OPEN ACCESS
Vietnam Journal of Computer Science
Vol. 10, No. 2 (2023) 117–134

DOI: 10.1142/S2196888822500269

World Scientific
www.worldscientific.com

Tweets Topic Classification and Sentiment Analysis Based on Transformer-Based Language Models*

Ranju Mandal
School of Information and Communication Technology
Griffith University
Brisbane, Australia
ranjumandal@gmail.com

Jinyan Chen† and Susanne Becken‡
Griffith Institute for Tourism
Griffith University, Brisbane, Australia
†*Jinyan.Chen@griffith.edu.au*
‡*S.Becken@griffith.edu.au*

Bela Stantic§
School of Information and Communication Technology
Griffith University
Brisbane, Australia
B.Stantic@griffith.edu.au

Received 14 December 2021
Accepted 7 May 2022
Published 2 September 2022

People provide information on their thoughts, perceptions, and activities through a wide range of channels, including social media. The wide acceptance of social media results in vast volume of valuable data, in variety of format as well as veracity. Analysis of such 'big data' allows organizations and analysts to make better and faster decisions. However, this data had to be quantified and information has to be extracted, which can be very challenging because of possible data ambiguity and complexity. To address information extraction, many analytic techniques, such as text mining, machine learning, predictive analytics, and diverse natural language processing, have been proposed in the literature. Recent advances in Natural

*This paper is an extended version of our ACIIDS 2021 paper "Empirical Study of Tweets Topic Classification Using Transformer-Based Language Models", and it is compiled on invitation from Professor Ngoc Thanh Nguyen. the paper is significantly extended both in length and content by providing more details of work related to topic classification and by extending and providing details of method and experimental results for sentiment analysis.

§Corresponding author.

Language Understanding-based techniques more specifically transformer-based architectures can solve sequence-to-sequence modeling tasks while handling long-range dependencies efficiently. In this work, we applied transformer-based sequence modeling on short texts' topic classification and sentiment analysis from user-posted tweets. Applicability of models is investigated on posts from the Great Barrier Reef tweet dataset and obtained findings are encouraging providing insight that can be valuable for researchers working on classification of large datasets as well as large number of target classes.

Keywords: Transformer; natural language processing; topic classification; target classification; deep learning.

1. Introduction

We witness a steady increase worldwide in a number of social media users, which is also reflected in an increase of number of social media posts. This in turn expands the scope of commercial activities by enabling the users to discuss, share, analyze, criticize, compare, appreciate and research about products, brands, and services. Social media platforms like Twitter, Facebook, Pinterest, Reddit, LinkedIn, Amazon Review, IMDB, and Yelp have become popular sources for retrieving public opinions and sentiment.[1] 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. Similarly, this pool of information can be explored for the mutual benefit of both the user and the organization. Identifying topics of discussions and analyzing sentiments of this extremely large corpus of opinions can help organizations in realizing the public opinion and user experiences of products or services.[2] There is also an increasing interest in using social media data for the monitoring of nature experience and environmental changes. Sometimes this is coupled with citizen science where common citizens are specifically encouraged to contribute data.[3] While further work is needed to assess how well citizen science, collective sensing and social media data integrate with professional monitoring systems[4] there are evidences that social media data is often only indirectly relevant to a particular research question, for example, how people perceive a natural phenomenon, where they go or what they do. However, by applying appropriate methods, it is possible to convert social media unstructured data into a useful information that provides insights into people's opinions and activities.[5] Using Twitter as a source of data, Daume and Galaz[6] concluded that Twitter conversations represent "embryonic citizen science communities". There are many other areas where Twitter demonstrated to be valuable source of data.[7,8]

Text classification is the process of assigning labels or categories to text according to its content. It is one of the fundamental tasks in Natural Language Processing (NLP) with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. Sentiment analysis has been studied to assess the polarities of user opinion in a particular context, and many publications can be found in the literature.[9–11] However, for social media platforms, such as Twitter, information filtering is a necessary step to select the relevant tweets as the analysis may

become overwhelmed by the raw data.[12] One solution to this problem is automatic classification of tweet text. Automatic detection of target or topic of tweets enables applications featuring opinion target/aspect extraction, polarity assessment as well as target/aspect-oriented sentiment analysis. Target-specific polarity detection is a key challenge in the field of sentiment analysis, as the sentiment polarity of words and phrases may depend on the aspect.[13] Other reasons for tweet classification include identifying trending topics.[14–16] Real-time detection of breaking news or relevant target and tracking them is essential due to the real-time nature of Twitter data.

