‘Human sensors’ for monitoring Great Barrier Reef environmental changes and quality of marine waters through harnessing Big Data analysis

Professor Susanne Becken

Professor Bela Stantic

Professor Rod Connolly

Ms Jinyan Chen

Griffith Institute for Tourism Research Internal Report

December 2017

Griffith University, Queensland, Australia

About this report:

This report is a final output from project 2.3.2: ‘Human sensors’ for monitoring Great Barrier Reef environmental changes and quality of marine waters through harnessing Big Data analysis. The project is funded by the National Environment Science Program (NESP) Tropical Water Quality Hub.

Disclaimer:

By using this information you acknowledge that this information is provided by the Griffith Institute for Tourism (GIFT). You agree to release and indemnify GIFT for any loss or damage that you may suffer as a result of your reliance on this information. The producing University does not represent or warrant that this information is correct, complete or suitable for the purpose for which you wish to use it.

Organisations involved:

Griffith University is a top ranking University based in South East Queensland, Australia. Griffith University hosts the Griffith Institute for Tourism, a world-leading institute for quality research into tourism. The Institute links university-based researchers with the business sector and organisations, as well as local, state and federal government bodies. Professor Bela Stantic leads the ‘Big Data and Smart Analytics’ lab within the Institute for Integrated and Intelligent Systems - IIIS, Griffith University. Professor Rod Connolly is affiliated with the Australian Rivers Institute at Griffith University.

Executive Summary

The Great Barrier Reef (GBR) is a heavily visited area (about 4 million per year) and in addition, about one million local residents live on the nearby coast. Both visitors and residents are using social media to share content about their surroundings, perceptions and experiences with the Reef. This data was shown to be useful for monitoring of human sentiment and to some extent the perception of environmental conditions.

This final report summarise research on whether informal information from social media can complement existing citizen science approaches and biophysical monitoring. The report details the information technology architecture developed for this project, the specific data sources that were examined and key results that illustrate the range of findings and insights. Where possible, the report draws on and provides reference to already published work where more detail can be found.

These social media data were analysed in terms of overall volumes, frequency of particular keywords, and sentiment. A comparison between Twitter feeds and Facebook comments/posts is made and differences are identified, for example the more emotional and experience-focused nature of Facebook communications compared with the more factual tweets. It was found that certain keywords attract negative sentiment, although the frequency was relatively low, even for major events (e.g. ‘bleaching’).

A comparison of different data sources along the spectrum of collective sensing, citizen science, and professional monitoring provides evidence that a portfolio approach can be beneficial. Early analysis of correlating sentiment and weather is promising.

Finally, a proof of concept web platform has been developed and tested with key stakeholders. The platform visualises information extracted from Twitter, for example what activities Twitter users engaged in, what sentiment they displayed, and where they were from. Further work would be useful to turn this concept into a functioning tool.

TABLE OF CONTENTS

1. Introduction

1.1. Background

This project explores the potential of using ‘human sensors’ to improving monitoring of environmental conditions in real time at the Great Barrier Reef (GBR). The data mining integrates human sensing data (e.g. from social media) with existing monitoring data, meteorological data, tourism statistics, and others data sources. This project aims to demonstrate how citizen/visitor data can complement other relevant data to explore new ways of monitoring environmental change

The project was undertaken against the background of increasing pressure on the Great Barrier Reef Marine Park, including deteriorating water quality and two consecutive coral bleaching events (GBRMPA, 2017). The GBR is a UNESCO World Heritage Area stretching over 2,000 kilometres along the Queensland coast. The Great Barrier Reef (GBR) is integral to how Australians define their identity and an iconic tourism destination (Becken et al., 2014). Over 2.2 million international and 1.7 million domestic visitors travel to the region every year. Not surprisingly then, the largest economic benefit associated with the GBR comes from tourism. A Deloitte Access Economics (2017) study revealed that tourism generates an estimated AU$6.4 billion per year and sustains over 64,000 jobs.

In addition to regional economic benefits, tourism contributes directly to the environmental management of the GBR through an Environmental Management Charge (EMC). Collecting the EMC also provides important visitor statistics to Reef stakeholders. Accordingly, the Great Barrier Reef Marine Park Authority (GBRMPA, 2016) reported 2.62 million visitor days spent on the GBR for the financial year ending 30 June 2016. Visitor statistics include trips to the Reef on commercial vessels of various forms, as well as scenic flights.

1.2. Research aim

This research project examined whether visitors to the Reef and other users (i.e. residents) talk about the marine environment in their social media interactions, and whether information contained in such posts is useful for GBR managers for the benefit of monitoring environmental change or alerts.

The aim of this research therefore was to assess whether people use social media to talk about the GBR, what the topic of their posts is, and whether messages reflect a positive or negative sentiment. Further, the researcher identified correlations of social media data with other forms of monitoring, including those related to citizen science and professional monitoring.

2. Different types of data

The underlying assumption of this project was that the many different types of Internet-based data sources (e.g. micro blogs) contain some useful information about the Great Barrier Reef. It was recognised early on that these types of crowd sensed data are in contrast to more structured data generated through citizen science programs and professional monitoring (Figure 1). Importantly, though a comparison of the data might enable an evaluation and verification of different data sources, and ultimately allow a portfolio approach to monitoring.

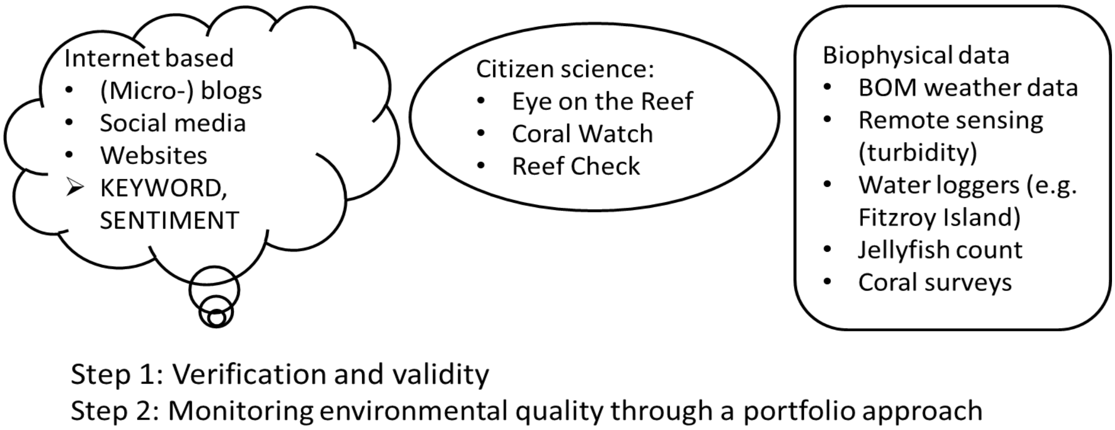


Figure 1 Overview and examples of different data sources

More specifically, the research revealed that different data sources have different characteristics. Figure 2 summarises the differences between collective sensing, human sensing, citizen science and professional monitoring.

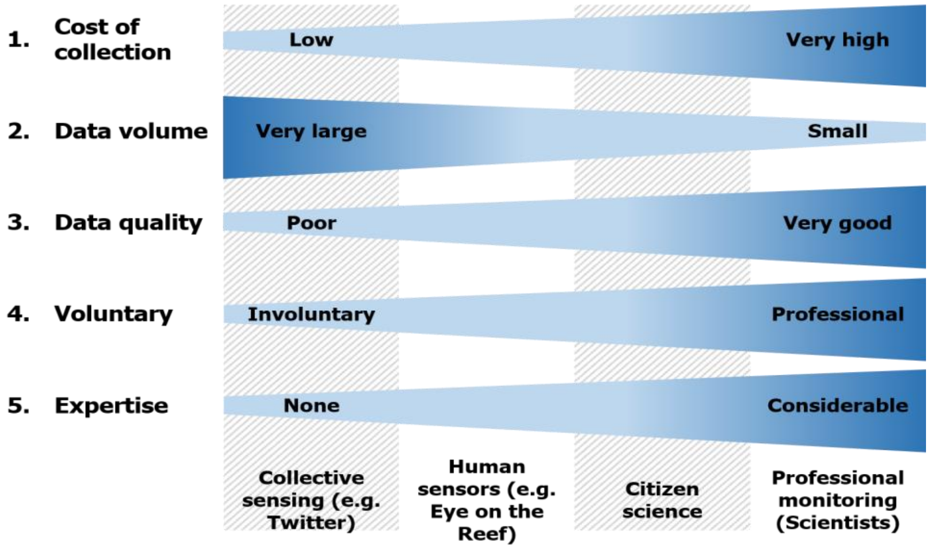
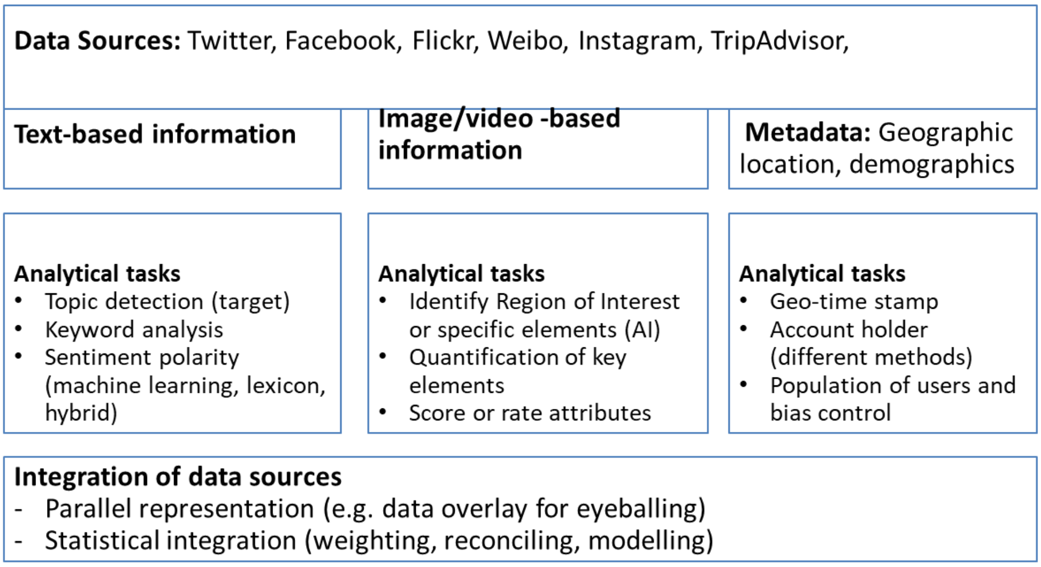


