Computer Science and Information Systems 16(2):639655 https://doi.org/10.2298/CSIS181015013C

Lexicon Based Chinese Language Sentiment Analysis Method

Jinyan Chen[1], Susanne Becken[1], and Bela Stantic[2]

[1] Griffith Institute for Tourism
Griffith University, Queensland, Australia
{Jinyan.Chen@griffithuni.edu.au
[2] Griffith Institute for Tourism
Griffith University, Queensland, Australia
s.becken@griffith.edu.au
[3] School of Information and Communication Technology
Griffith University, Queensland, Australia
B.Stantic@griffith.edu.au

Abstract. The growing number of social media users and the vast volume of posts could provide valuable information about the sentiment toward different locations, services as well as people. Recent advances in Big Data analytics and natural language processing often means to automatically calculate sentiment in these posts. Sentiment analysis is a challenging and computationally demanding task due to the volume of data, misspelling, emoticons as well as abbreviations. While significant work was directed toward the sentiment analysis of English text there is limited attention in the literature toward the sentiment analytic of the Chinese language. In this work, we propose a method to identify the sentiment in Chinese social media posts, and to test our method we rely on posts sent by visitors of the Great Barrier Reef by users of the most popular Chinese social media platform Sina Weibo. We elaborate on the process of capturing of Weibo posts, describe the creation of lexicon as well as develop and explain the algorithm for sentiment calculation. In the case study, related to sentiment toward the different GBR destinations, we demonstrate that the proposed method is effective in obtaining the information and is suitable to monitor visitors' opinion.

Keywords: Sentiment Analysis, Social Media, Natural Language Processing.

1. Introduction

Recent advances in computer science, technology, and communication, in combination with advanced equipment and services, reinvented the channels through which people collect and generate information. The fundamental concept of generating data has changed. In the past, a small number of main sources have been generating data and all other actors have been consuming data. However, today all of us are both generating data and we are also consumers of this shared data. This is particularly evident in relation to social media platforms, which are attracting more and more users who talk about diverse topics. Despite concerns, related to the privacy and credibility of the posted information, this new method of obtaining relatively independent information about the quality of a product or service has benefits as it is limiting the ability of businesses to control and influence

customers. Feedback about a wide range of services and products can now be acquired from independent reviews by consumers themselves. These types of reviews are known as user-generated content, and they have been found to play an important role in customers' future behaviors due to the electronic word-of-mouth [38].

The growing role of social media is attracting increasing research interest. Social media plays a significant role in many aspects of life including retails, politics, tourism, and decision-making behavior. A large number of users actively engage with social media, for example, Twitter has 313 million monthly active users worldwide [27]. Furthermore, over 1.94 billion people monthly use Facebook [39] and 700 Million users use Instagram [15], and the Chinese Social Media platform Sina Weibo has over 400 million active users [11]]. People use social media to post stories from their daily life, and they are particularly likely to share impressions or emotions related to their travel experience. For the tourism industry, it is, therefore, possible to examine social media and understand what visitors share and how they perceive destinations, attractions, and products.

The Great Barrier Reef (GBR) is the world's largest coral reef system stretching over 2,600 kilometers along the coast of Queensland and it attracts over two million visitors each year from all around the world [12]. A significant proportion of GBR visitors are from China. While Chinese visitors are very active users of social media they typically use Chinese platforms such as Sina Weibo rather than those commonly used by other English-speaking visitors. To take advantage of this huge number of users and media posts on Weibo in this work we decided to capture those posts which specifically talk about the GBR.

Advances in Big Data analytic and natural language processing provide the opportunity to analyse visitor experience and their sentiment toward certain services or places [26]. Sentiment analysis is a method that can be used for analyzing social media content. It basically converts social media post text into quantitative data, whereby it can extract information about special events and identify patterns. Sentiment analysis is a challenging and computationally demanding task not just because of the vast volume of data but also due to the common grammatical errors and misspelling, slang, and abbreviations. Additionally, social media posts can contain sentiment emoticons that carry valuable information for calculating the sentiment scores and should not be discarded.

Social media posts have been harnessed for different purposes including monitoring environmental changes [4], [5] as well as sentiment analysis in tourism [19], [22], [37]. In most cases concept relied on sentiment derived from the short text of social media posts and analytics of posts' metadata along with other available online or scientific data. Different statistical and machine learning methods have been used, such as Support Vector Machines (SVM) and Naive Bayes. Recent literature also addresses the issues with regard to trust and reputation measures in social network systems, especially in presence of thematic social groups [10].

There are many sentiment analysis methods for English text presented in literature such as the Valence Aware Dictionary for Sentiment Reasoning (VADER) approach, which is purposely developed for sentiment analysis of short text found in social media posts [14]. VADER relies on a dictionary but also has a set of rules, which takes into consideration punctuation, emoticons, and many other heuristics to compute sentiment polarity. Dictionary contains items with associated sentiment intensities that are annotated by humans. For instance, a dictionary may contain words, such as excellent, better,

good, bad, worse, terrible, with their respective sentiment intensity and polarity. While the sentiment analysis is matured for English language there is limited work for Chinese language sentiment analysis and was mostly directed toward lexicon creation.

