

Target Sentiment and Target Analysis

Bela Stantic, Ranju Mandal and Emily Chen



Target Sentiment and Target Analysis

Bela Stantic¹, Ranju Mandal¹ and Emily Chen²

¹ School of Information and Communication Technology, Griffith University

² Griffith Institute for Tourism, Griffith University



Australian Government



National
**Environmental
Science**
Programme

Supported by the Australian Government's
National Environmental Science Program
Project 5.5 Measuring aesthetic and experience values using Big Data approaches

© Griffith University, 2020



Creative Commons Attribution

Target Sentiment and Target Analysis is licensed by Griffith University for use under a Creative Commons Attribution 4.0 Australia licence. For licence conditions see: <https://creativecommons.org/licenses/by/4.0/>

National Library of Australia Cataloguing-in-Publication entry:
978-1-925514-46-9

This report should be cited as:

Stantic, B., Mandal, R. and Chen, E. (2020) *Target Sentiment and Target Analysis*. Report to the National Environmental Science Program. Reef and Rainforest Research Centre Limited, Cairns (25pp.).

Published by the Reef and Rainforest Research Centre on behalf of the Australian Government's National Environmental Science Program (NESP) Tropical Water Quality (TWQ) Hub.

The Tropical Water Quality Hub is part of the Australian Government's National Environmental Science Program and is administered by the Reef and Rainforest Research Centre Limited (RRRC). The NESP TWQ Hub addresses water quality and coastal management in the World Heritage listed Great Barrier Reef, its catchments and other tropical waters, through the generation and transfer of world-class research and shared knowledge.

This publication is copyright. The Copyright Act 1968 permits fair dealing for study, research, information or educational purposes subject to inclusion of a sufficient acknowledgement of the source.

The views and opinions expressed in this publication are those of the authors and do not necessarily reflect those of the Australian Government.

While reasonable effort has been made to ensure that the contents of this publication are factually correct, the Commonwealth does not accept responsibility for the accuracy or completeness of the contents, and shall not be liable for any loss or damage that may be occasioned directly or indirectly through the use of, or reliance on, the contents of this publication.

Cover photographs: Front cover, Christopher Lowe; Back cover, www.vpnrus.com

This report is available for download from the NESP Tropical Water Quality Hub website:
<http://www.nesptropical.edu.au>

CONTENTS

Contents.....	i
List of Tables.....	ii
List of Figures.....	ii
Acronyms	iii
Abbreviations	iii
Acknowledgements	iv
Executive Summary	1
1.0 Introduction.....	2
1.1 Recent Context	2
1.2 Aim of this Technical Report	3
2.0 Methodology.....	4
2.1 Sentiment Analysis	4
3.0 Research Process - methodology.....	7
3.1 Downloading and Storing Tweets.....	7
3.2 Human Annotation	7
3.3 Pre-processing and Cleaning of Data.....	9
3.4 Text Embedding.....	9
3.5 Methodology (Language Modelling)	10
4.0 Results	14
4.1 Results from the human annotation	14
Tweet classification experiment – sentiment scores	14
Tweet classification experiment – target identification	15
4.2 Training and evaluating the sentiment analysis systems.....	17
4.3 Tweet Target Classification	17
Recap on specific method	17
Findings	18
5.0 Concluding Remarks	20
References.....	21
Appendix 1: Setup for the Experiments	24

LIST OF TABLES

Table 1:	Targets for human annotation.....	8
Table 2:	Annotation Sentiment scores with number of tweets for each score	14
Table 3:	Targets with number of annotated tweets for each target	15
Table 4:	Results from the target classification experiment.	19
Table 5:	Packages used to create an Anaconda virtual environment.....	25

LIST OF FIGURES

Figure 1:	Sentiment Classification techniques (Medhat et al. 2014).....	5
Figure 2:	Data Scope from Griffith Bigdata server's tweets repository.	8
Figure 3:	Three stages of text input embeddings.	9
Figure 4:	The two important steps of BERT NLP framework (Semi-supervised and Supervised training, image source: Alammar, 2018). The diagram on left shows Step 1 (trained on un-annotated data), and diagram on the right shows adapt and fine-tuning the model with a problem specific application data.	11
Figure 5:	Proposed context diagram for target classification – a supervised model. The system takes a sample tweet as an input and predicts the target label. The NLP framework converts a sample tweet input into a feature tensor, which is next classified using a Neural Network to determine its target label (for example if does not fit to any, “other”).....	11
Figure 6:	Stacked Encoders in BERT base (12) and BERT large (24) models.....	12
Figure 7:	The Transformer Model Architecture, originally from paper Attention is All You Need (Vaswani et al., 2017).....	13
Figure 8:	An encoder module (Alammar, 2018).	13
Figure 9:	Annotation Sentiment score distribution. X-axis shows sentiment score values and Y-axis shows frequency of tweets with respective scores.	15
Figure 10:	Target distribution. X-axis shows target name and Y-axis shows samples of tweets available with respective targets.	16
Figure 11:	Annotation results for sentiment by target.....	16
Figure 12:	Context diagrams of sentiment assessment pipeline, showing that: (a) our sentiment analysis system uses a tweet-text as input to produce a score (-5 to +5). (b) our system takes a tweet-text as input and produce a class label (i.e. positive or negative).....	17
Figure 13:	Context diagram of target classification pipeline. The proposed system takes a tweet text as input and returns a label from 11 target categories.	18

ACRONYMS

AI	Artificial Intelligence
API	Application programming interface
DL	Deep Learning
DoEE	Department of the Environment and Energy
GBR	Great Barrier Reef
GBRMPA	Great Barrier Reef Marine Park Authority
GPU	Graphics Processing Unit
ML	Machine learning
NESP	National Environmental Science Program
NLP	Natural Language Processing
RIMReP	Reef 2050 Integrated Monitoring and Reporting Program
RNN	Recurrent Neural Network
RRRC	Reef and Rainforest Research Centre
SANN	Sentiment Aware Nearest Neighbor
SELTMP	Social and Economic Long Term Monitoring Program
SVM	Support Vector Machine
TWQ	Tropical Water Quality
VADER	Valence aware dictionary for sentiment reasoning

ABBREVIATIONS

fp	false positives
tp	true positives

ACKNOWLEDGEMENTS

This report provides the technical background for research underpinning the assessment of experiences people have in association with the Great Barrier Reef. More specifically, social media (Twitter) data are used to establish sentiment and main topics of conversation, so-called targets. The work presented here relates to the development of a natural language-based model to automatically detect sentiment and targets in tweets. The authors are from the Griffith University School of Information and Communication Technology and the Griffith Business School, Griffith University.

The project is funded through the Australian Government's National Environmental Science Program (NESP) Tropical Water Quality (TWQ) Hub and addresses the aesthetic value of the Great Barrier Reef. It is administered by the Griffith Institute for Tourism.

We also thank Dr Ross Westoby, Ms Dung Le, Ms Arghavan Hadinejad and Prof Susanne Becken for their help with the human annotation of Twitter posts. Thank you to Ms Dung Le for proof reading.

EXECUTIVE SUMMARY

A significant investment has been made by the Australian Government in improving reef health and monitoring changes of the Great Barrier Reef (GBR) (through the Great Barrier Reef Marine Park Authority (GBRMPA) and the Reef 2050 Integrated Monitoring and Reporting Program (RIMReP). Monitoring environmental change, however, is costly and the notion of citizen science has attracted increasing attention. In addition to drawing on ‘people power’ for data collection, understanding the human dimensions of the GBR, including perceptions of natural beauty as an integral part of heritage value, is critical in enhancing support for conservation.

People provide information on their thoughts, perceptions and activities through a wide range of channels, including social media. Online generated content from social media can be used to better understand how people perceive or interact with nature. Analysis of such ‘big data’ allows organizations and analysts to make better and faster decisions, especially when such data was not previously available, accessible or usable. In this project, we capitalise on Big Data and Artificial Intelligence (AI) approaches to assess how social media data sources, namely Twitter, can be used to better understand human experiences related to the Reef.

The aim of this project is to develop and test advanced analytics techniques such as text mining, machine learning, predictive analytics, natural language processing, and data visualization to capture user experiences and potentially changing environmental conditions at the GBR. In summary, the project findings will deliver an innovative basis for enhancing current management systems of the GBR by measuring experience value, expressed through visitors’ sentiment and emotions contained in social media platforms. An automated tool based on machine learning could be used as part of a long-term monitoring plan, possibly situated within or alongside the Social and Economic Long Term Monitoring Program (SELTMP) for the Great Barrier Reef.