Figure 2 Comparison of data in terms of key characteristics.

This research helped classify social media data into different types of information they provide, namely text-base, images and metadata. Figure 3 shows the classification and provides examples of the types of analysis that are possible. This particular project did not include image-based analyses, but a a related NESP-funded project (3.2.3) is investigating the value of user supplied images and videos.



Data Sources: Twitter, Facebook, Flickr, Weibo, Instagram, TripAdvisor,

Text-based information

Image/video -based information

Metadata: Geographic location, demographics

Analytical tasks
- Topic detection (target)
- Keyword analysis
- Sentiment polarity (machine learning, lexicon, hybrid)

Analytical tasks
- Identify Region of Interest or specific elements (AI)
- Quantification of key elements
- Score or rate attributes

Analytical tasks
- Geo-time stamp
- Account holder (different methods)
- Population of users and bias control

Integration of data sources
- Parallel representation (e.g. data overlay for eyeballing)
- Statistical integration (weighting, reconciling, modelling)

Figure 3 Different types of information that can be extracted from social media sources.

3. Social media and environmental monitoring

The social media landscape has evolved rapidly and this has been documented in more detail in Becken et al. (2017 a,b). However, the use of social media data in the context of environmental monitoring is limited. Social media data are used in disaster and crisis management (Vivacqua & Borges, 2012; Steiger, de Albuquerque & Zipf, 2015). Capitalising on the real-time spread of online information via such channels, the U.S. Geological Service now monitors seismological activity by data mining of Twitter feeds in addition to its network of sensors (Meyer, 2015).

The advantages of accessing large numbers of evidenced in social media observations or ‘measurements’ on specific phenomena are beginning to be recognised in the ecological domain. Recent research in the United States, for example, used photo imagery uploaded on Flickr, a photo-sharing website, to replace costly visitor surveys for monitoring the number of recreational visitors to lakes (Keeler et al., 2015). Building on this research, a team of scientist working for The Nature Conservancy used Flickr photos to determine tourist visitation to coral reefs, and to estimate the economic value of reefs globally (The Nature Conservancy, 2017).

Another recent example of researchers using Twitter data for conservation purposes is noteworthy. Daume (2016) analysed close to 3,000 tweets that made references to invasive alien species of interest. The findings showed that Twitter can provide useful information on species occurrence, as well as on human perceptions of species and their distribution.

Other approaches to utilising citizens for recording environmental changes have followed a more structured approach, for example through a bespoke mobile phone app. A wide range of citizen science platforms encourage people to engage in a process of voluntary information provision on specifically designed web sites. GBRMPA has developed such a platform to collect data and ‘sightings’ from visitors to the Reef.

The Eye on the Reef program enables both visitors and operators to contribute information about reef health, marine animals and incidents. Several platforms form part of this program. At the least formal level, visitors to the Reef can provide information through a mobile app or online system. The app is used to report observations of particular species. It also facilitates the upload of photos. As with other programs involving people from the general population as “sensors”, the information provided describes the particular subject of interest, the time and the particular location it relates to. In addition to the mobile app, Reef tourism operators contribute to monitoring through the Rapid Monitoring Survey or the Tourism Operators Weekly Monitoring Survey. The latter survey demands ongoing commitment to the monitoring of environmental indicators in the same location (i.e. where dive operators have a license to operate).

This research explored whether more generic and informal information from social media can complement the targeted approach of citizen science, as evidenced in the Eye on the Reef program.

4. Method

4.1. IT requirements

Figure 4 visualises the underpinning information system architecture, including data storage, management and analysis:

1) Configuration of the architecture required for handling vast volume of streaming and batching social media and other data relevant to the project.
2) Due to the unstructured nature and volume of data, a Hadoop cluster with share-nothing architecture along with NoSQL databases is most suitable for storing and analysing data[i].

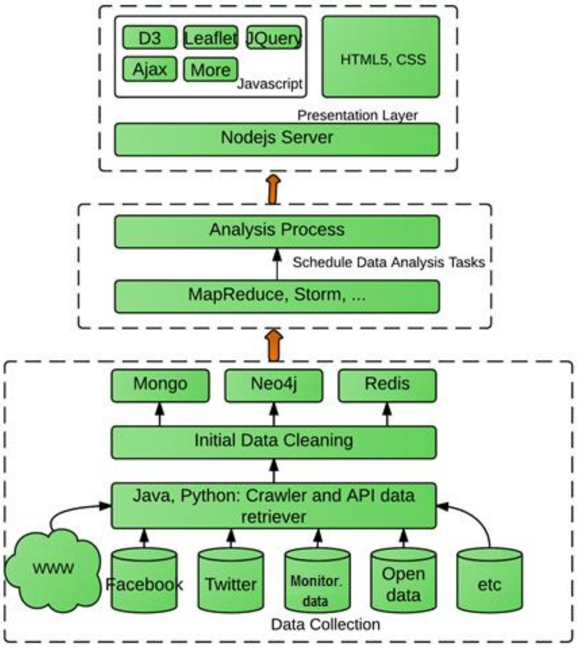


Figure 4 Overview of IT architecture developed for this project.

A hybrid open data approach was taken (Boroujeni et al., 2017). 'Hybrid' refers to the combination of periodic download of open source data (e.g. weather data from BOM) and data queries on demand (Franciscus et al., under review).

More specifically, to be able to take into consideration continuous open access data such as weather data we proposed a method that relies on API and periodically accesses the data source to download data into our Big Data lab-based local NoSQL database. Data are in JSON format. We chose that the periodic download is every 30 minutes, but in some cases data are needed instantly. For this case we introduced online access and developed a code, which queries the open data. For more detail see Chen et al. (2017a).

4.2. Data sources

This project draws on existing third-party data to generate new variables. The different data sources are discussed in more detail in the following. In addition to the sources below, we have begun to analyse China Sina Weibo data that are stored on a server in a Chinese partner University. For more information please refer to Chen et al., (2017b).

4.2.1. Twitter

Twitter is a publicly available source of information. Twitter releases at least 1% of the total tweets free of cost to the users, who can chose a random sampling approach for data collection. We used a public Twitter API with restrictions to capture geo-tagged tweets posted from polygon defined by a bounding box. Geo-tagged tweets are a subsample of tweets associated with explicit geographic coordinates measured by either an exact coordinate or an approximate coordinate (polygon).

To determine an approximate region of the GBR for data collection a rectangular bounding box was defined for filtering data (Southwest coordinates: 141.459961, -15.582085 and Northeast coordinates: 153.544922, -10.69867). For the sub-set of tweets that is associated with an exact location, the coordinates are obtained either based on GPS embedded in mobile devices, or on the IP location of the computer located to the nearest address. In the case of a tweet associated with a polygon, the polygon is created based on either the place that the sender of the tweet explicitly specified when the tweet was posted, or on the default place chosen by Twitter from the user profile location.

We started to collect freely/publicly available Twitter data from the beginning of the project. In total, for the period from 18/03/2016 to 20/11/2017 we collected 592,691 tweets. The types of metadata stored for each tweet are shown in Table 1.