The majority of proposed sentiment analysis methods for Chinese language are machine learning based [41], for example, Xu and colleagues looked into to classify sentiments of microblogs from Sina Weibo. They relied on labeled emoticons as a training corpus and built a fast Bayesian classifier based on assumption that both smiley and emoticons are strongly related to typical sentiment words and are viewed as convincing indicators of emotions [36]. Authors of [21] claimed that they improved sentiment analysis by identify- ing features with SVM as a learning engine, which they named the global optimization-based sentiment analysis approach. However, the authors pointed out that if the parameters are not selected well the result will not be accurate and that the sentiment feature subset choice influences appropriate kernel parameters, and vice versa. On the other hand authors of [41] relied on lexicon from the National Taiwan University Sentiment Dictionary (NTUSD) [32], which has both traditional Chinese and simplified Chinese characters. NTUSD contains 2810 positive words (which are assigned +1 as sentiment intensity) and 8276 negative words (assigned -1 as sentiment intensity). Also [33] adopted NTUSD lexicon to calculate a sentiment score to monitor the public opinion and forecast election [6], however, they also rely on lexicon which has a limited number of words and additionally has only two levels of sentiment intensity +1 or -1, associated with positive and negative words.

To overcome the above-mentioned shortcomings of existing approaches, in this work we propose and test method to identify sentiment in Chinese social media posts and we rely on the most popular Chinese social media Sina Weibo [25]. We developed a method to crawl the web and collect Weibo posts that mentioned specific words (in this case word ”Great Barrier Reef” in Chinese language). We elaborate the process of capturing, managing, we describe the creation of a comprehensive Chinese lexicon with sentiment intensity and we also propose an algorithm for sentiment calculation, which takes into consideration the length of the post as well as the number of matching words with a lexicon. Additionally, as proof of concept, we provide a sample of additional information, which can be extracted from the social media post metadata; such as where from Chinese people who have interest and comment on GBR originate from as well as what is average sentiment depending on province poster originate from.

2. Background

Social media has influenced the way people search for information and make decisions. Tourism organizations and destination marketing organizations are aware of ongoing trends and thus try to explore the opportunities to use tourist-generated content for their own marketing and include it as an integral part of their branding and positioning.

2.1. Social Media Data

The emergence of user-generated content (UGC) on the Internet in the mid-2000s has provided a new source of data and enabled millions of tourists to exchange content on popular platforms for mutual benefits, such as social networking (e.g., Facebook), micro-blog

(e.g., Weibo, Twitter), multimedia sharing (e.g. YouTube), location sharing (FourSquare), and review forums (TripAdvisor).

Social media data has many forms, Figures 1 and 2 show several samples only to demonstrate the diversity, while below we list and highlight the main characteristics of social media data:

1. It generates a massive volume of data.
2. Data are generated at unprecedented speed.
3. Data are complex and high-dimensional, attribute-value data such as text, comments, and other metadata about the poster, Figure 1.
4. Data exist in different forms: text, emotions, images, videos, etc. See Figure 2
5. Additionally, data can be noisy as well as incomplete and contain a lot of misspelling, slang, and abbreviations.


```
"_id" : ObjectId("5b337062af543679e3c87d03"),
"reposts_count" : 0,
"text" : ": 六月马拉松 黄金海岸在等你
"geo_enabled" : 1,
"geo" : {
        "coordinate" : "-35.008038,138.571808",
        "location" : $$$
},
"id" : "2202125923M_G8ylSwhnP",
"likes_count" : 0,
"imgurl_list" : [
        "$$$"
],
"created_at" : "2018-03-22 15:01",
"userid" : NumberLong("2202125923"),
"mid" : "$$",
"terminal" : "none",
"source" : "$$",
"comments_count" : 3,
"userinfo" : {
        "gender" : "女",
        "region" : "海外 澳大利亚",
        "name" : " $$$",
        "birthdate" : "01-01"
}
```


Fig. 1. Sample Weibo data indicates complex high-dimensionality of data

Social media platforms including Twitter, Flickr, and Sina Weibo also offer the possibility of geo-referencing the content shared by users. This enables the opportunity to trace social media users and compare findings with other sources of data.

Sina Weibo is a Chinese micrologic website launched by Sina Corporation in 2009, it is similar to Twitter but from 2016 with some flexibility with regard to the length of the post [3]. Sina Weibo is the first and the most popular Chinese social media platform [17], it has more than 411 million monthly active users in the first quarter of 2018 [28].

Fig. 2. Several samples of data formats.

2.2. Sentiment Analysis

Sentiment analysis or opinion mining relates to the study of opinions, attitudes, and emotions toward entities, individuals, issues, or events. Sentiment analysis dates back to the 1970s and is conceptually grounded in the work of Osgood and his associates on content analysis of people's judgments by evaluating text [24]. Osgood and colleagues distinguish between three dimensions of meaning: *Evaluation*, *Potency*, and *Activity*. Assessment good-bad or favorable-unfavorable are along the *Evaluation* dimension of meaning, while the intensity of these evaluations such as strong/weak or powerful/powerless represents *Potency*, and fast/slow or active/passive comprise the *Activity* dimension.

Today we talk about sentiment analysis as a process of computationally identifying and categorizing opinions to determine the writer's attitude toward a particular issue. This can be achieved by employing computer-based natural language processing which aims to detect sentiment by relying on an Artificial Intelligence system that would be able to reason about emotion [18].