First, Twitter API with restrictions to capture geo-tagged tweets posted from the GBR geographic area was used to collect tweets data from the Twitter social media platform. Before large scale analytics can be performed in an automated fashion, it is important to develop and train algorithms. In our case, this was achieved by human annotation. To the end, a total of 13,006 tweets relevant to our project were selected for annotation. A label score was assigned to each tweet based on a scale from -5 to +5 along with a target label (e.g. travel, GBR, climate). The labels were derived with stakeholder input and reflected the most important and common topics of Twitter conversations. In total, there were 11 labels, or so-called targets.

Building on the human annotation, we implemented and validated a workflow for different models based on the language modelling approach to address the two different research problems, namely sentiment analysis and target classification. The models are based on deep learning-based techniques (precisely Deep Bidirectional Encoder or BERT). All downloaded tweets go through a pre-processing stage involving multiple steps. The processed texts are then represented by features using the language modelling approach. Tweets are then allocated to the 11 different targets, depending on their content. In addition, tweets are classified into positive and negative sentiment categories. Results reveal a relatively high accuracy, confirming the suitability of the developed algorithms for future implementation.

1.0 INTRODUCTION

1.1 Recent Context

In recent times, there has been a considerable increase in social media websites, blogs, and personalized websites, which in turn has popularized the platforms/forums for the expression of public opinions. Social media websites like Facebook, Twitter, Pinterest, and LinkedIn, and review websites like Amazon, IMDB, and Yelp have become popular sources for retrieving public opinions (Tang et al., 2015). The influence of social media, Mobile, Analytics and Cloud has offered the new technology paradigm and has transformed the operative environment and user engagement on the web. It has expanded the scope of commercial activities by enabling the users to discuss, share, analyse, criticize, compare, appreciate and research about products, brands, services through various social media platforms. Similarly, this pool of information can be explored for the mutual benefit of both the user and the organization. Analysing sentiments of this extremely large corpus of opinions can help an organization in realizing the public opinion and user experiences of the products or experiences (Kumar & Jaiswal, 2019).

There is also increasing interest in using social media data for the monitoring of nature experience and environmental change (Becken et al., 2018). Sometimes this is coupled with citizen science where members of the public are specifically encouraged to contribute data (Lodi & Tardin, 2018). Further evidence is needed to assess how well citizen science, collective sensing and social media data integrate with professional monitoring systems (Becken et al., 2018). Social media data are often only indirectly relevant to a particular research question, for example, the way people perceive a natural phenomenon, where they go or what they do. However, with appropriate filtering rules, it is possible to convert these unstructured data into a more useful set of data that provide insights into people's opinions and activities (Becken et al., 2017). Using Twitter as a source of data, Daume and Galaz (2016) concluded that Twitter conversations represent "embryonic citizen science communities" (p. e0151387).

The opportunity to collect insights into 'the public' or 'users' of the environment is pertinent for places that face rapid change and potentially decline. The Great Barrier Reef (GBR) is an example of a natural environment and an iconic visitor destination that faces significant environmental challenges. Nevertheless, the GBR is visited by over 2 million people each year (Becken et al., 2017). Also, a considerable number of local residents enjoy the natural environment of the Reef for recreation and other activities (including those which deliver economic returns). In addition to ecological changes, it is the aesthetic value that has been of growing interest. In their earlier report on the 'beauty of the GBR' Becken et al. (2018) reported on research that identified colourful fish as key attributes of aesthetic value. An Artificial-Intelligence-based system was developed to automatically identify a wide range of fish species and score the beauty of an image. The research concluded with recommendations on the future use of Big Data and artificial intelligence for cost-effective and real-time monitoring of visitor experiences and their aesthetic perceptions of the Reef. Drawing on big data approaches could enhance existing research on the social and economic dimensions of the GBR, as implemented through the Social and Economic Long Term Monitoring Program (SELTMP) for the Great Barrier Reef.

1.2 Aim of this Technical Report

Building on the existing research, and using publicly available Twitter data, a first useful step is to understand what people are talking about when they are in a particular location around the GBR, and whether their tone is positive or negative. Target-detection and sentiment analysis are suitable methods to generate insights into these questions. Both methods have benefitted from a range of recent developments, including (1) an escalation of web- and social-media-based information, (2) evolution of new technologies, especially machine learning approaches for text analysis, and (3) shifts in business models and applications that make use of this information. Despite its popularity, sentiment analysis is still in its infancy compared to earlier technologies, such as data mining and text summarization (Pan et al., 2007).

This technical report presents the development of a methodology and algorithm that enables the large-scale scoring of sentiment in tweets. It also identifies specific targets and can derive the sentiment by target.

Specifically, the aim of this report is to:

1. Provide a synopsis of relevant literature/technologies.
2. Explain the process, including human annotation, that was developed to automate sentiment and target detection for user experiences at the GBR.
3. Provide results from the testing to evaluate the algorithms to determine the accuracy of the developed algorithms.

The findings from this work will help assess the suitability of machine-learning-based systems for monitoring the experience value of the GBR, based on automated scoring of Twitter feeds. Such a system could form part of the existing human dimension monitoring as implemented through the wider Reef 2050 Integrated Monitoring and Reporting Program.

2.0 METHODOLOGY

2.1 Sentiment Analysis

Opinion mining based on sentiment orientation was studied in recent years to understand perceptions and characteristics of population or market groups and to determine the credibility of content and motivations for posting reviews (Ribeiro et al. 2016). Data collection, data cleaning, mining process, and then evaluation and understanding of the results are the major steps used in most of the applications in relation to social media data analysis (Hippner & Rentzmann, 2006). Text summarization aims to transform lengthy documents into shortened versions using natural language processing (NLP), and text classification also uses NLP along with machine learning technologies to facilitate information processing and data analysis (Cantalops & Salvi, 2014; Gerdes et al., 2015). Sentiment analysis basically refers to the use of computational linguistics and NLP to analyze text and identify its subjective information. Different sentiment analysis methods were developed in various domains (Alaei et al., 2017).

Support Vector Machine (SVM) and Naïve Bayes are the key machine learning-based classification methods used for sentiment analysis in the literature (Brob, 2013; Kang et al., 2012; Markopoulos et al., 2015; Shi & Li, 2011; Shimada et al., 2011; Ye et al., 2009), as these two methods were conventionally designed for two-class classification problems. A SVM classifier uses annotated data for training to obtain an optimal separating line to accurately categorize new samples into different groups. A Naïve Bayes classifier is a probabilistic classifier, which uses Bayes's theorem in the classifier's decision rule, with an assumption that features are independent. SVM and Naïve Bayes methods need comparably less annotated data for model training compared with the Neural Network-based approach. Neural network and deep learning models (Irsoy & Cardie, 2014; Socher et al., 2013) and the K-nearest neighbour method (Schmunk et al., 2014) are alternative methods for semantic analysis. K-means clustering techniques (Xiang et al., 2015) and statistical models based on the probability distribution of sentiment of reviews (Rossetti et al., 2015) were proposed for sentiment analysis of short length text data. Elsewhere, Naïve Bayes models were also adapted in an unsupervised fashion for sentiment analysis by Shimada et al. (2011).

Dictionary-based systems rely on the use of comprehensive sentiment lexica and sets of fine-tuned rules. A sentiment lexicon can be created either by humans, by machine or by both (semi-automatically). For instance, a dictionary may contain words such as "good", "fantastic," "bad" or "ugly," with their associated values of polarity (Alaei et al., 2017). Few methods were published for dictionary-based approaches (Bjorkelund et al., 2012; Bucur, 2015; Garcia et al., 2012; Hutto & Gilbert, 2014; Levallois, 2013). One is SentiWordNet which has been used on its own (Bucur, 2015; Garcia et al., 2012), or in combination with a simplified Lesk algorithm, (Bjorkelund et al., 2012). Valence aware dictionary for sentiment reasoning (VADER) method has provided promising results on Twitter data (Hutto & Gilbert, 2014; Becken et al., 2017). VADER combines a lexicon and a series of intensifiers, punctuation transformation, and emoticons, along with heuristics to compute sentiment polarity of text. Umigon is another dictionary-based method that uses heuristics for Twitter sentiment analysis (Levallois 2013).