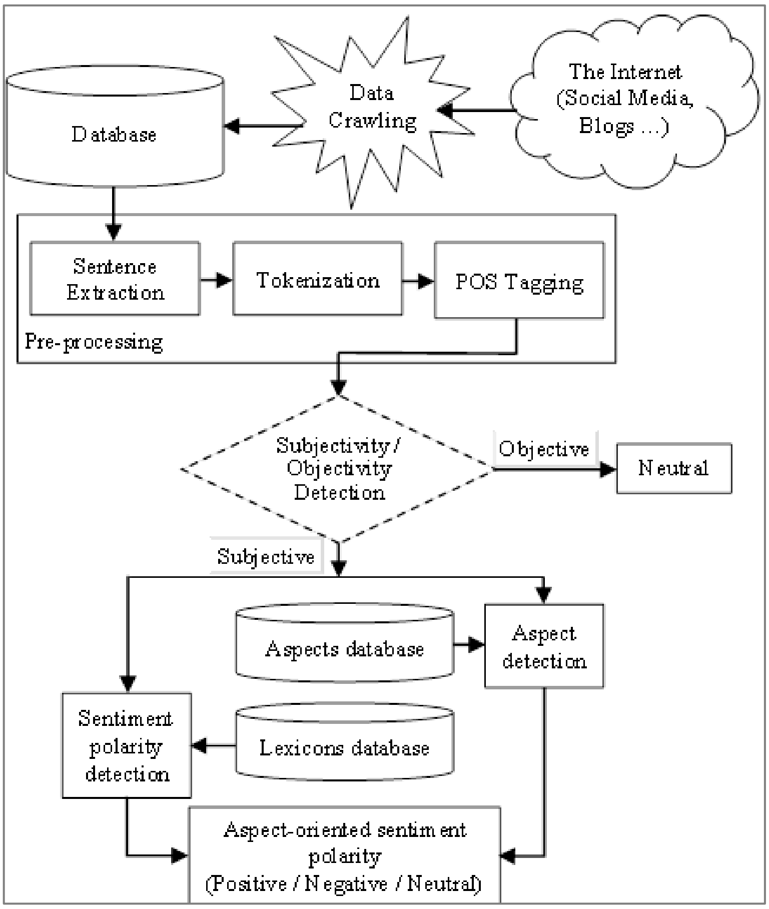
Table 1 Relevant variables provided in the Twitter database.

Variable Name	Variable Label
Username	User name
Id	Respondent ID
userstatuses_count	Count of User Statuses
Text	Tweet text
Lang	Language of Tweet
timestamp_ms	Time stamp of tweet
created_at	Time tweet created (e.g. Tue Mar 29 22:57:46 +0000 2016)
Placename	Place tweet created (short)
placefull_name	Place tweet created (full name, location hierarchy)
usertime_zone	Users time zone setting (Account details)
Userlocation	Users specified location (Account details)
usercreated_at	When user created Twitter account (Account details)
userfollowers_count	Count of Users Followers (Account details)
Userlang	Users specified language (Account details)

The analysis involved several steps. First, the total volume of tweets was filtered for those posts that were deemed relevant. This process required several iterations to identify a suitable range of keywords (e.g. diving, snorkelling). Appendix A shows the complete list of keywords. The Twitter data were analysed by using Natural Language Processing (NLP) and sentiment analysis technologies. Keyword search and keyword count were used to compute key statistics (e.g. frequency). More detail can be found in Becken et al. (2017a,b).

For sentiment analysis we employed an existing algorithm that was specifically developed for the analysis of social media text. Valence Aware Dictionary for Sentiment Reasoning

(VADER) is a rule-based algorithm that combines a general lexicon / dictionary and a series of intensifiers, punctuation transformation, emoticons, and many other heuristics to compute sentiment polarity of a review or text. For a detailed review of sentiment research please refer to Alaei et al. (2017). An overview of the process is presented in Figure 5.



Database
Data Crawling
The Internet (Social Media, Blogs ...)
Sentence Extraction
Tokenization
POS Tagging
Pre-processing
Subjectivity / Objectivity Detection
Objective
Neutral
Subjective
Aspects database
Aspect detection
Sentiment polarity detection
Lexicons database
Aspect-oriented sentiment polarity (Positive / Negative / Neutral)

Figure 5 General framework of a sentiment analysis system (Source: Alaei et al., 2017).

In order to improve accuracy as well as performance we have made several modifications to the original VADER algorithm and we have also started a dedicated lexicon suitable for environmental changes. In addition, we have developed a target-specific lexicon for tourism tweets. The lexicon recognises 12 targets, of which two (Environmental Risk Lexicon, and Marine Species Lexicon) are specific to the marine environment. In addition, we have developed a semi supervised machine learning method for target/aspect identification.

4.2.2. Facebook

Facebook is globally the most prominent social media platform. Its market share is considerable at 18% of social media users, but the content is not publicly available. Commercial Facebook pages, or those maintained by public organisations, however, can be accessed. Thus, this project identified Facebook pages of interest to test the notion that environmentally relevant information could be extracted

We have created a method to identify public Facebook pages within the defined bounding box, which is aligned with the same geographic area used for Twitter discussed above. Out of all public pages in the region, we identified those relevant for tourism and environmental changes. We used Facebook-sdk to develop programs which downloads Facebook posts, comments as well as likes. This client library is designed to support the Facebook Graph API and the official Facebook JavaScript SDK, which is the canonical way to implement Facebook authentication.

Raw Facebook data is stored in the JSON format. Data are stored in NoSQL MongoDB database, which is located on same cluster computer as twitter data, however stored in different database. Facebook data storage began on the 1st of June 2016. By the 17th of March 2017, a total of 1,870 posts and 4,948 comments were saved for analysis. At this moment we do not collect Facebook data. However, we have developed an API and as data in Facebook are quite stable, data collection from Facebook can be performed any time based on needs.

Facebook data are analysed by keyword to measure frequencies and calculate sentiment. For most analysis, posts and comments were combined. The same method that was used for scoring sentiment of tweets was used for Facebook text. Again, the modified VADER algorithm has been employed for the analysis (see Becken et al., 2017a, b).

The Facebook database is divided into 'posts' and 'comments'. Posts originate from those who own and operate the page and comments come from the public who like or follow the page or who visited a particular post and chose to provide an opinion or comment. Table 2 summarise the variables that are extracted for Facebook posts and Table 3 shows variables related to comments.

Table 2 Relevant variables provided in the Facebook database for posts

Variable Name	Variable Label/Description
_id	The post ID
Time: created_time	A timestamp of when this message was created
Message	The text of the message
Shares Count	How many people shared the post
Reactions (LIKE, LOVE, HaHa, SAD, ANGRY)	People who have reacted to this post

Table 3 Relevant variables provided in the Facebook database for comments

Variable Name	Variable Label/Description
Id	The comment ID
Time: Created_time	The time this comment was made
Message	The comment text
From.name	Who comments

4.2.3. Eye on the Reef

The Eye on the Reef program is a citizen science platform of GBRMPA that enables visitors and operators to contribute information about reef health, marine animals and incidents. The program facilitates the contribution of data through various mechanisms that are targeted at different types of sensors or citizens. Visitors to the Reef can provide information through a mobile APP or online system. The APP helps to report sightings of particular species and upload photos (the "Sightings" data) (Figure 6).

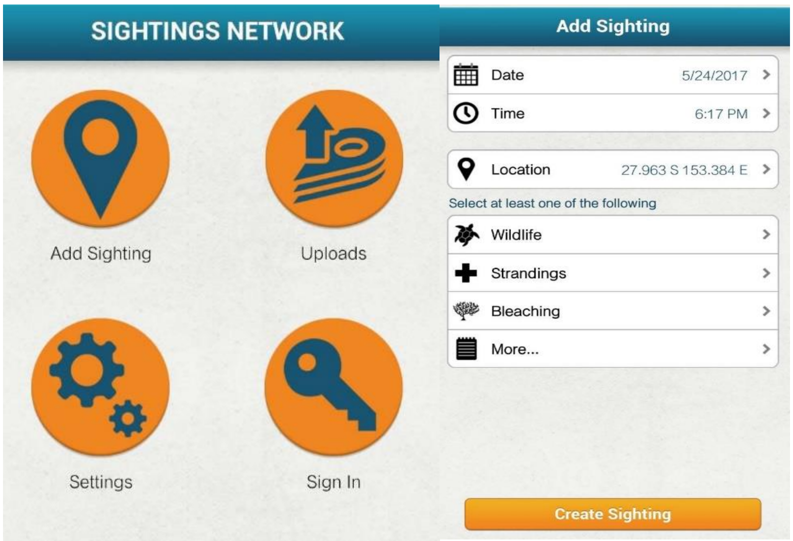


Figure 6 Screenshot of the Eye on the Reef App used by visitors to the Reef.

Reef tourism operators are contributing through the Rapid Monitoring Survey. This part of the Eye on the Reef program requires that an underwater monitoring slate is completed and submitted to the database afterwards through an online portal. Furthermore, the Tourism Operators Weekly Monitoring Survey demands ongoing monitoring of environmental indicators in the same location. The Tourism Weekly method is a 30 minute swim (dive or snorkel) over the same general area once per week, taking mental estimates of about 25 different variables which includes quantitative coral bleaching. There is no qualitative metric (severity) so is just binary yes/no with categories of quantity (number of colonies impacted).

Finally, the Reef Health Impact Survey report produces a Google Earth KML (with associated Excel files) summarising the impact and severity of either bleaching OR damage OR disease events at the Management Area level across the Great Barrier Reef Marine Park. More information on the program can be found on GBRMPA's website at http://www.gbrmpa.gov.au/managing-the-reef/how-the-reefs-managed/eye-on-the-reef.