Comprehensive sentiment analysis also considers metadata such as, who provided the information and at what time. Literature elaborates methods of sentiment analysis which in general falls into one of three categories [1]:

– **Machine learning**: Machine learning approaches involve creating a model by using human-annotated data. Mostly supervised machine learning approaches have been mentioned in literature and these methods are composed of preprocessing, feature extraction, learning, and classification steps.

- **Dictionary-based**: Also referred to as Lexicon based, as there is associated polarity and weight on each word in the dictionary. These systems mainly rely on the use of dictionaries containing comprehensive sentiment lexicons and sets of fine-tuned rules. A sentiment dictionary can be created either by humans (manually), by machine (automatically), or by humans and machine (semi-automatically).
- **Hybrid approaches**: These methods combine dictionary-based and machine learning-based approaches and they work in parallel to compute sentiment polarities.

As mentioned earlier, the typical representative of Dictionary-based methods is VADER, it is suited for sentiment analysis of short text [14]. A clear advantage of dictionary-based methods for sentiment analysis is that there is no need for annotation of the text for training. Another advantage is the possibility to create domain-specific dictionaries which ensure higher accuracy of calculated sentiment scores. But there is still a need to create a dictionary and annotate the polarity of words, however, it needs to be done only once. This is obviously less costly considering that the annotation for a supervised machine learning method needs to be undertaken for each new context.

In hybrid approaches both dictionary and machine learning are used to independently compute sentiment and then individual results are combined to provide sentiment intensity and polarity [16]. A specific hybrid model which is using keywords and the Naive Bayes algorithm has been recommended to calculate sentiment polarities of social media tweets [8].

3. Methodology

A sentiment analysis process is composed of several independent steps as can be seen in Figure 3. It starts with data collections which in the case of social media data most often rely on the utilization of a dedicated Application Programming Interface (API). In cases when purpose-built API is not available web crawling is needed. Crawling initially involves the identification of a data source, for example, a commercial website or a social media network. A dedicated web crawling code needs to be developed and used to collect data from these sources. Considering that a huge amount of posts have been generated by users there is a need to filter and acquire only relevant posts, most often filtering is related to particular geographic areas or particular keywords. After the initial cleaning, which discards data that do not contain full records, data are saved in the database.

Once data have been stored in a database, it is possible, for example, to calculate the sentiment from the content of the text within a set of data fields in particular posts. In addition to the actual text of the post additional metadata fields, if publicly available can also be captured. These fields can be used to analyze locations where from posts have been sent or where users originate from, which is based on place users indicated as their location when they created accounts.

Due to the limited options with public Weibo API, we developed in Python programming language dedicated method to crawl the Weibo website and to collect relevant social media posts. Considering that data is in Java Script Object Notation (JSON) format we stored it locally in MongoDB NoSQL database, which was chosen because it has been shown that the relational databases are struggling in handling a large amount of unstructured data [26]. NoSQL database is able to store and efficiently access a diversity of unstructured data including text, emotions, and media files.

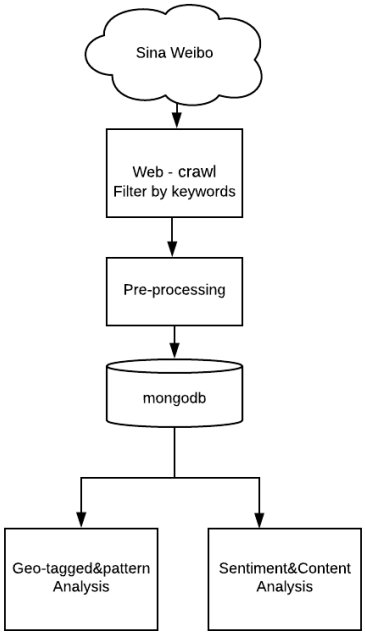

Fig. 3. Global Sentiment analysis process of Weibo data

The following procedure was implemented: we extracted only posts that contained the keywords ”Great Barrier Reef”, in Chinese language. The extraction process is shown in Figure 3. We identified that about 15% of posts have exact geolocation (Latitude, Longitude), which can help in the analysis of metadata and for example provide information about visited places. In addition, the analysis involves assessment of the ’emotional’ tone of the text; which will be calculated with the method proposed in this paper.

Our experiments were conducted on an in-house Big Data cluster running Hadoop (2.6.0) and MongoDB (3.2.9).

3.1. Creation of Chinese Lexicon

Considering that there is limited work on sentiment for Chinese text, it is also reflected in the shortage as well as the quality of available Chinese language lexicons. When looking into available dictionaries we have identified that despite there are some duplications of words these dictionaries complement each other. However the biggest disadvantage was that they had no intensity of sentiment. The only lexicon from Bo Yuan, Tinghua University [7] has intensity associated with specific words, however, it has a limited number of words. Another approach such as HowNet [40] has intensity associated with different dictionaries such as ’most’, ’more’, etc. Also, dictionaries from Dalian University has a simple label for positive and negative polarity. All other dictionaries that we found are formed and grouped by positive or negative words.