The dictionary-based approach was improved by introducing semantic-based analysis methods (Tsytarau & Palpanas, 2012) that require a dictionary of domain-specific terms and their associated polarity values. In hybrid approaches, dictionary, and machine learning-based approaches can work in parallel to compute two sentiment polarities. Several sentiment analysis model architectures that incorporate dictionary-based and machine-learning-based methods at different stages of the model are available (Kasper & Vela 2011; Claster et al. 2013; Pappas & Popescu-Belis 2013; Schmunk et al. 2014; Chiu et al. 2015). The Sentiment Aware Nearest Neighbor (SANN) model is one example. SANN initially classifies a text as either a subjective or objective (Pappas & Popescu-Belis 2013). If it is classified as subjective, further classification into either positive or negative sentiment is undertaken.

Figure 1 presents different sentiment analysis methods, identifying three categories, namely machine learning-based approach, lexicon-based approach and hybrid approach. The Machine learning-based approach (ML) applies diverse machine learning algorithms and uses linguistic features. The *Lexicon-based approach* relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approaches and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The *Hybrid approach* combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods (Medhat et al. 2014). Machine learning methods are further categorized into supervised and unsupervised approaches. The dictionary-based approach also includes a subcategory called semantic-based approach (Tsytarau & Palpanas, 2012). With regards to target identification only machine learning and natural language processing methods have been considered in this present research.

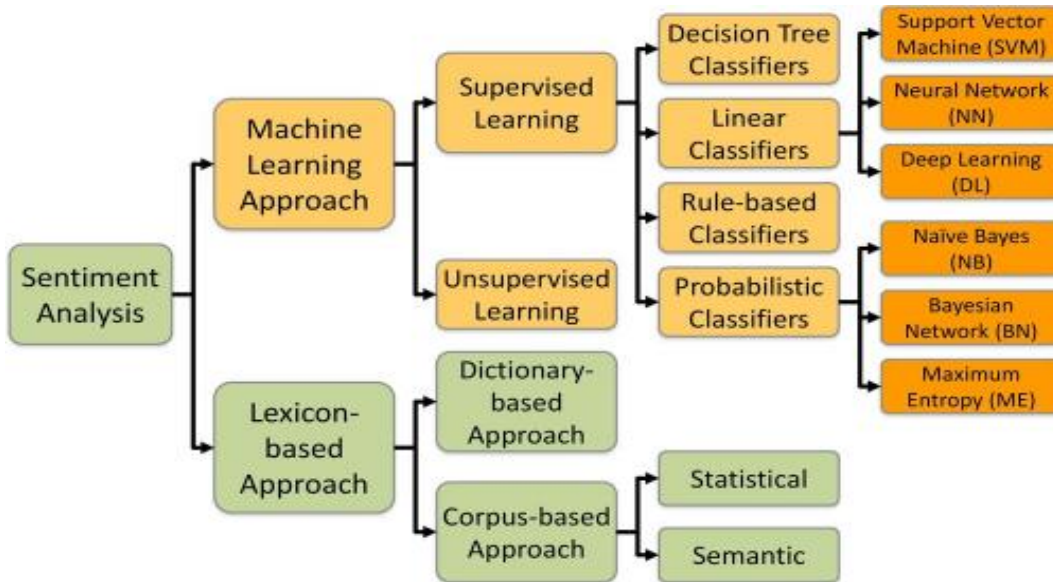


Figure 1: Sentiment Classification techniques (Medhat et al. 2014).

For the task of analysing GBR-related Twitter posts, we propose deep learning-based techniques (deep learning is a subfield of machine learning) (precisely Deep Bidirectional Encoder or BERT, Devlin et al., 2018). At the initial stage, we accepted language modelling-based sentiment analysis approach because of its state-of-the-art performance on a range of

Natural Language Processing (NLP) tasks. Also, language modelling-based models outperformed human coding on various datasets (Devlin et al., 2018). All downloaded tweets go through a multi-step pre-processing stage, after which they are processed via a language model for feature representation. Finally, a classifier is employed for classification. We implemented and validated a workflow for three independent systems:

- i) Sentiment analysis (i.e. binary classification of positive and negative)
- ii) Sentiment analysis with intensity (i.e. score between -5 to +5 range)
- iii) Target or topic-level classification for aspect or topic-level sentiment analysis.

3.0 RESEARCH PROCESS - METHODOLOGY

The proposed work for developing sentiment analysis and target classification algorithms is based on deep learning techniques and sophisticated language modelling. The following sections detail the steps involved in the experimentation from obtaining social media data (Twitter) to producing results.

3.1 Downloading and Storing Tweets

As outlined in earlier publications (Becken et al., 2017, 2018), the research team drew on a public Twitter Application programming interface (API) with restrictions to capture geo-tagged tweets posted from the GBR geographic area. A rectangular bounding box was defined (Southwest coordinates: 141.459961, -25.582085 and Northeast coordinates: 153.544922, -10.69867) that broadly represents the GBR region.

Data are stored in a NoSQL MongoDB database, which is located on a cluster computer with a Hadoop architecture at Griffith University. Each tweet in the database contains additional information (i.e. metadata) such as language, location where the account was registered, and location from where the tweet was posted. A data management plan has been compiled and shared with eAtlas and NESP.

3.2 Human Annotation

To train and validate an algorithm, it is important to have an annotated dataset that prescribes sentiment and target for each tweet. This approach falls under the supervised learning path, and it involves manual annotation by humans.

For this research, we extracted 13,000 tweets relevant to our research work for human annotation. More specifically, the following steps were put in place:

- Agree on a preliminary set of targets (i.e. topics) that are of relevance to this research, and that reflect the general coverage of themes prevalent in tweets. Discuss among the research team.
- Obtain feedback and input from key GBR stakeholders on the proposed targets and make adjustments to ensure the selected topics are useful for future decision making.
- Finalise the targets and agree on a scale for scoring sentiment. In this case, tweets were classified based on a scale from -5 to +5 (-5 for highly negative to +5 highly positive).
- A team of six researchers coded 13,000 tweets, whereby a minimum of 2000 tweets were allocated per researcher.

Since the first set of 11,800 tweets randomly selected tweets mainly contained tweets that fell into the category 'other', it was important to extract additional tweets that specifically addressed issues directly related to the GBR. To this end, and using GBR-related keywords as shown in Table 1 below, an additional 1,206 tweets were extracted from the database. These were annotated by a team of six researchers. Figure 2 visualises the sampling procedure for the annotation phase.

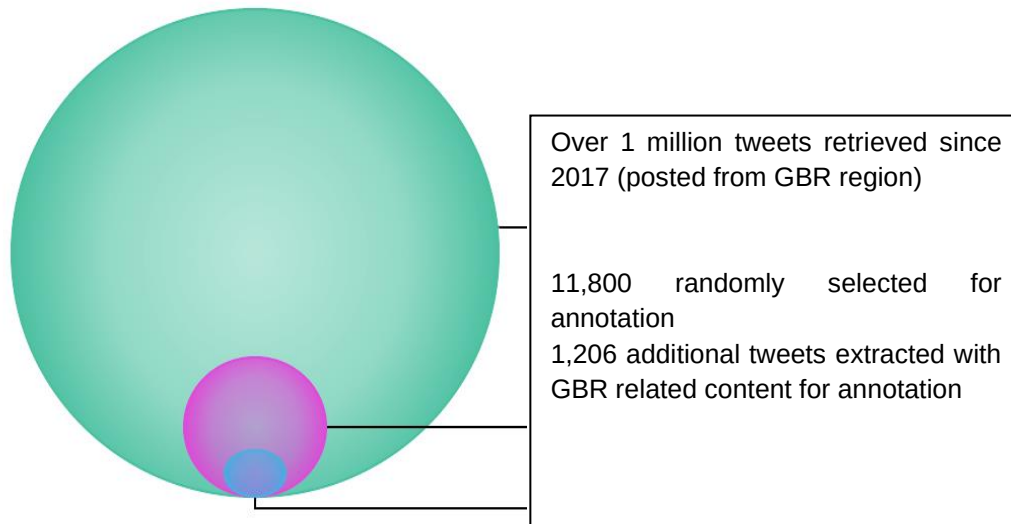


Figure 2: Data Scope from the Griffith Big data server's tweets repository.