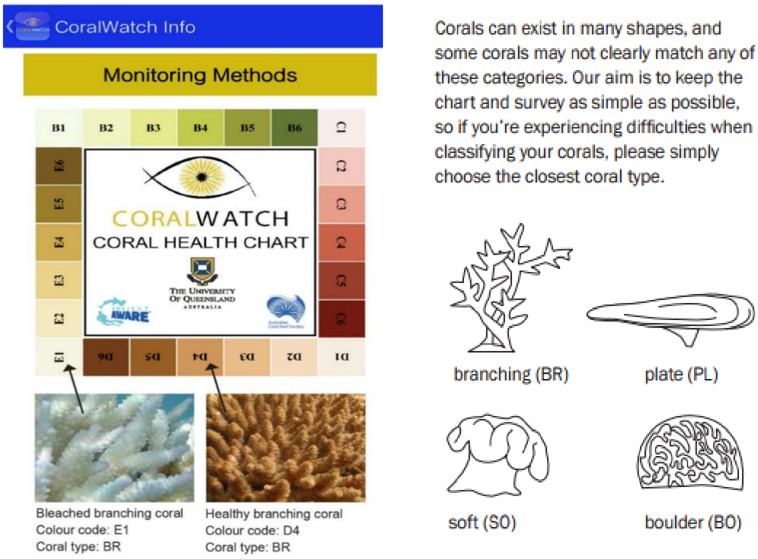
The Eye on the Reef data were obtained from the Great Barrier Reef Marine Park Authority for this project. Three sets of data were provided: the 'Sighting data', the 'Tourism Weekly', and the 'Reef Health Impact Survey' data. The data were emailed in an Excel spreadsheet and contained sightings of species and incidents from 2009 to 1st of April 2017. The RHIS data were for 2016 and focused on coral bleaching.

To understand spatial coverage, the data are visualised on a map to identify locations where observations were made. For this purpose, coordinates were truncated after one decimal, roughly reflecting a 11-kilometre grid. Other spatial resolutions are possible, but initial testing showed that a 1-kilometre grid did not produce many locations with joint (with other data sources) observations. Second, the value (e.g. extent of coral bleaching) was visualised on a map using colour coding. The data was then integrated with other data sources (e.g. Twitter) to explore correlations.

4.2.4. CoralWatch

CoralWatch is a citizen science project based at the University of Queensland. It has been developed to engage non-scientists in Australia and elsewhere to not only understand and appreciate coral reef management, but to contribute by adding data into the tailored system. More information can be found at www.coralwatch.org.

CoralWatch provides tools for collecting scientific data about the health of corals. The Coral Health Chart (see Figure 7) provides assistance to make decisions about the state and type of coral. The chart standardises changes in coral colours, and provides a simple way for people to quantify coral health and contribute to the CoralWatch global database.



Bleached branching coral
Colour code: E1
Coral type: BR

Healthy branching coral
Colour code: D4
Coral type: BR

Corals can exist in many shapes, and some corals may not clearly match any of these categories. Our aim is to keep the chart and survey as simple as possible, so if you're experiencing difficulties when classifying your corals, please simply choose the closest coral type.

Figure 7 "Coral Health Chart" provided by the CoralWatch[1].

The CoralWatch data were obtained personally from CoralWatch. We have data from 2016 and 2017 and agreement was reached to mention the source of data in any research related outputs. The data were emailed in the format of an Excel spreadsheets. CoralWatch data have been imported into MongoDB databases to simplify collection with human sensors, however to enable more powerful analysis by utilizing Structure Query Language (SQL) we also imported CoralWatch data into MySql relational database.

This data contain details related to cover and state of the corals. Specifically it includes the following details of interest to this project:

- Date and Time of observation
- Reef name, and geographic location (Longitude, Latitude)
- Light condition (Full sunshine, Broken cloud, Cloud)
- Coral type (Branching, Boulder, Plate, Soft)
- Coral Colour (colour type and shade measured as 1 to 6 with 6 being darkest)

4.2.5. Weather data

Official weather data in Australia come from the Bureau of Meteorology (BOM). BOM generally provides four types of data including "Weather and Climate", "Rainfall",

[1] http://www.coralwatch.org/c/document_library/get_file?uuid=29755135-527f-4e6a-88c7-ec28958e2e45&groupId=10136

“Temperature”, and “Solar Exposure” in relation to each region extracted from a particular meteorology station in daily or monthly basis. For example, monthly “Rainfall” data collected from the Amberley Station (AMO) contains amount of daily rainfall in every month with some statistics, such as the total volume of rainfall at every month and the highest daily rain at every month. Data fields of “Temperature” data are almost the same as “Rainfall”, with the name temperature instead of rainfall. The source is open data that are publicly available for GBR region.

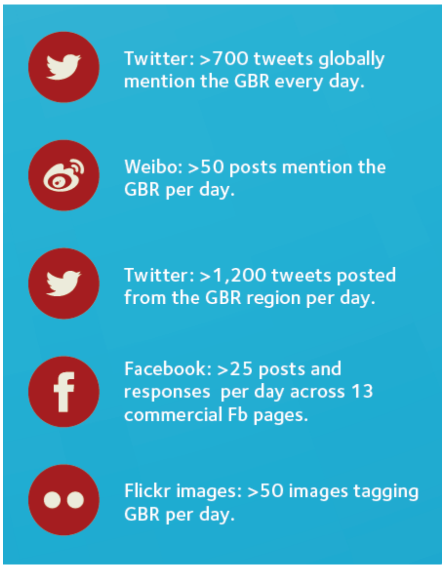
Weather data were accessed via API directly from BOM. We developed a program that utilizes the available API and accesses it every 30 minutes to download relevant data from the GBR region. We have also developed method to access and query weather data on demand. Data is in Jason format and are stored in MongoDB NoSql database in a dedicated collection.

The weather data can be integrated with the social media sentiment scores to examine possible correlations. This work is to be seen as a proof of concept and future research can correlate weather data with other environmental or social media data sources.

5. Results

5.1. What does social media tell us?

An analysis of several social media platforms gives us an idea of the volume of data. As can be seen in Figure 8, there are about 700 tweets globally that mention the GBR. Within the region, we captured about 1,200 tweets per day; however, after the filtering process it emerged that in the order of 50 tweets per day were relevant to the marine environment.



Twitter: >700 tweets globally mention the GBR every day.

Weibo: >50 posts mention the GBR per day.

Twitter: >1,200 tweets posted from the GBR region per day.

Facebook: >25 posts and responses per day across 13 commercial Fb pages.

Flickr images: >50 images tagging GBR per day.

Figure 8 Number of social media data captured for the different platforms.

As expected, social media data tend to be geographically concentrated. The heat maps presented in Figure 9 visualise that the majority of tweets that are posted from with the GBR bounding box come from the major population centres and tourist destinations.

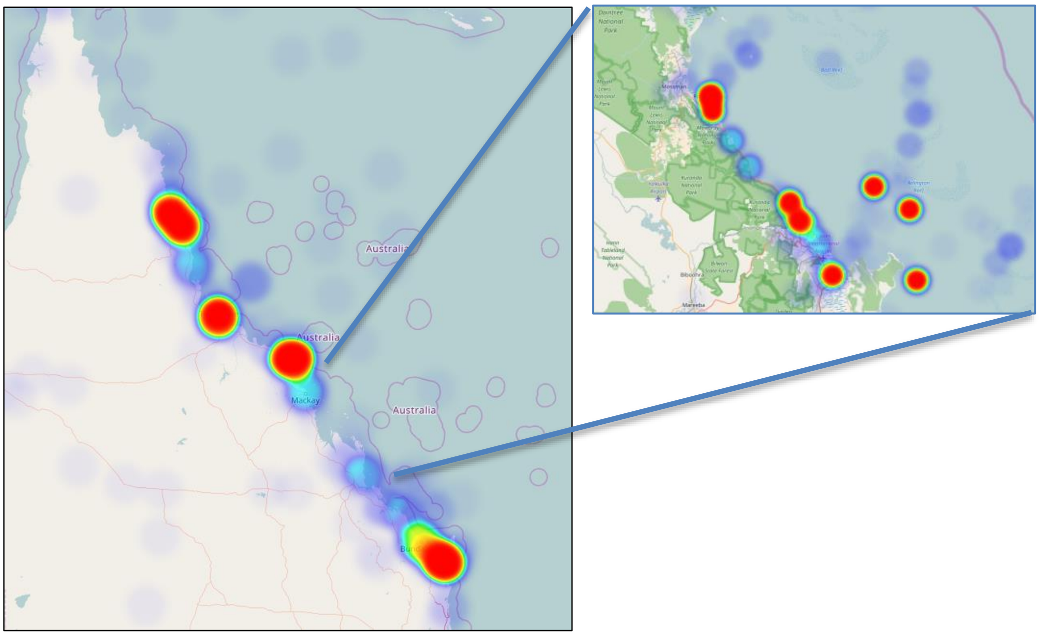


Figure 9 Twitter heat map showing where the tweets were posted (Insert showing the Cairns region). Note: red reflects higher number and purple lower number of tweets.

The analysis of social media that we undertook as part of this project focused on two aspects:

- Keyword analysis
- Sentiment of tweets and Facebook posts

The full results of the above analysis are presented in Becken et al. (2017a); however a selection of results are presented here to illustrate the type of insights.