In creating our lexicon we considered the following dictionaries:

- Chinese Sentiment Word Weight Table from BoYuan in Tinghua University: It contains 23,419 phrases with a positive or negative weight [7].
- How Net: has 17,887 phrases which are divided into 6 groups based on the emotional tendency, which are: ”Positive Evaluation”, ”Negative Evaluation”, ”Positive Emotion”, ”Negative Emotion”, ”perception” and ”Adverb of degree” [40].
- Chinese emotional vocabulary from Dalian University of Technology Information Research Laboratory. It contains 22,012 phrases [35].
- National Taiwan University Sentiment Dictionary - NTUSD: has both Simplified and Traditional Chinese Characters. It has 2,810 positive words and 8,276 Negative words [31], [32].
- Tsinghua University Positive and Negative Dictionary which contains 4,468 negative words and 5,567 positive words [20].

Apart from being limited in content and number of words these dictionaries have shortcoming as there is no sentiment intensity measures of individual words. Weight has been mostly addressed by creating dedicated dictionaries for specific words that are grouped based on association with: most, very, more, half, etc. This causes lower accuracy of sentiment score and also requires dedicated code in programming for calculation of sentiment, which significantly slows down the sentiment calculation.

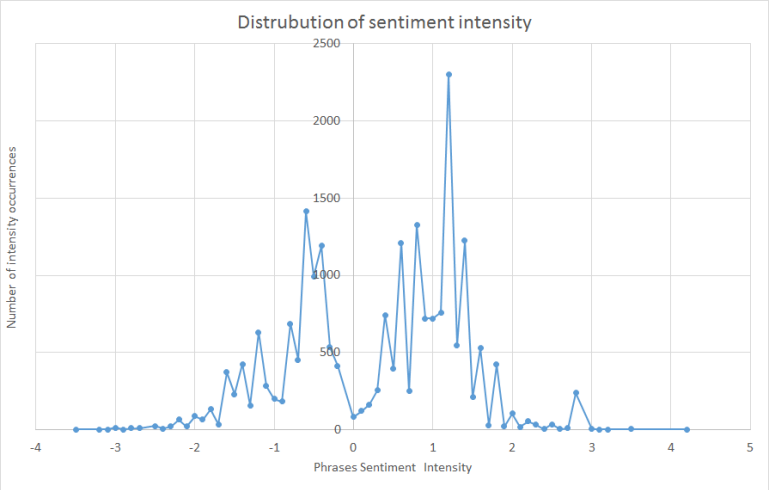

Fig. 4. Distribution of sentiment intensity of BoYuan Tinghua University lexicon

Therefore we created our lexicon which combined existing dictionaries and added weight intensity in accordance with meaning. In absence of human annotation, for example, we associated weight 3 times for words associated with most while 2 times for more, etc. Also, we assigned weight +1 to words in positive dictionaries add -1 for words in a negative dictionary. Words that existed in lexicons and had sentiment intensity validated by humans we retained, such as from HowNet lexicon [9], [40].

The result was a comprehensive lexicon for Chinese language with over 40,000 terms with associated weighting. We realize that certain words still have only weight +1 or -1, however, it is at this stage sufficient and more accurate than any existing method and any existing lexicon. To further improve accuracy it is possible to engage people to do the annotation of words sentiment weight, however, considering the fact that there are over 40,000 words and every individual could have different sentiment weight it is required to survey several people for the same terms, therefore this is a lengthy and expensive task.

In Figure 4 is shown the distribution of BoYuan Tinghua University lexicon and Figure 5 presents the distribution of lexicon we compiled as part of this work, which is the combination of several existing dictionaries.

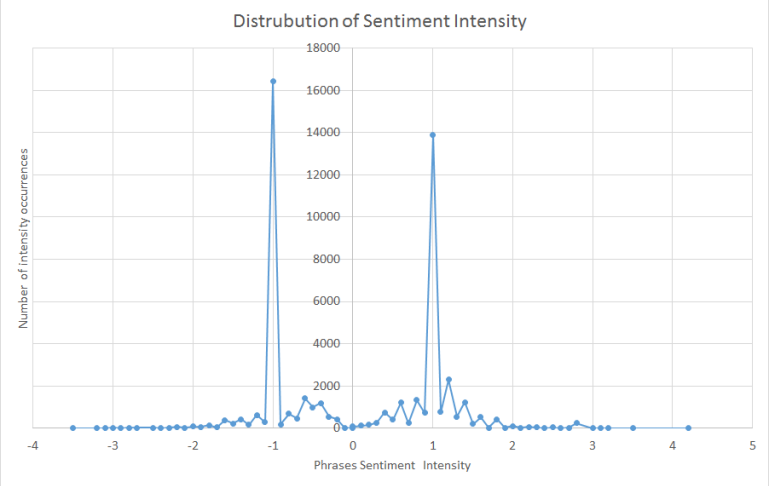

Fig. 5. Distribution of sentiment intensity of our Lexicon

3.2. Sentiment Analysis of Chinese Short Text

As indicated in Section 2 there are many sentiment analysis methods proposed for English language and they are mostly lexicon-based. With regard to sentiment analysis of short Chinese language text, there is limited work, and the process itself is more complicated when compared to the sentiment analysis of the English language because text also requires segmentation into meaningful terms. In Chinese language there are no spaces between characters and specific meaning is often defined by a combination of characters or words. Encouraged by the work of our Big data group on sentiment analysis for English language we decided to follow the same direction and propose a lexicon-based sentiment analysis method for Chinese language. In Figure 6 we show the concept of sentiment analysis of Chinese text that has been proposed and developed in this work.