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 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. In this work, we capitalize on social media data and state-of-the-art NLP approach and in a case study on the Great Barrier Reef (GBR), Australia, assess how social media data sources, namely, Twitter, can be used to better understand human opinions and experiences.

The motivation behind our work has many folds, all of these recent state-of-the-art Natural Language Understanding (NLU) models[17–20] can be applied on a variety of NLU tasks such as Sequence Classification, Token Classification, and Question Answering on a large dataset of English text corpus that contains a lot of tokens in each sequence. Therefore, at first, we investigate the recent NLU models to find out the best suitable and effective solution on short tweet text classification problem, such as found in social media posts. Our dataset has a limited number of sequences with a small number of tokens (i.e. most samples have less than 32 tokens and few samples contain more than 32 tokens but less than 64 tokens) in each sequence in contrast to the popular NLU dataset such as GLUE and RACE. A detailed description of data collection, preliminary data analysis, annotation procedure and the steps involved in the experimentation on social media data (Twitter) to producing results are presented in respective sections.

The proposed work for developing target classification method is based on deep learning techniques and sophisticated language modeling. To address this problem, transformer-based Bidirectional Encoder Representations from Transformers (BERT)[17] encoder model, and other three more improved BERT models[18–20] are used for feature extraction and classifications. We have opted for a transfer learning technique where instead of training a model from scratch, models used in experiments are pre-trained on a large dataset and then the specific (in this case GBR tweeter dataset) is used to fine-tune all the models for specific text classification task. During the investigation, a comparative study is pursued based on the model's

Fig. 1. Flow of target classification pipeline. The system takes a tweet text as input and returns a label from 11 predefined target categories.

performance on prediction accuracy and computational complexity. Figure 1 shows the diagram of our classification pipeline which takes a tweet text as input and produces a class label as output. The proposed approach effectively classifies the text to a predefined set of generic classes such as politics, travel, safety, culture, climate, terrestrial, coastal, etc. Similar approach is used for sentiment polarity calculation to calculate polarity on scale between −5 for highly negative and +5 for highly positive posts.

The rest of this paper is organized as follows. Section 2 provides a review of the relevant works of literature on text classification and sentiment analysis. In Sec. 3, we provide details of methodology along the transformer-based architectures used in our experiment, while in Sec. 4, we present experimental results and analyzes of the performance. Finally, conclusions are presented in Sec. 5.

2. Related Work

Significant attention in literature was devoted to opinion mining based on sentiment orientation to understand perceptions and characteristics of population or market groups and to determine the credibility of content and motivations for posting reviews. 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. Text summarization aims to transform lengthy documents into shortened versions using NLP, and text classification also uses NLP along with machine learning technologies to facilitate information processing and data analysis. 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.[13] Broadly speaking, three main categories, namely, machine learning-based approach, lexicon-based approach, and hybrid approach. The machine learning-based approach applies diverse machine learning algorithms and uses linguistic features. The lexicon-based approach relies on a sentiment lexicon, a collection of known and precompiled terms with polarity weights. The hybrid approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

Research articles on the problem of text classification can be found decades back,[12,21] and until recently it transformed[17] by a giant stride in terms of efficiently handling challenging not only the classification problem but of a wide variety of NLU tasks. Nigam *et al.*[21] proposed a text classification algorithm based on the combination of Expectation–Maximization and a naive Bayes classifier for learning from

labeled and unlabeled documents. Sriram *et al.*[12] proposed method classifies incoming tweets into multiple categories such as news, events, opinions, deals, and private messages based on the author's information and seven other binary features. Chen *et al.*[22] produced discriminatively features by leveraging topics at multiple granularities, which can model the short text more precisely. Vo and Ock[23] proposed topic models-based feature enhancing method which makes the short text seem less sparse and more topic-oriented for classification. The topic model analysis was based on latent Dirichlet allocation, and finally, features were enhanced by combining external texts from topic models that make documents more effective for classification, and a large-scale data collection consisting of Wikipedia, LNCS, and DBLP were explored for discovering latent topics. In another method, a cluster-based representation enrichment strategy was adopted to obtain a low-dimensional representation of each text for short text classification.[24]