The keyword analysis highlighted that Twitter feeds contained more factual information compared with Facebook, and also was more descriptive of particular locations and activities, such as diving. Facebook posts and comments focused more on experiences and emotional states, although more factual discussions (e.g. on the state of the Reef) were observed (Table 4).

Table 4 Top 10 keywords mentioned in tweets and Facebook text (see Becken et al., 2017a)

Activity	Tweets N	Face book N	Marine Species	Tweets N	Face book N	Environment / impact	Tweet s N	Face book N
Diving	876	1357	Fish	1023	475	White	709	73
Swim	753	96	Coral	434	355	Bleach	94	74
Water	590	245	Shark	404	281	Storm	85	5
Boat	515	225	Turtle	378	334	Oil	27	68
Snorkel	564	95	Cod	303	46	Dead	40	31
Sail	382	16	Dolphin	230	45	Coal	49	20
Scuba	300	256	Nemo	177	123	Mud	24	1
Marine	160	251	Whale	163	105	Algae	12	14
Paddle	61	9	Ray	119	77	Damage	12	8
Goggle	8	2	Crown	73	12	Died	13	6

Sentiment can be measured from different angles, including:

- Across time
- For specific locations
- For keywords of interest
- By specific markets (i.e. who is posting it).

Overall, it is important to note that social media conversations are biased towards positive content and language (see Alaei et al., 2017). The findings from this research showed, for example that only 9.8% of tweets posted in the GBR region were negative, that means they had a score between zero and minus one. For Facebook posts the share of negative comments was 5%. Second, the analysis showed that a considerable number of tweets were classified as neutral (50.2%). There are several reasons for this, including the occurrence of both negative and positive words in the tweet that balance each other, or the language being different from English, in which case the VADER algorithm automatically assigns a neutral score. Further refinement of the algorithm will enable greater recognition of polarised tweets (positive or negative).

To illustrate the differences in sentiment scores, Figure 10 presents the analysis of sentiment for key marine species. It can be seen that several species attracted negative comments, for example those that related to dugongs. For example, one tweet reads "*It is a tragedy that this cruelty to turtles and dugongs is allowed to happen. The hypocrisy of the Greens plain…*" (Sentiment score: -0.8225). Other species attract mainly positive sentiment scores, for example anemones or starfish.

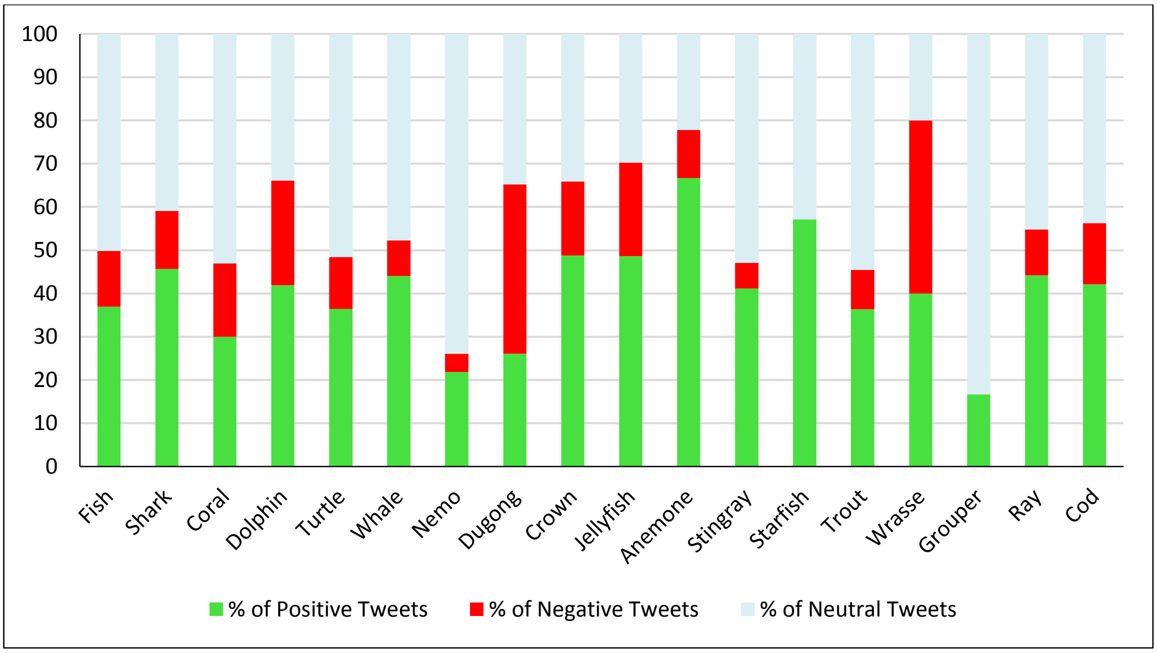


Figure 10 Percentage of tweets with positive, negative and neutral polarities collected from the Twitter in relation to different GBR marine park species (Becken et al., 2017a).

5.2. Can social media data be usefully integrated with other data?

To assess the value of social media 'observations', we compared the Twitter sentiment with other data sources that provide information on environmental conditions. More specifically, we used data from:

- CoralWatch
- Eye on the Reef Sightings
- Eye on the Reef Tourism Weekly
- Reef Health Impact Survey

The number of observations and unique locations differ for the datasets and one challenge was to geographically match locations to compare what had been recorded within a meaningful perimeter (Table 5). For Twitter only those tweets were used that contained the world coral. Flickr images have yet to be manually coded to extract those that contain imagery of coral. This will reduce the number of observations significantly.

Table 5 Overview of the data sources used for a comparison in relation to coral bleaching observations.

Data set	Twitter	Flickr	Eye on the Reef Sighting	CoralWatch	Tourism Weekly	Reef Health
Number of observations	435	6390	259	6118	665	1840
Observations with exact location	395	1440	259	6118	665	1840
Unique locations	47	78	50	39	19	85

The full details of this analysis can be found in Connolly et al. (2018 in preparation), but the following maps provide some initial insight into the potential computability of different data sources. This particular analysis focused on the incident of coral bleaching, as this phenomenon was captured in all five data sources.

Figures 11 and 12 provide a visualisation of the number of observations by location. CoralWatch observations, for example, are largely centred around the Whitsunday Islands, with a second conglomeration around Cairns. However, a relatively large number of observations has also been provided by divers in the Northernmost part of the GBR. The Eye on the Reef Sightings are mostly concentred around Cairns. Not surprisingly, tweets related to coral (or bleaching) were mostly posted from key population centres, i.e. Townsville and Cairns, possibly reflecting mobile phone coverage and opportunity to tweet.

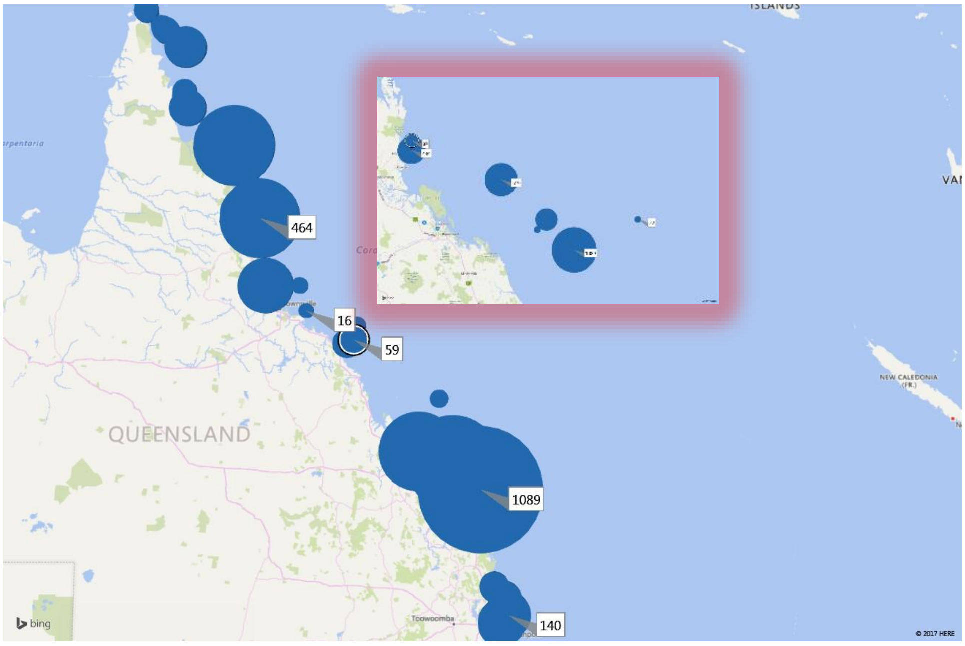


Figure 11 Number of observations by location in the CoralWatch citizen science data base for 2016 (the insert is showing Cairns).