Calculation of sentiment was done according to Algorithm 1. Once *WEIBO* Chinese social media posts have been collected and stored in MongoDB databases in JSON format, with regard to sentiment calculation of posts, the first step is to only consider the content of

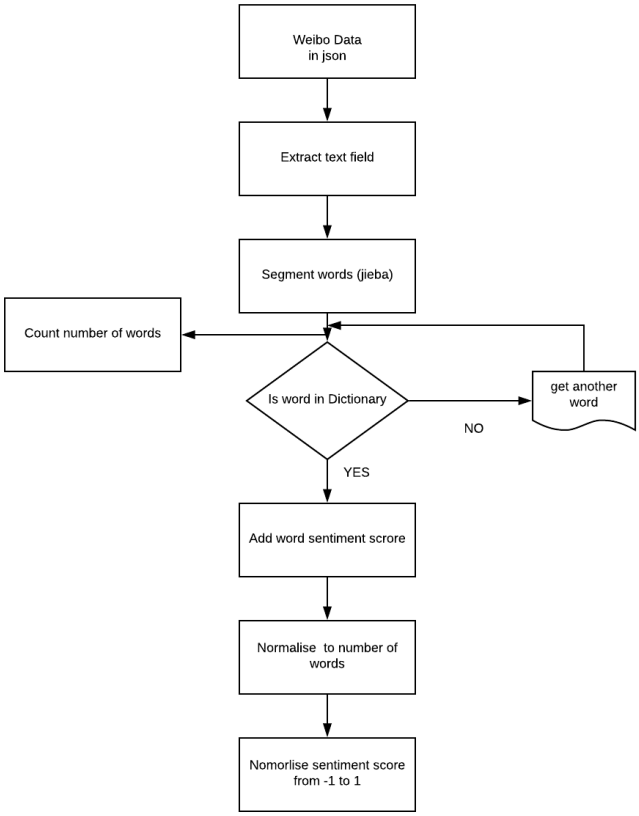

Fig. 6. Sentiment analysis concept

the posted text within the post *TEXT* and encode text with UTF-8 encoding *D*, suitable for Chinese language. Individual encoded text *d* is then considered for sentiment calculation. In the next step, considering that there is no space to separate characters and words in Chinese language, there is a need to do the segmentation to meaningful words. In Chinese language, one character usually does not have meaning, which is a challenge itself to segment text written in Chinese into meaningful terms. For example, in English ”I love traveling”, ”traveling” is one word, however, in Chinese language, it needs two characters ”lv” and ”you” to have the meaning of ”traveling”. Additionally, as there is no space to separate words in a sentence, if we need to find ”traveling” in Chinese language it is required to take into consideration both ”lv” and ”you” in order to have the meaning of ”traveling”. Segmentation of Chinese language is a research topic itself, and it is outside of the content of this work, it attracted significant attention in the literature [13], [29], [2], [30], [23].

In this work on developing sentiment analysis of Chinese short text we accepted and followed *jieba* segmentation method [29]. Apart from being widely used, jieba has an advantage because it has libraries for python, which we could simply plug into python

code for sentiment analysis. Jieba can search the maximum probability path and most probable combination based on the word frequency. It supports three word-segmentation models (accurate mode, full mode, and search engine mode), can process the traditional Chinese word segmentation, and supports custom dictionaries [34]. *Jieba* segmentation library files are publicly accessible at [29].

With regard to Algorithm 1, all segments from posts (*SEG*) are lopped and segments (*seg*) are matched with lexicon (*LEX*). If a segment exists in the lexicon associated intensity (*wordsentiment*) is obtained and added to the Sentence sentiment ($sen\_sentiment$). Because post in Weibo can be short or very long, we count the number of segments (*words*) as well as the number of matching segments with lexicon (*dictwords*), which we take into consideration when calculating the sentiment of an individual post. This is required because there are more positive than negative words in Lexicon and therefore it is more likely that the lengthy posts have more matching words with the lexicon and therefore result in higher positive sentiment if the adjustment was not performed.

Algorithm 1: SENTIMENT CALCULATION OF SHORT CHINESE TEXT

Input: Weibo posts in JSON format
Output: Text and Sentiment scores
WEIBO = Input Weibo posts in JSON format
LEX = Import Lexicon with sentiment intensity
$TEXT = Extract\_only\_text\_field\_from\_post(WEIBO)$
$D = encode\_with\_UTF8(TEXT)$
$sen\_sentiment = 0$
for $d \in D$ **do**
  $words = 0$
  $dictwords = 0$
  $sentiment = 0$
  $SEG = Perform\_Segmentation\_of\_text\_with\_jieba\ (d)$
  for $seg \in SEG$ **do**
    $words = words + 1$
    if $seg \in LEX$ **then**
      $wordsentiment = LEX(seg)$
      $sen\_sentiment = sen\_sentiment + wordsentiment$
      $dictwords = dictwords + 1$
  $adjusted\_sent = sentiment * dictwords \,/\, words$
  $normalize\_to\_one = adjusted\_sent \,/\, \sqrt{adjusted\_sent^2 + 2}$
  $R = append(d,\ normalize\_to\_one)$
  return R

When all segments from the individual post are taken into consideration at first adjusted sentiment ($adjusted\_sent$) was calculated with the following equation:

$$adjusted\_sent = sentiment * dictwords \,/\, words$$

where *dictwords* represent the number of words in a post which has been matched with lexicon and *words* is the total number of words in the post (to avoid the influence of the length of the post to sentence sentiment score).