Recently proposed methods, such as BERT,[17] universal language model fine-tuning (ULMFiT),[25] GPT-2,[26] and ELMo,[27] attain a significant milestone on the NLP domain, outperforming the state of the art on several NLP tasks including text classification. ULMFiT[25] is an effective transfer learning method that can be applied to any task in NLP. GPT-2[26] is a large transformer-based language model with 1.5 billion parameters, trained on a large dataset of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The introduction of a new dataset which emphasizes the diversity of content, and has been curated/filtered manually is also a significant contribution. ELMo[27] is a deep contextualized word-embeddings technique where representation for each word depends on the entire context of the text unlike using a fixed embedding for each word. A bidirectional LSTM model is trained to be able to create those embeddings to accomplish a specific task. These methods also have revolutionized the field of transfer learning in NLP by using language modeling during pre-training. Further enhancement of Google BERT[17] and many improved versions of BERT inspired transformer-based sequence modeling techniques such as XLNet,[19] RoBERTa,[20] ALBERT[18] significantly improved on the variety of tasks in NLU. Recent works on BERT performance enhancement are achieved by increasing training data, computation power, or training procedure. Further advancement that can enhance performance while using fewer data and compute resources will be a step forward.

2.1. *BERT*

It stands for Bidirectional Encoder Representations from Transformers[28] pre-trained over a large volume of unlabeled textual data to learn a language representation that can be used to fine-tune for specific machine learning tasks. The general transformer represents an encoder–decoder network architecture, but BERT is a pre-training model, that uses the encoder to learn a latent representation of the input text. It builds upon recent work in pre-training contextual

representations such as GPT,[26] Elmo,[27] and ULMFit[25] (these three methods were significant milestones before the BERT method had been introduced). 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 unlabeled text by jointly conditioning on both the left and right contexts. It is considered a key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling. 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. This bidirectional Transformer model redefines the state of the art for a range of NLP tasks (e.g. text classification), even surpassing human performance in the challenging area.[17]

2.2. *XLNet*

XLNet is a large bidirectional transformer that uses improved training methodology, larger data and more computational power to achieve better than BERT prediction metrics on 20 language tasks.[19] To improve the training, XLNet introduces permutation language modeling, where all tokens are predicted but in random order. This is in contrast to BERT's masked language model where only the masked (15%) tokens are predicted. This is also in contrast to the traditional language models, where all tokens were predicted in sequential order instead of random order. This helps the model to better handle dependencies and relations between words. Transformer XL was used as the base architecture, which showed good performance even in the absence of permutation-based training. XLNet was trained with over 130 GB of textual data and 512 TPU chips running for 2.5 days.

2.3. *RoBERTa*

RoBERTa modifies the BERT[17] pre-training procedure that improves end-task performance and these improvements were aggregated and evaluated to obtain combined impact.[20] The authors coined this modified configuration as RoBERTa for robustly optimized BERT approach. Few research works have been published as an improvement of BERT method performance on either its prediction accuracy or computational speed. RoBERTa improves the BERT performance on prediction accuracy. It is a replication study of BERT pre-training with a robust and improved training methodology training as BERT was significantly undertrained according to

the authors. The modifications include hyperparameter tuning such as bigger batch sizes (8k) with 10 times more data compare to BERT training, longer training sequences, dynamic masking strategy during training step, etc. RoBERTa's experimental dataset consists of 160 GB of text for pre-training, including 16 GB of books corpus and English Wikipedia used in BERT. The additional data included CommonCrawl News dataset (63 million articles, 76 GB), Web text corpus (38 GB), and Stories from Common Crawl (31 GB). RoBERTa outperforms both BERT[17] and XLNet[19] on the GLUE benchmark dataset.

2.4. *ALBERT*

ALBERT stands for A Lite BERT for self-supervised learning of language representations.[18] This method outperformed BERT on both metrics, its prediction accuracy, and computational complexity. The performance of the ALBERT language model superior to BERT on memory optimization and training time by parameter-reduction techniques and it also achieved state-of-the-art results on the three most popular NLP datasets (i.e. GLUE, RACE, and SQuAD 2.0). Recent state-of-the-art language models have hundreds of millions of parameters which make these model less convenient on low or limited hardware (i.e. low GPUs or TPUs computing) resource environments. Besides, researchers also have investigated that stacking more layers in the BERT-large model can lead to a negative impact on overall performance. These obstacles motivated Google to thorough analysis into parameter reduction techniques that could reduce the size of models while not affecting their performance. ALBERT is an attempt to scale down the BERT model by reducing parameters. For example, the ALBERT-large model has about 18 million parameters which are 18× fewer compared to BERT-large (334 million).