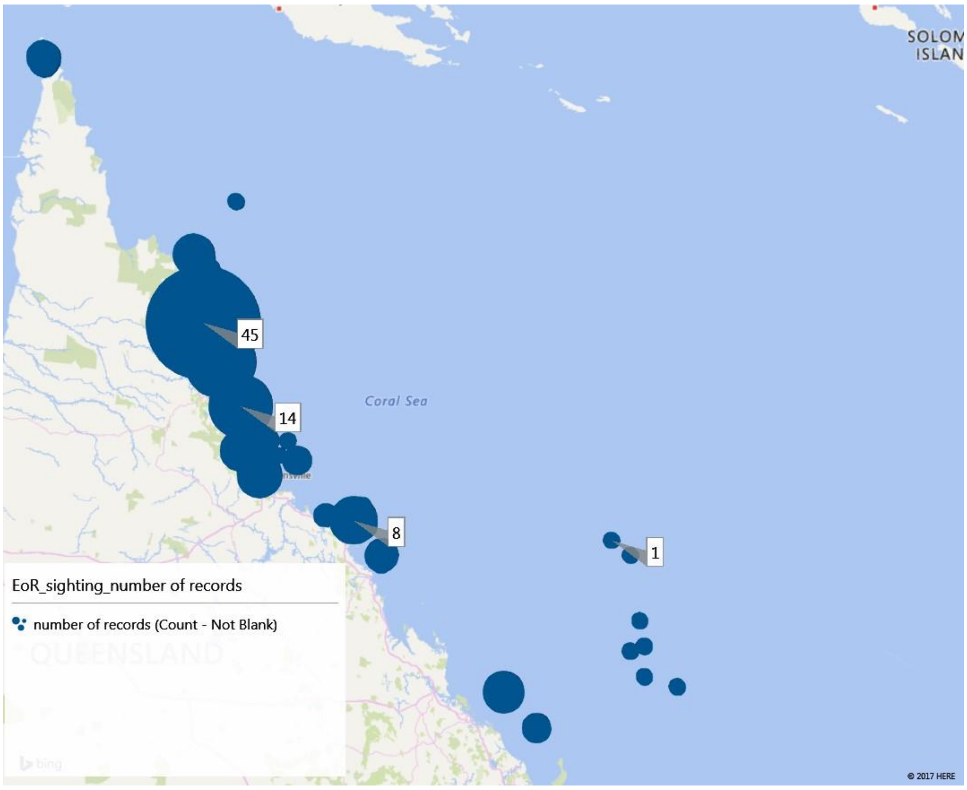

SOLOM
ISLAN
45
14
Coral Sea
8
1
EoR_sighting_number of records
number of records (Count - Not Blank)
QUEENSLAND
© 2017 HERE

Figure 12 Number of observations by location in the Eye on the Reef Sightings data base for 2016.

Considering now what each data source provides in terms of coral assessment, we compared the scores for the sources. Figure 13, for example, visualises the Eye on the Reef Sighting data on incident of coral bleaching (1 to 5, whereby 1 indicates totally bleached and 5 refers to healthy coral).

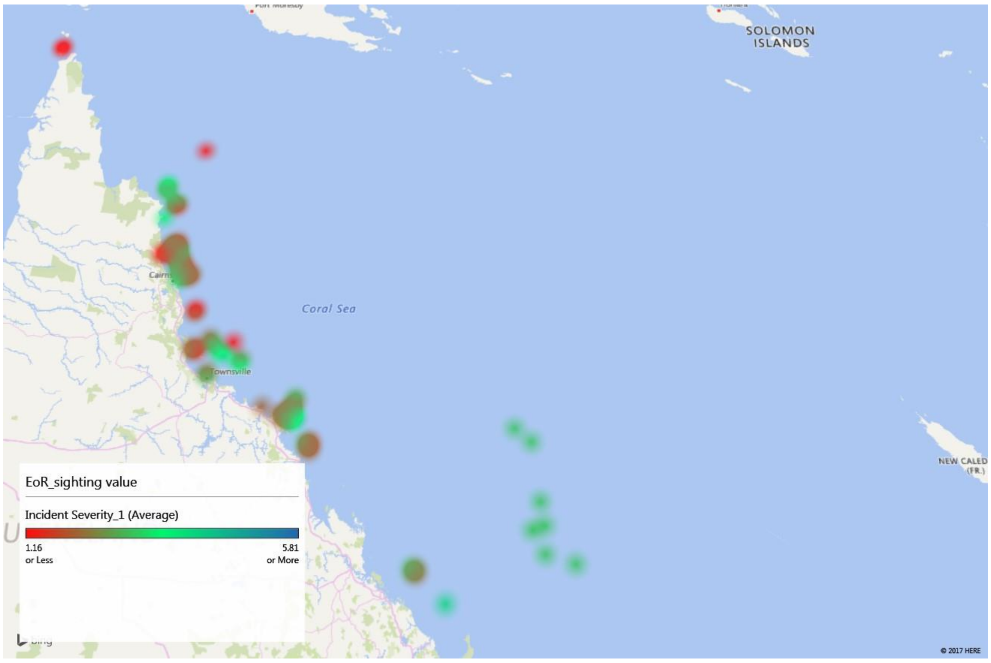


Figure 13 Coral bleaching severity as recorded in the Eye on the Reef Sightings for 2016.

The stronger bleaching in the North of the Reef is confirmed in the Reef Health Impact Survey, shown in Figure 14. Here, for better comparability, we have converted Yes (bleaching) into a score of 1, and No (bleaching) into a score of 5. Average scores are calculated for those locations with more than one observation.

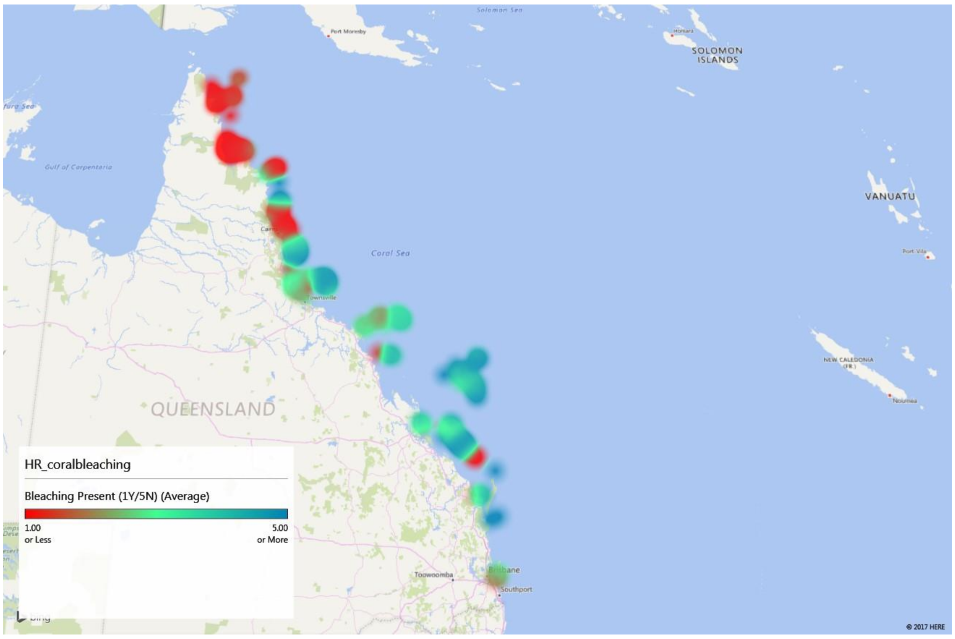


Figure 14 Coral bleaching severity as recorded in 2016 Reef Health Impact Survey.

5.3. Integration with weather data

While the individual social media posts may be trivial, the accumulated data can provide valuable information on diverse topics. To examine possible relationships between tweets and biophysical variables, we have looked into is possible correlations between human sentiment and different weather conditions. Specifically, we analysed the relationship between temperature, humidity, wind, rainfall, and Twitter sentiment polarity in the Cairns area.

The results revealed that the Twitter-based sentiment analysis demonstrated a fairly close relationships between different weather conditions and users' sentiments. A machine learning method based on artificial neural network (ANN) was developed for the Twitter dataset and the accuracy of the predictions was tested. The results show that the Big Data analysis and machine learning techniques can be used to analyse and predict sentiment to different weather conditions. It can also learn what weather conditions humans perceive as comfortable (H. Li et al., 2017).

Figure 15 shows the relationship between temperature, humidity, wind, rainfall, and sentiment in our Twitter dataset collected from Carins Australia. The x-axes in all graphs were rescaled considering the maximum value for the variables in the dataset. This helps to convert x-axis to a dimensionless value in the [0 1] range and compare the impact of different variables on sentiment. According to Figure1, the concerned tweets were in the range of 22 – 32°C (temperature), 47-87% (humidity), 10-31km/h (wind speed) and rain smaller than 14mm. The averaged sentiment values for this ranges are positive. It shows that these ranges were desirable conditions for users and made them feel positive. Although there are some negative sentiments in these ranges, the averaged sentiment for each weather variable is positive.