Finally post sentiment is normalized to 1 to ensure that the sentiment is always between -1, for negative, and +1 for positive post. The following equation is used to normalize sentiment to 1:

$$normalize\_to\_one \; = \; adjusted\_sent \; / \; \sqrt{adjusted\_sent^2 \; + \; 2}$$

The normalized sentiment score is stored in the database along with the associated text and the loop continues to the next post until all posts are finished.

4. Experimental Evaluation

In our experimental evaluation, we considered Sina Weibo posts posted in 2016 and 2017 that mentioned the' Great Barrier Reef', in Chinese language. A total of 24,308 relevant posts have been captured. As a demonstration of sentiment calculation, we relied on our previous experience on the Great Barrier Reef project [4], and selected several relevant keywords including key locations (Townsville, Cairns, etc), food (Seafood), hotel, and some relevant activities such as 'snorkeling' and calculated overall sentiment for these keywords. In Table 1 we provide some sentiment scores calculated by the method proposed in this work. It can be seen, for example, that all related posts are positive in these 2 years, however, the overall sentiment about the travel destinations was increasing, while only the seafood sentiment score dropped from 0.4474 to 0.4043.

Table 1. Sentiment Score of GBR related posts

Key words	SentimentScore2016	SentimentScore2017
Cairns	0.3828	0.4042
Townswille	0.385	0.4484
Whitehaven Beach	0.3713	0.423
Whitsunday	0.3626	0.3473
Hamilton Island	0.354	0.343
Green Island	0.4398	0.4406
Heart Reef	0.481	0.4747
Hotel	0.3261	0.4314
Seafood	0.4474	0.4043
Snorkeling	0.3551	0.3525
Coral	0.3326	0.3766
Heart Reef	0.481	0.474
Fish	0.250	0.328

Considering that apart from text social media posts can contain other relevant metadata we investigated what additional information metadata can provide. We noticed that out of 24,308 captured posts, almost all users (99%) provided their location at the time of registration, which could be a good indication of where the users are coming from. Figure 7 shows a heat map to illustrate where in China the Weibo users who talked about the GBR originate from at the time of registration of their accounts. Based on these details we could identify the number of users from different provinces, for example, Beijing is

top-ranked (a total of 5,091 users) that mentioned the Great Barrier Reef, followed by Guangzhou Province (2,389) and Shanghai (1,456). It is important to mention that this analysis result correlates well with the results released by Queensland Tourism Industry Council which announced that "Markets in Chinas first-tier cities (Beijing, Shanghai, Guangzhou and Shenzhen"). It has been also previously demonstrated that social media data correlates well with scientific observations [5]. Therefore, considering also observation in this work related to first-tier cities it is evident that the metadata can be used to gain valuable information. It is interesting to note that there are posts from all provinces in China, indicating that the GBR is very popular in Chinese social media.

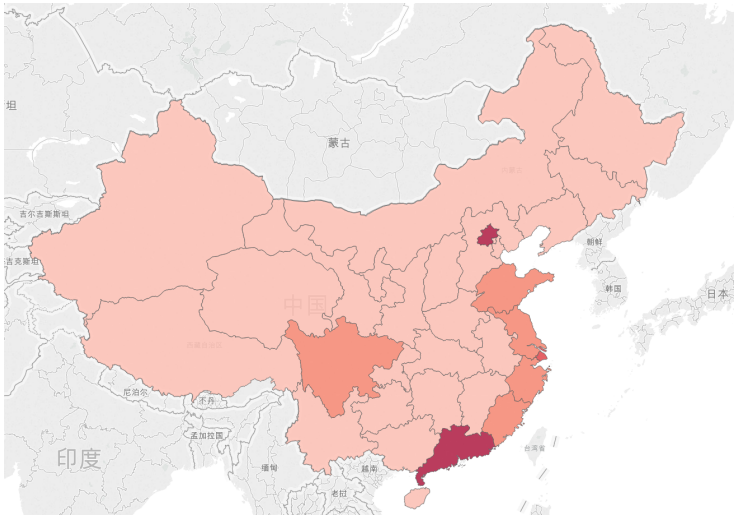

Fig. 7. Heat map of user locations that mentioned GBR from 2016 to 2017

Another analysis, shown in Figure 8, presents average sentiment in all posts that mentioned GBR depending on the province where the user opened their accounts. It is possible to see that some provinces have higher or lower sentiment, which could be investigated and found the reason for higher or lower sentiment toward the GBR and findings can be used for change in marketing strategy. This is another sample of how metadata can be used.

Also, we noticed that about 15% of posts have exact geographic locations (longitude, latitude) of posts, which can be used to identify the points of interest.

Figure 9 shows the distribution of posts in 2016. It is interesting to see that two peak times of posting about GBR are February and September with an increasing trend. The Chinese visitor satisfaction report says "Chinese holiday tourists arriving in Australia for leisure purpose have featured high seasonality". Australia summer tends to be the peak season for Chinese tourists to avoid freezing winter. The peak season lasts for four months from November to February, covering several important holidays such as Christmas and New Year's Eve. Our finding correlates with this observation.