More detail on the identified targets is shown in Table 1, alongside with indicator or keywords that researchers agreed on to represent a particular target. The keywords were designed to guide the decision making process, but were not exclusive to the respective targets.

Table 1: Targets for human annotation.

SHORT	Target	Includes indicator words, such as...
Accom	Accommodation, food, hospitality	Hotel, motel, AirBnB, sleep, Restaurant, meal, bar, drinks, hospitality, staff, welcoming, service, clean, dirty, friendly, master reef guides, knowledgeable, guide
GBRact	GBR-related Activities	Diving, dive, snorkelling, snorkel, swimming, swim, ocean swim, aquarium, divemaster, dive instructor
Landact	Events, Activity, Attractions	Museum, shopping, casino, skyrail, city, landscape, photograph, competition, race, celebration, game, team, birthday, play, party, regatta
Climate	Climate/weather	Rain, sun, forecast, storm, humidity, warm, cold, barometer, windy, gale, knots, cyclone
Terrestrial	Terrestrial environment	Rainforest, waterfall, creek, trees, park, World Heritage Area, crocodile
Coastal	Coastal, Reef, Marine animals	Beach, sand, bay, coast, island, strand, esplanade, jetty, reef, coral, shark, coral, fish, turtle, humpback, whale, nemo, cod, ray, eel, seabird, osprey, clam, anemone
Culture	Culture	Dance, heritage, Aboriginal, indigenous, performance, show, music, art, festival
Safety	Safety, health	Cyclone, flooding, impact, risk, damage, Hospital, sick, pharmacy, monsoon, stranded, evacuate
Trave	Transport/travel/ travel business/ infrastructure	Plane, bus, taxi, uber, travel, delayed, boat, charter, chopper, helicopter, tour operator, company, business, Airport, marine, road, information centre, port, charter
Politics	Politics	Adani, election, policy, government, council, auspol, news
Other	Other	Anything else

3.3 Pre-processing and Cleaning of Data

Twitter data are often messy and contain a lot of redundant information. Also, there are several other steps that need to be put in place to make subsequent analysis easier. In other words, to eliminate text/data which is not contributing to the assessment it is important to pre-process the tweet. Initial data cleaning involved the following:

- a) **Removing Twitter Handles (@user):** The Twitter handles do not contain any useful information about the nature of the tweet, so they can be removed.
- b) **Removing Punctuations, Numbers, and Special Characters:** The punctuations, numbers and even special characters are removed since they typically do not contribute to differentiating tweets.
- c) **Tokenization:** We split every tweet into individual words or tokens which is an essential step in any NLP task. The following example shows a tokenization result,

Input: [sunrise was amazing]

After tokenization: [sunrise, was, amazing]

- d) **Stemming:** It is a rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word. For example: “play”, “player”, “played”, “plays” and “playing” are the different variations of the word – “play”. The objective of this process is to reduce the total number of unique words in our data without losing a significant amount of information.

3.4 Text Embedding

The BERT-based (Devlin et al., 2018) method added a specific set of rules to represent the input text for the model. Many of these are designer choice that makes the model even work better. The input embedding is a combination of 3 embedding techniques (see Figure 3):

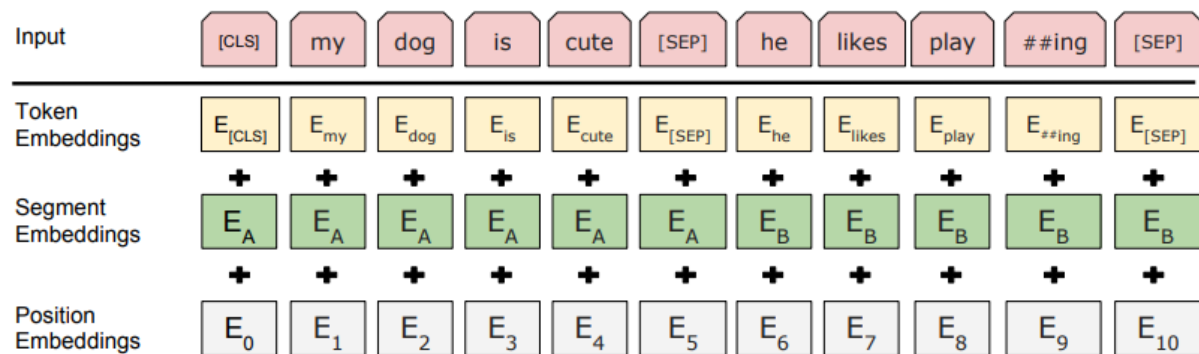


Figure 3: Three stages of text input embeddings.

Position Embeddings: Model learns and uses positional embeddings to express the position of words in a sentence. These help overcome the limitation of *Transformer* which, unlike an RNN (Recurrent Neural Network), is not able to capture “sequence” or “order” information.

Segment Embeddings: Model can also take sentence pairs as inputs for tasks (Question-Answering). That's why it learns a unique embedding for the first and the second sentences to help the model distinguish between them. In the Figure 2, all the tokens marked as EA belong to sentence A and similarly, EBs belong to sentence B.

Token Embeddings: These are the embeddings learned for the specific token from the WordPiece token vocabulary. The [CLS] token added at the beginning and [SEP] tokens in the right place.

3.5 Methodology (Language Modelling)

We have adapted BERT model, whereby BERT stands for Bidirectional Encoder Representations from Transformers (Vaswani et al., 2017). It builds upon recent work in pre-training contextual representations such as GPT, Elmo, and ULMFit (these three methods were significant milestones before the BERT method). Pre-trained representations can either be context-free or contextual, and contextual representations can further be unidirectional or bidirectional. It is the first deeply bidirectional (learn sequence from both ends), unsupervised language representation, pre-trained using only a plain text corpus (Wikipedia). It is designed to pre-train deep bidirectional representations from an unlabelled text by jointly conditioning on both the left and right context. It is considered a key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling.

This is in contrast to previous efforts that looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The experimental outcomes of BERT show that a language model that is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models. As a result, the pre-trained BERT model can be fine-tuned with an extra additional output layer to create state-of-the-art models for a wide range of NLP tasks (Figures 4 and 5). This bidirectional Transformer model that redefines the-state-of-the-art for a range of natural language processing tasks (e.g. Question answering), even surpassing human performance in the challenging area.

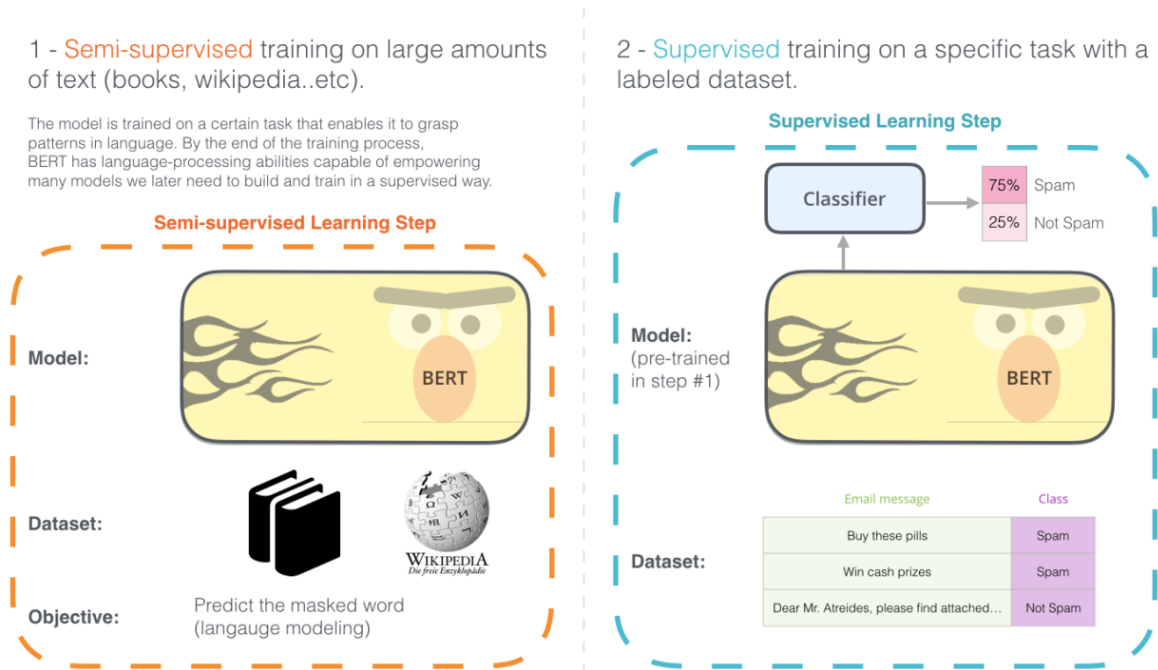


Figure 4: The two important steps of BERT NLP framework (Semi-supervised and Supervised training, image source: Alammur, 2018). The diagram on left shows Step 1 (trained on un-annotated data), and diagram on the right shows adapt and fine-tuning the model with a problem specific application data.