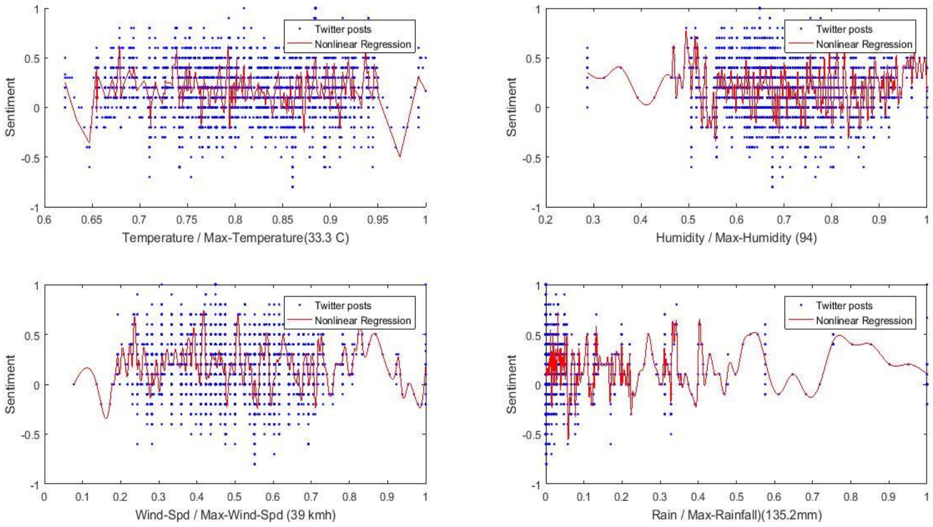


Figure 15 Comparison of the sentiments.

5.4. Online platform to visualise Twitter based information

The option to access the insights available from Twitter is important to a range of stakeholders, including from Reef management agencies and the tourism industry. An online tool was developed in the web-based platform with HTML and JavaScript language. The main purpose of this tool was to establish a proof of concept of what is possible and how data can be visualised, so that stakeholders can provide feedback for future developments.

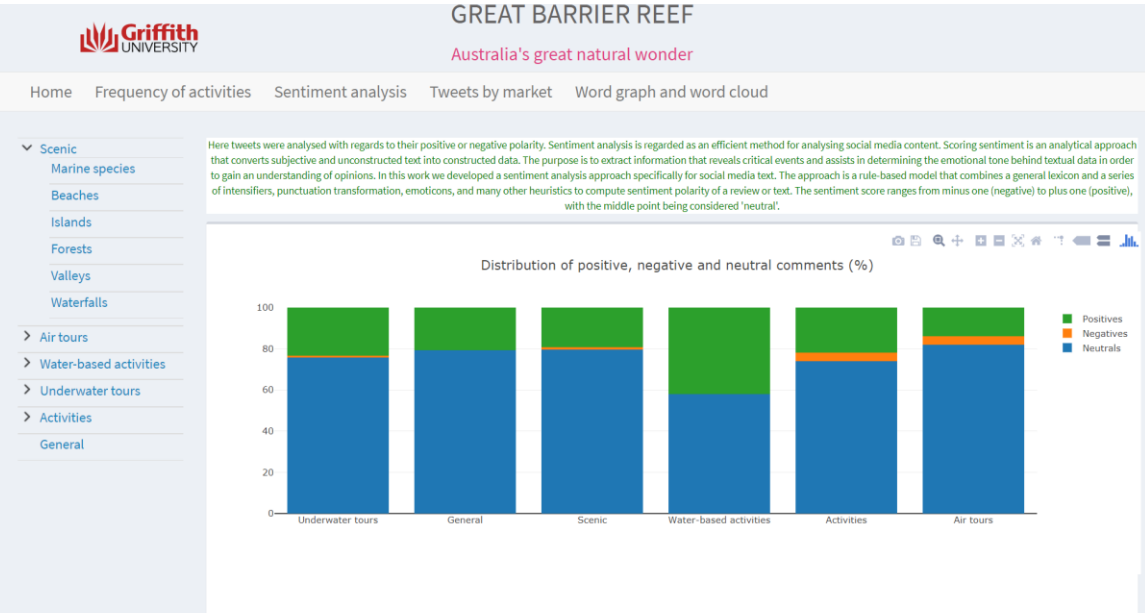
We accessed to GBR twitter data stored in MongoDB in JSON format stored on the Big Data and Smart Analytics Lab cluster by using Python. Each tweet was then automatically annotated into predefined categories. To complete the annotation, we proposed a 3-step approach including feature extraction, building training set, and learning the classification model. In detail, we extracted numerical feature vectors of each tweet by addressing the occurrences of tokens along with the weight of importance of tokens that occur in the majority of samples.

The training set, which was generated by gathering the extracted features and their manually assigned labels, was trained by a Support Vector Classifier method to obtain the classification model. The annotation module was developed in the form of web service by using the Python programing language. Hence, it is an independent tool for any web applications. Based on these annotations, we can understand the interesting aspects insides the GBR twitter data.

The statistics of data were visualized by Plotly, an open source tool for composing, editing, and sharing interactive data visualization via the Web. The word cloud and word graph were also obtained from Twitter data and visualized by two libraries named ZingChart and ViS.js. Both are libraries that have been used successfully in previous projects in the Griffith Big Data lab (Franziscus et al., 2018).

Finally, and as discussed further above, the sentiment associated with each tweet was extracted by using a modified VADER algorithm (Valence Aware Dictionary and Sentiment Reasoner). Modifications were related to speeding up the process of scoring, as well as improvements to the underpinning lexicon enhanced by domain specific terms. As discussed in Alaei et al. (2017), VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media data.

In the following Figures 16 to 18, a number of screenshots are shown from online tool to provide an impression of the type of information provided in the proof of concept application.



Here tweets were analysed with regards to their positive or negative polarity. Sentiment analysis is regarded as an efficient method for analysing social media content. Scoring sentiment is an analytical approach that converts subjective and unconstructed text into constructed data. The purpose is to extract information that reveals critical events and assists in determining the emotional tone behind textual data in order to gain an understanding of opinions. In this work we developed a sentiment analysis approach specifically for social media text. The approach is a rule-based model that combines a general lexicon and a series of intensifiers, punctuation transformation, emoticons, and many other heuristics to compute sentiment polarity of a review or text. The sentiment score ranges from minus one (negative) to plus one (positive), with the middle point being considered 'neutral'.

Figure 16 Distribution of sentiment.

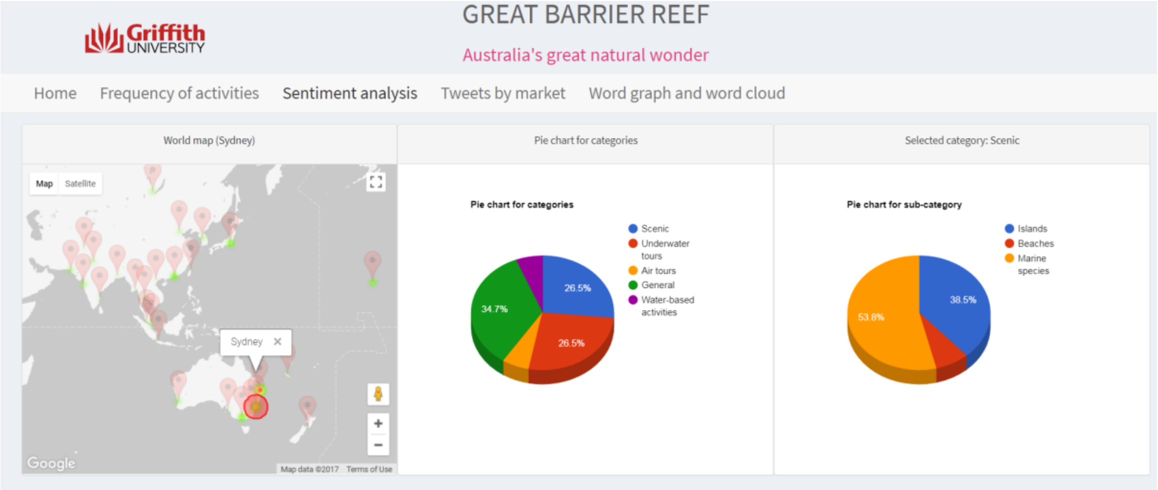


Figure 17 Interest in particular activities from different regions of origin.

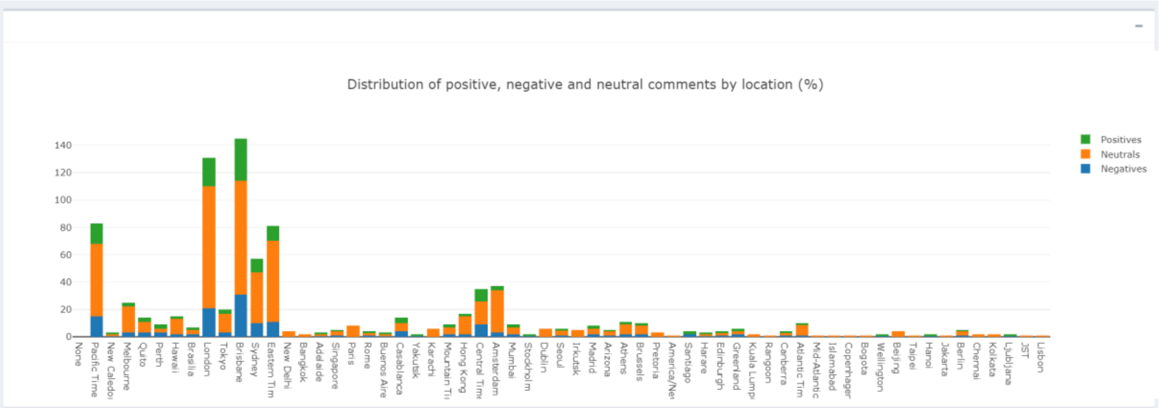


Figure 18 Sentiment polarity depending on location of residence/origin.