3. Methodology

For the task of analyzing short text messages such as found in Twitter posts, we propose a deep learning techniques and sophisticated language modeling algorithms. In the first stage, all downloaded tweets are pre-processed, after they are processed via a language model for feature representation. Finally, a classifier is employed for classification. We implemented and validated a workflow for two independent systems:

(1) Sentiment analysis with intensity (i.e. polarity score between −5 and +5 range).
(2) Target or topic-level classification for aspect or topic-level sentiment analysis.

3.1. *Dataset*

The research team drew on earlier work on data collections and relied on public Twitter application programming interface with filtering to capture geo-tagged

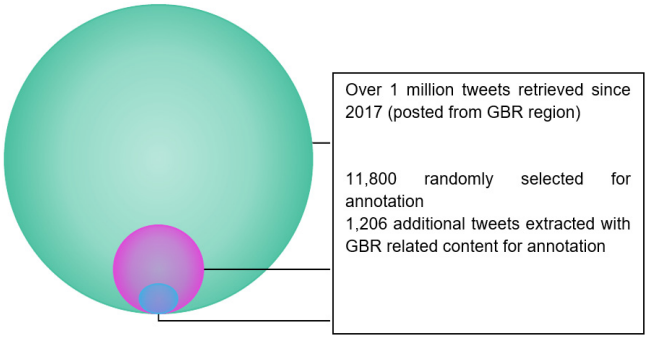


Fig. 2. Data scope for tweets repository.

tweets posted from the GBR geographic area.[29] This area is represented as a rectangular bounding box with coordinates: 141.459961, −25.582085 and 153.544922, −10.69867, which broadly represents the GBR region. Data are stored in a NoSQL MongoDB database located on a Big Data cluster at the Griffith University. The scope for data repository is shown in Fig. 2, over one million tweets were collected and 11,800 were randomly selected for annotation. Since the first set of randomly selected 11,800 tweets mainly contained tweets that fell into the category 'other', to diversify the topics it was important to extract additional tweets that specifically addressed issues directly related to the GBR. By using GBR-related keywords, as shown in Fig. 3, as a result, additional 1206 tweets were extracted from the database and added to annotation set.

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. More specifically, the following steps were put in place:

- Agree on a preliminary set of targets (Fig. 3) that reflect the general coverage of themes prevalent in tweets.
- 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 13,006 tweets, whereby a minimum of 2000 tweets were allocated per researcher.

More details on the identified targets are shown in Fig. 3, 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.

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

Fig. 3. Targets for human annotation.

3.3. *Pre-processing and cleaning of data*

Twitter data is often messy and contains a lot of redundant information which is not contributing to the assessment. Also, there are several other steps that need to be put in place to make subsequent analysis easier. Initial data cleaning involved the following aspects:

(1) Removing Twitter Handles (@user): The Twitter handles do not contain any useful information for target classification or sentiment polarity calculation.
(2) The punctuations, numbers, and even special characters are removed since they typically do not contribute to differentiating tweets.
(3) 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], Output: [sunrise, was, amazing]. The respective models' tokenizers (e.g. BertTokenizer, XLNetTokenizer, RobertaTokenizer, AlbertTokenizer) are used in this preprocessing step.
(4) Stemming: It is a rule-based process of stripping the suffixes such as 'ing', 'ly', 'es', and 's' from a word. The objective of this process is to reduce the total number of unique tokens in our data without losing amount of information.

3.4. *Text embedding*

To represent the input text for the model better, there is a need to add a specific set of rules. Many of these rules are designer's choice to make the model work better. The

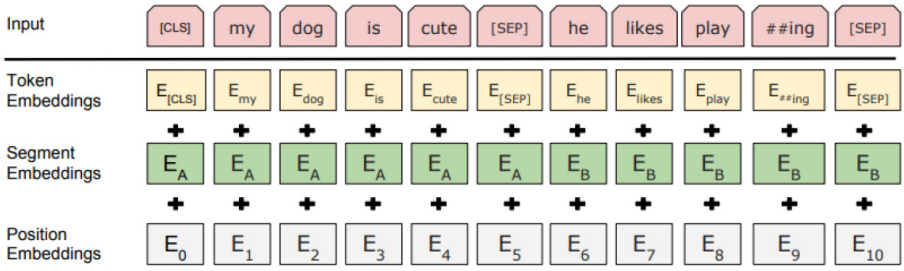


Fig. 4. Three stages of text input embeddings.

input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings, as shown in Fig. 4.[17]

Token Embeddings: These are the embeddings learned for the specific token from the WordPiece token vocabulary.[30] The [CLS] token added at the beginning and [SEP] tokens in the right place.