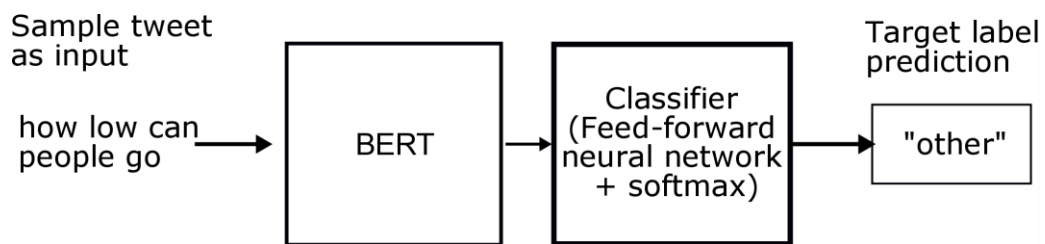


Figure 5: Proposed context diagram for target classification – a supervised model. The system takes a sample tweet as an input and predicts the target label. The NLP framework converts a sample tweet input into a feature tensor, which is next classified using a Neural Network to determine its target label (for example if does not fit to any, "other").

Trained Transformer Encoder stack and the transformer encoder and the transformer in Figure 6 and detail descriptions are given below. There are two variants of presents and two model sizes for the BERT framework. Both BERT models have a large number of encoder layers (this encoder layer is presented in the model as Transformer Blocks) twelve for the Base version and twenty-four for the Large version. BERT BASE has 12 encoder stacks and it is comparable in size to the OpenAI Transformer to compare performance. BERT LARGE model has 24 encoder stacks which achieved the state-of-the-art results on many NLP tasks. The model is pre-trained on over a 3.3-billion-word corpus, including Books Corpus (800 million words) and English Wikipedia (2.5 billion words).

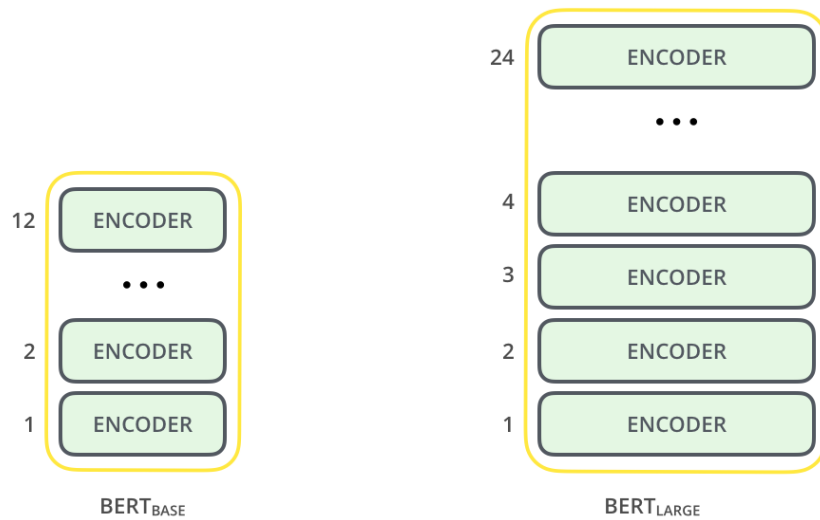


Figure 6: Stacked Encoders in BERT base (12) and BERT large (24) models

Each encoder in the model takes a sequence of words as input which keeps flowing up the stack. Each layer applies self-attention and passes its results through a feed-forward network, and finally passes it to the next encoder layer.

Transformer: We relied on Transformer model which was proposed in the paper ‘Attention is All You Need’ (Vaswani et al., 2017). Figure 7 presents the full architecture of this model. The Encoder is on the left and the Decoder is on the right. Both Encoder and Decoder are composed of modules that can be stacked on top of each other multiple times.

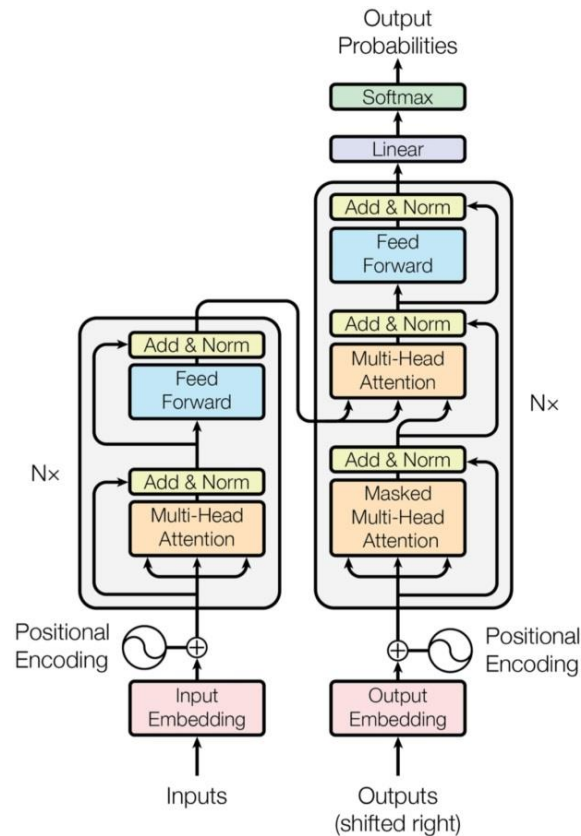


Figure 7: The Transformer Model Architecture, originally from paper Attention is All You Need (Vaswani et al., 2017).

The *encoding* component is a stack of encoders (whereby the stacks involve six of them on top of each other). The *decoding* component is a stack of decoders of the same number. The encoders are all identical in structure, but they do not share weights. Each encoder is broken down into two sub-layers. As shown in Figure 8, the encoder's inputs first flow through a self-attention layer – a layer that helps the encoder look at other words in the input sentence as it encodes a specific word. The outputs of the self-attention layer are fed to a feed-forward neural network. The exact same feed-forward network is independently applied to each position.

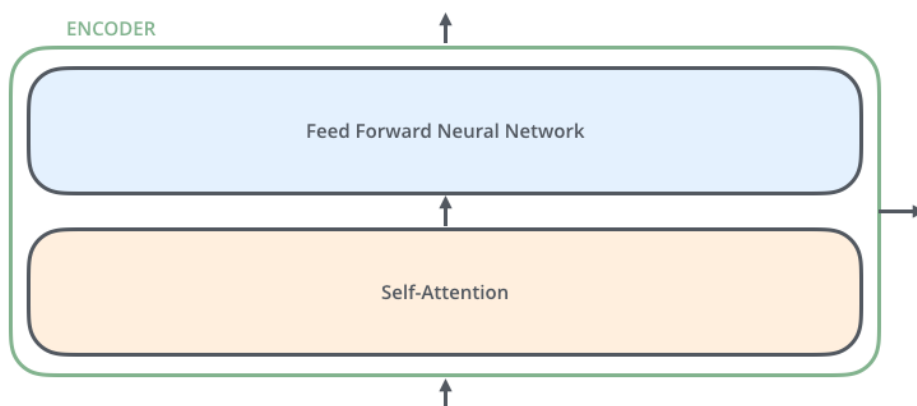


Figure 8: An encoder module (Alammar, 2018).

4.0 RESULTS

The following section presents the outcomes of the human annotation, followed by results from the testing of algorithms using a fresh data set of tweets.

4.1 Results from the human annotation

The proposed sentiment analysis method takes a pre-processed tweet as an input and returns a class label (positive or negative class) or a score ranges between -5 to +5 (according to our design) based on the sentiment assessment model and annotated values obtained in the training. Since the literature presents approaches that consider three sentiment classes (positive, negative and neutral), we have also included this option as a potential alternative. A tweet is considered as a neutral class when the tweet score is given 0 (neither positive nor negative).