6. Conclusion

The aim of this research was to examine whether social media contain useful information on the Great Barrier Reef. The research found that volume of social media data is considerable; however, the examination of tweets in particular has also highlighted the critical need for filtering procedures. For the Twitter analysis, two steps of unsupervised (i.e. computer automated) filtering were implemented; one related to the geographic location from where tweets were posted, and a second one based on a relatively large number of keywords that sought to capture those tweets that actually talked about the marine environment.

A keyword analysis of most frequent words (through word count) and the numbers of tweets retrieved that contained a priori specified keywords gave insights into what mattered to people, but also how the two social media channels differ in terms of focus. Further analysis of other data sources (e.g. Weibo Sina, Instagram) will be useful, as it enhances the volume of data but also taps into different markets and user groups.

The existing VADER sentiment algorithm was initially applied and then improved and extended to suit the particular context of the Reef environment. Further work, including manual annotation to compare the performance of the algorithm, will be necessary should this system be implemented for monitoring.

The sentiment around some of the environmental keywords was negative, although it was surprising that the frequencies were quite low. Even coral bleaching, which has received substantial media attention, was not mentioned very often. Those people who discussed bleaching (and other issues associated with Reef health) showed concern, and shared their views, in particular on Facebook where space is not limited. Several Facebook pages, especially GBRMPA's page, used this platform to engage and educate followers and share some useful information. Some pages tried to mobilise people to act to protect the GBR. More could be done to specifically engage Reef users and advocates in social media campaigns. The Citizens of The Reef initiative could be an important allay in this journey of citizen engagement and science (see here http://citizensgbr.org/).

The research has also explored possible correlations between Twitter, other citizen science data sets and professional monitoring. From the initial results, it appears that due to the different type of information provided and geographic spread, as well as a common trend of recognising environmental change (in our case coral bleaching) there is potential for an integrated monitoring portfolio. Early attempts to link human sentiment with weather data also provide promising findings and further analysis of linking social media data with environmental data should be encouraged.

Finally, communicating findings is key to research uptake by decision makers. Thus, a proof of concept web platform was developed to visualise selected aspects of Twitter-based information. Testing with stakeholders is now underway and future modifications, and professional development, are entirely feasible.

Stretching into image-based social media (i.e. Flickr in our case) is an important extension of this project. Images provide more and higher quality information on the marine environment, and for that reason might be a better source for environmental decision makers. It is suggested that further analysis into downloading, processing and interpreting (manually or automatically) imagery should be undertaken to extract the full value of these media.

7. References

Alaei, A.R., Becken, S. & Stantic, B. (2017). Sentiment analysis in tourism: Beginning of a new research paradigm? Accepted Journal of Travel Research.

Becken, S., Alaei, A., Chen, J., Connolly, R. & Stantic, B. (2017a). The role of social media in sharing information about the Great Barrier Reef. August 2017. Griffith Institute for Tourism Research Report No 14. https://www2.griffith.edu.au/institute-tourism/publications/research-report-series

Becken, S., Stantic, B., Chen, J. Alaei, A.R. & Connolly, R. (2017b). Monitoring the environment and human sentiment on the Great Barrier Reef: assessing the potential of collective sensing. Journal of Environmental Management. 203, 87-97.

Boroujeni, F.R., Stantic, B. & Wang, S. (2017). An Embedded Feature Selection Framework for Hybrid Data, 28th Australasian Database Conference, 138-150.

Can Wang, Chi-Hung Chi, Zhong She, Longbing Cao and Bela Stantic, (2017), Coupled Clustering Ensemble by Exploring Data Interdependence, TKDD Journal (ACM Transactions on Knowledge Discovery from Data)

Chen, J., Stantic, B., Wang, S. (2017a), Connecting Social Media Data with Observed Hybrid Data for Environment Monitoring. International Symposium on Intelligent and Distributed Computing, (IDC) 2017. DOI 10.1007/978-3-319-66379-1

Chen, J., Becken, S. & Stantic,B. (2017b). Harnessing Chinese Social Media to Analyze Tourists at Great Barrier Reef. Chinese Dream Conference, 23-25 November, Gold Coast, Australia.

Connolly, R., Becken, S., Chen, E. & Stantic, B. (2018). A hybrid is born: integrating collective sensing, citizen science and professional monitoring of environmental health. In preparation.

Deloitte Access Economics (2017). At what price? The economic, social and icon value of the Great Barrier Reef. Brisbane, Australia.

Franciscus, F., Ren, X. & Stantic, B. (2018). Far Beyond Word-Cloud: A Graph Model Derived from Beliefs. 10th Asian Conference Intelligent Information and Database Systems (accepted).

Franciscus, F., Ren, X. & Stantic, B. (under review). Answering Temporal Analytic Queries Over Big Data Based on Precomputing Architecture. Asian Conference on Intelligent Information and Database Systems – ACIIDS 2017, pp 281-290.

GBRMPA (2016). Great Barrier Reef Tourist Numbers. Available (08/05/17) http://www.gbrmpa.gov.au/visit-the-reef/visitor-contributions/gbr_visitation/numbers

GBRMPA (2017). Significant coral decline and habitat loss on the Great Barrier Reef (29/05/2017). Available (09/07/17) http://www.gbrmpa.gov.au/media-room/latest-news/coral-bleaching/2017/significant-coral-decline-and-habitat-loss-on-the-great-barrier-reef

Li, H., Jadidi, Z., Chen, J. & Jo, J. (2017), Machine Learning Method for Correlation of Sentiment and Weather data , RITA 2017, Korea.

Keeler, B.L., Wood, S.A., Polasky, S., Kling, C., Filstrup, C.T. & Downing, J.A. (2015). Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. Frontiers in Ecology and the Environment, 13(2), 76-81.

Meyer, R. (2015). How the USGS Detects Earthquakes Using Twitter. The Atlantic. Available (15/05/16) http://www.theatlantic.com/technology/archive/2015/10/how-the-usgs-detects-earthquakes-using-Twitter/409909/

Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. Transactions in GIS, 19(6), 809-834. doi: 10.1111/tgis.12132

The Nature Conservancy (2017). Mapping Ocean Wealth. Available (12/05/17) http://oceanwealth.org/

Vivacqua, A. S., & Borges, M. R. S. (2012). Taking advantage of collective knowledge in emergency response systems. Journal of Network and Computer Applications, 35(1), 189-198. doi: http://dx.doi.org/10.1016/j.jnca.2011.03.002

Appendix A: List of keywords to filter relevant tweets

Fish	Species	Togs	Swimsuit		Shells	Mussel	Died-off		
Anemone	Seahorse	Swim	Swimming	Swam	Reef	Reefs	Bleached	Bleaching	Bleach
Angelfish	Porcupinefish	Snorkel	Snorkelling	Snorkelled	Sand	Sandy	Pristine		
Barracuda	Boxfish	Fins	wetsuit	Goggles	Island	Islands	Colourful		
Clownfish	Puffer	Dive	Diving	Diver	Beach	Beaches	Murky		
Cod	Triggerfish	Scuba	Dived		Bay		Turbid	Turbidity	
Cots	Trumpetfish	Marine	Marina		Sea		Visibility		
Crown	Flutefish	Boat	Boating		Ocean				
Dolphin	Razorfish	Sail	Sailing	Sailed	Paradise				
Dory	Goatfish	Paddle	Paddling	Paddled					
Dugong	Eel	Fishing	Fished	Fishes					
Emperor	Seasnake	Fishable							
Grouper	Barramundi								
Lionfish	Damselfish								
Nemo	Rabbitfish								
Parrotfish	Batfish								
Shark	Unicornfish								
Snapper	Butterflyfish								
Starfish	Bannerfish								
Surgeonfish	Rockcod								
Tang	Stonefish								
Thorn	Crocodile								
Trevally	Marlin								
Trout	Mackerel								

Tuna	Stingray			
Turtle	Sawfish			
Whale	Hammerhead			
Wrasse	Wobbegong			
Coral	Flatworms			
Algae	Cucumber			
Plankton	Crown-of-thorns			
Jelly	Squid			
Jellyfish	Octopus			
Stinger	Cuttlefish			
Irukandji	Crabs			
Jellyfishes	Sponge			
Boxjelly				

Note: the list of species was informed by GBRMPAs Eye on the Reef app and other keywords were identified by manually analysing a subsample of tweets.

[i] IT components:

- apache hadoop (ver. 2.7); apache storm – parallel management of steaming data; apache kafka - fault-tolerant stream processing; apache spark - Streaming Processing Engine; apache zookeeper - provides communication bridge
- ganglia (for cluster monitoring); nodejs 0.10.42 (for front end development and visualisation); mongodb 2.6.7 (storing of unstructured data in share nothing concept); redis 3.0.1 (in memory NoSQL database); neo4j 2.3.3 (storing and accessing graph data); python 3.4.3 (programming language for algorithms development, some of required libraries (pymongo, nltk)); java (jre) 1.8.0 and VaderSentiment - Lexicon based sentiment analysis with Python + Pymongo libraries