Segment Embeddings: Model can also take sentence pairs as inputs for tasks (Question Answering). That is why it learns a unique embedding for the first and the second sentence to help the model distinguish between them. In Fig. 4, all the tokens marked as EA belong to sentence A and similarly, EB belongs to sentence B.

Position Embeddings: Model learns and uses positional embeddings to express the position of words in a sentence. This helps to overcome the limitation of Transformer which, unlike an recurrent neural network, is not able to capture 'sequence' or 'order' information.

In the transfer learning technique, the pre-trained weights of an already trained model on a large dataset can be used to fine-tune the model for a specific NLP tasks. We adopted a transfer learning strategy to leverage from pre-trained models that use language models pre-trained on exceptionally large curated datasets and the models have demonstrated state-of-the-art performance in text classification tasks. The pre-trained model weights already have enormous encoded information of English language, and it takes much less time to fine-tuned the model with the new dataset to obtain the features required for classification. Semantic understanding of language has been significantly improved by the language modeling approaches.

3.5. *Implementation details*

All experiments have been conducted on Linux cluster with nodes having Intel(R) Xeon(R) CPU E5-2609 v3 @ 1.90 GHz and a GeForce GTX 1080 GPU installed. The PyTorch-based Transformers library[a] was used to develop code for all the experiments. The models are trained with the AdamW optimizer,[b] learning rate of $4e-5$,

[a] https://huggingface.co/.
[b] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html.

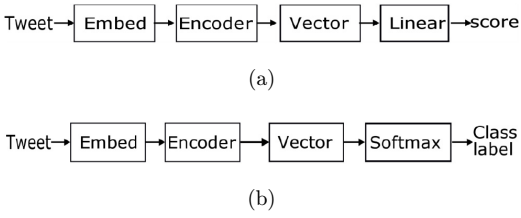
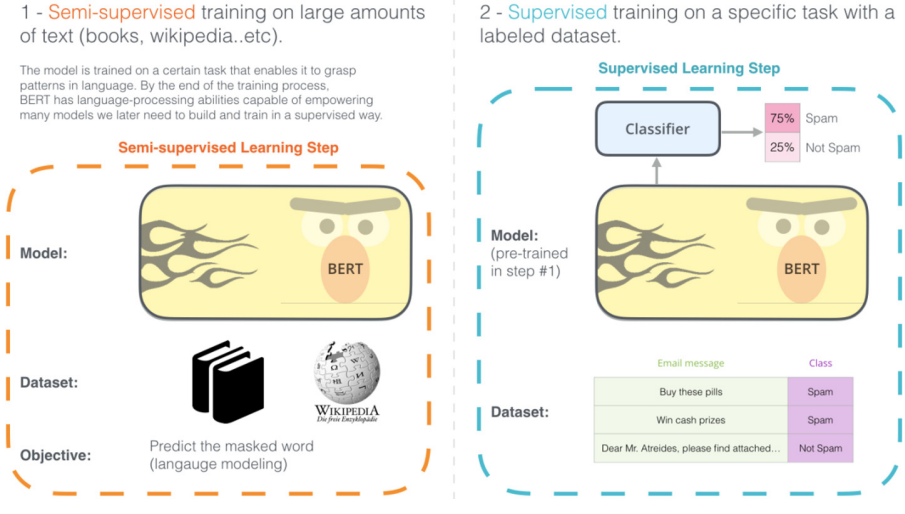


Fig. 5. Context diagrams of sentiment assessment pipeline, showing that: (a) our sentiment analysis system uses a tweet text as input to produce a score (−5 to +5). (b) Our system takes a tweet text as input and produce a class label (i.e. positive or negative).

and token size of 32 or 64. We train each model for 2–5 epochs with a batch size of 32. The results for each of the models are reported in Tables 2 and 3.

3.6. *Training and evaluating the sentiment analysis systems*

The proposed regression model is designed to generate a score for each tweet between −5 and + 5. The language modeling-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 Fig. 5, the pipeline diagrams are shown, whereby 'Embed' represents embeddings and the 'Encoder' stands for the language model.



Source: http://jalammar.github.io/illustrated-bert/.

Fig. 6. The two important steps of BERT NLP framework (semi-supervised and supervised training). 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.