The network models and training approaches are different for these different systems. Annotators annotated each tweet with a score (ranges between -5 to +5) from our training and test dataset. These scores have been used to train the system in order to produce a score within the same range (-5 to +5) on a new dataset. In the second method, we are using the annotated score to compute a label for each tweet into a positive or negative class label for the binary classification task.

Tweet classification experiment – sentiment scores

In total, 13,006 tweet samples from the GBR area (defined by a bounding box) have been downloaded using tweeter API. In order to perform machine learning-based sentiment analysis, we needed to manually annotate tweets. 2,285 tweets have been annotated with a Negative sentiment (-5 to -1), 3360 were recognised as a Positive sentiment (+1 to +5), and 7061 tweets were annotated as Neutral (0). The findings are shown in Table 2 and Figure 9, showing a relatively even distribution with a slight bias towards positive tweets.

Table 2: Annotation Sentiment scores with number of tweets for each score

Sentiment score	Number of Tweets	Percentage
-5	200	1.54%
-4	338	2.60%
-3	795	6.11%
-2	665	5.11%
-1	287	2.21%
0	7061	54.29%
1	343	2.64%
2	810	6.23%
3	1212	9.32%
4	774	5.95%
5	521	4.01%

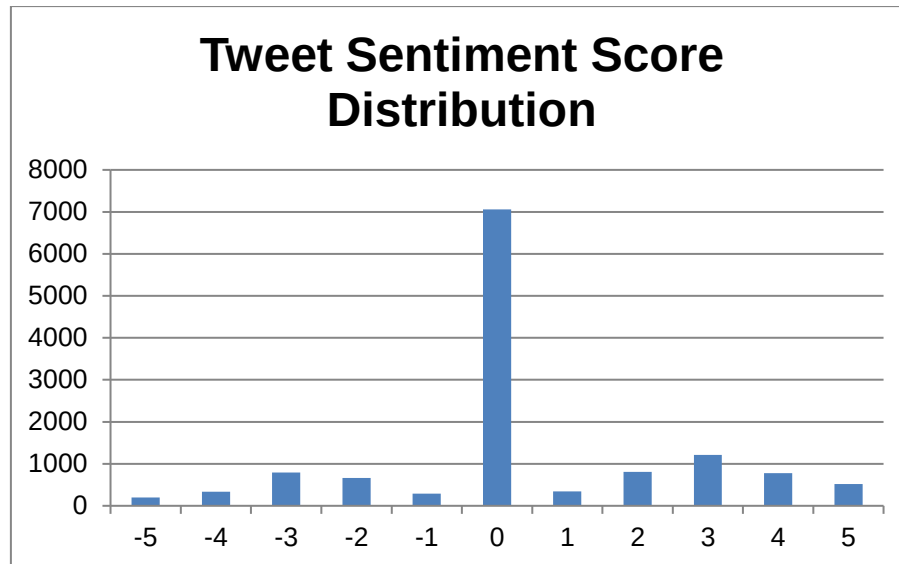


Figure 9: Annotation Sentiment score distribution. X-axis shows sentiment score values and Y-axis shows frequency of tweets with respective scores.

Tweet classification experiment – target identification

To perform machine Learning-based target analysis, each tweet was annotated and classified according to one of the 11 target categories (Table 3).

Table 3: Targets with number of annotated tweets for each target

Target	Number of Tweets	Percentage
Accom	292	2.25%
Climate	204	1.57%
Coastal	511	3.93%
Culture	127	0.98%
GBRact	185	1.42%
Landact	907	6.97%
Politics	903	6.94%
Safety	181	1.39%
Terrestrial	196	1.51%
Travel	336	2.58%
Other	9164	70.46%

Figure 10 visualises the distribution of targets within the sample of tweets, highlighting that the vast majority of tweets talk about topics that are neither relevant to the travel experience around the GBR, nor the environmental condition of the Reef itself. Tweets that contain unidentifiable content (e.g. a list of hashtags, nonsensical comments, and other) were also included in this target. Twitter users in the GBR region appear to talk more about land-based activities compared with water-based ones.

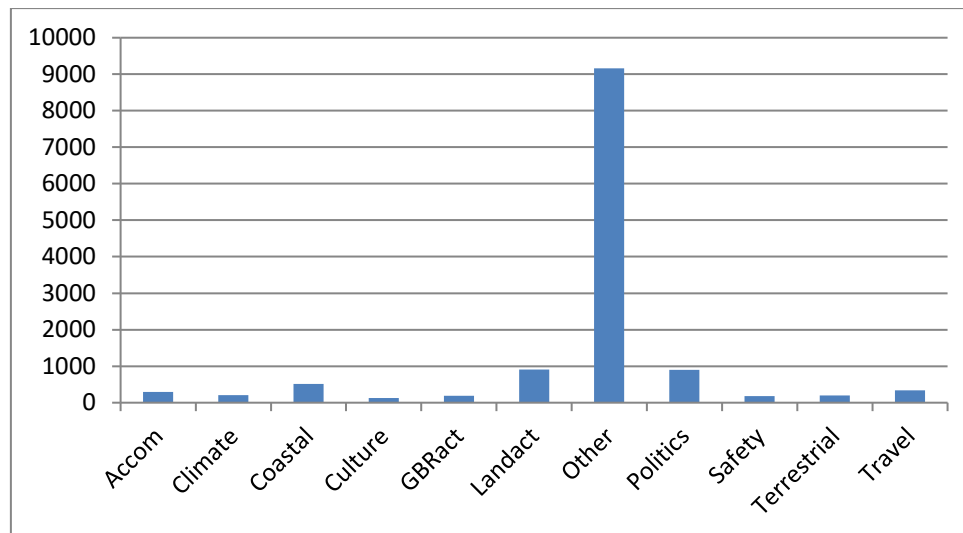


Figure 10: Target distribution. X-axis shows target name and Y-axis shows samples of tweets available with respective targets.

Finally, the results were assessed to determine the sentiment evident in tweets by target. Figure 11 shows that positive tweets are slightly more abundant, and this seems particularly evident for targets that relate to the coastal environment and land-based activities.

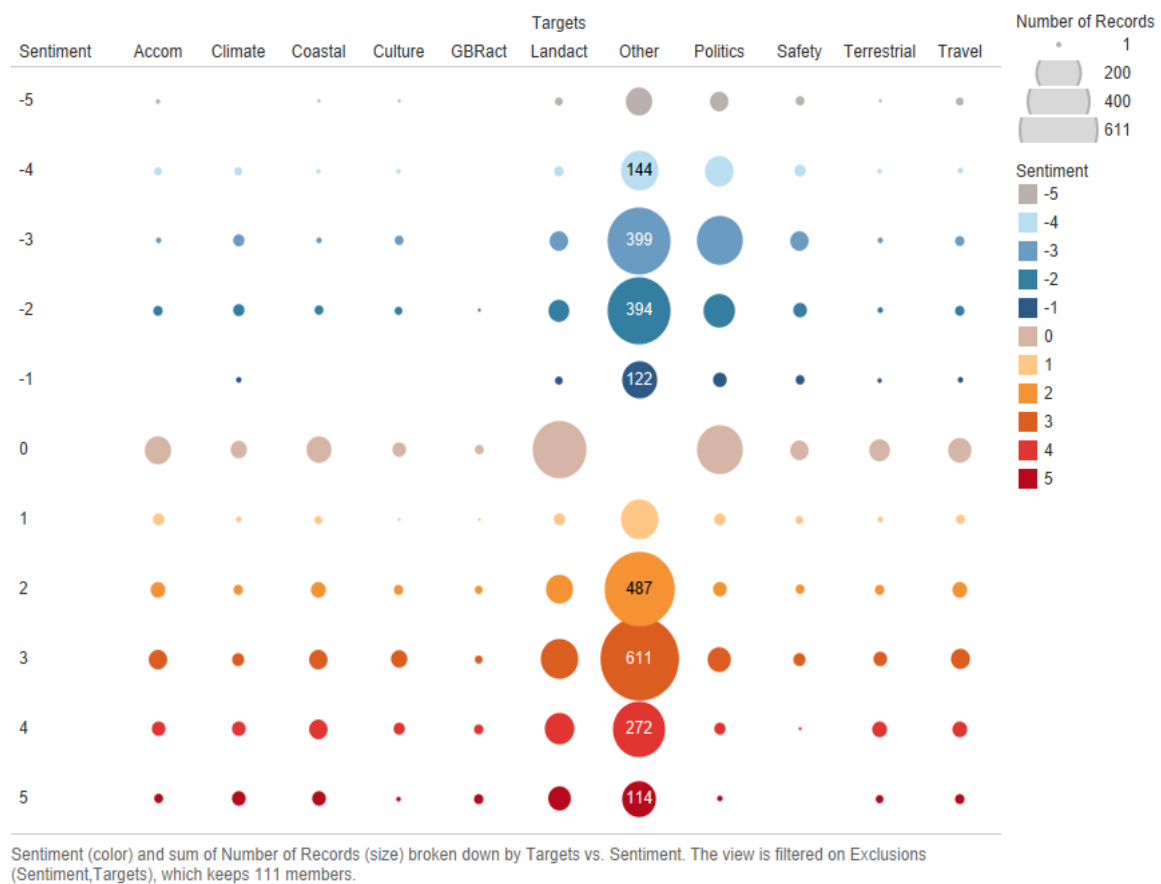


Figure 11: Annotation results for sentiment by target.

