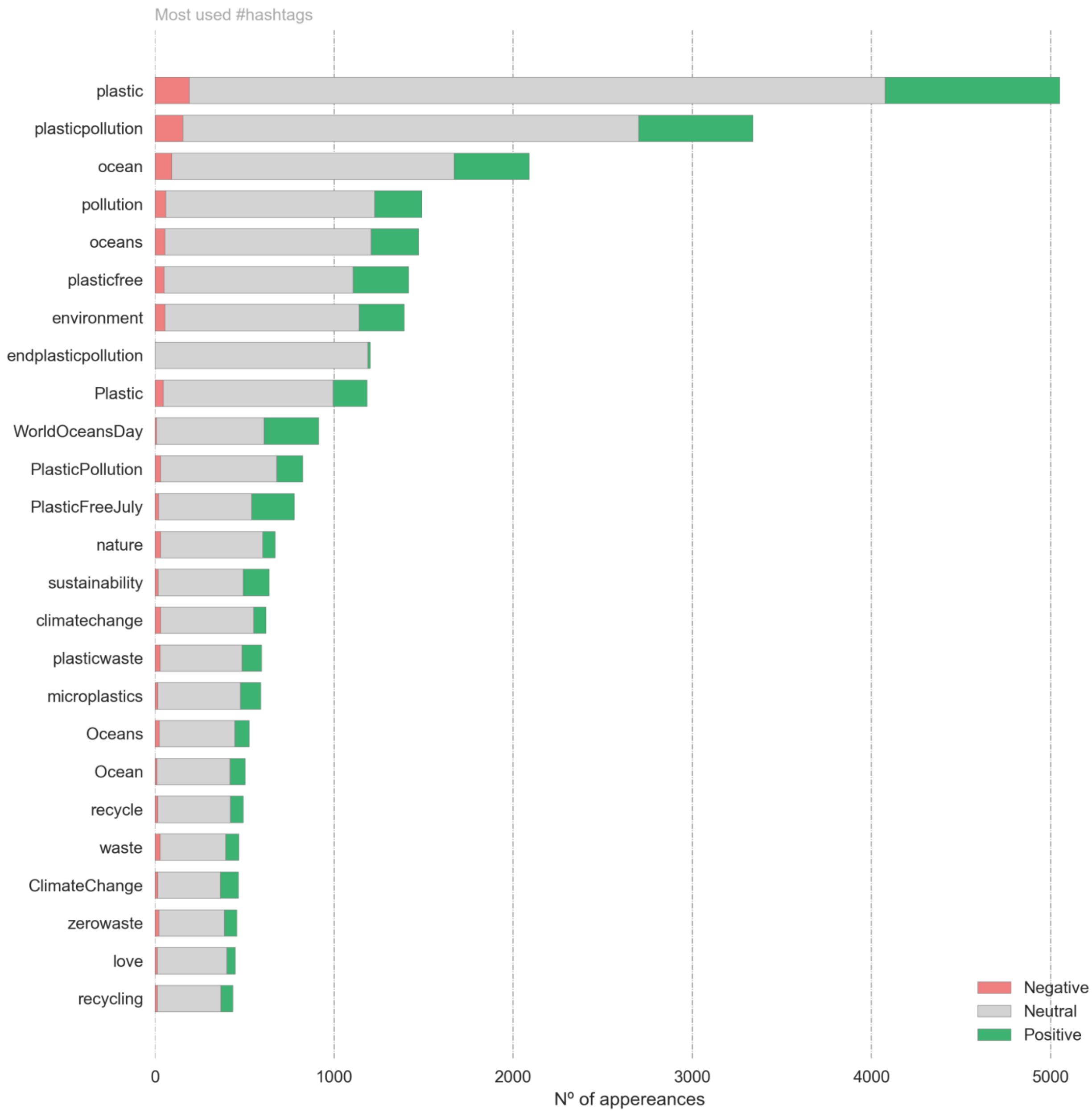

Figure 9: Networks of the 100 most frequent co-occurring words in tweets on marine pollution by plastics from 19th March to 1st June 2020 (time of the greatest lockdowns world-wide due to the COVID-19 pandemic, just before the summer relaxations). Nodes with the same colour belong to the same community. Node size represents eigenvector centrality normalized by the maximum centrality value in the network landscape (corresponding to the word *plastic*).