3.7. *Language modeling*

We have adapted BERT model,[28] 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 unlabeled 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 modeling.

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 (Fig. 6). This bidirectional Transformer model that redefines the state of the art for a range of NLP tasks (e.g. Question Answering), even surpassing human performance in the challenging area.

Some details related to Trained Transformer Encoder stack and the transformer encoder and the transformer are shown in Fig. 6. There are two variants 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) 12 for the Base version and 24 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). 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. We built upon Transformer model which was proposed in the paper 'Attention is All You Need'.[28]

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. The encoder's inputs first flow through a self-attention layer, which is 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.

4. Results and Discussions

In order to perform machine learning-based sentiment analysis, we needed to manually annotate tweets. 2285 tweets have been annotated with a Negative sentiment (-5 to -1), 3360 were recognized as a Positive sentiment ($+1$ to $+5$), and 7061 tweets were annotated as Neutral (0). The findings are shown in Fig. 7, showing a relatively even distribution with a slight bias towards positive tweets.

Table 1 shows the number of tweets and percentages for each of the 11 target classes present in our dataset. It can be observed that 70.46% of tweets are labeled as 'other' category. From the distribution of targets within the sample of tweets it can be seen that the vast majority of tweets talk about topics that were not relevant to intended annotation related 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. We created a smaller dataset with 4342 tweet samples where the 'other' category samples are randomly reduced to only 500 posts to have a balance in topic numbers, and we refer that dataset as a 'balanced' dataset. The experimental results of this two different datasets are shown in Tables 2 and 3.

Finally, the results were assessed to determine the sentiment evident in tweets by target. Figure 8 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.

The classification results on GBR full and balanced datasets from a range of experiments along with the hyperparameters are detailed in Tables 2 and 3. We present our test's accuracy using $F1$ score or F-measure and the Matthews correlation coefficient (MCC) score. $F1$ score is based on precision (p) and the recall (r)

$$F1 = \frac{2 * p * r}{p + r}.$$

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%

Fig. 7. Annotation sentiment scores with number of tweets for each score.

Table 1. Target classes with the number of annotated tweets for each class.

Target class	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

The precision p is the ratio TP/(TP + FP) where TP is the number of true positives and FP is the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. While the recall r is the ratio TP/(TP + FN) where FN is the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

The values of MCC are calculated from the confusion matrix as follows (TN is true negative):

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

In our experiments, based on empirical evaluation, we have found that of split data into 85% for training and 15% for testing is performing the best. We have considered seven different models with a parameter range of 11 million (i.e. "albert-base-v1") to 340 million (i.e. "bert-large"). In Table 2, the best MCC of 0.549 and $F1$ score of 0.788 are obtained from the "albert-base-v1" model after two epochs. In Table 3, which represents the result from a smaller balanced dataset, the "bert-large" model obtained the best MCC score of 0.660 and $F1$ score of 0.705 after four epochs.

Table 2. Results obtained from the experiment on the GBR full dataset. This dataset contains 13006 tweet samples (15% test and 85% train) and 11 classes.

Arch.	Model	Param.	Seq.	Epoch	MCC	$F1$
BERT	bert-base	110M	32	2	0.555	0.786
BERT	bert-base	110M	64	2	0.545	0.780
BERT	bert-large	340M	32	3	0.561	0.781
XLNet	xlnet-base	110M	32	3	0.555	0.777
RoBERTa	roberta-base	125M	32	2	0.548	0.774
ALBERT	albert-base-v1	11M	32	2	0.549	0.788
ALBERT	albert-base-v2	11M	32	3	0.530	0.768

Table 3. Results obtained from the experiments on balanced GBR dataset. This dataset contains 4342 tweet samples (3690 and 652 for train and test, respectively) and 11 classes.

Arch.	Model	Param.	Seq.	Epoch	MCC	*F*1
BERT	bert-base	110M	32	3	0.634	0.682
BERT	bert-base	110M	64	3	0.633	0.680
BERT	bert-large	340M	32	4	0.660	0.705
XLNet	xlnet-base	110M	32	5	0.628	0.676
RoBERTa	roberta-base	125M	32	5	0.646	0.693
ALBERT	albert-base-v1	11M	32	4	0.587	0.642
ALBERT	albert-base-v2	11M	32	5	0.630	0.679