4.2 Training and evaluating the sentiment analysis systems

The three different tasks for Twitter sentiment analysis are summarized in the following three categories:

- Classifying tweets into **two** sentiment classes (Positive and negative). A system using the BERT language model has been developed for this binary classification task. A total of 5,945 tweets with sentiments score of both positive and negative are available in our annotated dataset, whereby we used tweets with positive and negative sentiments only for the experiments. An accuracy of 86.25% was achieved on 20% of test data. We are working on even further improving the binary classification results.
- Classifying tweets into **three** sentiment classes (Positive, negative and Neutral). A system using a language modelling technique has been developed for this ternary classification task. The final classification layer and training technique is slightly different from the above binary classification model as we consider one more additional classes (i.e. Neutral).
- The proposed regression model is designed to generate a score for each tweet between -5 to + 5. The language modelling-based model has been employed to generate the feature vector from a tweet text and the vector is passing through a linear regression function (it applies a linear transformation to the incoming data) instead a softmax regression (i.e. generalization of logistic regression) to generate the score. In Figure 12 the bottom part of the pipeline diagrams provides some more detail, whereby 'Embed' represents embeddings and the 'Encoder' stands for the language model.

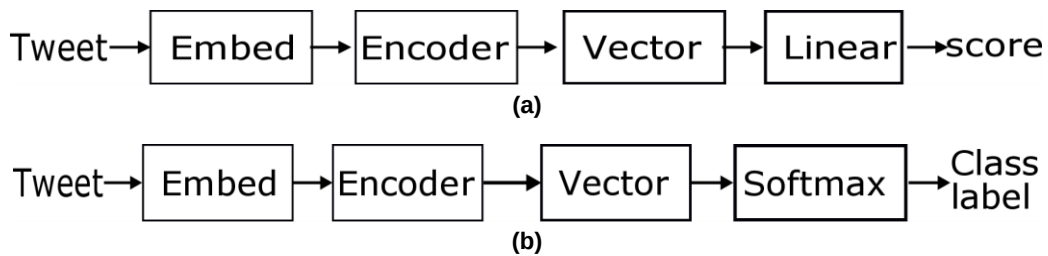


Figure 12: Context diagrams of sentiment assessment pipeline, showing that: (a) our sentiment analysis system uses a tweet-text as input to produce a score (-5 to +5). (b) our system takes a tweet-text as input and produce a class label (i.e. positive or negative).

4.3 Tweet Target Classification

Recap on specific method

A system has been developed for multi-class twitter text classification based on sequence representation-based model and experimented with our dataset that contains 13,006 (training and test data) tweets categorised by 11 targets. In the pipeline diagram shown below (Figure 13), 'Embed' represents embeddings and the 'Encoder' represents the language model. The

text embeddings and the language model that have been used in our multi-class classification task have been described earlier.

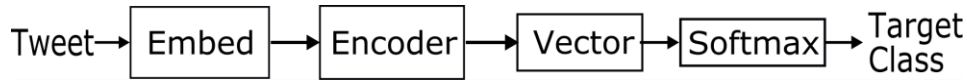


Figure 13: Context diagram of target classification pipeline. The proposed system takes a tweet text as input and returns a label from 11 target categories.

The following terms are of importance to understand the findings:

Evaluation Method: We compute precision, recall, F-measure and support metrics for each class (scikit-learn).

Precision: The precision is the ratio $tp / (tp + fp)$ where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

Recall: The recall is the ratio $tp / (tp + fn)$ where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

F-measure: The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.

Micro: Calculate metrics globally by counting the total true positives, false negatives and false positives.

Macro: Calculate metrics for each label, and find their unweighted mean. This does not take into account any label imbalance.

Weighted: Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance; it can result in an F-score that is not between precision and recall.

Samples: The samples are the number of occurrences of each class in test dataset.

Findings

An overall precision of 82% and 0.76 f-score have been achieved on the experimental dataset (Table 4). A detailed result of the target classification (test dataset contains 1301 samples) has been presented below. For a few targets, due to having low samples of data from a few target-classes, the accuracy is low. To improve the result in such cases, a more detailed analysis is underway to find out how the accuracy can be improved. However, incorporating more data into the dataset will be useful to create a balanced dataset.

Table 4: Results from the target classification experiment.

Target class	precision	recall	f1-score	samples
Accom	0.5	0.04	0.08	24
Climate	0	0	0	27
Coastal	0.62	0.24	0.35	62
Culture	0	0	0	15
Gbract	0	0	0	22
Landact	0.38	0.31	0.34	83
Other	0.87	0.89	0.88	896
Politics	0.75	0.6	0.67	105
Safety	0	0	0	21
Terrestrial	1	0.14	0.25	14
Travel	0.58	0.22	0.32	32
<i>Micro Avg</i>	0.82	0.7	0.76	1301
<i>Macro Avg</i>	0.43	0.22	0.26	1301
<i>Weighted Avg</i>	0.75	0.7	0.71	1301

5.0 CONCLUDING REMARKS

Big Data and AI-based approaches can help organizations to discover more insight from different aspects of Big Data that will eventually make the decision making improved and faster. Research-driven approaches and data-driven practices can support domain experts to understand or explain phenomena as well as to realize new dimensions. In this work, we designed efficient systems based on deep learning to analyze public sentiments from a large set of social media twitter data from the Great Barrier Reef region, Australia. We have achieved encouraging results (accuracy) on sentiment analysis and target classification tasks that are reported in this document in finer details. However, our next target is to improve the dataset by adding more topic relevant data and fine-tuning the systems to achieve better accuracy.

REFERENCES

- Alaei, A.R., Becken, S. & Stantic, B. (2017). Sentiment analysis in tourism: Beginning of a new research paradigm? *Journal of Travel Research*, <http://journals.sagepub.com/doi/pdf/10.1177/0047287517747753>
- Alammar, J. (2018). The Illustrated BERT, ELMo, and co." Available (11/12/19) <http://jalammar.github.io/illustrated-bert/>.
- Becken, S., Stantic, B., Chen, J. Alaei, A.R. & Connolly, R. (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.
- Becken, S., Connolly, R., Stantic, B., Scott, N., Mandal, R. & Le, D. (2018). Monitoring aesthetic value of the Great Barrier Reef by using innovative technologies and artificial intelligence. Report to the National Environmental Science Program. Reef and Rainforest Research Centre Limited, Cairns (48pp.) Available (11/07/18) <http://nesptropical.edu.au/wp-content/uploads/2018/07/NESP-TWQ-Project-3.2.3-Final-Report.pdf>
- Becken, S., Connolly, R.M., Chen, J. & Stantic, B. (2018). A hybrid is born: integrating collective sensing, citizen science and professional monitoring of the environment. *Ecological Informatics*, 52: 35-45.
- Bjorkelund, E., Burnett, T. H. & Norvag, K. (2012). A Study of Opinion Mining and Visualization of Hotel Reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications and Services*, 229-38. New York: ACM.
- Bucur, C. (2015). Using Opinion Mining Techniques in Tourism. In *Proceedings of the 2nd Global Conference on Business, Economics, Management and Tourism*, 1666-73. Amsterdam: Elsevier.
- Brob, J. (2013). Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques. PhD diss., University of Berlin.
- Chen, S. (2018). Attention Is All You Need — Transformer, Available (20/12/19) <https://medium.com/towards-artificial-intelligence/attention-is-all-you-need-transformer-4c34aa78308f>.
- Claster, W. B., P. Pardo, M. Cooper, & Tajeddini. K. (2013). Tourism, Travel and Tweets: Algorithmic Text Analysis Methodologies in Tourism. *Middle East Journal of Management* 1 (1): 81-99.
- Chiu, C., Chiu, N.H., Sunga, R.J. & Hsieh, P.Y. (2015). Opinion Mining of Hotel Customer-Generated Contents in Chinese Weblogs. *Current Issues in Tourism* 18 (5): 477-95.
- Daume, S. & Galaz, V. (2016). Anyone Know What Species This Is? – Twitter Conversations as Embryonic Citizen Science Communities. *PLoS ONE* 11(3): e0151387. doi:10.1371/journal.pone.0151387
- Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, <http://arxiv.org/abs/1810.04805>.