Fig. 8. Heat map of average sentiment in posts that mentioned GBR, depending on where users originate from

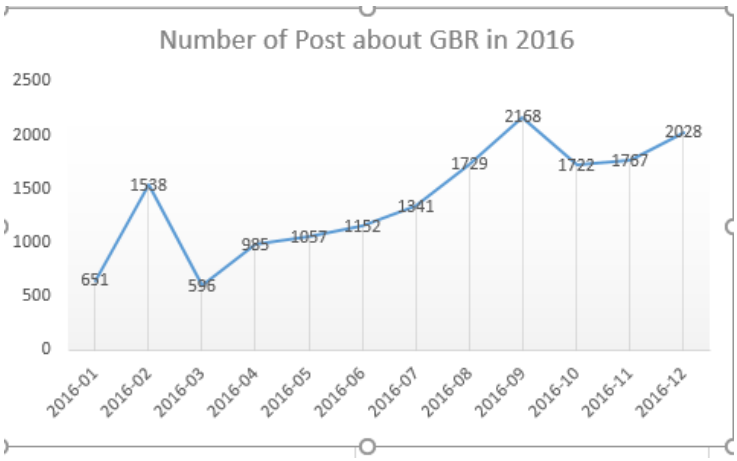

Fig. 9. Number of posts per month in 2016

5. Conclusions

Consumer perception assessments often rely on existing approaches to data collection such as surveys and opinion polls. However, these methods have a range of limitations both in terms of sample size and bias. There is also a risk that the traditional methods are unable to capture opinions and behaviors due to the relatively small sample size, as the sample size is limited due to the cost of the survey. To overcome these limitations, we proposed to tap into available social media posts and perform Big Data analytics. Due

to the fact that there are many Chinese tourists in Australia, specifically in the area of Great Barrier Reef, we decided to gain insight into tourists with Chinese background by capturing and calculating the sentiment of Sina Weibo Chinese social media posts.

Despite sentiment analysis is a well-researched and matured area for the English language there is limited attention for Chinese language and is mostly concentrated on lexicon creation, which itself is small and does not have word sentiment intensities. To overcome the shortcomings of existing approaches, in this work we proposed and tested a method to identify sentiment in Chinese social media posts. We developed a method to crawl the web and specifically collected Weibo posts that mentioned the word Great Barrier Reef both in the Chinese language and the English language. We also elaborated on the process of capturing, managing, described the creation of a comprehensive Chinese lexicon with sentiment intensity, and proposed an algorithm for sentiment calculation. In contrast to all other existing methods proposed method takes into consideration the length of the text as well as the number of identified words in the lexicon. To the best of our knowledge, this is the very first method which avoids bias because there are more negative than positive words in the lexicon and therefore longer posts tend to be more negative since longer posts are more likely to have more matching words in the lexicon.

Additionally, we provided samples of other valuable information, as proof of concept, which can be extracted from social media posts metadata; such as where from Chinese people who have interest and comment on GBR originate from as well as what is the average sentiment depending on locations users indicated as their place of residence. In a case study related to sentiment toward the different GBR destinations, we demonstrated that the proposed method is effective to obtain information and to monitor visitors' opinions.

As of future work to further improve the accuracy of sentiment analysis a more comprehensive lexicon is needed, especially a lexicon that can include words that people like to use in social media, as well as emoji, need to be taken into consideration.

References

1. Alaei, A.R., Becken, S., Stantic, B.: Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research 58(2), 175–191 (2017)
2. Ansjsun: http://www.cnblogs.com/en-heng/p/6274881.html
3. BBC: Sina weibo ends 140-character limit ahead of twitter (2016), https://www.bbc.com/news/technology-35361157
4. Becken, S., Stantic, B., Chen, J., Alaei, A., Connolly, R.: Monitoring the environment and human sentiment on the great barrier reef: Assessing the potential of collective sensing. In: Journal of Environmental Management. vol. 203, pp. 87–97 (2017)
5. Chen, J., Wang, S., Stantic, B.: Connecting social media data with observed hybrid data for environment monitoring. In: Proceedings of the 11th International Symposium on Intelligent Distributed Computing - IDC 2017. pp. 125–135 (2017)
6. Chen, S., Ding, Y., Xie, Z., Liu, S., Ding, H.: Chinese weibo sentiment analysis based on character embedding with dual-channel convolutional neural network. In: 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). pp. 107–111. IEEE (2018)
7. Chinese Sentiment Word Weight Table, BoYuan, T.U.: https://github.com/bung87/bixin/tree/master/dictionaries