However, model "bert-large" has 31 times more parameter compared to "albert-base-v1". We adapted the sliding window-based approach for long sequences during the evaluation step. It is an experimental feature moves a sliding window over each sequence and generates sub-sequences with length equal to the "max-seq-length" parameter. The model output for each sub-sequence is averaged into a single vector before the classification step. We obtained a 0.01 accuracy boost using this method. Also, we observed from our experiments that setting Max Sequence length to 64 tokens does not improve performance as we use a sliding window-based evaluation method. The full dataset achieved overall the higher accuracy, which is due to the

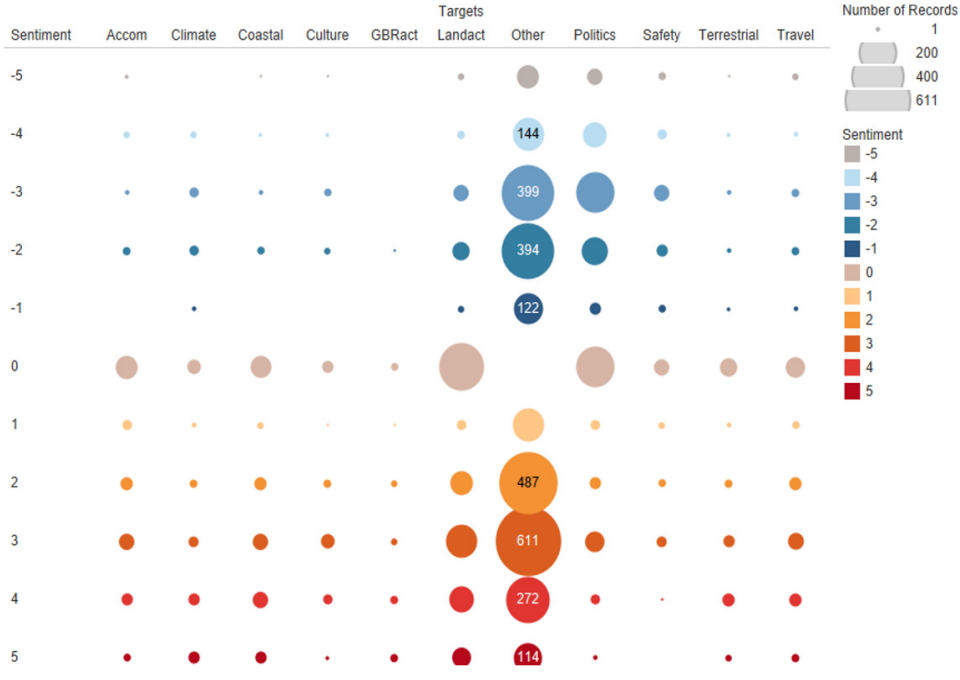


Fig. 8. Annotation results for sentiment by targets.

bigger training set and higher accuracy of 'other' target class. However, balanced dataset was introduced to balance number of targets and achieved better accuracy for targets with smaller sample sizes.

5. Conclusion

In this work, we developed a multi-class tweet (i.e. short text) classification system to identify the pre-defined topic and sentiment polarity of a tweet and as case study we used tweets sent from the GBR region. The importance of language modeling on text classification systems is studied in detail from the literature, and state-of-the-art transformer-based sequence modeling architectures are adapted to for the modeling of our GBR tweet text classification. A wide range of experiments and comprehensive analysis of performance is presented on short text sequence classification. A number of parameters have been varied and the findings can be valuable for researchers working on classification with large datasets and a large number of target classes. We concluded that the proposed system is suitable for target classification and sentiment polarity identification of short text messages such as found in social media Twitter posts.

Acknowledgments

This research is supported by the National Environmental Science Programme (NESP) — Tropical Water Quality Hub: Project 5.5.

References

1. J. Chen, S. Becken and B. Stantic, Harnessing social media to understand tourist mobility: The role of information technology and big data, *Tour. Rev.* (2021), ahead of print, doi: 10.1108/TR-02-2021-0090.
2. A. Kumar and A. Jaiswal, Systematic literature review of sentiment analysis on twitter using soft computing techniques, *Concurrency Comput. Pract. Exp.* **32**(1) (2019) e5107.
3. L. Lodi and R. Tardin, Citizen science contributes to the understanding of the occurrence and distribution of cetaceans in southeastern Brazil — A case study, *Ocean Coast. Manage.* **158** (2018) 45–55.
4. S. Becken, R. M. Connolly, J. Chen and B. Stantic, A hybrid is born: Integrating collective sensing, citizen science and professional monitoring of the environment, *Ecol. Inform.* **52** (2019) 35–45.
5. S. Becken, B. Stantic, J. Chen, A. Alaei and R. M. Connolly, Monitoring the environment and human sentiment on the Great Barrier Reef: Assessing the potential of collective sensing, *J. Environ. Manage.* **203** (2017) 87–97.
6. S. Daume and V. Galaz, "Anyone know what species this is?" — Twitter conversations as embryonic citizen science communities, *PLOS One* **11** (2016) 1–25.
7. C. Prentice, J. Chen and B. Stantic, Timed intervention in COVID-19 and panic buying, *J. Retail. Consum. Serv.* **57** (2019) 102203.
8. S. Becken, B. Stantic, J. Chen and R. M. Connolly, Twitter conversations reveal issue salience of aviation in the broader context of climate change, *J. Air Transp. Manag.* **98** (2021) 102157.