- Garcia, A., Gaines, S. & Linaza, M.T. (2012). A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain. *e-Review of Tourism Research* 10 (2): 35-38.
- Hippner, H., & R. Rentzmann. (2006). *Text Mining*. Ingolstadt: Springer-Verlag.
- Cantallos, A. S., and F. Salvi. 2014. "New Consumer Behavior: A Review of Research on eWOM and Hotels." *International Journal of Hospitality Management* 36: 41-51.
- Hutto, C. & E. Gilbert. (2014). Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 216-25. Palo Alto, CA: AAAI.
- Irsoy, O. & Cardie, C. (2014). Opinion Mining with Deep Recurrent Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 720- 28. New York: ACM.
- Kasper, W., & Vela, M. (2011). Sentiment Analysis for Hotel Reviews. In *Proceedings of the Computational Linguistics-Applications Conference*, 45-52. New York: IEEE.
- Kumar, A., & Jaiswal, A. (2019). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*. e5107. 10.1002/cpe.5107.
- Kang, H., S. J. Yoo, & D. Han. (2012). Senti-lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews. *Expert Systems with Applications* 39 (5): 6000-6010.
- Levallois, C. (2013). Umigon: Sentiment Analysis for Tweets Based on Terms Lists and Heuristics. In *Second Joint Conference on Lexical and Computational Semantics. Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, 414-17. Stroudsburg, PA: ACL.
- Lodia, L. & Tardin, R. (2018). Citizen science contributes to the understanding of the occurrence and distribution of cetaceans in South-eastern Brazil – A case study. *Ocean & Coastal Management*, 158: 45-55.
- Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, pp. 1093-1113, Vol. 5.
- Markopoulos, G., G. Mikros, A. Iliadi, & Lontos, M. (2015). Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features of Cultural Tourism in a Digital Era. In *Cultural Tourism in a Digital Era*, 373-83. New York: Springer.
- Pan, B., T. MacLaurin, & J. C. Crotts. (2007). Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research* 46 (1): 35-45.
- Pappas, N., & A. Popescu-Belis. (2013). Sentiment Analysis of User Comments for One-Class Collaborative Filtering over TED Talks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 773-76. New York: ACM.
- Rossetti, M., F. Stella, L. Cao, & Zanker, M. (2015). Analysing User Reviews in Tourism with Topic Models. In *Information and Communication Technologies in Tourism 2015*, edited by I. Tussyadiah and A. Inversini, 47-58. New York: Springer.

- Ribeiro, F. N., M. Araujo, P. Goncalves, M. A. Goncalves, & Benevenuto, F. (2016). SentiBench—A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods. <https://arxiv.org/abs/1512.01818>2015arXiv151201818N.
- Shi, H.-X., & Li, X.-J. (2011). A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, 950-54. New York: IEEE.
- Shimada, K., S. Inoue, H. Maeda, & Endo, T. (2011). Analyzing Tourism Information on Twitter for a Local City. In *Proceedings of the First ACIS International Symposium on Software and Network Engineering*, 61-66.
- Socher, R., A. Perelygin, J. Y. Yu, J. Chuang, C. D. Manning, A. Y. Ng, & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1631-42. Stroudsburg, PA: ACL.
- Schmunk, S., W. Höpken, M. Fuchs, Lexhagen, M. (2014). Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC. In *Information and Communication Technologies in Tourism (2014)*, edited by P. Xiang and I. Tussyadiah, 253-65. New York: Springer.
- Tang, D., Qin, B. & Liu, T. (2015). Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisc Rev: Data Mining Knowledge Discovery* 5 (6): 292–303.
- Tsytsarau, M., & T. Palpanas. (2012). Survey on Mining Subjective Data on the Web. *Data Mining and Knowledge Discovery* 24 (3): 478-514.
- Vaswani, N., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need, <https://arxiv.org/abs/1706.03762>.
- Xiang, Z., Schwartz, Z., Gerdes, J. H. Jr. & Uysal, M. (2015). What Can Big Data and Text Analytics Tell Us about Hotel Guest Experience and Satisfaction? *International Journal of Hospitality Management* 44:120-30.
- Ye, Q., Zhang, Z. & Law, R. (2009). Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches. *Expert Systems with Applications* 36:6527-35.

APPENDIX 1: SETUP FOR THE EXPERIMENTS

System configurations: Data processing and all the experiments related to this work are conducted on the Bigdata servers at Griffith University, which is enhanced by multiple Graphics Processing Unit (GPU) units GEFORCE GTX 1080 with 12GB of RAM and over 3,500 CUDA Cores each. GPU is considered as the heart of Deep Learning (DL), and DL is a variant of machine learning. It is a single-chip processor used for extensive graphical and mathematical computations which are necessary for large deep learning models along with large datasets. In our projects, GPUs are mandatory because we are developing and experimenting with large deep learning models and using large size datasets for our experiments. Besides, we use the GPU-accelerated Python library for deep learning like PyTorch and TensorFlow. A total of 5 powerful GPUs are installed in our server which provides fast GPU-enabled computation power in our deep learning-based training and testing experiments. The servers' configurations and GPU configurations are as follows.

Intel(R) Xeon(R) CPU E5-2609 v3 @ 1.90GHz, AMD Ryzen Thread ripper 1900X 8-Core Processor, 8234 MB Memory, a total of 5 GPUs (4 GEFORCE GTX 1080 Ti, 12GB memory each along with NVIDIA CUDA Cores 3584 and one GeForce GTX 1080 GPU with 2560 NVIDIA CUDA Cores).

Software requirements: Anaconda Distribution is a free, easy-to-install package manager, environment manager, and Python distribution with a collection of 1,500+ open source packages with free community support. Anaconda is platform-agnostic, so user or programmer can use it on Windows, macOS, or Linux operating systems. Using Anaconda we can easily deploy our projects into interactive data applications, live notebooks, and machine learning models with APIs. We can also manage our data science assets: notebooks, packages, environments, and projects in an integrated data science experience (source: Anaconda distribution document). An Anaconda virtual environment was created for the experiments with the following packages (Table 5).

Table 5: Packages used to create an Anaconda virtual environment.

Package Name	Version	Package Name	Version
Cudatoolkit	10.1.168	pytorch-pretrained-bert	0.6.2
ipykernel	5.1.2	qt	5.9.7
ipython	7.8.0	qtconsole	4.5.5
ipython_genutils	0.2.0	readline	7
ipywidgets	7.5.1	regex	2019.08.19
jupyter	1.0.0	scikit-learn	0.21.3
jupyter_client	5.3.4	scipy	1.3.1
jupyter_console	6.0.0	sentencepiece	0.1.83
jupyter_core	4.6.0	setuptools	41.4.0
nbconvert	5.6.0	six	1.12.0
notebook	6.0.1	sklearn	0
numpy	1.17.2	sqlite	3.30.0
numpy-base	1.17.2	terminado	0.8.2
pandas	0.25.2	testpath	0.4.2
pandoc	2.2.3.2	tk	8.6.8
pandocfilters	1.4.2	torchtext	0.4.0
pillow	6.2.0	torchvision	0.4.1
pip	19.3.1	tornado	6.0.3
pyqt	5.9.2	tqdm	4.36.1
python	3.7.4	traitlets	4.3.3
python-dateutil	2.8.0	transformers	2.1.1
pytorch	1.3.0	wheel	0.33.6