8. Claster, W.B., Cooper, M., Sallis, P.: Thailand–tourism and conflict: Modeling sentiment from twitter tweets using naïve bayes and unsupervised artificial neural nets. In: Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on. pp. 89–94. IEEE (2010)
9. Dong, P.Z.: Hownet (2000), http://www.keenage.com/html/e_index.html
10. Duong, C.T., Nguyen, Q.V.H., Wang, S., Stantic, B.: Provenance-based rumor detection. In: 28th Australasian Database Conference - ADC. pp. 125–137 (2017)
11. Fan, B.: Weibo data report in 2016 (2016), http://data.weibo.com/report/reportDetail?id=346
12. Greatbarrierreef: www.greatbarrierreef.org, http://www.greatbarrierreef.org/about-the-reef/great-barrier-reef-facts/
13. Hua-ping, Z., Qun, L.: Ictclas (2013)
14. Hutto, C., Gilbert, E.: 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 (2014)
15. instagram: instagram (26 April,2017), http://blog.instagram.com/post/160011713372/170426-700million
16. Kasper, W., Vela, M.: Sentiment analysis for hotel reviews. In: Computational linguistics-applications conference. vol. 231527, pp. 45–52 (2011)
17. Kim, S.E., Lee, K.Y., Shin, S.I., Yang, S.B.: Effects of tourism information quality in social media on destination image formation: The case of sina weibo. Information & Management 54(6), 687–702 (2017)
18. Kirilenko, A.P., Stepchenkova, S.O., Kim, H., Li, X.: Automated sentiment analysis in tourism: Comparison of approaches. Journal of Travel Research p. 0047287517729757 (2017)
19. Leung, D., Law, R., Van Hoof, H., Buhalis, D.: Social media in tourism and hospitality: A literature review. Journal of Travel & Tourism Marketing 30(1-2), 3–22 (2013)
20. Li, J.: Chinese positive and negtive lecxion (22 January,2011), http://nlp.csai.tsinghua.edu.cn/site2/index.php/resources/13-v10%20Seeaustralia,%202017
21. Li, X., Li, J., Wu, Y.: A global optimization approach to multi-polarity sentiment analysis. PloS one 10(4), e0124672 (2015)
22. Liu, Z., Shan, J., Balet, N.G., Fang, G.: Semantic social media analysis of chinese tourists in switzerland. Information Technology & Tourism 17(2), 183–202 (2017)
23. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
24. Osgood, C.E.: The nature and measurement of meaning. Psychological bulletin 49(3), 197 (1952)
25. Sina: (2009-2018), https://weibo.com
26. Stantic, B., Pokornỳ, J.: Opportunities in big data management and processing. Databases and Information Systems 270, 15–26 (2014)
27. statista: Number of monthly active twitter users worldwide from 1st quarter 2010 to 4th quarter 2017 (in millions) (28 January,2014), https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
28. statista.com: Number of monthly active users (2018), https://www.statista.com/topics/1164/social-networks/
29. Sun, J.: jieba chinese word segmentation tool. available at https://github.com/fxsjy/jieba (2012)
30. Sun, M., Chen, X., Zhang, K., Guo, Z., Liu, Z.: Thulac: An efficient lexical analyzer for chinese. Tech. rep., Technical Report (2016)
31. University, N.T.: Http://academiasinicanlplab.github.io/
32. University, N.T.: Ntusd: National taiwan university semantic dictionary (22 January,2011), https://rdrr.io/rforge/tmcn/man/NTUSD.html

33. Wang, M.H., Lei, C.L.: Boosting election prediction accuracy by crowd wisdom on social forums. In: Consumer Communications & Networking Conference (CCNC), 2016 13th IEEE Annual. pp. 348–353. IEEE (2016)
34. Xiao, K., Zhang, Z., Wu, J.: Chinese text sentiment analysis based on improved convolutional neural networks. In: Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on. pp. 922–926. IEEE (2016)
35. Xu, L., Lin, H., Pan, Y., Ren, H., Chen, J.: Chinese emotional vocabulary. ???? 27(2), 180–185 (2008)
36. Xu, Y., Liu, Z., Zhao, J., Su, C.: Aweibo sentiments and stock return: A time-frequency view. PloS one 12(7), e0180723 (2017)
37. Yin, F., Zhang, B., Su, P., Chai, J.: Research on the text sentiment classification about the social hot events on weibo. In: IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). pp. 1537–1541 (2016)
38. Zeng, B., Gerritsen, R.: What do we know about social media in tourism? a review. Tourism Management Perspectives 10, 27–36 (2014)
39. zephoria: Valuable facebook statistics (2017), hhttps://zephoria.com/top-15-valuable-facebook-statistics/
40. Zhendong, D., Qiang, D.: Hownet And The Computation Of Meaning. World Scientific (2006)
41. Zhuo, S., Wu, X., Luo, X.: Chinese text sentiment analysis based on fuzzy semantic model. In: Cognitive Informatics & Cognitive Computing (ICCI* CC), 2014 IEEE 13th International Conference on. pp. 535–540. IEEE (2014)

Jinyan Chen is a doctoral student from the Department of Tourism, Sport and Hotel Management at Griffith University. Her research interest is related to Big Data Analysis in tourism and tourists travel pattern and sentiment. Jinyan has been also working as research assistant on different projects related to social media and tourists behavior.

Dr Susanne Becken is the Director of the Griffith Institute for Tourism and a Professor of Sustainable Tourism at Griffith University, Australia. Susanne has published widely in the field of sustainable tourism, is on the editorial board of 10 journals, and works closely with Government, businesses and international organisations on issues related to tourism management.

Bela Stantic is Professor in Computer Science and Director of Big Data and Smart Analytics Lab within the Institute of Integrated and Intelligent Systems at Griffith University. Bela is internationally recognized in field of Big Data analytics and efficient management of complex data structures. He successfully applied his research interdisciplinary to many areas including health, environment and tourism.