9. C. J. Hutto and E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in *Proc. 8th Int. AAAI Conf. Weblogs and Social Media* (AAAI Press, 2014), pp. 216–225.
10. F. N. Ribeiro, M. Araújo, P. Gonçalves, F. Benevenuto and M. A. Gonçalves, A benchmark comparison of state-of-the-practice sentiment analysis methods, preprint (2015), arXiv:abs/1512.01818.
11. J. Chen, S. Becken and B. Stantic, Lexicon based Chinese language sentiment analysis method, *Comput. Sci. Inf. Syst.* **16**(2) (2019) 639–655.
12. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu and M. Demirbas, Short text classification in twitter to improve information filtering, in *Proc. 33rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (Association for Computing Machinery, New York, 2010), pp. 841–842.
13. A. R. Alaei, S. Becken and B. Stantic, Sentiment analysis in tourism: Capitalizing on big data, *J. Travel Res.* **58**(2) (2019) 175–191.
14. K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A. Choudhary, Twitter trending topic classification, in *2011 IEEE 11th Int. Conf. Data Mining Workshops* (IEEE, 2011), pp. 251–258.
15. J. Allan, Introduction to topic detection and tracking, *Inf. Retr. Ser.* **12** (2012) 1–16.
16. A. E. Yüksel, Y. A. Türkmen, A. Özgür and B. Altınel, Turkish tweet classification with transformer encoder, in *Proc. Int. Conf. Recent Advances in Natural Language Processing (RANLP 2019)* (INCOMA Ltd., 2019), pp. 1380–1387.
17. J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, preprint (2018), arXiv:abs/1810.04805.
18. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, Albert: A lite bert for self-supervised learning of language representations, preprint (2019), arXiv:1909.11942.
19. Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov and Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, preprint (2019), arXiv:abs/1906.08237.
20. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, preprint (2019), arXiv:abs/1907.11692.
21. K. Nigam, A. K. Mccallum and S. Thrun, Text classification from labeled and unlabeled documents using EM, *Mach. Learn.* **39**(2) (2000) 103–134.
22. M. Chen, X. Jin and D. Shen, Short text classification improved by learning multi-granularity topics, in *Proc. Twenty-Second Int. Joint Conf. Artificial Intelligence (IJCAI)* (AAAI Press, 2011), pp. 1776–1781.
23. D.-T. Vo and C.-Y. Ock, Learning to classify short text from scientific documents using topic models with various types of knowledge, *Expert Syst. Appl.* **42** (2015) 1684–1698.
24. Z. Dai, A. Sun, X.-Y. Liu, J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu, Crest: Cluster-based representation enrichment for short text classification, in *Advances in Knowledge Discovery and Data Mining* (Springer, Berlin, 2013), pp. 256–267.
25. J. Howard and S. Ruder, Universal language model fine-tuning for text classification, preprint (2018), arXiv:abs/1801.06146.
26. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, Language models are unsupervised multitask learners, Technical Report, OpenAi http://www.persagen.com/files/misc/radford2019language.pdf, pp. 1–24.
27. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualized word representations, in *Proc. NAACL* (Association for Computational Linguistics, 2018), pp. 2227–2237.

28. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, preprint (2017), arXiv:abs/1706.03762.
29. R. Mandal, J. Chen, S. Becken and B. Stantic, Empirical study of tweets topic classification using transformer-based language models, in *Asian Conf. Intelligent Information and Database Systems - ACIIDS 2021* (Springer, Cham, 2021), pp. 340–350.
30. Y. Wu *et al.*, Google's neural machine translation system: Bridging the gap between human and machine translation, preprint (2016), arXiv:1609.08144.